

Article

CBERS 4A/WPM Image Classification for Mapping Land Use and Land Cover in TCRAs

Carla Rodrigues Santos ^{1,*}, Francisco Salazar ^{2,*} , Fernanda Beatriz Jordan Rojas Dallaqua ²  and Bruno Schultz ² 

¹ Grupo de Pesquisa DEGET da Unemat, Sinop 78555-000, Mato Grosso, Brazil

² Visiona Tecnologia Espacial S.A., São José dos Campos 12247-016, São Paulo, Brazil; fernanda.dallaqua@visionaespecial.com.br (F.B.J.R.D.); bruno.schultz@visionaespecial.com.br (B.S.)

* Correspondence: carlageors@gmail.com (C.R.S.); fsalazar.ext@visionaespecial.com.br (F.S.)

Abstract

Tropical restoration areas present high spectral heterogeneity and spatial fragmentation, making land use and land cover (LULC) classification challenging, particularly when using medium-resolution satellite imagery. Although spectral mixture analysis, object-based image analysis, and machine learning approaches have been widely explored individually, their integrated application using freely available China–Brazil Earth Resources Satellite 4A Wide-field Panchromatic and Multispectral (CBERS-4A/WPM) imagery for monitoring Environmental Restoration Commitment Agreements (TCRAs) remains limited. This study investigated whether the integration of Linear Spectral Mixture Model (LSMM) fractions, Geographic Object-Based Image Analysis (GEOBIA), and machine learning algorithms could improve LULC classification for mapping TCRAs in the state of São Paulo, Brazil. Spectral, textural, and sub-pixel attributes derived from pan-sharpened CBERS-4A/WPM imagery were combined into a unified feature dataset, and Random Forest, XGBoost, Multilayer Perceptron, and Support Vector Machine classifiers were evaluated. Random Forest achieved the best classification performance, with an overall accuracy of approximately 0.66, demonstrating superior predictive capability compared with the other tested algorithms. Feature importance analysis indicated that LSMM-derived vegetation and shadow fractions contributed significantly to class discrimination. The integration of spectral, spatial, and machine learning approaches provides an interpretable and cost-effective framework for LULC mapping in highly fragmented tropical restoration landscapes, supporting environmental monitoring and restoration assessment.

Keywords: image classification; machine learning; CBERS 4A



Academic Editor: Alexander N. Fedorov

Received: 3 June 2026

Revised: 16 August 2026

Accepted: 19 August 2026

Published: 1 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Environmental restoration has become a global priority to mitigate biodiversity loss, climate change, and ecosystem degradation, supporting international initiatives such as the Bonn Challenge and the UN Decade on Ecosystem Restoration. Large-scale restoration contributes to biodiversity conservation, carbon sequestration, erosion control, hydrological regulation, and the recovery of ecosystem services, while increasing landscape resilience to environmental change [1,2]. In Brazil, Environmental Restoration Commitment Agreements (TCRAs), originally named *Termos de Compromisso de Recuperação Ambiental*, are legal instruments established under environmental legislation to regulate the restoration of degraded areas. Particularly in the state of São Paulo, these agreements define mandatory restoration actions, implementation schedules, and monitoring requirements to ensure environmental

compliance. Beyond their legal function, TCRAAs play a fundamental ecological role by promoting native vegetation regeneration, biodiversity conservation, landscape connectivity, erosion control, and the recovery of ecosystem services in degraded landscapes [2,3].

Consequently, systematic monitoring of restoration trajectories is essential to evaluate restoration success, quantify vegetation recovery, support adaptive management, and guide environmental policies. Reliable monitoring methods also enable environmental agencies to assess compliance with restoration agreements and improve decision-making processes at regional scales [1,2]. However, tropical landscapes represent a challenge for remote sensing applications due to their high spatial fragmentation, spectral heterogeneity, and the coexistence of vegetation, exposed soil, urban areas, and water bodies within small spatial units. These characteristics increase pixel mixing, reduce class separability, and can decrease the performance of conventional pixel-based classification methods, particularly when medium-resolution satellite imagery is used. In fragmented tropical environments, the spectral response of a single pixel frequently represents a combination of multiple land cover components, limiting the discrimination of complex classes and increasing classification uncertainty.

In this context, monitoring highly fragmented restoration areas requires approaches capable of integrating complementary information derived from satellite imagery, including spectral, sub-pixel, spatial, and textural attributes. The integration of these sources of information may provide a more comprehensive representation of heterogeneous landscapes and support land use and land cover (LULC) classification in restoration areas. Previous studies have demonstrated the potential of spectral mixture analysis to represent mixed pixels and improve the characterization of heterogeneous land cover [4,5]. However, relatively few studies have evaluated the complementary effects of integrating LSMM-derived fractions, GEOBIA attributes, and multiple machine learning algorithms within a unified classification framework using freely available medium-resolution CBERS-4A/WPM imagery, particularly for highly fragmented Environmental Restoration Commitment Agreement (TCRA) areas. This gap motivates the investigation of integrated classification approaches to support environmental restoration monitoring and decision-making.

In this context, this study investigates whether the integration of LSMM-derived fractions, GEOBIA attributes, and machine learning algorithms can improve land use and land cover classification in Environmental Restoration Commitment Agreements (TCRAAs) in the state of São Paulo, Brazil, using CBERS-4A/WPM imagery. Specifically, the study evaluates the performance of Random Forest, XGBoost, Multilayer Perceptron, and Support Vector Machine classifiers based on spectral, fractional, and textural attributes.

This study makes four main contributions to the remote sensing literature. First, it integrates LSMM-derived fractions, GEOBIA attributes, and four machine learning algorithms within a unified framework for LULC classification. Second, it evaluates the applicability of freely available CBERS-4A/WPM imagery for monitoring highly fragmented environmental restoration areas. Third, it investigates feature importance using SHAP to improve model interpretability. Finally, it provides practical evidence supporting cost-effective remote sensing approaches for monitoring TCRAAs in tropical landscapes. The remainder of this paper is organized as follows. Section 2 presents the literature review, focusing on LULC classification in heterogeneous tropical landscapes, LSMM, GEOBIA, machine learning, and their integration. Section 3 describes the study area, datasets, preprocessing procedures, classification methods, and accuracy assessment. Section 4 presents and discusses the classification results and model performance. Finally, Section 5 summarizes the main findings, limitations, implications, and directions for future research.

2. Literature Review

2.1. Land Use and Land Cover Classification in Heterogeneous Tropical Landscapes

Land use and land cover (LULC) classification in tropical landscapes remains challenging because of high spatial and spectral heterogeneity, land cover fragmentation, and the coexistence of vegetation, agricultural areas, exposed soils, urbanized areas, and water bodies within relatively small spatial units. These conditions can reduce class separability and increase classification uncertainty, particularly when medium-spatial-resolution imagery is used, because individual pixels may represent more than one land cover component [6]. In addition, within-class spectral variability and spectral similarity among different classes can further complicate discrimination in fragmented environments [7,8].

Classification performance is influenced by several factors beyond the choice of algorithm, including training data quality, feature selection, parameter optimization, and class distribution [9,10]. Ghorbanzadeh et al. [10], for example, demonstrated that spectral heterogeneity and class imbalance can affect machine learning performance, whereas Sovann et al. [11] showed that the integration of spectral, seasonal, and topographic information can improve classification in tropical forest environments. These findings indicate that classification performance depends not only on the classifier itself but also on how landscape characteristics are represented in the predictor space.

Recent LULC studies have therefore explored complementary spatial, textural, temporal, and ancillary information. Oliveira et al. [12] demonstrated the applicability of Sentinel-2/MSI imagery combined with Geographic Object-Based Image Analysis (GEOBIA) for LULC mapping in a seasonally dry tropical forest in Brazil. Lasko et al. [13] integrated spectral and textural attributes, distance information, ancillary geospatial data, and Random Forest, whereas Subedi et al. [14] combined Sentinel-2 time-series imagery, high-spatial-resolution data, and machine learning for LULC mapping in heterogeneous landscapes. Collectively, these studies demonstrate the value of representing land cover using complementary sources of information rather than relying exclusively on conventional spectral features [9–14].

This issue is particularly relevant to Environmental Restoration Commitment Agreements (Termos de Compromisso de Recuperação Ambiental—TCRAs), where small agricultural fields, forest remnants, riparian vegetation, exposed soils, and anthropogenic disturbances may occur in close proximity. Although multisource and object-based approaches have increasingly been investigated, the combined use of spectral, spatial, textural, and sub-pixel information for LULC classification using CBERS-4A/WPM imagery remains comparatively less explored in highly fragmented restoration landscapes. This context provides a basis for investigating integrated classification approaches for TCRA areas in the state of São Paulo, Brazil.

2.2. LSMM and GEOBIA as Complementary Approaches for LULC Classification

Spectral mixture analysis addresses the sub-pixel composition problem by decomposing pixel spectra into physically interpretable components. Its theoretical foundation was established by Shimabukuro and Smith [4], while Roberts et al. [5] demonstrated the potential of multiple endmember spectral mixture analysis to improve the representation of mixed pixels by allowing different endmember combinations to be modeled for individual pixels. Ponzoni and Shimabukuro [15] further emphasized the relevance of spectral mixture approaches for characterizing vegetation structure and spectral behavior in heterogeneous environments. The Linear Spectral Mixture Model (LSMM) has subsequently been applied to environmental monitoring in tropical landscapes.

Hasan et al. [16] used spectral unmixing to derive vegetation, soil, and non-photosynthetic vegetation fractions for assessing cumulative forest disturbances in the

northeastern Amazon, while Cao et al. [17] applied multiple-criteria spectral mixture analysis to characterize succession in tropical dry forests. These studies indicate that fraction-based representations can provide information that complements conventional spectral features when different land cover components coexist within individual pixels.

Whereas spectral mixture analysis characterizes the internal composition of pixels, Geographic Object-Based Image Analysis (GEOBIA) represents land cover through spatially meaningful image objects. By grouping neighboring pixels according to their spectral and spatial characteristics, GEOBIA enables the incorporation of contextual, geometric, and textural information that is not directly represented by individual pixel values [6,8]. Its relevance to tropical LULC mapping has been demonstrated by Oliveira et al. [12], who combined Sentinel-2/MSI imagery with object-based analysis in a seasonally dry tropical forest in Brazil. Carvalho et al. [18] demonstrated the potential of GEOBIA using CBERS-4A/WPM imagery for land cover classification, while Magalhães et al. [19] combined CBERS-4A/WPM imagery with spectral indices, principal component analysis, pseudo-labeling, and deep learning for LULC mapping.

LSMM and GEOBIA consequently provide complementary representations of heterogeneous landscapes: LSMM describes the proportional composition of mixed pixels, whereas GEOBIA incorporates their spatial organization and contextual characteristics. However, GEOBIA alone does not fully address sub-pixel composition, and fraction-based representations do not explicitly describe object-level spatial structure. The combination of these approaches therefore provides an opportunity to represent both the internal composition and spatial organization of land cover units. Although previous studies have demonstrated the potential of CBERS-4A/WPM, LSMM, and GEOBIA individually or in other combinations [12,18,19], their integration with textural attributes and machine learning remains comparatively less explored, particularly in fragmented tropical restoration landscapes.

2.3. Machine Learning, Training Data, and Class Imbalance in LULC Classification

Machine learning algorithms have been widely applied to LULC classification because of their ability to model nonlinear relationships and integrate multiple predictors [9,10]. Random Forest [20], XGBoost [21], Support Vector Machine [22], and Multilayer Perceptron [23] represent different machine learning paradigms that have been applied to complex classification problems involving high-dimensional remote sensing data. However, the literature does not indicate that a single classifier consistently outperforms the others across different datasets and environmental conditions. Maxwell et al. [9] showed that classification performance depends on training data, feature space, parameter tuning, and landscape complexity, while Belgiu and Drăguț [24] highlighted the robustness of Random Forest in remote sensing while emphasizing its dependence on the characteristics of the input data and mapped classes.

Comparative studies have similarly demonstrated context-dependent performance. Tan et al. [25] compared Random Forest and Support Vector Machine for land cover classification, whereas Kupidura et al. [26] evaluated Random Forest, XGBoost, and Support Vector Machine under different training sample sizes. These findings support comparative experiments in which different machine learning paradigms are evaluated using the same input features and validation framework.

The quality, representativeness, and distribution of training samples are also important determinants of classification performance. Huang et al. [27] demonstrated that training sample size can substantially influence classification accuracy, while Foody and Mathur [28] emphasized that insufficient or poorly distributed samples can reduce model generalization. Maxwell et al. [9] likewise identified training data quality, feature selection, parameter

optimization, and class distribution as relevant factors affecting classification performance. These considerations become particularly important in fragmented landscapes, where land cover classes may occupy markedly different spatial extents.

Class imbalance can cause dominant classes to disproportionately influence model training, whereas minority classes may be more difficult to discriminate. Ma et al. [8] and Ghorbanzadeh et al. [10] highlighted the importance of landscape structure, object characteristics, and class distribution in remote sensing classification, while recent studies have shown that imbalanced class distributions may bias classifiers toward majority classes and reduce the representation of minority categories [29–31]. Consequently, Overall Accuracy alone may not adequately describe classification performance, particularly when dominant classes are prevalent. Class-specific measures, including F1-score and confusion matrix analysis, provide complementary information for evaluating discrimination among dominant and minority categories [29–32]. Sampling strategies such as undersampling, oversampling, and synthetic sample generation have also been investigated to reduce the effects of class imbalance [29–31].

Taken together, these findings indicate that classifier comparison should be conducted within a consistent experimental framework that accounts not only for algorithmic differences but also for the characteristics and distribution of the training data. This is particularly relevant when spectral, sub-pixel, spatial, and textural predictors are combined, as their relative contribution to classification performance may vary according to the classifier and the composition of the training dataset.

2.4. Integrated Approaches and Research Gap

The literature indicates that different methodological approaches address distinct aspects of the LULC classification problem in heterogeneous landscapes. Spectral mixture analysis provides information about the proportional composition of mixed pixels [4,5,17], GEOBIA incorporates spatial, textural, geometric, and contextual characteristics [6,8,12,18,19], and machine learning methods provide flexible frameworks for integrating heterogeneous predictors and modeling nonlinear relationships with LULC classes [9,20–26]. Previous studies have consequently explored combinations of spectral, textural, object-based, temporal, ancillary, and machine learning information [12–14,18,19], demonstrating the potential of complementary representations for improving LULC characterization.

Nevertheless, important methodological gaps remain. Existing studies have generally focused on individual components or selected combinations, whereas the combined use of LSM-derived sub-pixel fractions, GEOBIA attributes, textural information, and multiple machine learning algorithms within a common CBERS-4A/WPM-based classification framework remains comparatively less explored. This gap is particularly relevant in fragmented restoration landscapes, where small land cover units and heterogeneous spectral responses can simultaneously affect pixel-level and object-level classification [6–8,10]. Evaluating these complementary representations under the same training and validation framework can therefore provide a more systematic assessment of their suitability for discriminating heterogeneous LULC classes.

Accordingly, the present study investigates the combined use of CBERS-4A/WPM imagery, LSM-derived fractions, GEOBIA attributes, and four machine learning algorithms—Random Forest, XGBoost, Support Vector Machine, and Multilayer Perceptron—for LULC classification in TCRA areas in the state of São Paulo, Brazil. By evaluating these components within a common experimental framework, the study assesses their suitability for discriminating heterogeneous land cover classes and addresses the identified methodological gap.

2.5. Explainable Artificial Intelligence and SHAP for LULC Classification

The integration of spectral, fractional, spatial, and textural variables increases the complexity of machine learning-based LULC classification and highlights the importance of model interpretability. Explainable Artificial Intelligence (XAI) provides approaches for interpreting model predictions and identifying the contribution of individual variables to model outputs. Among these approaches, SHapley Additive exPlanations (SHAP) quantifies the contribution of individual features to predictions based on Shapley values, providing a theoretically grounded framework for interpreting complex machine learning models [33,34]. Recent studies have demonstrated the applicability of SHAP to remote sensing classification by identifying the relative contribution of spectral and environmental variables to LULC predictions [35].

For tree-based models such as Random Forest, SHAP can complement conventional feature-importance measures by providing an interpretable assessment of individual predictors [34]. This is particularly relevant when the input dataset combines different types of information, such as spectral bands, LSMM-derived fractions, and GEOBIA-based attributes. In the present study, SHAP analysis was therefore used to complement conventional accuracy assessment by identifying the relative contribution of spectral, fractional, and object-based features to model predictions. This analysis provides an additional interpretative dimension to the proposed LULC classification framework and supports the identification of the information most relevant for discriminating land cover classes in fragmented TCRA landscapes.

3. Materials and Methods

3.1. Study Area and Data

Environmental Restoration Commitment Agreements (TCRAs) are distributed throughout the state of São Paulo, Brazil, encompassing diverse environmental conditions, restoration stages, and land cover configurations. For this study, a test area was selected based on the high concentration of TCRAs and the occurrence of complex landscape conditions, including pronounced spatial fragmentation and spectral heterogeneity, which represent major challenges for land use and land cover (LULC) classification using remote sensing data [10].

The selected area comprises 47 municipalities located in the state of São Paulo (Figure 1). These municipalities were identified from the official TCRA database provided by the São Paulo State Secretariat for Environment, Infrastructure and Logistics (SEMIL), prioritizing regions with a high concentration of restoration polygons and heterogeneous environmental conditions [36]. Together, the 47 municipalities encompass a substantial proportion of the TCRAs considered in the study and provide a spatial context for evaluating the proposed LULC classification framework.

The study region is predominantly located within the Atlantic Forest biome, including Dense Ombrophilous Forest and Seasonal Semideciduous Forest formations, as well as other native vegetation types. The coexistence of native vegetation, agricultural fields, pastures, urban areas, and restoration sites results in pronounced spatial and spectral heterogeneity. Such landscape complexity increases the occurrence of mixed pixels and class confusion, representing an important challenge for LULC classification in heterogeneous environments [7,8,10].

Part of the selected area corresponds to locations previously investigated by the research team. The accumulated field knowledge available for these sites supported the interpretation of satellite imagery and contributed to the validation of the resulting LULC maps.

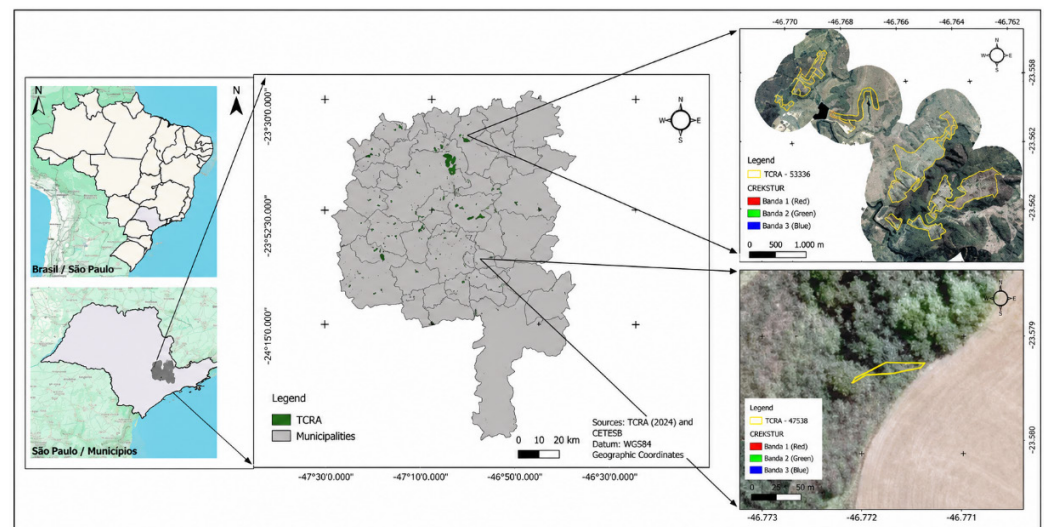


Figure 1. Location of the study area.

3.2. Data Preprocessing and Feature Extraction

Vector datasets containing the boundaries of Environmental Restoration Commitment Agreements (TCRAs) were obtained from the São Paulo State Secretariat for Environment, Infrastructure, and Logistics (SEMIL), with reference to June 2025 [36]. Multispectral satellite imagery from the China–Brazil Earth Resources Satellite 4A Wide-field Panchromatic and Multispectral (CBERS-4A/WPM) sensor was acquired through the DGI/INPE catalog between February and September 2025, covering orbit/path combinations 141/202, 142/203, and 143/204.

The WPM sensor provides four multispectral bands (blue, green, red, and near-infrared) at 8 m spatial resolution, together with a 2 m panchromatic band. The spectral characteristics and derived variables used in the analysis are presented in Table 1.

Table 1. Spectral bands and derived variables used in the analysis.

Source/Sensor	Description/ Application	Resolution (m)	Spectral Range (μm)	Variable/Band	Category
CBERS-4A/WPM	Sensitive to reflectance of water bodies and urban areas	2	0.45–0.52	Blue (B1)	Spectral
CBERS-4A/WPM	Enhances vegetation and exposed soils	2	0.52–0.59	Green (B2)	Spectral
CBERS-4A/WPM	Useful for vegetation discrimination and agricultural use	2	0.63–0.69	Red (B3)	Spectral
CBERS-4A/WPM	High spatial resolution at the pixel level	2	0.77–0.89 μm	NIR	Spectral
Derived from multispectral bands	Proportion of exposed soil in each pixel	–	–	Soil fraction	Linear Spectral Mixture Model (LSMM)
Derived from multispectral bands	Proportion of green vegetation cover	–	–	Vegetation fraction	Linear Spectral Mixture Model (LSMM)
Derived from multispectral bands	Represents shaded areas caused by terrain or dense vegetation	–	–	Shadow fraction	Linear Spectral Mixture Model (LSMM)

Note: NIR = Near infrared.

Preprocessing steps included radiometric and geometric corrections prior to spectral decomposition. A pan-sharpening procedure was then applied to fuse the spectral information from the four multispectral bands with the spatial detail of the panchromatic band, producing multispectral images with a final spatial resolution of 2 m. The resulting images were used for subsequent object-based segmentation and feature extraction.

The pan-sharpened multispectral images constituted the primary input for the subsequent processing steps. Spectral variables were derived directly from the four multispectral bands, while sub-pixel information was obtained using the Linear Spectral Mixture Model (LSMM). The LSMM was applied to estimate the proportions of soil, vegetation, and shade within each pixel, following the approach described by Vasconcelos and Novo [37] and the theoretical foundations of spectral mixture analysis [4,5]. The resulting fraction images were incorporated into the feature extraction procedures.

Geographic Object-Based Image Analysis (GEOBIA) was subsequently applied to the pan-sharpened CBERS-4A/WPM imagery. Image segmentation was used to generate spatially homogeneous objects from which spectral, fractional, and textural attributes were extracted. These object-based attributes complemented the original spectral bands and the LSMM-derived fractions [6,8,18,19].

Following feature extraction, the spectral, fractional, and object-based attributes were integrated into a unified feature dataset for subsequent classification. Particular attention was given to the discrimination between consolidated native vegetation and areas under natural regeneration, given their relevance to the assessment of restoration dynamics [1,3]. In addition, a temporal image mosaic was adopted to minimize seasonal effects and provide a more stable representation of LULC classes, particularly in agricultural and transitional environments [11,14].

The overall methodological workflow, including image acquisition, preprocessing, spectral mixture analysis, GEOBIA segmentation, feature extraction, and integration of the resulting variables, is presented in Figure 2.

3.3. Machine Learning Classification

The land use and land cover (LULC) classes used in the classification were defined based on the classification system adopted by the MapBiomas Project, which served as the reference framework for class definition in this study [37]. This approach ensured compatibility with a nationally recognized LULC classification system. The selected classes were defined according to their occurrence and relevance within the analyzed Environmental Restoration Commitment Agreements (TCRAs), considering the predominance of forest formation, pasture, and agricultural areas, as well as the presence of urban areas and water-related classes. These classes represent the main land cover conditions and restoration contexts observed within the study area.

A visual interpretation key was subsequently developed to support the recognition of LULC patterns within the TCRA areas using CBERS-4A/WPM imagery in natural color composition (R: Band 3, G: Band 2, B: Band 1). The key considered spectral response, color, texture, shape, spatial arrangement, and landscape context to guide the identification and characterization of the main LULC patterns prior to the classification procedure. The visually identified patterns were subsequently consolidated into the final classification scheme according to their thematic relevance, spectral distinguishability, and the characteristics of the available imagery. The consolidated classes were used consistently as the target classes for the machine learning algorithms, and the corresponding interpretation criteria are presented in Table 2.

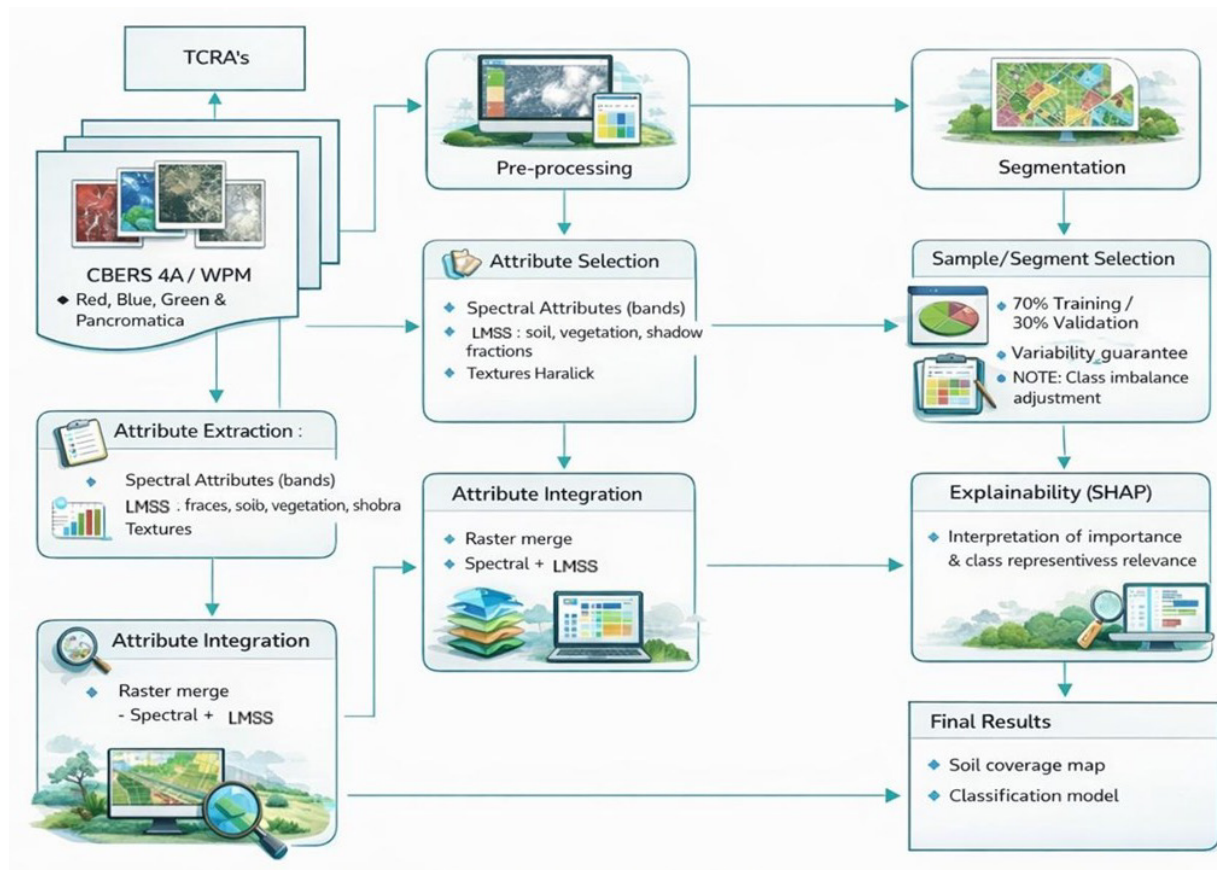
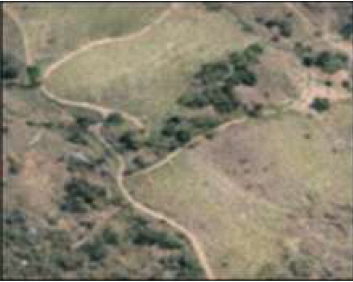
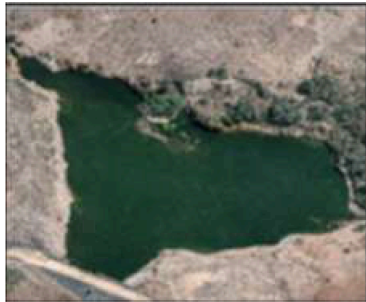


Figure 2. Steps in the methodology.

Table 2. Visual Interpretation Key for Land Use and Land Cover Classes.

WPM Image	Criteria	Characteristics	Classes
	Color: dark green. Texture: rough and granular, adjacent to an area of exposed soil. Shape: sinuous and irregular, identified as an area of native vegetation.	Area characterized by Dense Ombrophilous Forest, Semideciduous Forest, and Pioneer Formation, composed of distinct vegetation strata, including the understory and herbaceous layers.	Forest formation
	Color: shades of green and light brown. Texture: exhibiting minor variations potentially caused by shrubs, small trees, or the uneven distribution of ground vegetation. Shape: irregular, with the presence of trails or linear roads.	Planted pasture areas directly associated with livestock farming. Natural pasture areas predominantly feature grassland or flooded field formations and may or may not be subject to grazing practices.	Pasture

Table 2. *Cont.*

WPM Image	Criteria	Characteristics	Classes
	Color: shades of white and pink, predominant in built-up areas. Texture: rough and heterogeneous, with the presence of various built and natural elements. Shape: irregular and fragmented, typical of an urban environment with houses, streets, and other constructions.	Areas with a significant density of buildings and roads, including infrastructure and open spaces.	Urban area
	Color: light to medium green, or light brown for exposed soil. Texture: medium to smooth, without abrupt variations in tone. Shape: regular and geometric, characterized as a planned agricultural area.	Temporary, perennial, and semi-perennial crops intentionally cultivated by humans.	Crop
	Color: dark blue to black. Texture: smooth, formed by a dyke or artificial. Shape: linear. The color can range from black to blue; the lighter the color, the higher the sediment presence. It shows little variation between the dry and rainy seasons.	Rivers, lakes, dams, reservoirs, and other water bodies.	Flooded areas

The classification of the image objects was preceded by GEOBIA segmentation using the Orfeo Toolbox 9.1.1 (OTB) integrated within the QGIS 4.0 environment. The objective was to generate spatially and spectrally homogeneous image objects for subsequent feature extraction and classification [6,8,38–40]. The segmentation parameters were set to scale = 10, shape = 0.05, and compactness = 0.5. These parameters were selected to preserve fine-scale spatial variability, which is particularly important in restoration areas characterized by small and heterogeneous land cover patches. The low shape weight prioritizes spectral information over geometric regularity [6,8,38]. Parameter selection was performed through iterative visual assessment to reduce over-segmentation and under-segmentation and ensure spatial consistency of the generated image objects [6,8,38].

Following segmentation, spectral, fractional, and textural attributes were extracted from the image objects [6,8]. Descriptive statistics were calculated for the multispectral bands and LSMM-derived fractions, including minimum, maximum, mean, standard deviation, and the 5th, 25th, 50th, 75th, and 95th percentiles. A total of 63 attributes were generated from the four multispectral bands (Blue, Green, Red, and NIR) and the LSMM-derived soil, vegetation, and shadow fractions. The complete set of attributes is presented in Table 3.

Table 3. Table of Analyzed Attributes.

Description	Attribute *	Feature Group	Category
Minimum reflectance (Blue band)	blue_min	Blue	Spectral
Maximum reflectance (Blue band)	blue_max		
Mean reflectance (Blue band)	blue_mean		
Standard deviation (Blue band)	blue_std		
5th percentile	blue_p05		
25th percentile	blue_p25		
Median (50th percentile)	blue_p50		
75th percentile	blue_p75		
95th percentile	blue_p95		
Same statistics applied to Green band	green_*	Green	Fractional (LSMM)
Same statistics applied to Green band	red_*	Red	
Same statistics applied to Green band	nir_*	NIR	
Fraction of soil component	LSMM_solo_*	Soil	
Fraction of shadow component	LSMM_sombra_*	Shadow	
Fraction of vegetation component	LSMM_vegetation_*	Vegetation	

* A total of 63 attributes were derived from the multispectral bands (Blue, Green, Red, and NIR) and spectral mixture analysis (LSMM) components (soil, shadow, and vegetation). For each variable, nine statistical descriptors were computed: minimum, maximum, mean, standard deviation, and percentiles (5th, 25th, 50th, 75th, and 95th). NIR = Near infrared, LSMM = Linear Spectral Mixture Model.

The resulting feature dataset combined spectral, LSMM-derived fractional, and GEO-BIA object-based attributes. Feature selection procedures were applied prior to classification to reduce redundancy among predictors and improve model efficiency. The resulting dataset was subsequently used as input for four supervised machine learning classifiers: Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) [9,21–24].

Random Forest (RF) is an ensemble learning algorithm that constructs multiple decision trees from random subsets of samples and predictor variables, with the final classification determined by majority voting [20,24]. Extreme Gradient Boosting (XGB) sequentially constructs decision trees, with each iteration focusing on errors from previous models [21]. Multilayer Perceptron (MLP) is a feedforward artificial neural network trained using back-propagation to model nonlinear relationships among predictor variables [23]. Support Vector Machine (SVM) identifies an optimal decision boundary between classes and uses kernel functions to model nonlinear relationships [22,27].

The Random Forest classifier was configured with 500 trees, with maximum depth optimized during training. XGBoost was parameterized with a learning rate of 0.1 and a maximum depth of 6. The SVM classifier used a radial basis function (RBF) kernel, while the MLP consisted of a single hidden layer with 100 neurons.

3.4. Experimental Design and Accuracy Assessment

The sampling strategy considered the internal spectral and spatial variability of the LULC classes to ensure representative samples across the dataset [9,11]. The dataset was divided into training and validation subsets, with approximately 70% of the samples allocated to model training and 30% reserved for independent validation. The training subset was used for model calibration, whereas the validation subset was retained for evaluating classification performance.

Class imbalance was observed in the dataset, particularly because Forest Formation and Pasture contained substantially more samples than the minority classes. To mitigate this imbalance, a balancing procedure was applied exclusively to the training dataset by reducing the number of samples from the majority classes while preserving the available

samples from the minority classes [30–32]. The validation dataset retained its original class distribution, allowing model performance to be evaluated under the natural class proportions observed in the study area. The resulting sample distribution and balancing procedure are presented in Table 4.

Table 4. Table showing class distribution, balancing strategy, and division between training and validation.

Validation Proportion	Training Proportion	Validation	Training	Class Weight	Final Samples	Balanced Samples	Initial Samples	Class *
0.7	0.3	700	300	0.0079	1	1	6033	Forest Formation
0.7	0.3	700	300	0.0079	1	1	4818	Pasture
0.7	0.3	65	28	0.0853	93	93	93	Agriculture
0.7	0.3	46	20	0.1203	66	66	66	Land Use Mosaic
0.69	0.31	45	20	0.1221	65	65	65	Urban Area
0.71	0.29	39	16	0.1443	55	55	55	Wetlands
0.71	0.29	22	9	0.256	31	31	31	Water
0.71	0.29	22	9	0.256	31	31	31	Silviculture

* The coefficient of variation (CV) across class sample size was reduced from 1.79 (initial distribution) to 0.49 (balanced dataset).

The coefficient of variation (CV) across class sample sizes was reduced from 1.79 in the initial dataset to 0.49 after the balancing procedure, indicating a reduction in disparities among class sample sizes in the training dataset.

Model performance was assessed using overall accuracy (OA), F1-score, and confusion matrix analysis at both the overall and class levels [9]. These metrics were selected to evaluate general classification performance and class-specific discrimination, particularly under the observed class imbalance [32]. In addition, SHAP (Shapley Additive Explanations) was applied to interpret model predictions by quantifying the contribution of individual features to the classification outcomes based on Shapley values [33,34].

A single train–validation split was adopted because of the large number of segmented objects and the computational cost associated with feature extraction and model training. Although this strategy enabled the evaluation of model performance using an independent validation subset while preserving the original class distribution, it may limit the statistical generalization of the results. Therefore, future studies should incorporate k-fold cross-validation and spatially stratified validation schemes to improve robustness and reduce potential sampling bias in heterogeneous restoration landscapes.

4. Results and Discussion

This section presents the results of the analysis of CBERS 4A/WPM images, taking into account spectral bands and LSMM fractions. It evaluates the performance of the classifiers and the importance of each attribute in class discrimination.

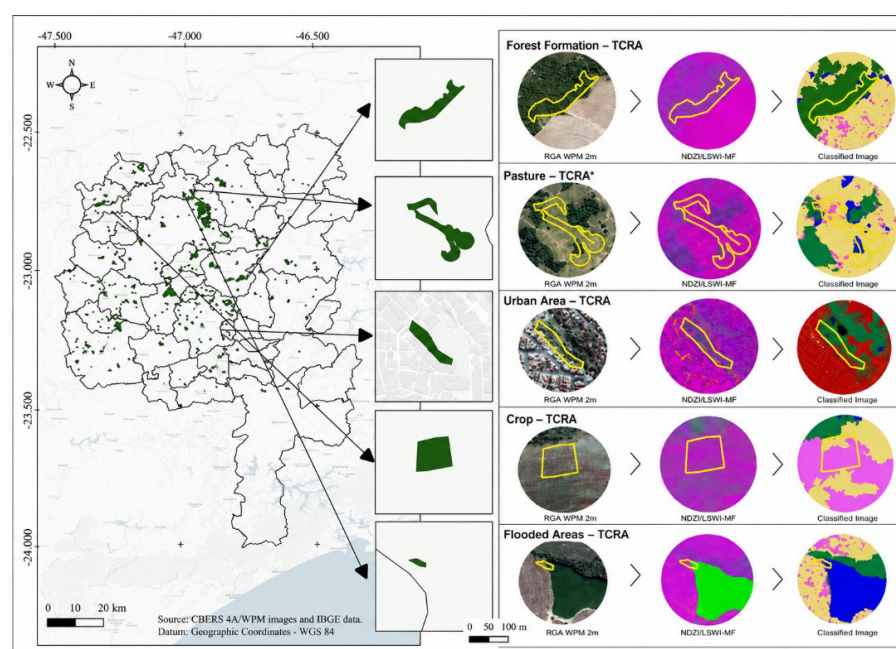
4.1. Identification of TCRAs and Land Use and Land Cover Map

Table 5 presents a summary of the areas (in hectares) of the TCRAs evaluated in this study, including descriptive statistics such as minimum, maximum, mean, standard deviation, and variance. These indicators characterize the spatial heterogeneity of the polygons analyzed, highlighting the coexistence of small-scale TCRAs—which may be more susceptible to edge effects, spectral mixing, and greater influence from the surrounding environment—and larger areas, which may exhibit greater internal homogeneity [4,16]. This variation in spatial scale reinforces the need to adopt methodological approaches capable of capturing patterns across both compact areas and extensive polygons.

Table 5. TCRA Data.

Value	Metric
902	Number of TCRA polygons
0.27 ha	Minimum Area
8.76 ha	Maximum Area
6.02 ha	Mean Area
1.86 ha	Area standard deviation
3.40 ha	Area variance

The land use and land cover map (Figure 3) shows a heterogeneous spatial distribution of TCRAs in the study area, characterized by polygons of varying sizes, shapes, and levels of fragmentation. Larger TCRAs, especially those associated with the Florestal Formation, exhibit more continuous and homogeneous cover, while smaller or more fragmented areas, such as urban areas, croplands, and pastures, display greater spatial complexity, marked by the coexistence of agricultural matrices, bare soil, and vegetation at different stages of succession.

**Figure 3.** Location and classification of land use and land cover.

This structural variability poses challenges for automatic classification, such as increased spectral confusion between classes and the underrepresentation of less frequent categories in classifier training. In scenarios with sample imbalance, these factors can compromise the generalization ability of models [30,31].

In the present study, this effect was mitigated by applying balancing techniques to the training set, reducing the influence of unequal class representation during model training and promoting a more balanced learning process [30–32]. Nevertheless, performance limitations persist for classes with greater internal heterogeneity, indicating that balancing, while essential, does not completely eliminate the effects of spectral complexity.

In this context, the integration of the Linear Spectral Mixture Model (LSMM) with supervised classification provided an additional source of subpixel information for class discrimination [4,15,41]. By estimating the relative contributions of soil, vegetation, and shade within individual pixels, LSMM can improve the representation of spatially heterogeneous landscapes and support the discrimination of land cover conditions affected by

spectral mixing [16,39,40]. This approach is particularly relevant for fragmented restoration landscapes, where multiple land cover components may occur within individual pixels.

4.2. Comparison of the Accuracy Results of the Classification Models Used

The Random Forest (RF) model showed the best overall performance, with an accuracy of approximately 0.66, indicating its suitability for classifying the heterogeneous land use and land cover patterns observed in TCRAs. This performance may be related to the ensemble structure of Random Forest, which combines multiple decision trees generated from different subsets of samples and predictor variables, increasing model robustness and reducing sensitivity to individual decision boundaries [20,24]. The result is also consistent with previous studies in heterogeneous remote sensing environments, in which tree-based ensemble methods have demonstrated strong capability to model nonlinear relationships and complex spectral responses [24,25].

The variation in accuracy as a function of the number of attributes (Top-k) exhibited a nonlinear pattern, with substantial gains during the initial iterations, increasing from approximately 0.40 ($k = 1$) to values above 0.60 at around $k \approx 10$ (Figure 4). This pattern indicates that a relatively small subset of variables contains a substantial proportion of the discriminatory information required for land use and land cover classification in fragmented tropical restoration landscapes, whereas the contribution of additional attributes becomes progressively smaller. Among these variables, LSMM-derived fractions may have contributed to the early increase in accuracy, particularly those related to vegetation abundance and shadow components, which provide information on sub-pixel composition and differences in vegetation cover and illumination conditions [16,39,40].

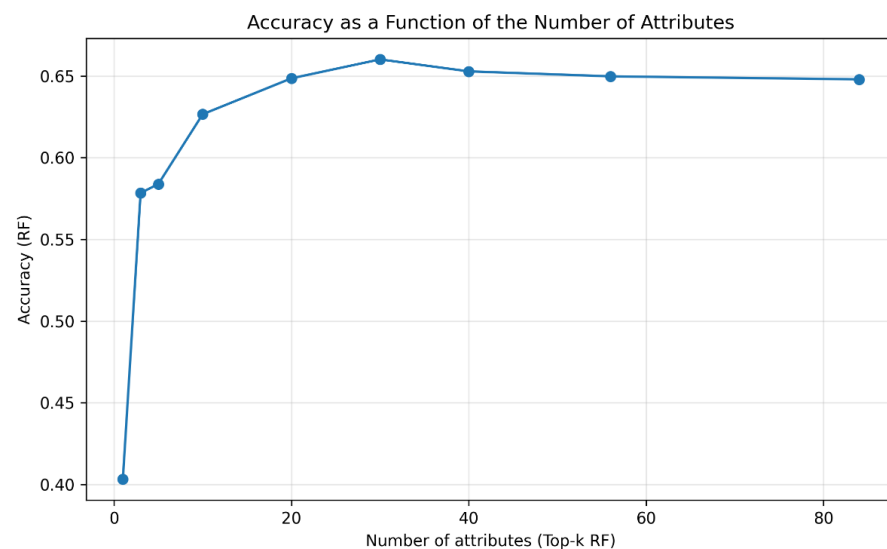


Figure 4. Accuracy Over Time by Number of Attributes.

Maximum performance was reached near $k \approx 30$, suggesting an optimal dimensionality for the dataset. Beyond this point, accuracy stabilized and subsequently decreased slightly, indicating that additional attributes contributed limited discriminatory information. This pattern reflects diminishing returns, whereby increasing the number of predictors may increase computational complexity without providing corresponding improvements in classification performance. In this context, the inclusion of highly correlated or weakly discriminative variables may add noise to the feature space and limit the effectiveness of the classifier, particularly in datasets characterized by substantial spectral variability [8,9].

Although the overall accuracy of approximately 0.66 indicates a moderate level of classification performance, it should be interpreted in relation to the characteristics of

the TCRAs analyzed. Small polygon sizes, fragmented land cover, and the coexistence of different cover components within individual pixels increase classification uncertainty and reduce class separability [4,5,16]. These characteristics are consistent with challenges reported in previous studies of heterogeneous tropical environments, particularly those involving mixed spectral responses, complex landscape patterns, and diffuse transitions between classes [5,8,12]. In addition, the medium spatial resolution of the satellite imagery may intensify mixed-pixel effects in areas undergoing secondary succession and presenting heterogeneous vegetation structures [4,16].

Despite these constraints, the integration of LSMM, GEOBIA, and machine learning provided complementary information for representing the spatial and sub-pixel characteristics of the study areas. LSMM contributed information on the relative composition of individual pixels, GEOBIA incorporated spatial context, and machine learning enabled the integration of these attributes for class discrimination. This combination therefore represents a suitable framework for land cover classification in complex tropical restoration landscapes [6,9,16,39,40].

The variable-importance analysis (Figure 5) further supports the contribution of derived attributes to classification performance. Variables associated with the Linear Spectral Mixture Model (LSMM) were among the most relevant predictors, with the SHAP analysis highlighting the minimum vegetation fraction (LSMM_vegetation_min) as particularly influential. This result may be explained by the relatively persistent vegetation signal in forested areas, which can facilitate their discrimination from pasture and agricultural mosaic classes. By estimating the relative contribution of vegetation and other spectrally distinct components within individual pixels, LSMM provides information that complements conventional spectral bands [4,16,39,40]. Previous studies have likewise demonstrated the usefulness of spectral mixture analysis for characterizing vegetation conditions and discriminating spectrally heterogeneous tropical environments [5,16,17].

Variables related to vegetation and shadow fractions were particularly relevant for distinguishing forest formations, whereas the water class showed a strong association with near-infrared attributes, consistent with its high spectral separability. In contrast, individual spectral variables, including the blue, green, and NIR bands, showed moderate importance when considered separately. These results indicate that combining original spectral information with derived attributes can provide a more comprehensive representation of land cover characteristics, particularly in heterogeneous landscapes [7,8,38,40].

Classes such as land-use mosaic, silviculture, and wetlands exhibited a more diffuse distribution of variable importance, suggesting that their discrimination depends on multiple attributes rather than on a single dominant predictor. Pasture also showed contributions distributed across several variables, consistent with its intermediate spectral characteristics and broad representation in the dataset. The lower contribution of individual attributes for urban areas and wetlands may similarly reflect the influence of multiple landscape components on their spectral responses.

The relevance of shadow fractions for distinguishing native vegetation formations is particularly noteworthy. Shadow components can capture information related to canopy structure, vertical stratification, and variations in illumination within vegetation canopies, thereby providing an additional dimension for differentiating forest environments from pasture and other anthropogenic land uses [16,39,40]. This result highlights the value of derived attributes for characterizing structural properties that may not be adequately represented by individual spectral bands.

The application of SHAP further improved model interpretability by quantifying the contribution of individual predictors to classification outcomes [33–35]. In addition to identifying the most influential variables at the global level, SHAP provides local explanations

of individual predictions, allowing model behavior to be examined in greater detail [33,34]. This approach was particularly useful for identifying the contribution of LSMM-derived and object-based attributes to the classification results.

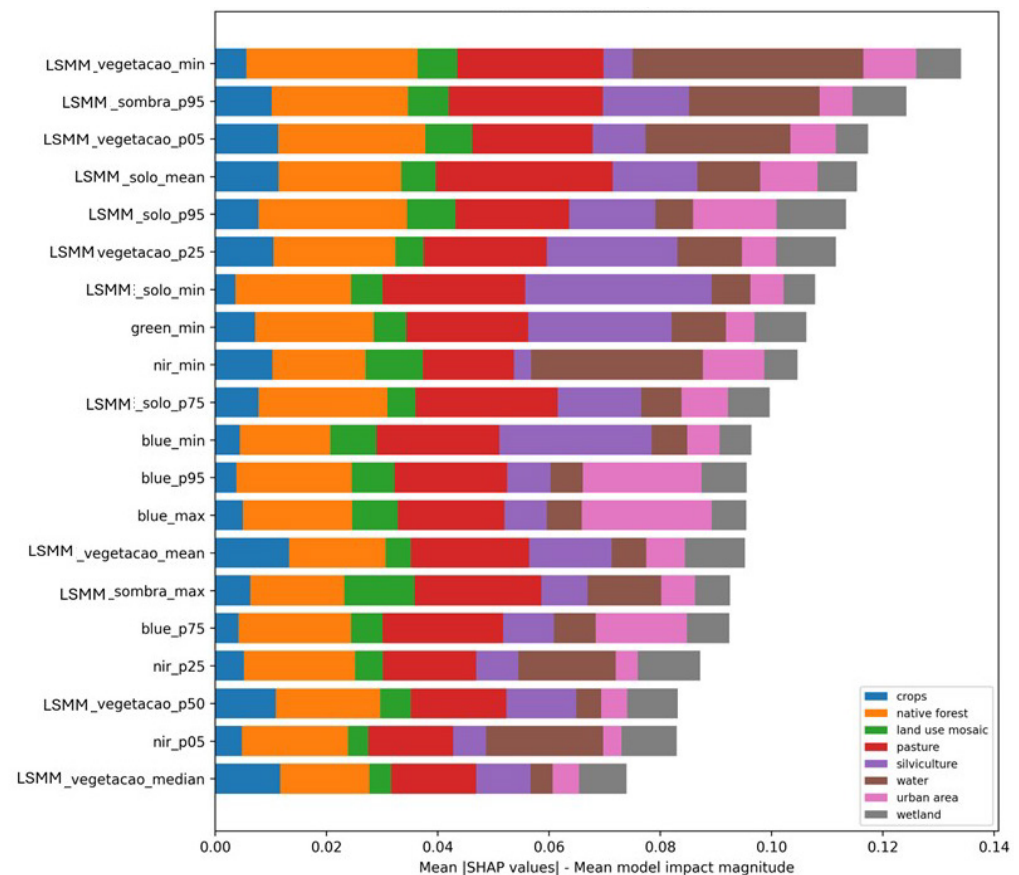


Figure 5. SHAP importance of attributes.

Figure 6 presents the correlation matrix of the 20 most relevant attributes (Top-20), revealing strong associations among statistical descriptors derived from the same variable groups, particularly spectral bands and LSMM-derived fractions. This pattern indicates that some predictors contain overlapping information, which can increase feature dimensionality and computational complexity without necessarily improving classification performance [7,9].

The observed correlations are partly explained by the feature extraction strategy adopted in this study. For each spectral band and LSMM-derived fraction, several object-level statistical descriptors were calculated, including minimum, maximum, mean, standard deviation, and percentiles. Because these descriptors summarize the same underlying spectral distribution, a certain degree of correlation is expected. The potential redundancy becomes more relevant when numerous spectral, fractional, and object-based attributes are simultaneously incorporated into machine learning models [8,9].

Nevertheless, the correlation between spectral bands and LSMM-derived fractions does not imply that these variables provide equivalent information. LSMM fractions offer a physically interpretable representation of sub-pixel composition by explicitly estimating the relative contributions of vegetation, soil, and shadow [4,16,39,40]. Their high importance in the SHAP analysis suggests that, despite partial correlations with the original spectral variables, they retain information relevant to class discrimination.

Among the evaluated classifiers, Support Vector Machine (SVM) presented the lowest performance. This result may be related to the high dimensionality of the feature space,

spectral overlap among classes, and the nonlinear variability of the study area. SVM performance depends on class separability, training-sample characteristics, and classifier and kernel configuration [22,27,28]. Under heterogeneous remote sensing conditions, overlapping class distributions can make the identification of effective decision boundaries more challenging [22,27,28].

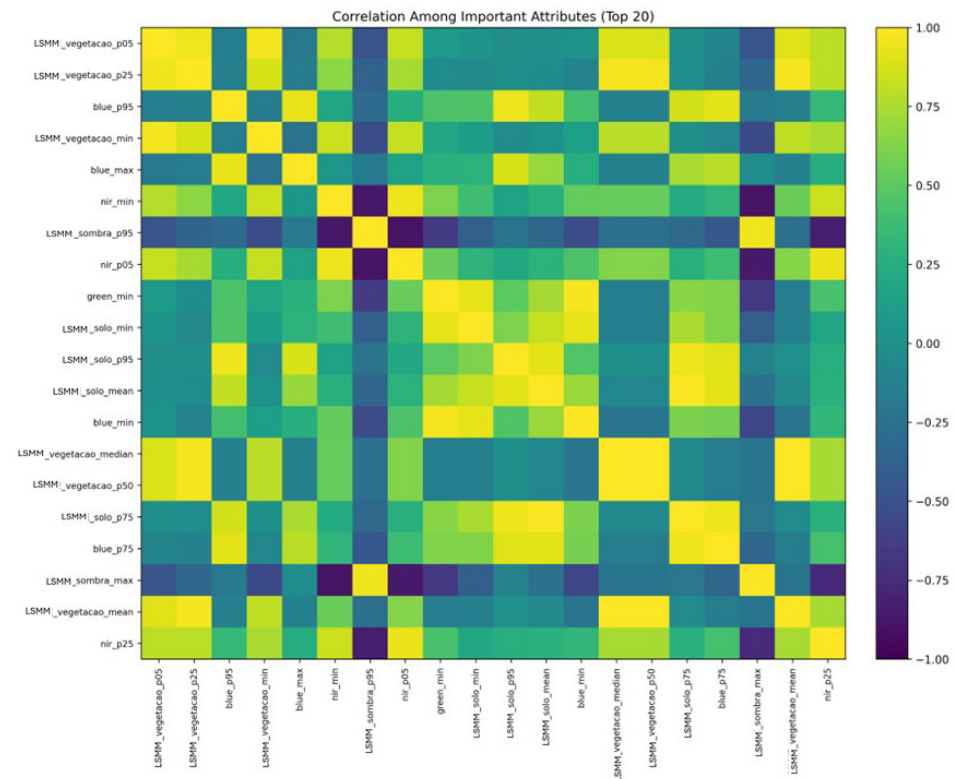


Figure 6. Correlation between attributes (Top 20).

These characteristics are particularly relevant to the TCRA dataset, in which small restoration areas, transitional vegetation, pasture, agricultural areas, and remnants of native vegetation may occur in close spatial proximity. Such spatial arrangements can increase spectral overlap and classification uncertainty [8,9]. Foody and Mathur [28] demonstrated the influence of mixed spectral responses and training-sample characteristics on supervised classification, while Mountrakis et al. [22] emphasized the dependence of SVM performance on feature-space characteristics and classifier configuration.

The lower SVM accuracy observed here should therefore not be interpreted as evidence that SVM is intrinsically unsuitable for remote sensing classification. Rather, it suggests that the specific combination of feature dimensionality, class overlap, mixed spectral responses, and unequal class representation in the TCRA dataset was less favorable to SVM than to the ensemble-based classifiers. This interpretation is consistent with previous studies showing that classifier performance depends on the interaction among the algorithm, feature representation, training data, and landscape characteristics [9,22,27,28].

Overall, the variable-importance and correlation analyses demonstrate that classification performance depends not only on the number of predictors but also on their discriminatory value, complementarity, and degree of redundancy. The results support the integration of original spectral variables with LSMM-derived fractions and object-based attributes, while emphasizing the importance of feature selection to retain informative predictors, reduce redundancy, and improve computational efficiency [7,9,16,38,40].

The LSMM components (vegetation, soil, and shadow) exhibited distinct correlation patterns with one another and with the multispectral bands, reflecting differences in

the physical properties and sub-pixel composition of the analyzed targets. In particular, shadow-related variables showed negative correlations with vegetation and near-infrared attributes, consistent with contrasting responses associated with vegetation structure and illumination conditions [4,16,39,40]. These relationships provide additional evidence that the fractional components capture aspects of pixel composition that are not directly represented by the original spectral bands.

The relationships among spectral variables, including *red_mean* and *green_mean*, also indicate differences in the internal spectral variability of land-cover classes. More homogeneous targets tend to present more consistent responses, whereas classes such as urban areas and wetlands may exhibit greater dispersion because of the coexistence of multiple materials and land-cover components within the same objects [7,8]. This variation supports the use of combined spectral, fractional, and object-based attributes, particularly for representing classes with complex internal composition and lower representation in the dataset [6–9].

Class imbalance may have further influenced classifier performance, particularly for less represented categories. Imbalanced training datasets can favor majority classes and reduce the ability to correctly identify minority categories, especially when class distributions overlap in high-dimensional feature spaces [30–32]. Although balancing was applied to the training dataset in this study, differences in class-level performance indicate that this procedure alone was insufficient to fully compensate for variation in class representation and separability.

Figure 7 shows that the tree-based ensemble models, particularly Random Forest (RF) and XGBoost (XGB), achieved stronger class discrimination, as indicated by the greater concentration of correct predictions along the main diagonal of the confusion matrices. In contrast, SVM presented a more dispersed pattern of errors among classes. The performance of RF and XGB is consistent with the ability of tree-based ensemble approaches to capture nonlinear relationships and interactions among multiple predictors, which can be advantageous when integrating spectral, fractional, and object-based information [20,21,24].

Overall, the classification accuracies were approximately 0.66 for RF, 0.63 for XGB, and 0.62 for MLP, indicating moderate performance across the evaluated models. Rather than reflecting a limitation of a single predictor or classifier, these results highlight the difficulty of simultaneously discriminating multiple land-cover classes with different levels of spectral variability, spatial complexity, and representation in the reference dataset [5,9,16]. The integration of multispectral information, LSMM-derived fractions, and object-based attributes therefore provides a useful framework for mapping complex restoration landscapes, while further gains may be obtained through more extensive spatial validation, expanded and better-balanced reference datasets, and the incorporation of complementary temporal or ancillary information.

The forest formation class showed comparatively high classification performance, although some confusion with pasture was observed, indicating partial spectral similarity between these vegetated classes [5,16]. Land-use mosaic and wetlands exhibited greater dispersion of classification errors, consistent with the greater diversity of land-cover components represented within these classes [8,16]. In contrast, water achieved comparatively high classification performance, likely because of its distinctive spectral response, particularly in the near-infrared region, which increases its separability from most terrestrial classes.

Despite the balancing of the training dataset, differences in class representation and variability remained relevant to model performance. In addition, the use of a single train-validation partition limits the assessment of performance variability and may reduce the statistical robustness and generalizability of the reported estimates. Cross-validation and

spatially stratified validation are therefore recommended for future studies to provide more robust and reproducible estimates, particularly when assessing classification performance across spatially heterogeneous landscapes. This consideration is especially important for environmental monitoring applications, in which differences in class-level accuracy can influence the detection and interpretation of land-cover changes.

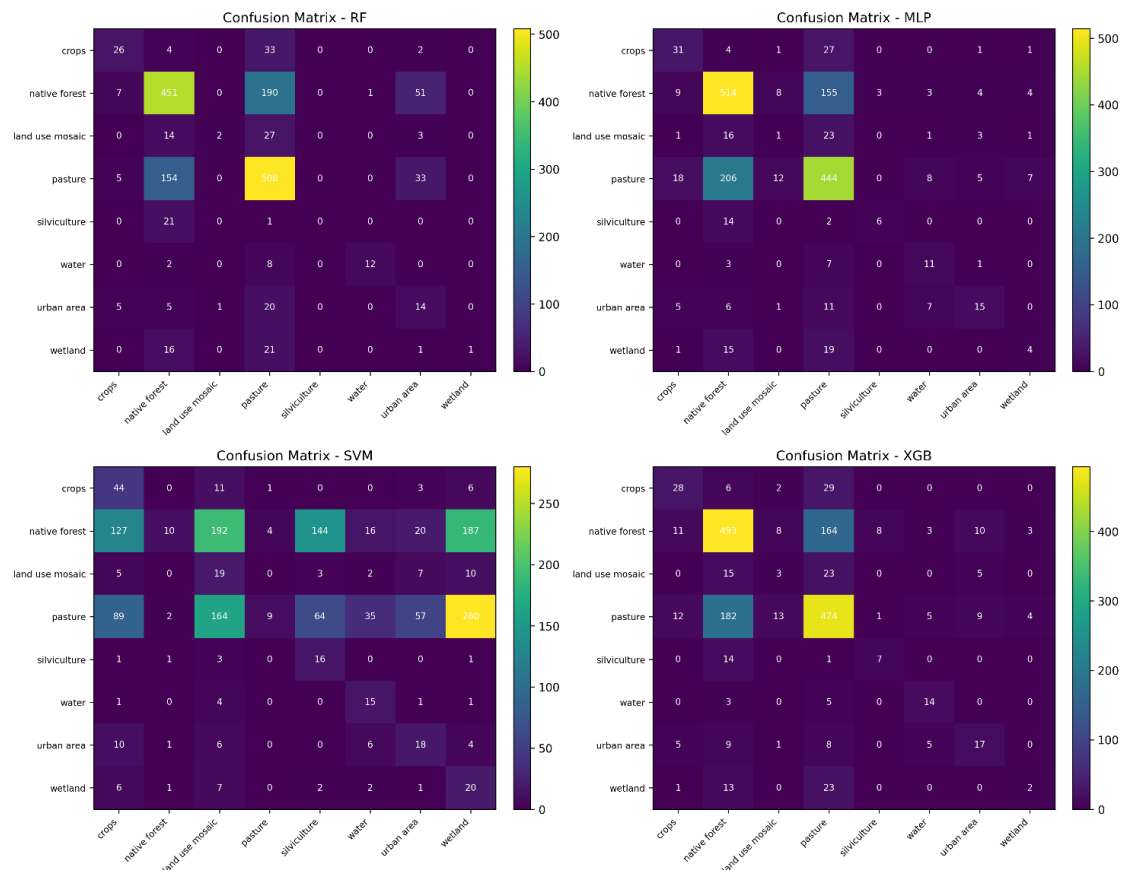


Figure 7. Confusion matrix for the MLP, RF, SVM, and XGB models.

Class imbalance was another relevant factor, with Forest Formation and Pasture substantially more represented than minority classes such as Water and Silviculture. Imbalanced datasets may bias machine learning models toward majority classes, reducing the identification of less represented categories and affecting class-specific performance [30–32]. Although the balancing strategy adopted in this study improved the distribution of samples in the training dataset, reducing the number of majority-class samples may also discard potentially informative observations. Future studies could therefore evaluate alternative approaches, such as class weighting and synthetic sampling, in combination with spatially stratified validation to assess their effects on model robustness and generalizability [30–32].

Figure 8 shows that Random Forest (RF) and XGBoost (XGB) achieved the highest overall accuracies, approximately 0.66 and 0.63, respectively, followed by Multilayer Perceptron (MLP), with approximately 0.62. Support Vector Machine (SVM) showed substantially lower performance, with an overall accuracy of approximately 0.10. The stronger performance of RF and XGB is consistent with the ability of tree-based ensemble methods to represent nonlinear relationships and interactions among high-dimensional predictors in heterogeneous remote sensing datasets [9,20,21,24].

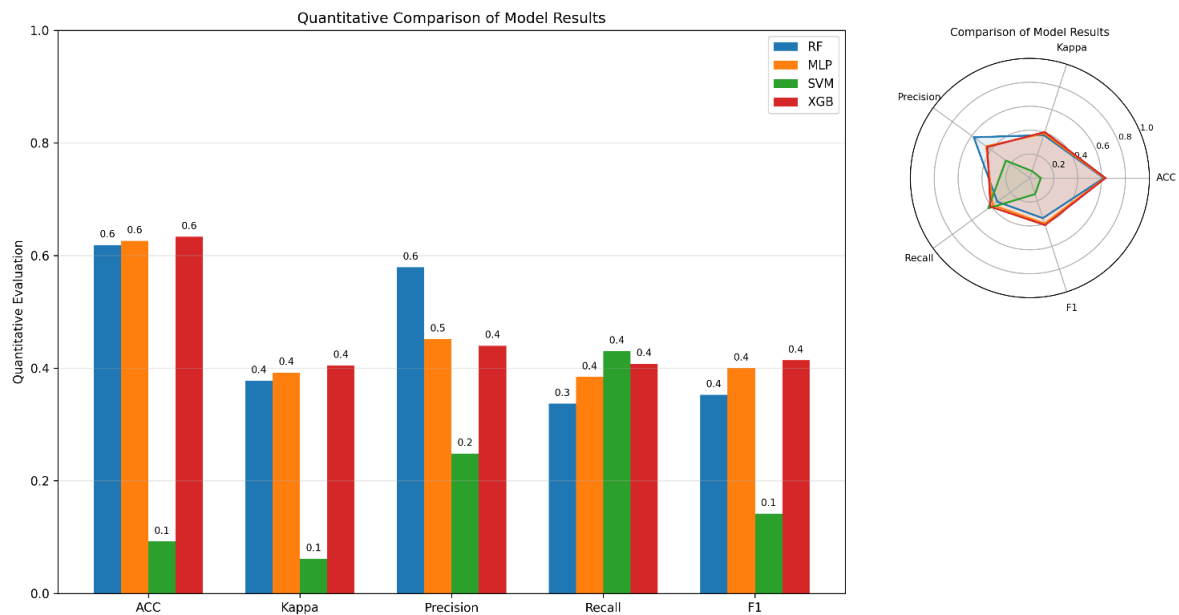


Figure 8. Quantitative Comparison of the Models.

The precision results showed a different pattern across the classifiers. RF achieved the highest precision (approximately 0.58), followed by XGB and MLP, whereas SVM presented a substantially lower value (approximately 0.24). Recall values were more similar among the models, ranging from approximately 0.33 for RF to 0.43 for SVM. Consequently, the F1 scores were approximately 0.38 for RF, 0.40 for MLP, 0.14 for SVM, and 0.41 for XGB, indicating a marked imbalance between precision and recall for SVM. Kappa values were approximately 0.38–0.41 for RF, MLP, and XGB, compared with approximately 0.06 for SVM. Taken together, these metrics indicate that RF and XGB provided a more balanced combination of classification performance, whereas the low precision and F1 score of SVM reveal greater difficulty in producing reliable class assignments despite its comparatively higher recall [22,27,28].

Regarding class balance, XGBoost showed the most consistent overall class-level performance, with an F1 score of approximately 0.41 and Recall of approximately 0.40, followed by MLP, whereas RF presented a higher Precision (approximately 0.58). Although RF achieved the highest accuracy during the Top-k feature-selection procedure (approximately 0.66), the final comparative evaluation showed slightly higher overall accuracy for XGBoost and MLP (approximately 0.63) than for RF (approximately 0.62). This difference indicates that overall accuracy alone may not fully represent classifier performance when class frequencies are unequal, as majority classes can have a greater influence on aggregate metrics [30–32]. In contrast, SVM exhibited substantially lower overall performance, with an F1 score of approximately 0.14 and Kappa of approximately 0.06, indicating greater difficulty in producing reliable class assignments across the evaluated categories [22,27,28].

The comparison of complementary metrics further highlights differences among the classifiers. RF showed the highest Precision (approximately 0.58), whereas XGB achieved the highest Kappa (approximately 0.41) and F1 (approximately 0.41), with MLP showing similar values (Kappa $\approx$ 0.39; F1 $\approx$ 0.40). SVM presented a comparatively high Recall (approximately 0.43), but this was not accompanied by comparable Precision (approximately 0.24) or F1 performance, indicating a less balanced classification outcome. These results emphasize the importance of considering multiple evaluation metrics when comparing classifiers under imbalanced conditions, since overall accuracy may not adequately reflect differences in class-specific recognition [30–32]. Thus, the final comparative

results do not indicate a single model as superior across all metrics; rather, XGBoost provided the most balanced overall performance, while RF showed the highest Precision and the strongest result during the Top-k feature-selection procedure [20,21,24].

Thus, although XGBoost exhibited greater balance across class-level performance metrics, it provided the strongest overall performance in the final comparative evaluation. Random Forest, however, achieved the highest Precision and the strongest result during the Top-k feature-selection procedure. These results highlight the importance of considering both overall and class-specific metrics when selecting a classifier for complex restoration landscapes [20,21,24].

The ROC curves shown in Figure 9 reveal differences in the discriminative capacity of the XGBoost model across the mapped classes. The highest AUC values were obtained for Water (0.99), Silviculture (0.93), and Crops (0.90), indicating strong discrimination for these classes. The high AUC for Water is consistent with its distinctive spectral response, particularly its low reflectance in the near-infrared (NIR) region, which facilitates its separation from most terrestrial land-cover classes. The high AUC for Silviculture may be associated with the relatively consistent spectral and structural characteristics of planted vegetation, whereas the high AUC for Crops indicates effective discrimination of this class within the integrated feature space.

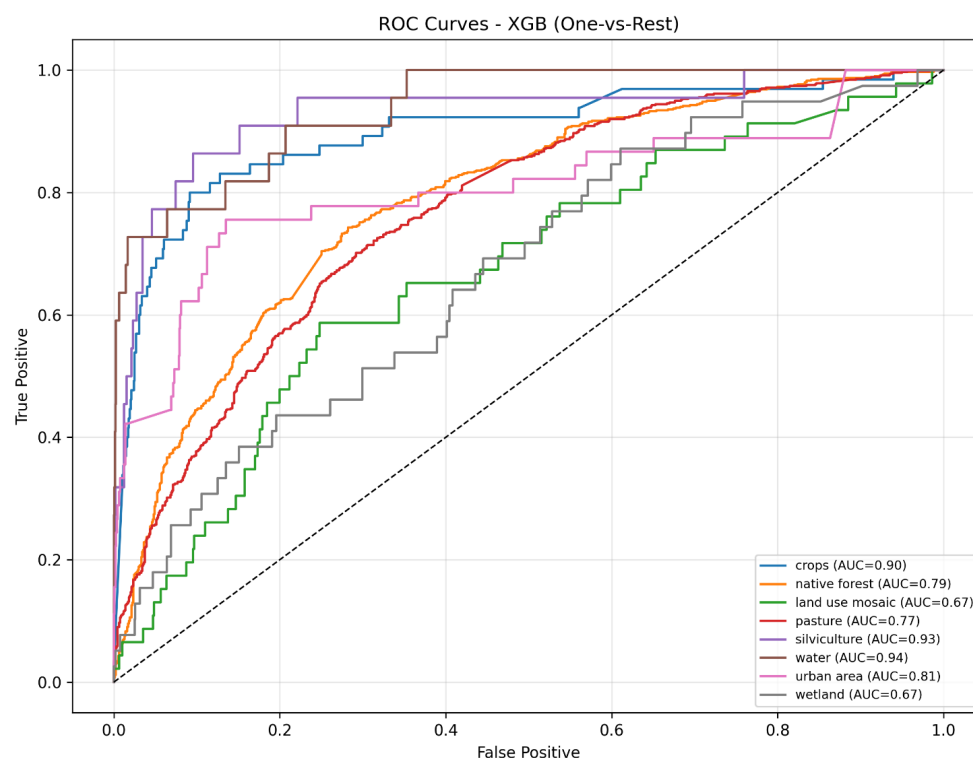


Figure 9. ROC Curves (One-vs-Rest).

Intermediate AUC values were observed for Urban area (0.81), Native forest (0.79), and Pasture (0.77), whereas Land use mosaic and Wetland showed lower values (0.67 for both classes). These differences indicate that the XGBoost model was more effective in distinguishing classes with comparatively distinctive spectral or structural characteristics, while classes represented by more diverse or overlapping land-cover conditions showed lower discriminative capacity. Overall, the ROC analysis complements the classification metrics by showing that model performance varied substantially among classes, with Water, Silviculture, and Crops being the most readily discriminated categories and Land use mosaic and Wetland presenting the greatest discrimination challenges.

The Urban area (AUC = 0.81), Forest Formation (0.79), and Pasture (0.77) classes showed intermediate discriminative performance, indicating greater overlap with other categories in the feature space. Forest Formation and Pasture may be particularly challenging to separate because of similarities in vegetation abundance, canopy conditions, and illumination effects, despite differences detectable in the red, green, and near-infrared (NIR) regions [16,39,40]. The performance of Urban area may likewise reflect the diversity of materials and surface conditions represented within this class [7,8].

The lowest AUC values were obtained for Land-use mosaic and Wetland (0.67 for both classes), indicating greater overlap with other categories. Their lower discriminative performance may be related to the diversity of land-cover components and transitional conditions represented within these classes [5,8,16]. Seasonal changes in vegetation condition, soil exposure, and water availability may further modify their spectral responses, increasing temporal variability [12,15]. These results indicate that Land-use mosaic and Wetland require greater caution when used in applications involving automated classification and change detection.

The differences in class-level discriminative performance have direct implications for environmental monitoring. Water and Silviculture, which showed the highest AUC values, can be mapped with greater confidence, whereas Forest Formation and Pasture require greater attention in multitemporal analyses. Land-use mosaic and Wetland represent the most challenging classes and may benefit from complementary temporal or ancillary information to improve classification reliability [12,15].

The integration of GEOBIA attributes with LSMM-derived fractions provided complementary spatial and sub-pixel information for the classification. LSMM characterized the relative abundance of vegetation, soil, and shadow components, whereas GEOBIA incorporated object-level spatial, textural, and contextual properties [4,6,16,39,40]. This combination is particularly relevant to TCRA environments, where restoration areas may encompass different vegetation stages and land-cover configurations within relatively small spatial units [1,3,17,18]. Object-based approaches can also reduce some limitations of pixel-level classification by incorporating spatial context into the analysis [6].

The class-specific SHAP analysis (Figure 10) revealed differences in the contribution of predictors among land-cover categories. For Forest Formation, LSMM-derived fractions, particularly vegetation and shadow, showed strong influence. Vegetation fractions represent the relative abundance of vegetated components within each pixel, whereas shadow fractions provide information related to canopy structure and illumination conditions [16,39,40]. Their contribution may therefore facilitate the separation of Forest Formation from Pasture and other classes with less complex vegetation structure [17,18].

In pasture areas, the combined contribution of soil and vegetation fractions reflects differences in vegetation abundance, soil exposure, and seasonal variations in herbaceous cover [12,16,17]. The land-use mosaic class shows a diffuse distribution of SHAP values, indicating strong spectral mixing associated with heterogeneous land-cover composition [8,16,18]. In contrast, the water class exhibits a well-defined pattern dominated by near-infrared attributes, consistent with the low reflectance of water in this spectral region and its high separability from most terrestrial classes. Silviculture shows a greater concentration of vegetation-related attributes, suggesting greater structural and spectral consistency than more heterogeneous vegetated classes.

The urban and wetland classes display greater dispersion in SHAP values, with no clear dominance of individual attributes, reflecting the coexistence of multiple spectral components within these classes [8,11]. Overall, the observed patterns highlight the influence of biophysical properties and spectral mixing on class-specific model responses [16,17,40].

Figure 11 presents the SHAP dependence plots for LSMM_vegetacao_min and LSMM_solo_p95, highlighting the influence of these variables on model predictions. LSMM_vegetacao_min shows an approximately monotonic increase in SHAP contribution, particularly at higher values, indicating that its effect varies across the feature space and that nonlinear relationships are captured by the tree-based model. Through recursive partitioning of the predictor space, tree-based ensemble methods can represent such complex relationships, which is particularly relevant for heterogeneous remote sensing datasets [21,25,26]. The non-uniform contribution of LSMM_vegetacao_min across its range further suggests that specific vegetation-fraction intervals may reflect differences in the sub-pixel composition of the mapped targets [4,16,39,40].

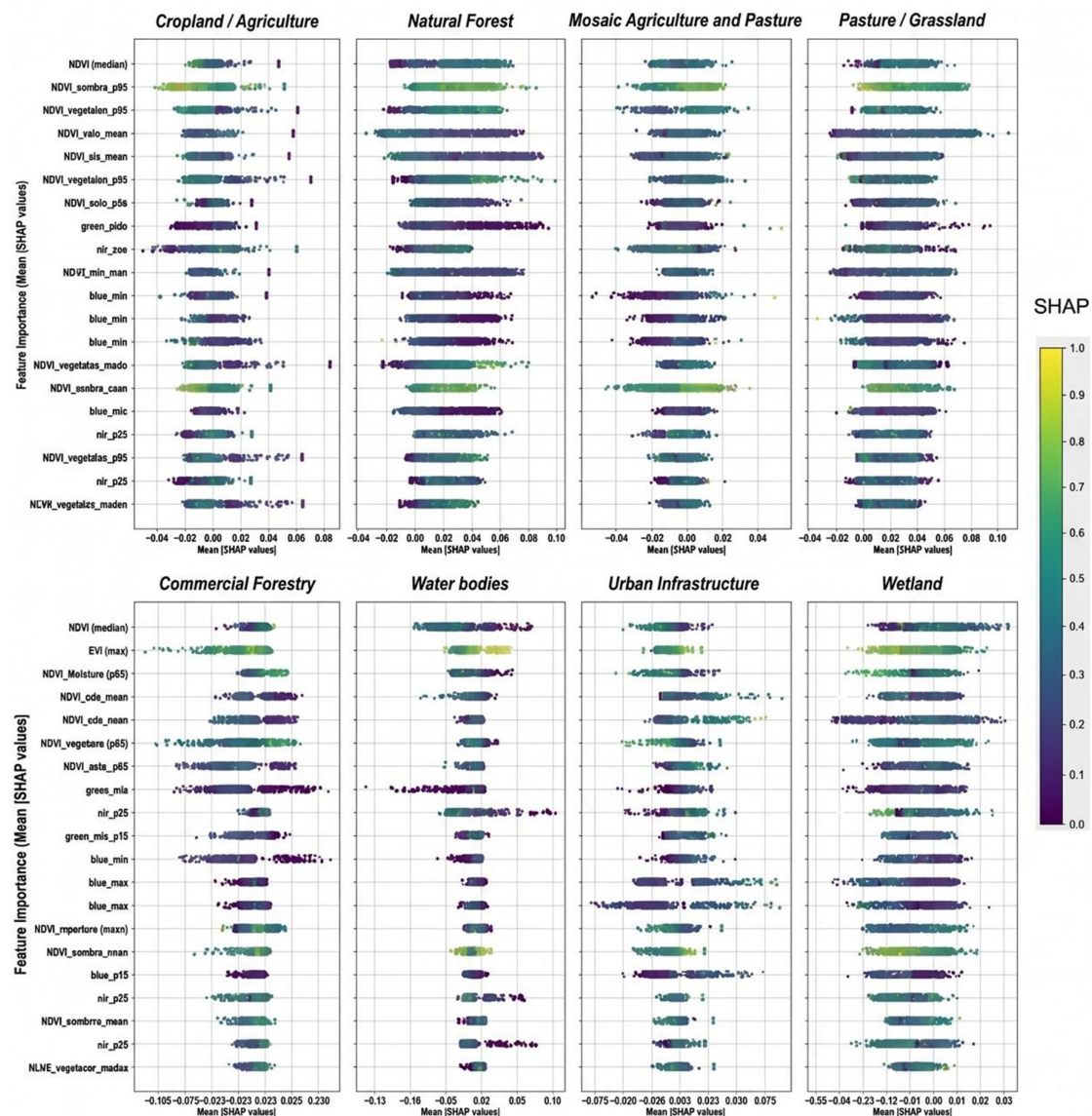


Figure 10. SHAP by Class.

On the other hand, the LSMM_solo_p95 attribute shows an inverse and nonlinear relationship with the model response, with lower values associated with higher positive SHAP contributions and increasing values leading to a progressive reduction in predictive influence. This pattern indicates varying sensitivity across the feature space, with stronger effects within specific value ranges and greater stabilization or dispersion at extreme values. Such behavior is consistent with the nonlinear structure of tree-based models, which represent predictor effects through recursive partitioning of the feature

space [20,21,25]. The influence of LSMM-derived soil fractions may also reflect variations in sub-pixel composition, particularly in heterogeneous landscapes where soil exposure and vegetation cover vary spatially [4,16,39–41]. Overall, the observed relationship suggests that the contribution of soil-related attributes is conditioned by interactions among biophysical properties, sub-pixel composition, and spectral variables rather than by a simple linear response [16,40,41].

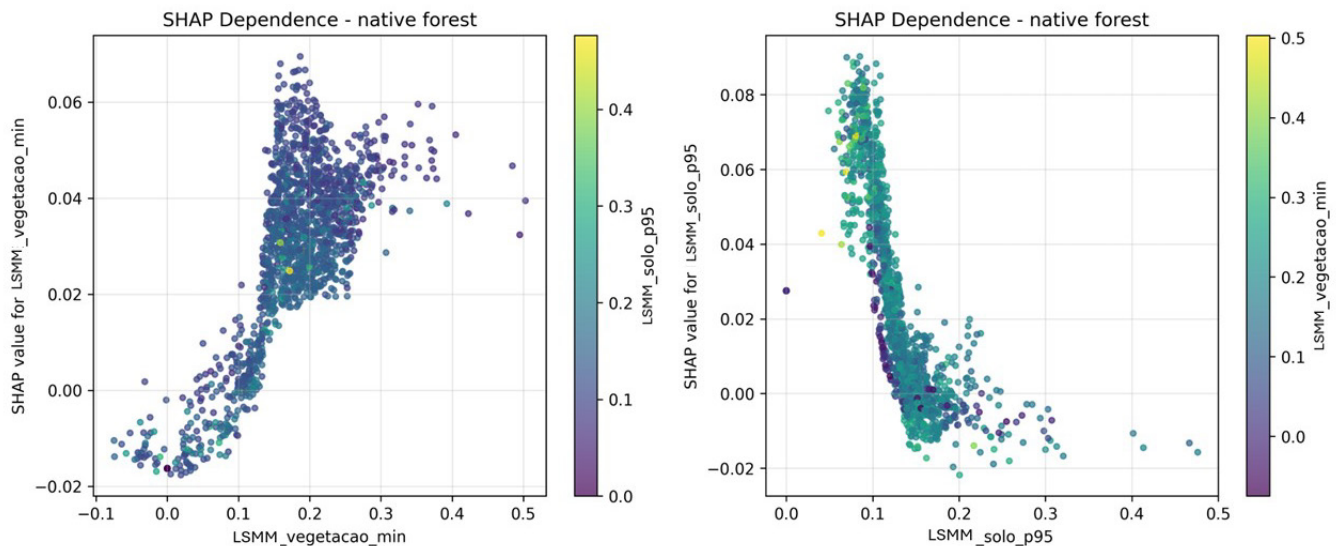


Figure 11. SHAP Dependency Plots.

5. Conclusions

This study evaluated the integration of Linear Spectral Mixture Model (LSMM) fractions, textural attributes, Geographic Object-Based Image Analysis (GEOBIA), and machine learning algorithms for land use and land cover (LULC) classification in Environmental Restoration Commitment Agreements (TCRAs) in the state of São Paulo, Brazil. Among the evaluated classifiers, Random Forest achieved the best overall performance (0.66), followed by XGBoost (0.63) and Multilayer Perceptron (0.62), whereas Support Vector Machine showed substantially lower performance (0.10). Water bodies and forest formation exhibited greater class separability, while more heterogeneous classes showed greater classification uncertainty. These results indicate that combining spectral, sub-pixel, textural, and object-based information can improve the representation of heterogeneous and fragmented restoration landscapes.

Classification performance was affected by spectral and structural variability, class imbalance, and landscape complexity. LSMM fractions provided information on sub-pixel composition, while GEOBIA incorporated spatial and textural characteristics, complementing the spectral information used by the classifiers. Nevertheless, the use of a single training–validation split and medium-resolution imagery limits the statistical robustness and generalization of the results and should be considered when interpreting the reported accuracies.

From an applied perspective, the proposed framework shows potential for supporting LULC monitoring in TCRAs using freely available CBERS-4A/WPM imagery. Its combination of complementary image attributes and machine learning offers a cost-effective approach for characterizing fragmented restoration areas and supporting environmental monitoring and decision-making. Future research should assess spatially stratified cross-validation, time-series imagery, higher spatial resolution data, and alternative machine

learning and deep learning methods to enhance the robustness and generalizability of LULC classification in heterogeneous restoration landscapes.

Author Contributions: Conceptualization, B.S.; Methodology, C.R.S. and B.S.; Software, C.R.S. and F.B.J.R.D.; Validation, C.R.S. and B.S.; Formal analysis, C.R.S.; Writing—original draft, C.R.S.; Writing—review & editing, C.R.S., F.S. and F.B.J.R.D.; Visualization, C.R.S.; Supervision, B.S. and F.B.J.R.D.; Project administration, B.S.; Funding acquisition, B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FAPESP (São Paulo Research Foundation, Brazil), grant number 2024/19134-1 (Innovative Research in Small Business).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: We thank Visiona Tecnologia Espacial S.A. for the technical support provided throughout this study. We also thank the technical recommendations given by Companhia Ambiental do Estado de São Paulo (CETESB). Finally, we thank the financial support of FAPESP.

Conflicts of Interest: Author Francisco Salazar was employed by the company Visiona Tecnologia Espacial S.A. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TCRA	Environmental Restoration Commitment Agreement
CV	Coefficient of variation
GEOBIA	Geographic Object-Based Image Analysis
LULC	Land use and land cover
LSMM	Linear Spectral Mixture Model
MLP	Multilayer Perceptron
OTB	Orfeo Toolbox
RF	Random Forest
SEMIL	São Paulo State Secretariat for Environment, Infrastructure, and Logistics
SHAP	Shapley Additive Explanations
SVM	Support Vector Machine
XGB	XGBoost

References

- Schweizer, D.; Petter, G.; César, R.G.; Ferraz, S.; Moreno, V.S.; Brancalion, P.H.S.; Bugmann, H. Natural Forest Regrowth under Different Land Use Intensities and Landscape Configurations in the Brazilian Atlantic Forest. *For. Ecol. Manag.* **2022**, *508*, 120012. [[CrossRef](#)]
- Brancalion, P.H.S.; Holl, K.D. Guidance for Successful Tree Planting Initiatives. *J. Appl. Ecol.* **2020**, *57*, 2349–2361. [[CrossRef](#)]
- Rodrigues, R.R.; Gandolfi, S.; Nave, A.G.; Aronson, J.; Barreto, T.E.; Vidal, C.Y.; Brancalion, P.H.S. Large-Scale Ecological Restoration of High-Diversity Tropical Forests in SE Brazil. *For. Ecol. Manag.* **2011**, *261*, 1605–1613. [[CrossRef](#)]
- Shimabukuro, Y.E.; Smith, J.A. The Least-Squares Mixing Models to Generate Fraction Images Derived from Remote Sensing Multispectral Data. *IEEE Trans. Geosci. Remote Sens.* **1991**, *29*, 16–20. [[CrossRef](#)]
- Roberts, D.A.; Gardner, M.; Church, R.; Ustin, S.; Scheer, G.; Green, R.O. Mapping Chaparral in the Santa Monica Mountains Using Multiple Endmember Spectral Mixture Models. *Remote Sens. Environ.* **1998**, *65*, 267–279. [[CrossRef](#)]
- Ma, L.; Li, M.; Ma, X.; Cheng, L.; Du, P.; Liu, Y. A Review of Supervised Object-Based Land-Cover Image Classification. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 277–293. [[CrossRef](#)]

7. Lu, D.; Weng, Q. A Survey of Image Classification Methods and Techniques for Improving Classification Performance. *Int. J. Remote Sens.* **2007**, *28*, 823–870. [[CrossRef](#)]
8. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Object-Based Image Analysis for Remote Sensing Applications: Status, Challenges, and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 249–268. [[CrossRef](#)]
9. Maxwell, A.E.; Warner, T.A.; Fang, F. Implementation of Machine-Learning Classification in Remote Sensing: An Applied Review. *Int. J. Remote Sens.* **2018**, *39*, 2784–2817. [[CrossRef](#)]
10. Ghorbanzadeh, O.; Blaschke, T.; Kerle, N.; Tiede, D. Tree-Based Ensemble Learning Methods and Class Imbalance Issues in Remote Sensing Image Classification. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 257–271. [[CrossRef](#)]
11. Sovann, C.; Olin, S.; Mansourian, A.; Sakhoen, S.; Prey, S.; Kok, S.; Tagesson, T. Importance of Spectral Information, Seasonality, and Topography on Land Cover Classification of Tropical Land Cover Mapping. *Remote Sens.* **2025**, *17*, 1551. [[CrossRef](#)]
12. Oliveira, J.S.; Lima, E.R.V.; Souza, B.I.; Seabra, V.S.; Santos, M.P. Sentinel-2/MSI Images and GEOBIA Methodology in Mapping Land Use and Land Cover in a Seasonally Dry Tropical Forest Area in Brazil. *Ra'e GA—O Espaço. Geogr. Anál.* **2024**, *59*, 3–22. [[CrossRef](#)]
13. Lasko, K.; O'Neill, F.D.; Sava, E. Automated Mapping of Land Cover Type within International Heterogenous Landscapes Using Sentinel-2 Imagery with Ancillary Geospatial Data. *Sensors* **2024**, *24*, 1587. [[CrossRef](#)] [[PubMed](#)]
14. Subedi, M.R.; Portillo-Quintero, C.; McIntyre, N.E.; Kahl, S.S.; Cox, R.D.; Perry, G.; Song, X. Ensemble Machine Learning on the Fusion of Sentinel Time Series Imagery with High-Resolution Orthoimagery for Improved Land Use/Land Cover Mapping. *Remote Sens.* **2024**, *16*, 2778. [[CrossRef](#)]
15. Ponzoni, F.J.; Shimabukuro, Y.E. Spectral Mixture Analysis of Landsat TM Images: A Tool for Vegetation Studies. *Int. J. Remote Sens.* **2007**, *28*, 685–698.
16. Hasan, A.F.; Laurent, F.; Messner, F.; Bourgoïn, C.; Blanc, L. Cumulative Disturbances to Assess Forest Degradation Using Spectral Unmixing in the North-Eastern Amazon. *Appl. Veg. Sci.* **2019**, *22*, 394–408. [[CrossRef](#)]
17. Cao, S.; Yu, Q.; Sánchez-Azofeifa, G.A.; Feng, J.; Rivard, B.; Gu, Z. Mapping Tropical Dry Forest Succession Using Multiple Criteria Spectral Mixture Analysis. *ISPRS J. Photogramm. Remote Sens.* **2015**, *109*, 17–29. [[CrossRef](#)]
18. Carvalho, W.S.; Ávila, M.E.B.; Felix, A.B.A.; Tognon, A.; Grau, P.F.G.; Nogueira, R.H.; Guilherme, D.O. Classificação Baseada em Objeto (GEOBIA) Utilizando Imagem de Satélite CBERS 4A WPM em uma Área Urbana. *Rev. Políticas Públicas Cid.* **2024**, *13*, e1051. [[CrossRef](#)]
19. Magalhães, D.M.; Lucas, M.M.P.; França, E.A.G. Pseudo-Labeling Strategy and U-Net for High-Resolution LULC Mapping Using CBERS-04A Imagery in the Servidão River Basin, Brazil. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2026**, *11*, 61–68. [[CrossRef](#)]
20. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
21. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [[CrossRef](#)]
22. Mountrakis, G.; Im, J.; Ogole, C. Support Vector Machines in Remote Sensing: A Review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [[CrossRef](#)]
23. Haykin, S. *Neural Networks and Learning Machines*, 3rd ed.; Pearson Education: Upper Saddle River, NJ, USA, 2009.
24. Belgiu, M.; Drăguț, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [[CrossRef](#)]
25. Tan, Y.-C.; Duarte, L.; Teodoro, A.C. Comparative Study of Random Forest and Support Vector Machine for Land Cover Classification and Post-Wildfire Change Detection. *Land* **2024**, *13*, 1878. [[CrossRef](#)]
26. Kupidura, P.; Kępa, A.; Krawczyk, P. Comparative Analysis of the Performance of Selected Machine Learning Algorithms Depending on the Size of the Training Sample. *Rep. Geod. Geoinform.* **2024**, *118*, 53–69. [[CrossRef](#)]
27. Huang, C.; Davis, L.S.; Townshend, J.R.G. An Assessment of Support Vector Machines for Land Cover Classification. *Int. J. Remote Sens.* **2002**, *23*, 725–749. [[CrossRef](#)]
28. Foody, G.M.; Mathur, A. The Use of Small Training Sets Containing Mixed Pixels for Accurate Hard Image Classification: Training on Mixed Spectral Responses for Classification by a SVM. *Remote Sens. Environ.* **2004**, *91*, 126–136. [[CrossRef](#)]
29. Chen, P.; Liu, Y.; Ren, Y.; Zhang, B.; Zhao, Y. A Deep Learning-Based Solution to the Class Imbalance Problem in High-Resolution Land Cover Classification. *Remote Sens.* **2025**, *17*, 1845. [[CrossRef](#)]
30. Deng, Y.; Chen, G.; Tang, B.; Duan, X.; Zuo, L.; Zhao, H. Study on Class Imbalance in Land Use Classification for Soil Erosion in Dry–Hot Valley Regions. *Remote Sens.* **2025**, *17*, 1628. [[CrossRef](#)]
31. Bae, Y.; Ban, Y. Class-Balanced Random Patch Training to Address Class Imbalance in Tiling-Based Farmland Classification. *Appl. Sci.* **2025**, *15*, 7056. [[CrossRef](#)]
32. Farhadpour, S.; Warner, T.A.; Maxwell, A.E. Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices. *Remote Sens.* **2024**, *16*, 533. [[CrossRef](#)]
33. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Long Beach, CA, USA, 2017.

34. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [[CrossRef](#)] [[PubMed](#)]
35. Temenos, A.; Temenos, N.; Kaselimi, M.; Doulamis, A.; Doulamis, N. Interpretable Deep Learning Framework for Land Use and Land Cover Classification in Remote Sensing Using SHAP. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 8500105. [[CrossRef](#)]
36. Secretaria de Meio Ambiente, Infraestrutura e Logística do Estado de São Paulo (SEMIL). *Painel Verde SP: Dados de Termos de Compromisso de Recuperação Ambiental (TCRA)*; Secretaria de Meio Ambiente, Infraestrutura e Logística do Estado de São Paulo: São Paulo, Brazil, 2026. Available online: <https://mapas.infraestruturameioambiente.sp.gov.br/portal/apps/experiencebuilder/experience/?id=ede801f60edc4586a8dcf40e10b7dde0/> (accessed on 1 October 2025).
37. MapBiomass. MapBiomass—Annual Land Cover and Land Use Mapping of Brazil. São Paulo, Brazil. Available online: https://brasil.mapbiomas.org/?utm_source=chatgpt.com (accessed on 1 October 2025).
38. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Feitosa, R.Q.; van der Meer, F.; van der Werff, H.; van Coillie, F.; et al. Geographic Object-Based Image Analysis—Towards a New Paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [[CrossRef](#)] [[PubMed](#)]
39. Ruggeri, S.; Henao Céspedes, V.; Garces Gomez, Y.A.; Parra, A. Optimized Unsupervised CORINE Land Cover Mapping Using Linear Spectral Mixture Analysis and Object-Based Image Analysis. *Egypt. J. Remote Sens. Space Sci.* **2021**, *24*, 1061–1069. [[CrossRef](#)]
40. Vasconcelos, C.H.; Novo, E.M.L.M. Mapping Land Use and Cover Based on Segmentation and Classification of Images—Soil, Vegetation and Shade Fractions Derived from the Linear Spectral Mixture Model Applied to Landsat TM Data in the Tucuruí Region, Brazil. *Acta Amaz.* **2004**, *34*, 487–493. [[CrossRef](#)]
41. Roberts, D.A.; Dennison, P.E.; Gardner, M.E.; Hetzel, Y.; Ustin, S.L.; Lee, C.T. Spectral Mixture Analysis: A Review of Algorithms and Applications. *Remote Sens. Environ.* **2017**, *191*, 110–131. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.