Data-centric approach for land use and land cover classification in Brazil

Glauber J. Vaz[1,2], João Francisco G. Antunes[1], Júlio C.D.M. Esquerdo[1], Alexandre C. Coutinho[1], André S. Tavares[1], Adriane Calaboni[1], Lídia S. Bertolo[1], Filipe C. Felix[1], Anderson Rocha[2]

[1] Embrapa Digital Agriculture, Campinas, SP, Brazil – {glauber.vaz, joao.antunes, julio.esquerdo, alex.coutinho}@embrapa.br, {andre.tavares, adriane.calaboni, lidia.bertolo, filipe.felix}@colaborador.embrapa.br
[2] Recod.ai, Institute of Computing, University of Campinas, Campinas, SP, Brazil – arrocha@unicamp.br

Keywords: Data quality, Deep learning, Expert analysis, Land use and land cover mapping, Sentinel-2, Multidimensional features.

Abstract

Land use and land cover (LULC) classification plays a crucial role in addressing numerous real-world challenges. Hence, we proposed methodological advances in LULC classification from a data-centric artificial intelligence perspective, which prioritizes data quality as a key factor in improving machine learning performance. The main contributions include evaluations of novel approaches for: (i) constructing an accurately labeled dataset based on agreement among existing reliable maps; (ii) curating remote sensing data to improve accuracy, consistency, unbiasedness, relevance, diversity, and completeness; (iii) generating training samples that capture the spatial, temporal, and spectral dimensions of remote sensing data; and (iv) developing a deep learning model designed to leverage multidimensional features. The study evaluates a sample generation method grounded in reference map agreement and multidimensional feature extraction, along with a deep learning model that leverages these features, attaining high accuracy across all LULC classes and providing a robust basis for large-scale, data-centric LULC mapping.

1. Introduction

Data-centric artificial intelligence (AI) is a paradigm emphasizing the systematic design and engineering of data as foundational for developing effective and efficient AI-based systems (Jakubik et al., 2024). This approach focuses on robust processes for data collection, labeling, and quality monitoring to enhance dataset quality and, consequently, improve machine learning performance. In contrast to the model-centric perspective, which concentrates efforts on optimizing model architectures and code, the data-centric approach targets improvements to the data itself (Polyzotis and Zaharia, 2021).

Roscher et al. (2024) described the machine learning lifecycle as comprising six stages: (i) problem definition, (ii) data creation, (iii) data curation, (iv) model training, (v) evaluation, and (vi) deployment with feedback. In the data-centric approach, particular emphasis is placed on data creation, data curation, evaluation, and deployment with feedback. This focus enables the training of models that generalize well and perform reliably under unforeseen circumstances. Additional advantages of data-centric systems include reducing training requirements for each dataset, developing robust and fair models, improving long-term efficiency and adaptability to changing conditions, and optimizing model performance. In many real-world applications, improving the quality of input data has proven more effective than experimenting with various algorithms (Kumar et al., 2024).

In the geospatial domain, land use and land cover (LULC) classification is essential for urban and rural planning, agricultural management, disaster prediction, environmental monitoring, and sustainable development. For instance, LULC maps are important for quantifying greenhouse gas emissions resulting from land-use change, particularly in Brazil, where such changes are the main source of national emissions (Zimbres et al., 2024). Nevertheless, several challenges persist, including data diversity, category imbalance, accurate sample labeling, limited model generalization, and contextual modeling to capture spatial correlations (Zhao et al., 2023). Hence, this study sought to address all of these aspects.

MapBiomas (Souza Jr. et al., 2020) and TerraClass (2024) are two prominent initiatives for LULC mapping in Brazil. MapBiomas offers nationwide LULC mapping, whereas TerraClass focuses on the Cerrado and Amazon biomes. Both rely on remote sensing data: MapBiomas accesses satellite imagery via the Google Earth Engine platform, while TerraClass primarily acquires data through the Brazil Data Cube (BDC), which offers large volumes of remote sensing imagery modeled as multidimensional data cubes that cover the entire Brazilian territory (Ferreira et al., 2020; Instituto Nacional de Pesquisas Espaciais, 2025).

Although the datasets from these initiatives are highly reliable and widely used by the scientific community and other stakeholders, LULC mapping remains a task with ongoing opportunities for improvement, such as increased classification accuracy or finer granularity in LULC classes. Deep learning methods offer a promising direction for such advancements. Despite growing research interest in applying deep learning techniques to LULC classification, their use in this domain remains insipient (Digra et al., 2022; Vali et al., 2020), offering further potential for progress in LULC classification and the exploration of practical remote sensing imagery applications (Ma et al., 2019).

However, fully harnessing deep learning’s potential depends on the availability of high-quality datasets. This requires selecting datasets well-suited to the specific problem and processing them to maximize their utility. Therefore, further research is required in data creation and curation to produce higher-quality datasets, as well as in developing more robust data-centric methods capable of addressing real-world challenges (Roscher et al., 2024). Developing novel approaches to fully exploit existing LULC maps and obtaining high-quality sample data for applying deep learning techniques to LULC mapping remain key challenges in the context of geospatial big data (Zhang and Li, 2022).

Roscher et al. (2024) identified five criteria of particular relevance to the geospatial domain: diversity and completeness, accuracy, consistency, relevance, and unbiasedness. Biases are inherent in machine learning algorithms and may occur, for example, when a model underfits. They also encompass systematic errors in algorithms or datasets, resulting in varying outcomes for specific subgroups within a class (Ghamisi et al., 2024). Suresh and Guttag (2021) proposed a framework for identifying various types of bias, and Ghamisi et al. (2024) analyzed them in Earth Observation applications.

In the data-centric AI approach, both data and models are iteratively evaluated and refined through three typical steps: error analysis to identify sources of mistakes based on data fit and consistency; the application of these findings to implement improvements; and systematic assessment of data quality. In this iterative process, involving increasingly complex, multidimensional, and problem-dependent data, the ongoing involvement of domain experts is essential. Their contributions to data collection, preparation, and result interpretation ensure that humans remain integral to the process (Jarrahi et al., 2023).

We aimed to present methodological advances in LULC classification by leveraging deep learning from a data-centric machine learning perspective. Additionally, we evaluated these advances according to the cited criteria, including different types of bias, while involving domain experts.

More specifically, this study’s main contributions include evaluations of novel approaches for:
a) creating an accurately labeled dataset for LULC classification based on agreement among existing reliable maps;
b) curating remote sensing data, with expert feedback, to enhance data quality;
c) generating training samples that exploit the spatial, temporal, and spectral dimensions of remote sensing data; and
d) developing a deep learning model that leverages these multidimensional features to achieve high accuracy in LULC classification.

2. Methodology

Figure 1 presents the components of the proposed method in alignment with the machine learning lifecycle. Problem definition, data creation, data curation, and model training are presented in this section, while evaluation, deployment, and feedback are described in the Results section.

2.1 Problem Definition

In this study, the LULC classification problem involves automatically identifying and categorizing the Earth’s surface into distinct LULC classes using remote sensing data. The ultimate goal is to produce large-scale LULC maps, for example, covering entire Brazilian biomes such as the Amazon and Cerrado. This constitutes a multiclass classification problem, as the goal is to categorize LULC into the following classes: body of water, urbanized area, mining, primary natural forest vegetation, secondary natural forest vegetation, silviculture, shrubby pasture, herbaceous pasture, single-cycle temporary agriculture, multi-cycle temporary agriculture, semi-perennial agriculture, and non-forest natural formation (NFNF). These classes are derived from the TerraClass project and harmonized with the MapBiomas classification.

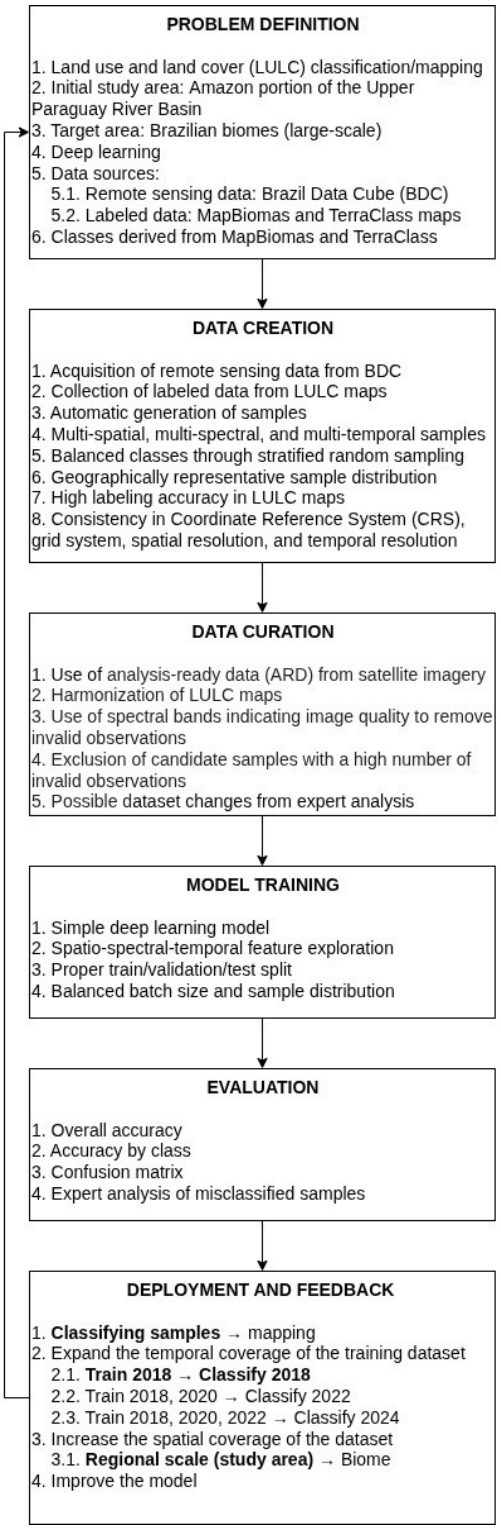

Figure 1. Components of the proposed method within the machine learning cycle.

Although random forest and other machine learning methods remain widely used for LULC classification, we employed deep learning due to its demonstrated advances across various domains—including this one—and its capacity to extract spatio-temporal features from data.

The LULC classification in the proposed method relies on TerraClass and MapBiomas maps to obtain labeled data. The

remote sensing data are provided by the BDC, selected primarily for its data quality, the reputation of its host institution, and its contribution to national data sovereignty. Specifically, the S2-16D-2 collection is utilized, comprising 16-day composites of Sentinel-2 images at a spatial resolution of 10 meters. Both the BDC remote sensing data and the LULC maps are distributed in GeoTIFF format.

Although the ultimate target is to cover all Brazilian biomes, we initially focused on a smaller area within the Amazon region of the Upper Paraguay River Basin (Figure 2). Starting at a regional scale enables testing and refining the proposed approach prior to large-scale applications using more robust computational infrastructure.

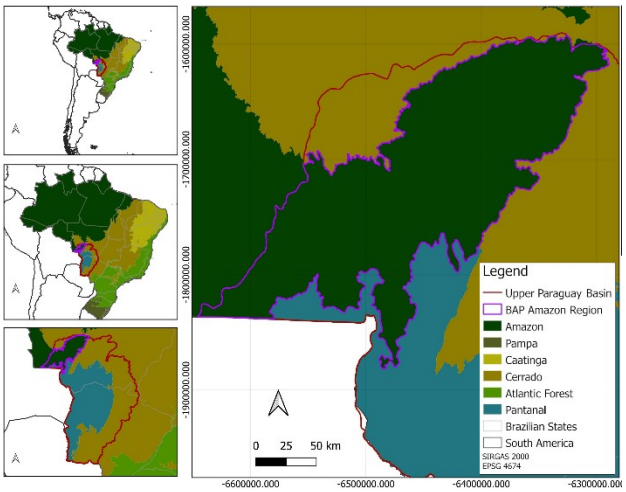

Figure 2. The study area located in the Amazon region of the Upper Paraguay River Basin.

2.2 Data Creation

The samples for the training, validation, and test sets were automatically generated from MapBiomas and TerraClass maps, as well as BDC images. Sample generation followed the method described by Vaz et al. (2025), which comprises two phases. The first phase involves collecting data from the BDC and processes it to improve quality, in addition to obtaining label information through harmonization and comparison of the LULC maps (Figure 3). In the second phase, samples are generated from the processed data. The result is a robust set of multi-pixel, multi-temporal, and multi-band samples, which are balanced, accurate, representative of the study area, and consistent. This is the first study in which this method has been applied and evaluated.

In this method, all candidate samples are generated under the condition that they include a minimum area of $p \times p$ pixels belonging to the same class. Considering all classes of interest, the study area, and with $p = 5$, at least 3,400 candidate samples were generated for each class, except for bodies of water and mining. Mining represented an activity with limited presence in the region of interest, while water areas exhibited a high proportion of invalid observations in the BDC images, thereby requiring special treatment.

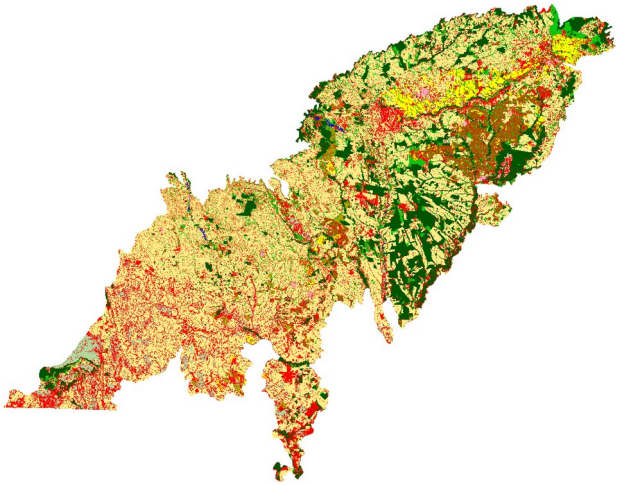

Figure 3. Comparison between the MapBiomas and TerraClass maps (Vaz et al., 2025). Red areas indicate no label agreement, while the remaining colors represent classes that are consistently identified in both maps.

The number of candidate samples varied substantially across classes. For example, over one million candidate samples were generated for herbaceous pasture. From this initial set, a final dataset comprising up to 3,400 samples per class was created through stratified random sampling.

2.2.1 Diversity and Completeness: To generate the samples, candidate locations were listed by tile and by position within each tile. Randomly selecting points from this ordered list ensures broader geographic coverage across the study area, and adopting a similar number of samples per class balances representation. Additionally, the dataset size—over three thousand samples per class—was deemed sufficient based on experimental verification. Thus, sample quantity was not a limitation in this case. In a relatively small region compared to future target areas, a satisfactory number of samples was achieved, enabling us to focus on quality rather than quantity.

Using all available BDC bands allows for the exploration of a wide range of features and facilitates multi-perspective analysis of the study region. Additional data, such as indices derived from these bands (e.g., vegetation, water, and soil indices) may be incorporated in future work.

The samples are multidimensional, as each sample contains, for each band in the data collection, a set of pixels representing a delimited region, and for each pixel, an associated time series covering the temporal range.

2.2.2 Accuracy: The method yields high-confidence samples, as agreement between different data sources increases the likelihood of label accuracy (Bratic et al., 2023). By selecting locations where TerraClass and MapBiomas agree, and given that each map independently exceeds 85% accuracy, the estimated accuracy of the samples exceeds 97% (Vaz et al., 2025).

2.2.3 Consistency: Consistency is ensured by the sample generation method for several reasons:
a) The coordinate reference system is unified as EPSG:4674 across all maps used.
b) The maps are divided into tiles, and a single reference grid is used (i.e., the BDC S2-16D-2 collection grid).
c) All samples share the same temporal resolution.

d) LULC maps are harmonized and compared to generate the sample labels. Both maps and satellite images are provided in GeoTIFF format and converted for compatibility.

Additionally, sample labels are derived from MapBiomas and TerraClass, involving a final map editing phase by experts following the automated classification. Both projects include detailed processes of validation and auditing processes (MapBiomas, 2025), requiring substantial involvement of domain experts.

2.2.4 Unbiasedness: Three potential sources of bias are associated with data generation:

a) Historical bias: In this study, the target is LULC classification using remote sensing data with retrospective labels. Therefore, such applications require one to account for significant changes that may occur, especially in the context of climate change. For example, extreme events not represented in the training data may lead to errors in future LULC classifications, consequently underscoring the importance of appropriately addressing such events. One possible approach is introducing a new phase for specifically handling exceptional conditions prior to general classification. However, each case should be analyzed individually when identified.

b) Representational bias: The sample generation method mitigates under-representation across the population, as stratified random sampling ensures all classes are represented. However, as some classes include diverse land use types, certain forms may remain underrepresented. For instance, the temporary crop class covers a diverse set of crops, some of which may not be well represented. The current label data are limited to general classes; future adoption of more detailed label datasets could improve both classification granularity and representation by capturing subclass diversity. Regarding the remote sensing data, areas with a high proportion of invalid observations risk under-representation; in this study region, this pertains primarily to the body of water class, which is treated individually before multiclass classification. While we focused on a limited geographical region, future research will expand to broader areas. Therefore, evaluating sample representativeness will be important when scaling the study region.

c) Measurement bias: The remote sensing imagery was sourced from an analysis-ready data (ARD) collection available in the BDC (Ferreira et al., 2020). ARD are produced through calibration, standardization, and other preprocessing steps applied to satellite images. Using ARD promotes interoperability, reduces user effort, enables large-scale analyses, and helps mitigate acquisition-related biases. However, labeling introduces bias due to ambiguities in class definitions (Roscher et al., 2024). Although the method here harmonizes MapBiomas and TerraClass, this process cannot ensure total class compatibility.

2.2.5 Relevance: The dataset is consolidated and up-to-date for both BDC and the LULC maps in the temporal range and region of interest. For the Amazon and Cerrado biomes, TerraClass and MapBiomas maps are available for 2018, 2020, and 2022. Expanding this approach to other biomes will require adjustments and potentially additional sources of ground truth data. For remote sensing, BDC provides the requisite Sentinel-2 imagery for these years.

2.3 Data Curation

The method employed involves several data curation processes to improve data quality for deep learning model training (Vaz et al., 2025). Data are adjusted using a BDC-provided band that indicates the quality of each pixel, accounting for issues such as cloud, cloud shadow, and missing data. Observations failing quality standards are removed and replaced through linear interpolation. Additionally, if a pixel’s time series contains many invalid observations, it is excluded from the training set. This process, together with ARD usage, helps address gaps in spatial and temporal coverage.

It is also important to understand why some samples are misclassified. Misclassification may arise from model errors or from samples that are incorrectly labeled, represent transitions between classes, or correspond to specific conditions or underrepresented subclasses.

Class selection was determined through initial harmonization between TerraClass and MapBiomas, then further refined through data analysis and preliminary results. These LULC mapping initiatives are the most widely used in Brazil, with TerraClass providing official national data. As previously explained, mining was excluded due to its rarity in the study area, and bodies of water were excluded owing to the high proportion of invalid observations in BDC. Primary forest was excluded from the final class set because its satellite imagery closely resembles that of secondary forest. The former designates areas of native vegetation never cleared, while the latter refers to regenerating vegetation following disturbance. In preliminary results, these classes were often confused; however, they can be differentiated using historical deforestation data.

TerraClass provides additional information on LULC in deforested areas identified by the Deforestation Monitoring Program by Satellite (PRODES) (de Almeida et al., 2016), the official deforestation data source in Brazil. Each year, PRODES maps newly deforested areas relative to the previous year (de Almeida et al., 2021). Therefore, forests can be categorized as primary or secondary based on PRODES data.

Using pixel quality information, along with the removal of low-quality observations and interpolation, ensures greater consistency in data retrieved from BDC. Nonetheless, consistency can also be affected by differing expert interpretations during labeling. For instance, experts may classify the same data differently, especially for transition between classes such as herbaceous and shrubby pastures. Although fully resolving this is challenging, the method facilitates identifying such cases by tracking misclassified samples for expert review and establishing consistent analysis patterns.

Collectively, these actions improve sample diversity, completeness, and accuracy, reduce systematic bias, and prioritize the most relevant information.

2.4 Model Training

Effective dataset utilization is crucial for model performance. Data samples are first split into training, validation, and test sets. Due to the sampling method, class sample sizes are not perfectly equal, but are closely matched. In this study, the maximum common value across all classes (n = 2,325) was calculated, ensuring that at least 30% of samples could be allocated to validation and test sets. For each class, n samples

were randomly selected for training set. The remaining samples for each class were then split evenly to form the validation and test sets. Table 1 shows the number of valid samples per class for the training, validation, and test sets, as well as the total sample size. This procedure results in a balanced dataset.

Class	Train	Valid.	Test	Total
1 - Secondary forest	2,325	499	498	3,322
2 - Silviculture	2,325	536	537	3,398
3 - Shrubby pasture	2,325	523	522	3,370
4 - Herbaceous pasture	2,325	526	526	3,377
5 - Semi-perennial agriculture	2,325	533	534	3,392
6 - 1-Cycle temporary agriculture	2,325	531	532	3,388
7 - +1-Cycle temporary agriculture	2,325	534	534	3,393
8 - Urbanized	2,325	519	519	3,363
9 - Non-forest natural formation	2,325	524	524	3,373

Table 1. Training, validation, and test sets (balanced dataset).

The class-balanced design is intended to mitigate learning bias toward dominant classes during training, thereby promoting improved representation of minority classes. The test set follows the same class-balance principle to enable fair, per-class evaluation; accuracy is computed separately for each class, ensuring that performance metrics are not disproportionately influenced by the majority classes.

The training set is reorganized so that the corresponding labels are interleaved. This approach ensures that, when divided into batches, each batch tends to represent the entire training set well. For instance, in a nine-class problem, the training labels are arranged as [1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 2, …, 8, 9, 1, 2, …].

Figure 4 presents the sample distribution across the study area, divided into training, validation, and test sets. These datasets originate from the same underlying distribution and exhibited nearly identical spatial structure. Sample density varied across the map, as certain classes were spatially concentrated. For example, pasture occupied a large portion of the study area, so its samples were more sparsely distributed, whereas urban areas were limited to a few regions, resulting in dense clusters of urban samples, as illustrated by the two urban regions in Figure 5. Some regions were not sampled because of a lack of agreement between the reference maps or because they represent classes not considered in this phase of the study, such as water and primary forest.

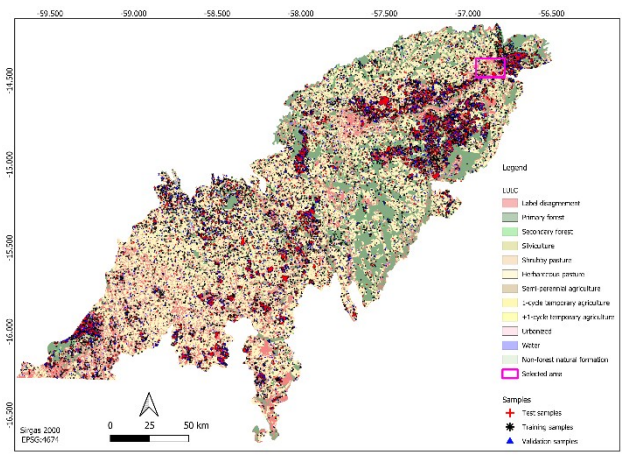

Figure 4. Distribution of samples across the study area, split into training, validation, and test sets.

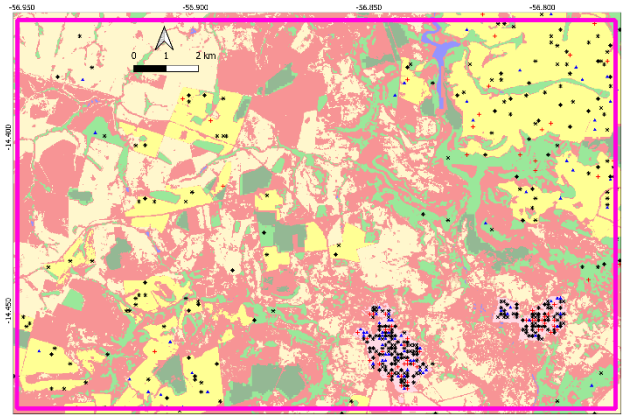

Figure 5. Sample distribution in a selected area, illustrating how sample density varies according to the spatial distribution of classes.

Spatial autocorrelation is inherent in remotely sensed data; nearby pixels tend to be more similar in both spectral features and class labels than distant ones. Ignoring spatial dependence between training and test sets can lead to overestimation of model performance, especially when the goal is to classify new sites. Karasiak et al. (2022) proposed two approaches to mitigate this issue: spatially segregating training and test sets during the data-splitting procedure, and defining distance-based buffers around hold-out samples to ensure that the model is trained only on spatially independent data with disjoint spectral and spatial characteristics.

However, the primary objective of this study was to predict LULC in unsampled locations within the same area where samples were collected, rather than extrapolating to independent geographic regions. While spatial segregation is suitable for assessing spatial transferability, such strategies may introduce unrealistic extrapolation effects in highly heterogeneous landscapes. As the ultimate goal is to map the entire region, the model should encounter examples from all geographic contexts to improve class representativeness and generalization within the region and avoid blind spots in rare classes. Using a distance-based buffer to avoid spatially dependent data would significantly reduce the sample count for less frequent classes, such as urban areas. Therefore, reference samples for both training and accuracy assessment were drawn throughout the study region to ensure full representation of landscape heterogeneity. Model evaluation thus focuses on within-domain map accuracy rather than spatial transferability to independent areas. In addition to per-class accuracy assessment, evaluation of results by domain experts is important.

The model architecture is intentionally simple, yet sufficient for extracting spatio-spectral-temporal features. Figure 6 presents the proposed model, which comprises a spatial-spectral module based on 3D convolutions, followed by a temporal module. This modular design is well suited to handling multi-spectral, multi-temporal data (Qiao et al., 2021).

Three-dimensional (3D) convolutional neural networks (3D CNNs) have demonstrated strong performance in jointly extracting spectral and spatial features, leveraging the structural characteristics of hyperspectral images (Li et al., 2017; Firat, 2022). In the proposed architecture, each convolutional layer is followed by 3D batch normalization and a Rectified Linear Unit (ReLU) activation, consistent with Qiao et al. (2021).

```
SpatioSpectralTemporalClassifier(
  (spat_spectr): Sequential(
  (0):Conv3d(15, 64, kernel_size=(1, 5, 5))
  (1):BatchNorm3d(64)
  (2):ReLU()
  (3):Conv3d(64, 128, kernel_size=(1, 3, 3))
  (4):BatchNorm3d(128)
  (5):ReLU()
  (6):MaxPool3d(kernel_size=(1, 2, 2))
  (7):Conv3d(128, 256, kernel_size=(1, 3, 3))
  (8):BatchNorm3d(256)
  (9):ReLU()
  (10):MaxPool3d(kernel_size=(1, 2, 2))
  )
  (avg_pool):
   AdaptiveAvgPool3d(output_size=(None, 1, 1))
  (rnn):GRU(256, 256, bidirectional=True)
  (proj):Sequential(
  (0):Linear(in_features=512, out_features=256)
  (1):ReLU()
  (2):LayerNorm(256)
  )
  (attn):Linear(in_features=256, out_features=1)
  (clas):Linear(in_features=256, out_features=9)
)
```

Figure 6. Proposed LULC classification model using 3D convolutions for spatio-spectral encoding and a bidirectional gated recurrent unit for temporal modeling.

By using 3D convolutions, the network can jointly model spectral-spatial or spatio-temporal information, depending on input dimension configuration (Gallo et al., 2023). Here, 3D convolutions were employed to extract spectral-spatial features, as convolutional operations are particularly effective at capturing spatial relationships within data (e.g., patterns and correlations across channels). The convolutional blocks extract fine-scale spatial and spectral patterns, while pooling and deeper layers capture broader spatial context. In contrast, recurrent mechanisms are better suited for modeling temporal dependencies (Foumani et al., 2024).

Garnot et al. (2019) investigated several deep learning models for crop type classification and provided evidence that, for this task, the temporal structure in Sentinel-2 data is richer than the spatial structure. They combined recurrent neural networks and CNNs to exploit both structures in image time series, achieving the best results with a hybrid approach in which a CNN-like network generates spatial embeddings that are then processed in chronological order by a gated recurrent unit (GRU).

The bidirectional GRU captures both past and future context at each time step, as the entire time series is available. A projection block subsequently reduces the dimensionality of the GRU output. An attention mechanism is then employed to assign different weights to each time step, enhancing interpretability. Finally, the resulting 256-dimensional feature representation is mapped to the target classes.

3. Results

Figure 7 shows accuracy and loss for the training and validation sets during model training. Validation accuracy reached 96% or higher for all classes, illustrating both model convergence and strong, consistent performance across classes, indicative of good generalization. This high accuracy further reinforced the consistency and diversity of the dataset. Completeness will be evaluated during mapping, when all pixels in the study area are classified, not only those covered by the samples.

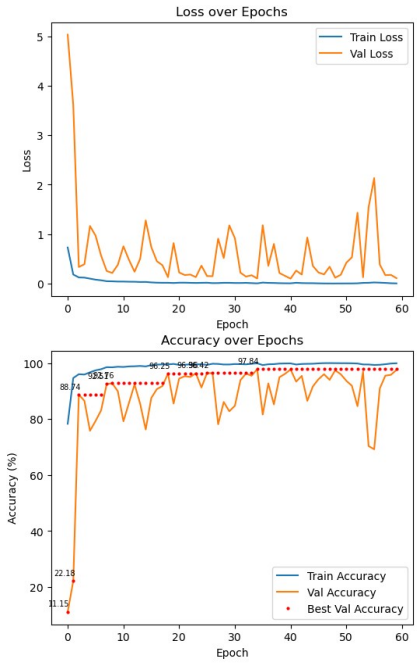

Figure 7. Accuracy and loss on the validation set during model training, demonstrating model convergence.

Bias in model training typically involves aggregation bias or learning bias (Suresh and Guttag, 2021). Regarding modeling choices, our findings indicated that learning bias was not significant, as evidenced by the high classification accuracy (Figure 7). Aggregation bias occurs when subgroups are combined into a single class, which is common in LULC classification due to the general nature of class definitions; for example, different crop types may be aggregated under a single class. Using subgroup labels within the considered classes could help reduce aggregation bias.

Instead of quantifying each data sample’s relevance, misclassified samples were analyzed by experts.

3.1. Evaluation

The samples were divided into training, validation, and test datasets. The overall accuracy on the test set was comparable to that of the validation set, both exceeding 96%. This confirms that the data quality criteria are met for the LULC classification problem.

Evaluation bias may occur from unequal representation of classes in the evaluation set or inappropriate performance metrics (Ghamisi et al., 2024). This is not the case in this study, as all classes were balanced. The selected metrics were appropriate and indicated that accuracy for every class exceeded 96%. Figure 8 presents the confusion matrix for the test set.

Misclassified samples were analyzed by three domain experts to identify causes and ways to potentially improve both the dataset and the classification model. The experts, each holding a PhD and possessing extensive experience in LULC analysis using satellite imagery, time series data, and complementary fieldwork in related projects, conducted the analysis, which revealed the following findings:

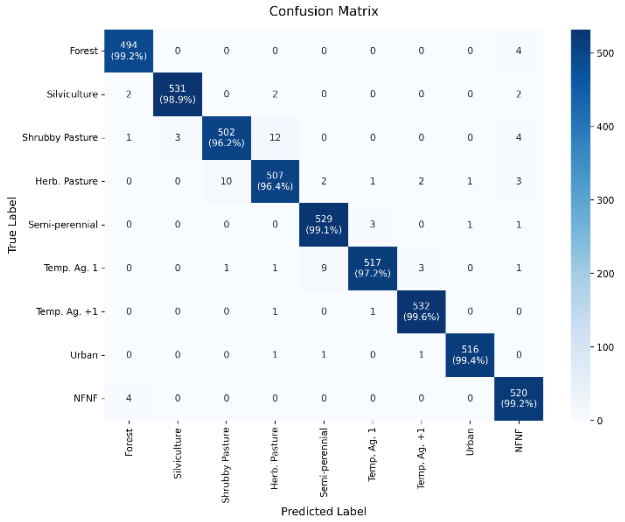

Figure 8. Confusion matrix for the test set. The accuracy for all classes is greater than 96%.

a) Incorrect labels: Nine samples labeled as single-cycle temporary agriculture but classified as semi-perennial were in fact sugarcane and should be labeled as semi-perennial. Additionally, all samples labeled as NFNF but classified as forest actually represent forest. This analysis helps identify incorrect labels in the source maps. Confident learning could be used to identify label errors prior to human verification (Northcutt et al., 2021a, 2021b). However, because the estimated label accuracy exceeded 97%, all classes achieved accuracies above 96% in the test set, and the absolute number of detected mislabeled samples was low, automated label error detection was deemed unnecessary. Moreover, deep learning models are generally robust to label noise. Given the high cost of label verification, these factors also motivated us to not review samples that were correctly classified according to their labels.

b) Specific conditions or subclasses: Silviculture illustrates cases where samples are correctly labeled, but the model failed to classify them accurately. For example, two samples containing teak (*Tectona grandis*) were classified as NFNF, while two samples containing eucalyptus (*Eucalyptus* spp.) were classified as forest due to the particular characteristics of these species. Two samples were misclassified as herbaceous pasture: one corresponds to an area during the teak harvesting period, and the other is a very narrow region surrounded by herbaceous pasture. Another factor contributing to misclassification is field renewal, which occurred with sugarcane in the semi-perennial class.

c) Transition between classes: Several samples of herbaceous and shrubby pasture represent transitional areas between these two classes, where even experts may disagree. There are twenty-two misclassifications involving these two pasture types. In eight of these, the model predictions match the consensus from the three experts, indicating incorrect labels in the source LULC maps. In nine cases, the model misclassified eight herbaceous pasture samples as shrubby pasture and one shrubby pasture sample as herbaceous pasture. For the remaining five, expert opinions diverged, with at least one expert’s label matching the model’s outcome. If the majority view among experts is considered correct, the model’s classification is accurate for four of these five samples. These divergences highlight the inherent difficulty of distinguishing these classes, particularly in transitional areas.

d) Areas close to the border: Three urban area samples were misclassified, all situated near class boundaries. This may result from the model’s consideration of the contextual pixels during training, which is challenging for heterogeneous classes such as urban areas. Green spaces within cities may also lead to such misclassification.

e) Uncertainty in class characterization: Four forest samples were misclassified as NFNF, despite being typical representations of forests. Other classes were also occasionally misclassified as NFNF. This class is particularly challenging because it encompasses a wide diversity of land covers, especially in the Upper Paraguay River Basin within the Amazon biome near the Pantanal and Cerrado biomes. In MapBiomas, NFNF includes wetlands and grasslands, while savanna formations are treated separately. In TerraClass for the Amazon biome, savanna is included in NFNF, as it is in this study. A more precise characterization of NFNF, explicitly listing its main subclasses, would likely improve classification performance for this class.

Although overall model accuracy was very high, these observations suggested directions for potential improvements.

3.2. Deployment and Feedback

This study provides the first iteration of an ongoing effort to achieve large-scale LULC mapping across Brazil. In this phase, 2018 samples were used to classify another set of 2018 samples within a defined study area. The next steps are as follows (Figure 1):

a) From sample classification to full mapping: Classifying all map pixels requires addressing certain specific classes before applying the proposed model. Although all classes were initially considered, data analysis and preliminary experimental results indicated that some classes require separate treatment. These include bodies of water, mining, and primary forest. Possible approaches for these cases include:

a.1) mapping primary forest using PRODES data, as in the TerraClass project;

a.2) identifying mining areas through expert visual interpretation and official datasets; and

a.3) delineating bodies of water using existing sources (e.g., PRODES, TerraClass, and MapBiomas) and official datasets provided by government entities (e.g., the Brazilian National Water Agency).

Once these specific classes are mapped, the multi-class model is applied to the remaining areas. The overall mapping process is thus guided by the data available and the specific characteristics of the classification problem.

b) Expanding the temporal coverage of training data: The model is intended to classify maps from different years using historical data. The following steps are proposed to progressively increase the temporal generalization capability of the model:

b.1) train a model with 2018 samples to classify 2020 data;

b.2) train a model with 2018 and 2020 samples to classify 2022 data; and

b.3) train a model with 2018, 2020, and 2022 samples to classify 2024 data.

c) Expanding the spatial coverage of the dataset: Broaden the study area to encompass larger regions at the biome scale (e.g., the Amazon and Cerrado). Additional attributes may be incorporated to capture broader contextual information, such as

including biome type as a categorical feature. At larger scales, the greater number of candidate samples enables the use of sample selection strategies to reduce spatial dependence among observations, for example, via the distance-based approach proposed by Karasiak et al. (2022).

d) Model improvement: Experiments under new conditions may necessitate refinements to the neural network architecture. Temporal convolutional neural networks may be explored, as demonstrated by Pelletier et al. (2019), who demonstrated the effectiveness of using such networks for satellite image time series that integrate spectral and temporal data. Although the spatial dimension was not addressed in their study, the authors recommended its integration in future research.

Deployment bias occurs when model results are misinterpreted or applied in unintended contexts (Ghamisi, 2024) or when models are inappropriately utilized. For example, in Brazil, PRODES is the official national deforestation monitoring program, identifying both primary forests and areas of deforestation. While this study considered PRODES data to be correct, the map produced using the proposed method will also contain primary forest areas. Since this primary forest mapping is already performed by PRODES, its data should not be used to evaluate the accuracy of the model considered in this method. Another case involves comparing the mapping results to other initiatives (e.g., MapBiomas). Hence, it is critical to consider the equivalence of classes to enable an appropriate comparison between two different maps, especially given that each uses a class structure with particularities.

Domain experts provided valuable feedback on the results, contributing to a deeper understanding and enabling necessary adjustments to the dataset or model. For instance:
a) NFNF class review: Although this class is currently included in the machine learning model, its characterization remains uncertain in some aspects. It was initially included for evaluation, but future work should test approaches such as mapping NFNF using PRODES data, as is the case in TerraClass.
b) Sample quality control: Samples with incorrect labels should be corrected or excluded from the dataset.
c) Addressing sample deficiencies: Misclassification occurs where there are insufficient samples representing specific conditions, necessitating the addition of samples related to cases such as field renewal and forestry harvest.
d) Increasing intra-class diversity: Comprehensive coverage of within-class diversity is also important. For example, for silviculture, including samples of different species (e.g., teak, eucalyptus, and less-common varieties such as mahogany) may improve model’s generalization. While current data lack these subclass annotations, their inclusion should be considered in future research.

This deployment and feedback phase encompasses analyzing the results and incorporating feedback to refine the dataset and model. This iterative process drives the adjustments necessary to improve the LULC classification method and, ultimately, yield higher-quality LULC maps.

4. Discussion

This study contributes to strengthening the data-centric AI approach by prioritizing data quality over model complexity. In doing so, it helps achieve a balance between data-centric and model-centric strategies, directly addressing research gaps related to data diversity, class imbalance, sample label accuracy, model generalization, and context-awareness in modeling.

In addition, this work also aligns with human-centered AI principles, which advocate for AI-based solutions that serve as responsible, productive agents in society, in alignment with the framework proposed by Xu (2019), which comprises:
a) ethically aligned design aimed at developing fair models;
b) technology enhancement that reflects human intelligence, by incorporating expert knowledge—for example, analyzing time series and contextual information to classify LULC and evaluating data and results; and
c) human factors design, ensuring the usefulness and usability of LULC maps for addressing real-world challenges, while pursuing simpler and more comprehensible deep learning models.

The proposed method for LULC classification supports the automated creation and curation of data, aiming to enhance dataset quality according to standards of diversity, completeness, accuracy, consistency, unbiasedness, and relevance. A key contribution of this work is the coordinated set of measures that improve data quality by leveraging the characteristics of the problem and dataset.

The neural network model developed combines 3D convolutions for spatio-spectral encoding, a bidirectional GRU for temporal dynamics, and an attention-based aggregation followed by a linear classifier for predicting output classes. Although existing model architectures could have been applied, the proposed model is intentionally simpler to focus on effectively extracting spatio-spectral-temporal features and to emphasize the importance of data quality and input attributes for LULC classification. This approach enabled us to explore the problem’s characteristics while preserving interpretability. As such, the modeling strategy remained consistent with the data-centric approach by prioritizing insights from the data over architectural complexity.

Emphasis on data preparation (e.g., data diversity, balanced classes, and LULC maps-based sample labeling) led to a highly accurate classification model. Still, in the evaluation phase, model performance was evaluated using samples taken from within the study region. This evaluation thus reflects predictive consistency within the region, rather than spatial generalization, as the training and test data are not spatially independent. Domain experts analyzed misclassified samples to infer their causes and obtain valuable observations.

Data-centric learning often benefits from automated tools, although it may also rely on expert-driven analysis when semantic or domain-specific decisions are required, thereby highlighting the relevance of feedback. This first iteration in the machine learning cycle (Figure 1) provided valuable insights and established a strong baseline for future work. Nevertheless, several limitations remain, mainly related to the current phase:
a) While high accuracy is achieved in sample classification, mapping the entirety of an area likely introduces new challenges that must be addressed;
b) The samples are generated to be of high-quality, but mapping regions with many invalid observations may require new approaches to achieve good results; and
c) Compared to the Brazilian biomes, the study area is small. Hence, expanding this regional scale to a larger scale while maintaining good results may depend on incorporating new features to the machine learning model, such as geographic coordinates and the identification of ecoregions.

5. Conclusions

This is the first study in which the proposed method—a deep learning model leveraging spatio-spectral-temporal features and consensus-based labels from reliable LULC maps—has been fully applied and evaluated, encompassing all phases of the machine learning cycle. Despite certain limitations, the method establishes a robust, high-quality baseline for LULC classification, achieving accuracies exceeding 96% across all classes, and supports large-scale LULC mapping from historical data within a data-centric paradigm.

Acknowledgments

This research was funded by The World Bank Group (Process number 0002010084) and registered in the Embrapa Management System (code SEG 10.25.00.078.00.00), with partial support from Horus FAPESP (Proc. 2023/12865-8), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil), and FAEPEX/UNICAMP (Fundo de Apoio ao Ensino, à Pesquisa e à Extensão).

References

Bratic, G., Oxoli, D, Brovelli, M.A, 2023. Map of land cover agreement: ensambling existing datasets for large-scale training data provision. *Remote Sens*., *15*(15), 3774. doi.org/10.3390/rs15153774.

de Almeida, C.A., Coutinho, A.C., Esquerdo, J.C.D.M., Adami, M., Venturieri, A., Diniz, C.G., Dessay, N., Durieux, L., Gomes, A.R., 2016. High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data. *Acta Amaz.*, *46*(3), 291-302. doi.org/10.1590/1809-4392201505504.

de Almeida, C.A., Maurano, L.E.P., de Morisson Valeriano, D., Camara, G., Vinhas, L., Gomes, A.R., Monteiro, A.M.V., de Almeida Souza, A.A., Rennó, C.D., Silva, D.E., Adami, M., 2021. Methodology for forest monitoring used in PRODES and DETER projects. INPE. urlib.net/8JMKD3MGP3W34R/443H3RE (22 October 2025).

Digra, M., Dhir, R., Sharma, N., 2022. Land use land cover classification of remote sensing images based on the deep learning approaches: a statistical analysis and review. *Arab. J. Geosci.*, *15*(10), 1003. doi.org/10.1007/s12517-022-10246-8.

Ferreira, K.R., Queiroz, G.R., Vinhas, L., Marujo, R.F., Simoes, R.E., Picoli, M.C., Camara, G., Cartaxo, R., Gomes, V.C., Santos, L.A., Sanchez, A.H., 2020. Earth observation data cubes for Brazil: Requirements, methodology and products. *Remote Sens*., *12*(24), 4033. doi.org/10.3390/rs12244033.

Fırat, H., Asker, M.E., Hanbay, D., 2022. Classification of hyperspectral remote sensing images using different dimension reduction methods with 3D/2D CNN. *Remote Sens. Appl.: Soc. Environ.*, *25*, 100694. doi.org/10.1016/j.rsase.2022.100694.

Foumani, N.M., Miller, L., Tan, C.W., Webb, G.I., Forestier, G., Salehi, M., 2024. Deep learning for time series classification and extrinsic regression: A current survey. *ACM Comput. Surv.*, *56*(9), 1-45. doi.org/10.1145/3649448.

Gallo, I., Ranghetti, L., Landro, N., La Grassa, R., Boschetti, M., 2023. In-season and dynamic crop mapping using 3D convolution neural networks and sentinel-2 time series. *ISPRS J. Photogramm. Remote Sens.*, *195*, 335-352. doi.org/10.1016/j.isprsjprs.2022.12.005.

Garnot, V.S.F., Landrieu, L., Giordano, S., Chehata, N., 2019. Time-space tradeoff in deep learning models for crop classification on satellite multi-spectral image time series. In *IGARSS 2019 - 2019 IEEE Int. Geosci. Remote Sens. Symp.,* 6247-6250. doi.org/10.1109/IGARSS.2019.8900517.

Ghamisi, P., Yu, W., Marinoni, A., Gevaert, C.M., Persello, C., Selvakumaran, S., Girotto, M., Horton, B.P., Rufin, P., Hostert, P., Pacifici, F., Atkinson, P.M., 2024. Responsible AI for Earth Observation. *arXiv preprint*, doi.org/10.48550/arXiv.2405.20868.

Instituto Nacional de Pesquisas Espaciais, 2025. Brazil Data Cube. https://data.inpe.br/bdc (3 October 2025).

Jakubik, J., Vössing, M., Kühl, N., Walk, J., Satzger, G., 2024. Data-centric artificial intelligence. *Bus. Inf. Syst. Eng.*, *66*(4), 507-515. doi.org/10.1007/s12599-024-00857-8.

Jarrahi, M.H., Memariani, A., Guha, S., 2023. The Principles of Data-Centric AI. *Commun. ACM*, *66*(8) (August 2023), 84–92. doi.org/10.1145/3571724.

Karasiak, N., Dejoux, J.F., Monteil, C., Sheeren, D., 2022. Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. *Mach. Learn*., *111*, 2715–2740. doi.org/10.1007/s10994-021-05972-1.

Kumar, S., Datta, S., Singh, V., Singh, S.K., Sharma, R., 2024. Opportunities and challenges in data-centric AI. *IEEE Access*, *12*, 33173-33189. doi.org/10.1109/ACCESS.2024.3369417.

Li, Y., Zhang, H., Shen, Q., 2017. Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sens.*, *9*(1), 67. doi.org/10.3390/rs9010067.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.*, *152*, 166-177. doi.org/10.1016/j.isprsjprs.2019.04.015.

MapBiomas, 2025. MapBiomas 10-meters Handbook: Algorithm Theoretical Basis Document (ATBD) Collection 2 beta, V1. doi.org/10.58053/MapBiomas/F1SN9W.

Northcutt, C.G., Athalye, A., Mueller, J., 2021a. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint*. doi.org/10.48550/arXiv.2103.14749.

Northcutt, C., Jiang, L., Chuang, I., 2021b. Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Intell. Res.*, *70*, 1373–1411. doi.org/10.1613/jair.1.12125.

Pelletier, C., Webb, G.I., Petitjean, F., 2019. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sens.*, *11*(5), 523. doi.org/10.3390/rs11050523.

Polyzotis, N., Zaharia, M., 2021. What can data-centric AI learn from data and ML engineering?. *arXiv preprint,* doi.org/10.48550/arXiv.2112.06439.

Qiao, M., He, X., Cheng, X., Li, P., Luo, H., Zhang, L., Tian, Z., 2021. Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3D convolutional neural networks. *Int. J. Appl. Earth Obs. Geoinf.*, *102*, 102436. doi.org/10.1016/j.jag.2021.102436.

Roscher, R., Russwurm, M., Gevaert, C., Kampffmeyer, M., Dos Santos, J.A., Vakalopoulou, M., Hänsch, R., Hansen, S., Nogueira, K., Prexl, J., Tuia, D., 2024. Better, Not Just More: Data-centric machine learning for Earth observation. *IEEE Geosci. Remote Sens. Mag.*, *12*(4), 335-355. doi.org/10.1109/MGRS.2024.3470986.

Souza Jr., C.M., Shimbo, J., Rosa, M.R., Parente, L.L., A. Alencar, A., Rudorff, B.F., Hasenack, H., Matsumoto, M., G. Ferreira, L., Souza-Filho, P.W., De Oliveira, S.W., 2020. Reconstructing three decades of land use and land cover changes in Brazilian biomes with Landsat Archive and Earth Engine. *Remote Sens.*, *12*(17). doi.org/10.3390/rs12172735.

Suresh, H, Guttag, J., 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In: *Proc. 1st ACM Conf. Equity Access Algorithms Mech. Optim. (EAAMO’21)*. ACM, New York, NY, USA, *17*, 1–9. doi.org/10.1145/3465416.3483305.

TerraClass, 2024. https://www.terraclass.gov.br (3 October 2025).

Vali, A., Comai, S., Matteucci, M., 2020. Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. *Remote Sens.*, *12*(15), 2495. doi.org/10.3390/rs12152495.

Vaz, G.J., Coutinho, A.C., Esquerdo, J.C.D.M., Antunes, J.F.G., Rocha, A., 2025. Multidimensional Sampling for Land Use and Land Cover Classification. In: *Proc. Latin Am. GRSS & ISPRS Remote Sens. Conf. (LAGIRS 2025)*. doi.org/10.1109/LAGIRS68367.2025.11414795.

Xu, W., 2019. Toward human-centered AI: a perspective from human-computer interaction. *Interactions*, *26*(4), 42-46. doi.org/10.1145/3328485.

Zhang, C., Li, X., 2022. Land use and land cover mapping in the era of big data. *Land*, *11*, 1692. doi.org/10.3390/land11101692.

Zhao, S., Tu, K., Ye, S., Tang, H, Hu, Y, Xie, C, 2023. Land use and land cover classification meets deep learning: a review. *Sensors*, *23*(21), 8966. doi.org/10.3390/s23218966.

Zimbres, B., Shimbo, J., Lenti, F., Brandão, A., Souza, E., Azevedo, T., Alencar, A., 2024. Improving estimations of GHG emissions and removals from land use change and forests in Brazil. *Environ. Res. Lett.*, *19*(9), 094024. doi.org/10.1088/1748-9326/ad64ea.