MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21d/2023/03.31.22.17-TDI

AI4LUC: PIXEL-BASED CLASSIFICATION OF LAND USE AND LAND COVER VIA DEEP LEARNING AND A CERRADO IMAGE DATASET

Mateus de Souza Miranda

Master's Dissertation of the Graduate Course in Applied Computing, guided by Drs. Valdivino Alexandre de Santiago Júnior, and Thales Sehn Körting, approved in March 28, 2023.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34T/48QQB65>

INPE
São José dos Campos
2023

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE
Coordenação de Ensino, Pesquisa e Extensão (COEPE)
Divisão de Biblioteca (DIBIB)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21d/2023/03.31.22.17-TDI

AI4LUC: PIXEL-BASED CLASSIFICATION OF LAND USE AND LAND COVER VIA DEEP LEARNING AND A CERRADO IMAGE DATASET

Mateus de Souza Miranda

Master's Dissertation of the Graduate Course in Applied Computing, guided by Drs. Valdivino Alexandre de Santiago Júnior, and Thales Sehn Körting, approved in March 28, 2023.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34T/48QQB65>

INPE
São José dos Campos
2023

Cataloging in Publication Data

Miranda, Mateus de Souza.

M672a AI4LUC: pixel-based classification of land use and land cover via deep learning and a Cerrado image dataset / Mateus de Souza Miranda. – São José dos Campos : INPE, 2023.

xx + 177 p. ; (sid.inpe.br/mtc-m21d/2023/03.31.22.17-TDI)

Dissertation (Master in Applied Computing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2023.

Guiding : Drs. Valdivino Alexandre de Santiago Júnior, and Thales Sehn Körting.

1. Pixel-based image classification. 2. Deep learning. 3. Cerrado. 4. CBERS-4A. I.Title.

CDU 004.932:528.8(213.54)

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

DEFESA FINAL DE DISSERTAÇÃO MATEUS DE SOUZA MIRANDA
BANCA Nº 044/2023, REG. 994107/2021

No dia 28 de março de 2023, as 14h, por teleconferência, o(a) aluno(a) mencionado(a) acima defendeu seu trabalho final (apresentação oral seguida de arguição) perante uma Banca Examinadora, cujos membros estão listados abaixo. O(A) aluno(a) foi APROVADO(A) pela Banca Examinadora, por unanimidade, em cumprimento ao requisito exigido para obtenção do Título de Mestre em Computação Aplicada, com a exigência de que o trabalho final a ser publicado deverá incorporar as correções sugeridas pela Banca Examinadora, com revisão pelo(s) orientador(es).

Novo título: "AI4LUC: PIXEL-BASED CLASSIFICATION OF LAND USE AND LAND COVER VIA DEEP LEARNING AND A CERRADO IMAGE DATASET"

Membros da Banca:

Dr. Élcio Hideiti Shiguemori – Presidente da Banca– INPE
Dr. Valdivino Alexandre de Santiago Júnior – Orientador– INPE
Dr. Thales Sehn Körting - Orientador- INPE
Dra. Maria Isabel Sobral Escada – Membro Interno – INPE
Dr. João Paulo Papa - Membro Externo - UNESP

Documento assinado eletronicamente por **Valdivino Alexandre de Santiago Júnior**, **Tecnologista**, em 29/03/2023, às 13:47 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Joao Paulo papa (E)**, **Usuário Externo**, em 29/03/2023, às 17:49 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thales Sehn Korting**, **Pesquisador**, em 29/03/2023, às 17:54 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Maria Isabel Sobral Escada**, **Tecnologista**, em 02/04/2023, às 23:23 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Elcio hideiti shiguemori (E)**, **Usuário Externo**, em 04/04/2023, às 10:50 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.mcti.gov.br/verifica.html, informando o código verificador **10937449** e o código CRC **6813F2FC**.

Referência: Processo nº 01340.002177/2023-57 SEI nº 10937449

"Now does matter. What happened before no longer exists. What will happen next has not yet been written. We have only now. That is our greatest advantage. What we do now, here, in this moment, has the power to determine the future".

Bryan Fuller and Alex Kurtzman
in *"Star Trek Discovery"*, 2017

To my family.

ACKNOWLEDGEMENTS

I appreciate God so much for this journey, I learned a lot with this one. My acknowledgments to my family for having them by my side, on behalf of my dear mother Nilma Miranda, who gave me all her love, inspiration, and strength; my advisers, Valdivino and Thales, for their wisdom, patience, dedication and encouragement; the curatorial team, members of the TerraClass and PRODES *Cerrado* projects, for their incredible support in the development of the research, Alana Souza, Magog Araujo, Taise Pinheiro, Andreia Scheide, Erison Monteiro, and Jadson Queiroz; to my dear friends Rafael Marinho and Pedro Brito, Renato Maximiano, Flávia Pacheco, Adriany Barbosa, Jonas Oliveira, Rafael Santos, Felix Beer, Hadassa Jácome, Baggio Silva, Marcos Rodrigues, Jéssica Barbosa, Fernanda Paiva, Daniele Medeiros, Edith, Nara Fagundes, Nicolau (Lilico), Nicole, Luan Orion, Marcelly Coelho, Eder Santos, Sabrina Marques, Isa Nogueira, Linda Micals, and Rahima Moked.

I would like to thank the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) for funding this research via the grant 88887.603929/2021-00. Finally, I would like to thank the *Laboratório Nacional de Computação Científica* (LNCC/MCTI, Brazil) for granting resources from the SDumont supercomputer (`http://sdumont.lncc.br`) to support the project ***Classificação de imagens via redes neurais profundas e grandes bases de dados para aplicações aeroespaciais*** (Image classification via Deep neural networks and large databases for aeroSpace applications - IDeepS). This research was developed within the IDeepS project (`https://github.com/vsantjr`).

ABSTRACT

The *Cerrado* biome is known for the biodiversity of flora, as well as for its potential in agricultural production. Its landscapes of land use and land cover (LULC) are monitored in order to analyze and understand the social, economic, and environmental aspects related to causative factors and impacts of these activities. There are many efforts by the Remote Sensing (RS) community for employing machine learning (ML) or deep learning (DL) techniques aiming to improve classification tasks, in terms of either pixel-based classification or contextual classification. However, a few datasets containing images with high spatial resolution, representativeness, and a huge number of samples about the *Cerrado* biome are available. For supervised learning of either DL or ML models, dataset samples must be labeled. This procedure currently relies on manual execution, demanding significant time and attention. For instance, it involves generating and labeling reference masks, where specific pixels indicate the class to which they belong in the segment. Driven by these motivations, this master's dissertation strives to make a valuable contribution to the field of pixel-based classification, specifically focusing on semantic segmentation of Land Use and Land Cover (LULC) using deep learning techniques applied to a dataset of satellite images from the *Cerrado* region. To achieve this objective, a novel approach named Artificial Intelligence for Land Use and Land Cover Classification (AI4LUC) is introduced. Thus, a dataset regarding the *Cerrado* biome was created, called CerraData, amounting to unlabeled 2.5 million patches with a height and width of 256 pixels, and two meters of spatial resolution. The spectral bands were obtained from the Wide Panchromatic and Multispectral Camera (WPM) of the China-Brazil Earth Resources-4A (CBERS-4A) satellite. From this dataset, two novel labeled versions were designed. Furthermore, a novel convolutional neural network (CNN) called CerraNetv3 has been developed to enhance the pixel-based classification task. CerraNetv3, along with Google DeepLabv3plus, collaboratively contributes to this endeavor. Additionally, an innovative technique has been introduced to automate the generation and labeling of reference masks. By leveraging the capabilities of CerraNetv3, these reference masks are utilized to facilitate the training process of DeepLabv3plus for pixel-based classification. AI4LUC was subjected to a comparative analysis with other related approaches in the domain of semantic segmentation and contextual classification to assess its viability. The findings revealed that CerraNetv3 achieved the highest performance in the contextual classification experiment, attaining an impressive F1-score of 0.9289. As for the automatic mask generation and labeling method, it yielded an overall score of 0.6738, with F1-score metrics. In contrast, DeepLabv3plus obtained significantly lower scores of 0.2805 for the same metric. The lower scores of the mask generation method can be attributed to occasional deficiencies in the quality of generated masks, resulting in mislabeling by the CerraNetv3 classifier. Consequently, DeepLabv3plus also exhibited suboptimal performance.

Keywords: Pixel-based image classification. Deep learning. *Cerrado*. CBERS-4A.

AI4LUC: CLASSIFICAÇÃO BASEADA EM PIXELS DO USO E COBERTURA DA TERRA CONSIDERANDO UM CONJUNTO DE IMAGENS DO CERRADO

RESUMO

O bioma *Cerrado* é conhecido pela biodiversidade da flora, bem como pelo seu potencial na produção agrícola. Suas paisagens de uso e cobertura da terra (LULC) são monitoradas a fim de analisar e compreender os aspectos sociais, econômicos e ambientais relacionados aos fatores causadores e impactos dessas atividades. Existem muitos esforços da comunidade de Sensoriamento Remoto (SR) para empregar técnicas de aprendizado de máquina (AM) ou aprendizado profundo (AP) com o objetivo de melhorar as tarefas de classificação, seja em termos de classificação baseada em pixels ou classificação contextual. No entanto, poucos conjuntos de dados contendo imagens com alta resolução espacial, representatividade e um grande número de amostras sobre o bioma *Cerrado* estão disponíveis. Para aprendizado supervisionado de modelos AP ou AM, as amostras de conjunto de dados devem ser rotuladas. Este procedimento atualmente depende de execução manual, exigindo muito tempo e atenção. Por exemplo, a geração e rotulagem de máscaras de referência, onde cada pixel indicam a classe a que pertencem no segmento. Impulsionada por essas motivações, esta dissertação de mestrado visa contribuir para o campo da classificação baseada em pixels, focando especificamente na segmentação semântica do uso e cobertura da Terra (LULC) usando técnicas de AP aplicadas a um conjunto de dados de imagens de satélite do *Cerrado*. Para alcançar este objetivo, uma nova metodologia, denominada Artificial Intelligence for Land Use and Land Cover Classification (AI4LUC), é apresentada. Assim, foi criado um conjunto de dados referente ao bioma *Cerrado*, denominado CerraData, totalizando 2,5 milhões de manchas não rotuladas com altura e largura de 256 pixels e dois metros de resolução espacial. As bandas espectrais foram obtidas da *Wide Panchromatic and Multispectral Camera* (WPM) do satélite CBERS-4A. A partir deste conjunto de dados, duas novas versões rotuladas foram projetadas. Além disso, uma nova rede neural convolucional (CNN) chamada CerraNetv3 foi desenvolvida para tarefa de classificação contextual. Esta rede foi introduzida a no método para automatizar a geração e rotulagem de máscaras de referência, as quais são utilizadas para o treinamento do DeepLabv3plus. AI4LUC foi submetido a uma análise comparativa com outras abordagens no domínio da segmentação semântica e classificação contextual para avaliar a sua viabilidade. Os resultados revelaram que o CerraNetv3 alcançou o melhor desempenho no experimento de classificação contextual, atingindo de 0,9289 com F1-score. Quanto à geração automática de máscara e ao método de rotulagem, obteve uma pontuação geral de 0,6738, com F1-score. As pontuações mais baixas desse método podem ser associadas a qualidade das máscaras geradas, resultando em rotulagem incorreta pelo classificador CerraNetv3. Consequentemente, o DeepLabv3plus obteve 0,2805, desempenho abaixo do ideal esperado.

Palavras-chave: Classificação de imagem baseada em pixel. Aprendizado profundo. Cerrado. CBERS-4A.

LIST OF FIGURES

Page

LIST OF TABLES

Page

CONTENTS

Page

1 INTRODUCTION

The *Cerrado* biome, being the second largest in Brazil, has been the focal point of countless studies that aim to understand the diverse impacts of human activities in this environmental resources. This biome is renowned for its diversity of flora and fauna, as well as its potential for agricultural production (RIBEIRO; WALTER, 2008). In order to monitor changes in land use and land cover (LULC), the remote sensing (RS) community has been using high temporal and spatial resolution satellite imagery (SIMOES et al., 2021; WANG et al., 2022a). The data extracted from these images are processed via computer vision (CV) and artificial intelligence (AI) techniques, which include the abstraction of features and the creation of statistical representations of elements present in the scene composition (NEVES et al., 2021).

The satellite imagery to observe the Earth's LULC has proven to be a valuable source of information pertaining to climate, occupation of natural areas, deforestation rates, and urban expansion (CÂMARA, 2020). Universities and research institutes, such as the *Instituto Nacional de Pesquisas Espaciais (INPE)*, have made significant efforts in developing and employing enhanced methodologies for processing remote sensing imagery (RSI), with emphasis on image classification, considering its context. It is important to note that LULC classification differentiates between land covers, such as savanna, and land use, such as farming (FONSECA et al., 2021).

Despite the existence of initiatives aiming to monitor the *Cerrado* in terms of methods and techniques for a data processing, there are few datasets that cover an extensive area of the biome and contain high spatial resolution images available and ready-to-use. Nogueira et al. (2016) put forward a dataset[1], comprising the *Serra do Cipó* region, Cerrado of *Minas Gerais* State. In total there are 1311 images distributed in four classes, Agriculture, Arboreal, Herbaceous, and Shrubby Vegetation, whose dimensions of the image are 64×64 pixels, and a spatial resolution of five meters. The samples are composed of near-infrared (NIR), green (G), and red (R) spectral bands of the RapidEye satellite's sensors.

However, most works collect their own datasets from satellite image catalogs, such as INPE's *Divisão de Observação da Terra e Geoinformática* [2] (DIOTG) and Brazil Data Cube[3] (BDC). For instance, Neves et al. (2021) have designed eight datasets about The Brasília National Park (BNP), in the Federal District. This study site

[1] https://bityli.com/rMQv0
[2] http://www2.dgi.inpe.br/
[3] http://www.brazildatacube.org/

has major physiognomies found in the Cerrado biome. The samples of each set were merged in different ways, since considering the true color composition to all spectral bands of the Worldview-2's camera satellite, with two meters of spatial resolution. These datasets contain three major classes and another ten subclasses, of which every dataset has 12285 samples of 160×160px, where 10530 of them produced from six data augmentation transformation techniques.

As noted by Fonseca et al. (2021), traditional approaches such as support vector machine (SVM) (MA et al., 2017), random forest (RF) (HÄNSCH; HELLWICH, 2018), and geographic object-based image analysis (GEOBIA) (ADORNO et al., 2023) have remained the most commonly used for supervised image classification. However, these methods are implemented for specific case studies and do not cover all dynamics and LULC, i.e. they are restricted to one type of land cover, even by a dataset. When it comes to a large dataset, it may not be sufficient to rely solely on traditional machine learning (ML) algorithms in terms of efficiency and effectiveness, given the dynamic nature and complexity of the objects in question and their contexts of LULC. Conversely, deep neural networks (DNN) possess the ability to generalize their learning to unknown datasets, compared to ML approaches (DU et al., 2021). Yet, the dataset must have sufficient quantity and diversity of data representations (PEDRAYES et al., 2021).

DNNs have been demonstrating remarkable progress and making noteworthy contributions, not only in the detection of patterns but also in the areas of regression, prediction, and clustering of data (PACIFICI F., 2008), as known as deep learning (DL) models. Among DL models, convolutional neural networks (CNN) are widely employed for various satellite image processing scenarios. These networks are inspired by the human visual cortex, capable to extract features and learning from patterns between the objects in the image (GÉRON, 2019; GOODFELLOW et al., 2016). Moreover, CNNs are capable of achieving image contextual classification and semantic segmentation[4] in a more automated manner compared to classical ML (SANTIAGO JÚNIOR, 2022; MIRANDA et al., 2022).

Contextual classification considers all spectral information of each pixel, whereas segmentation involves the division of the image into uniform regions of contiguous pixels (INPE, 1996). This can be accomplished by instance, which separates in different regions (CARVALHO et al., 2021), or semantic, which groups instances of regions that share similar patterns (NIU, 2021). In the case of supervised learning, the CNN

[4]Also known as pixel-based classification (ESRI, 2022).

network model must be trained using a labeled dataset. This means that each image, or each pixel within the image, is assigned a label. Additionally, there are methods for assigning labels to certain parts of the image, known as sparse training samples (ESRI, 2022), and for extracting and labeling time series from images to aid in the classification of agricultural areas in particular (FONSECA et al., 2021).

1.1 Motivation

Given the increase of studies on the *Cerrado* and the little availability of datasets encompassing a wide area of the biome, as well as containing high spatial resolution samples, it is imperative to have a comprehensive labeled dataset that offers different types of representation of forest formations, cultivated areas, savanna formations, water courses, wetlands, the rate of deforestation, and burned areas (NEVES et al., 2021). Additionally, it should comprise high-spatial resolution rasters (WANG et al., 2022b), which enable the observation of more nuanced details such as the differentiation between pasture and savanna formations. Moreover, another crucial aspect is ensuring a balance in the number of samples for each class, as seen in many works that introduce an unbalanced distribution of samples per class.

The INPE is responsible for a relevant role in monitoring Brazilian biomes, standing out three main projects: i) the Brazilian Amazon Forest Monitoring Program (PRODES), ii) Real-time Deforestation Detection System (DETER), and iii) Land Use and Land Cover Mapping System in the Brazilian Legal Amazon (TerraClass). However, the classification procedures are done manually by experts who perform image interpretation, draw polygons using software, and classify them based on the context, tone, and texture of the scene (INPE, 2019).

In preparing a dataset, most labeling or annotation of data is done manually and consequently requires significant time and attention. It is not a trivial task (FONSECA et al., 2021; CÂMARA, 2020). In line with this perspective, there have been many efforts to automate this process by combining techniques, methods, and different types of data. For instance, Silva et al. (2022) has adopted a different strategy, using self-organized maps for spatiotemporal segmentation using time series from satellite images. On the other hand, Han et al. (2022) has employed deep convolutional neural network (DCNN) models to learn how to perform semantic segmentation from training image sets containing multi-spectral patches followed by their respective masks.

With regard to the semantic segmentation approach, the context information is con-

sidered to classify each segment belonging to the mask. It means each mask's pixel represents a class to each image's pixel. Among several CNN models, DeepLabv3plus has been employed due to its better extractor of the image's context features. With regard to the semantic segmentation approach, the entire image's context information is considered to segment and classify each segment belonging to the mask. It means each mask's pixel represents a class to each image's pixel (ESRI, 2022). Among several CNN models, DeepLabv3plus and U-Net have been employed due to their better extractor of the image's context features, applied to several types of image processing tasks, since RSI till medical data (RONNEBERGER et al., 2015; CHEN et al., 2018).

1.2 Objective and hypotheses

This master dissertation aims to contribute to the pixel-based classification of LULC via deep learning (DL) and a *Cerrado* RSI dataset. To contemplate this goal, the Artificial Intelligence for Land Use and Land Cover Classification (AI4LUC) method was developed, based on the methodology shown in INPE (2019). Firstly, a dataset, called CerraData, was created amounting 2.5 million patches[5] with height and width of 256 pixels and two meters of spatial resolution. Afterward, two novel version were created. The satellite images were obtained from the Wide Panchromatic and Multispectral Camera (WPM) of the China-Brazil Earth Resources-4A (CBERS-4A) satellite. Secondly, a new convolutional neural network (CNN), known as CerraNet, was designed. CerraNet and Google DeepLabv3plus are jointly considered to support the pixel-based classification task. A new technique was also proposed to automatically generate the masks of the patches to support the pixel-based classification. This is an important step towards the automation of the process to accomplish semantic segmentation in projects such a TerraClass.

Thus, three hypotheses related to this master dissertation were defined:

- Hypothesis 1: The AI4LUC, based on the CerraNetv3 network, can assist in the automated labeling of mask's segments, generated from the satellite scene;
- Hypothesis 2: DeepLabv3plus, while trained with CerraDatav3 and the labeled masks obtained with the help of CerraNetv3, will perform better than U-Net, in terms of semantic segmentation;

[5]Small clippings of an image.

- Hypothesis 3: The AI4LUC method, based on the integration of two DCNNs, contributes to automate the pixel-based classification of remote sensing images.

This research was developed within the project ***Classificação de imagens via redes neurais profundas e grandes bases de dados para aplicações aeroespaciais*** (Image classification via Deep neural networks and large databases for aeroSpace applications - IDeepS) (SANTIAGO JÚNIOR et al., 2022). The IDeepS project aims to carry out a large-scale investigation of several existing DNNs in order to primarily automate and improve remote sensing image classification to support LULC analysis accomplished by INPE.

Other computational vision tasks, such as object detection and semantic segmentation, are also addressed in the project. Results of the IDeepS project may contribute to the information issued by INPE regarding deforestation and fire outbreaks. In addition, the project intends to identify the best DNNs to support autonomous-drone flights, for example, to improve the autonomy of these systems with regard to the response to disasters and emergency situations in areas of difficult access. Thus, this project may guide other actions for the increasing dissemination of low-cost unmanned aerial vehicles (UAVs) in civil and military applications.

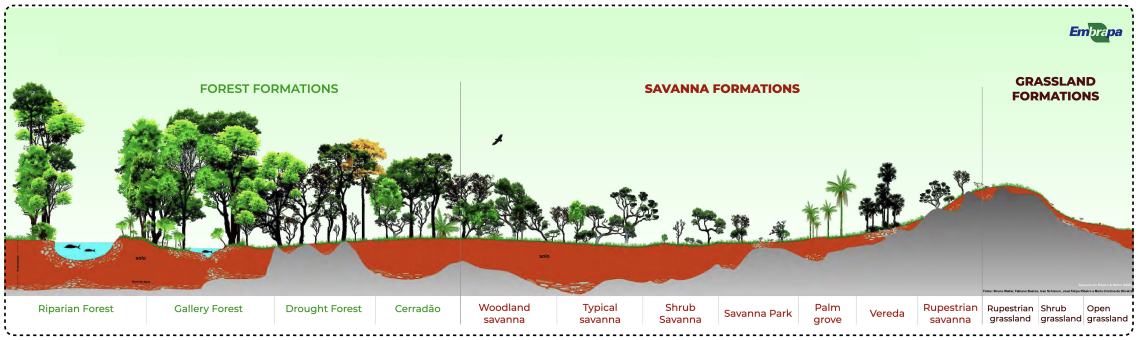
The IDeepS project is supported by the *Laboratório Nacional de Computação Científica* (LNCC/MCTI, Brazil) via resources of the SDumont supercomputer. Researchers, professors, and post-graduate students from the following organizations are involved in the project: INPE, *Instituto de Estudos Avançados* (IEAv), *Universidade Federal de São Paulo - Campus São José dos Campos* (UNIFESP), *Instituto Tecnológico de Aeronáutica* (ITA), and *Universidade Federal de São Carlos - Campus Sorocaba* (UFSCar)

2 THEORETICAL BACKGROUND

The *Cerrado* is the second-largest Brazilian biome, after the Amazon biome, corresponding to 23.9% of the national territory, extending over 2,036,448 km^2 (IBGE, 2023a). It covers continually the states of *Goiás, Tocantins,* and the Federal District; part of the states of *Bahia, Ceará, Maranhão, Mato Grosso, Mato Grosso do Sul, Minas Gerais, Piauí, Rondônia,* and *São Paulo*; and also occurs in disjunct areas to the north in the states of *Amapá, Amazonas, Pará* and *Roraima*, and to the south, in small regions in *Paraná* (IBGE, 2023b; EMBRAPA, 2014a). This term, origin in Spanish, has been used to refer to the biome, a set of vegetation features, as well as to a specific type of florists composition that occurs in the formation of savannas (SANO et al., 2008; RIBEIRO; WALTER, 2008).

This biome has a typical vegetation in which woody plants have thick stems, a dark tone, and are twisted, in other cases, the branches can be angled close to the ground and the tip facing upwards (EITEN, 1990).*Cerrado* contains a great diversity of vegetation, as illustrated in Figure 2.1. There are three main groups of phytophysiognomies and their sub-formations, i) forest formation, which includes riparian forests, gallery forests, drought forests, and *cerradão*; ii) savannas, assuming a Woodland savanna, typical savanna, shrub savanna, palm grove, savanna park, *veredas*, and rupestrian savanna; and iii) countryside formation, characterized as rupestrian, shrub, open, and humid grassland (RIBEIRO; WALTER, 2008).

Figure 2.1 - Phytophysiognomies of the *Cerrado* biome.



SOURCE: Adapted from EMBRAPA (2014a).

The Figure 2.1 shows that forest formation is dense and consists continuously of trees with a well-structured canopy. The savannas, however, have a low concentration of

herbaceous plants, with more significant space between trees or shrubs. Whereas in the countryside formation, there is an open landscape, containing a little or any trees or shrubs, and an abundance of herbaceous strata (COUTINHO, 2016). These structures are located in the tropical zone (EITEN, 1990), whose formations are influenced mainly by the type of soil and its geomorphology, climate, presence of water bodies, and other aspects.

Ribeiro and Walter (2008) say all vegetation is strong and adapted to the climate of each region of the biome, whose dynamics between soil, temperature, and precipitation influence and determine conditions for the development of the self-adaptation capacity of the flora. Proportionately almost the entire *Cerrado* climate is classified as rainy tropical, with two well-defined seasons, wet, occurring between the months of October and April; and drought, occurring between May and September (SANO et al., 2008).

The air temperature is correlated with regard to geographic location, since the maximun in the spring-summer period, can vary from 33°C to 36°C. In winter, for some locations can fluctuate between 20°C and 21°C. The minimums for both periods vary respectively from 16°C to 24°C degrees, in summer, and 8°C to 23°C, in winter (SILVA et al., 2008). Ab'Saber (2017) remarks that based on this well-defined seasonality, the humidity rate may be lower in the dry winter and higher in the rainy summer.

To Sano et al. (2008), the soil and its geology as well as topography are directly linked to the climate and contribute to its fertility characteristic, thereby, influencing the formation of vegetation, e.g., during the wet season, soil can lose important nutrients and become weathered. In line with Eiten (1990)'s studies, for certain regions where the soil is composed of acid clay aggregated with sand, rain penetrates easily, resulting in a deficit of water storage for the plants during the drought season. This can often happen on sites where the depth layers reach about two to three meters, in flat areas, making it difficult for the roots to absorb water.

Thus, thick-trunked trees or shrubs are able to fetch water in deeper layers that are still wet, and thus tend to spend more energy and absorb less water. Therefore, some plants need to get rid of all foliage, especially those with shallow roots, says the author in his work. According to Reatto et al. (2008) and Ab'Saber (2017), the soil is composed of latossol predominantly, found in sedimentary and crystalline areas, but also other different types of rocks, some richer or poorer in iron and magnesium, or derived from basic rocks, with low natural fertility, among other kinds of land.

Furthermore, *Cerrado* is home to several springs along the length of the biome, as it is located in eight of the twelve river Brazilian basin regions, specified by *Divisão Hidrográfica Nacional, instituída pelo Conselho Nacional de Recursos Hídricos* (CNRH) (ANA, 2014). A study by EMBRAPA (2008) highlights the flow of these hydrographic regions that bathe the *Pantanal* since the main rivers originate in the *Cerrado*. It is also responsible for supplying other regions, such as the hydrography of the *Parnaíba* River.

Faced with these diversities of natural resources, *Cerrado*'s landscape has been occupied by cultivated areas, pastures, hydropower plants, mining, and other anthropic activities (SANO et al., 2008). It is estimated that at least 40% of the entire range has been converted to important agricultural fields, specifically annual crops such as soybeans and corn (REATTO et al., 2008). Although agriculture is important and indispensable for the food supply in Brazil and in other countries, it is imperative to protect the use of land, water bodies, riparian forests, and other natural resources of *Cerrado* (AB'SABER, 2017).

2.1 Observation of land use and land cover in the Cerrado

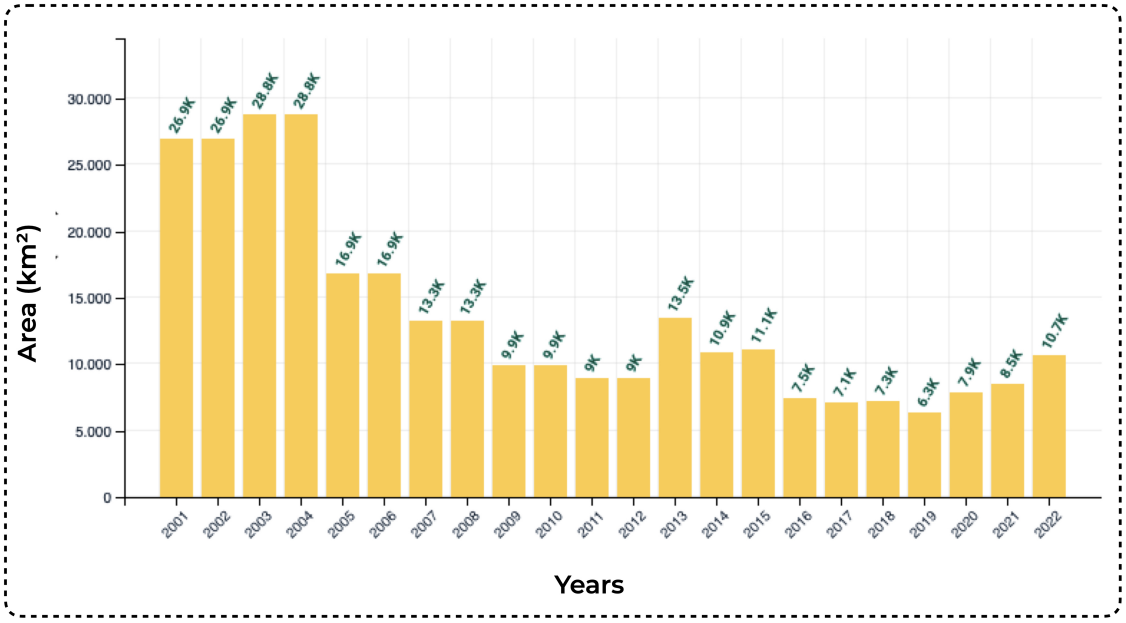
Given the progress in the production of grains and meat, untouched areas are deforested and with them, further landscapes are transformed (MAURANO et al., 2019). For this reason, the Cerrado observation is important. Several studies from the RS community use spatial and temporal data to identify deforestation areas or even follow a crop time series, for instance. It is very challenging since this biome has heterogeneous vegetation formation (NEVES et al., 2021), and the "spatial variability and spectral similarity among these phytophysiognomies" (FONSECA et al., 2021).

According to Fourest et al. (2012), the RS history was born in the 19^{th} century, with the emergence of aviation photography in military missions, which cameras needed a large amount of film to record. The main advance occurred with the adoption of cameras with digital image sensors, from which other forms of lenses, sensors and storage were improved and implemented. This term, therefore, refers to the methods used that determine the distance of properties of natural or artificial objects using the radiation reflected or emitted by them. Thus, there are two categories of sensors, passive and active. The first sensor type captures the reflectance of radiation emitted by the sun on objects on Earth. Whereas the second type of sensor emits radiation and captures the response produced by objects on Earth's surface.

At INPE, there are meaningful research projects accountable for aiming to monitor

LULC in Brazilian biomes, mainly concerning the deforestation increment. This is the mission of the PRODES and DETER[1]. According to INPE (2019), deforestation is mapped when there is an obliteration of the forested formation regions by anthropogenic actions, i.e., conversion of primary forest into another land use. It is occurring by clear-cutting when the whole forest covering is removed in a short time. In addition, there is deforestation due to forest degradation, which is done progressively. Initially, it is unsociable, because it happens slowly since the initial phase is the selective extraction of wood, removal of wood, burning, and clear-cutting. As a result of this, Figures 2.2 and 2.3 have introduced the increase in deforestation along the *Cerrado* biome from 2001 until 2023, considering clear-cutting cases.

Figure 2.2 - Deforestation increments in the *Cerrado*.



The x-axis shows the increments over the years, whereas y axes introduced the total area deforested for each year.

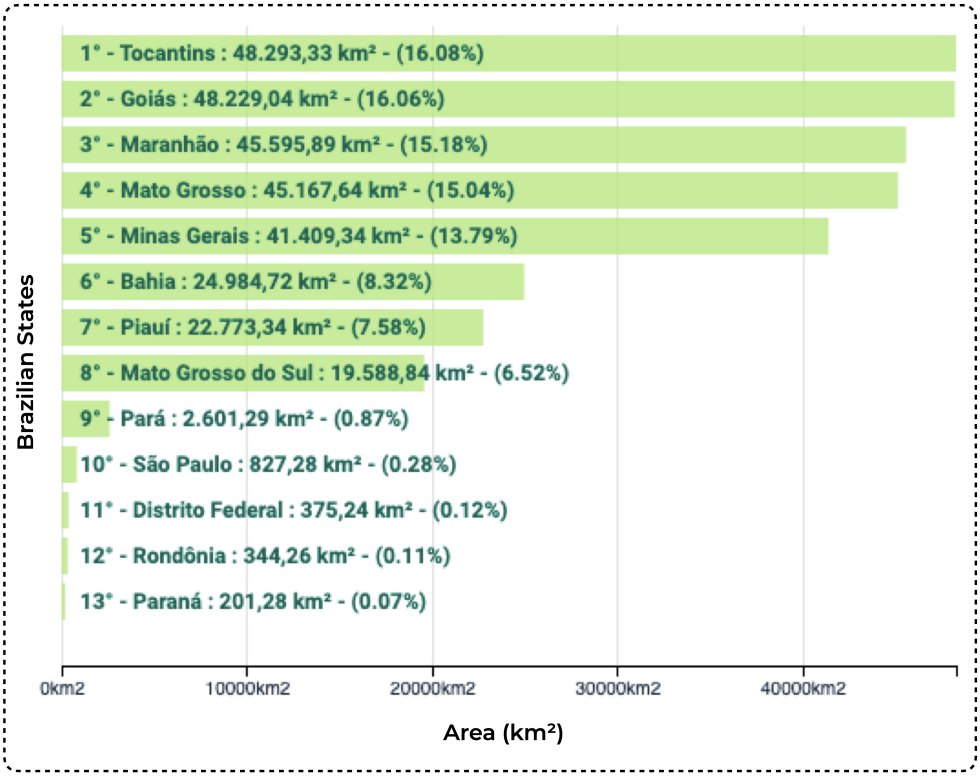
SOURCE: Adapted from INPE (2023).

Figure 2.2 shows that since 2019, when the biome registered a smaller deforested area, there was an increment of deforestation. In 2022, an area of 10.7 thousand square kilometers (km^2) was deforested, as reported in 2014, with a similar area of 10.9 km^2. In the last year, 2022, *Tocantins* registered the higher percentage of

[1]PRODES means Project for Monitoring Deforestation in the Legal Amazon by Satellite, and DETER means Deforestation Detection in Real Time Project (INPE, 2019).

deforestation, which correspond to 16.08% of the total increments, followed by *Goiás*, *Maranhão*, and *Mato Grosso*, as shown in Figure 2.3. It is important to remember that *Tocantins* belongs to an agricultural area well-known as MATOPIBA, which comprises the States of *Maranhão*, *Piauí*, and *Bahia*.

Figure 2.3 - Deforestation increments in the *Cerrado* per state.



The axes x and y contain respective data in regard to the deforested areas in square kilometers, for each Brazilian State. Green bars represent the percentage of annual increments compared among States.

SOURCE: Adapted from INPE (2023).

This area has great crop production of soy, corn and cotton. Also, this site is selected due to climatic factors, *Chapadas* geomorphological units and Depressions, and Soils of the Latosols and Neosols orders (EMBRAPA, 2014b). The suppression of native *Cerrado* vegetation is concentrated in extensive private areas (IPAM, 2022), despite this the conservation of the biome depends on actions originating from the public power that carries out monitoring, inspection, and penalization policies.

According to INPE (2019), these biomes monitoring projects use a set of remote

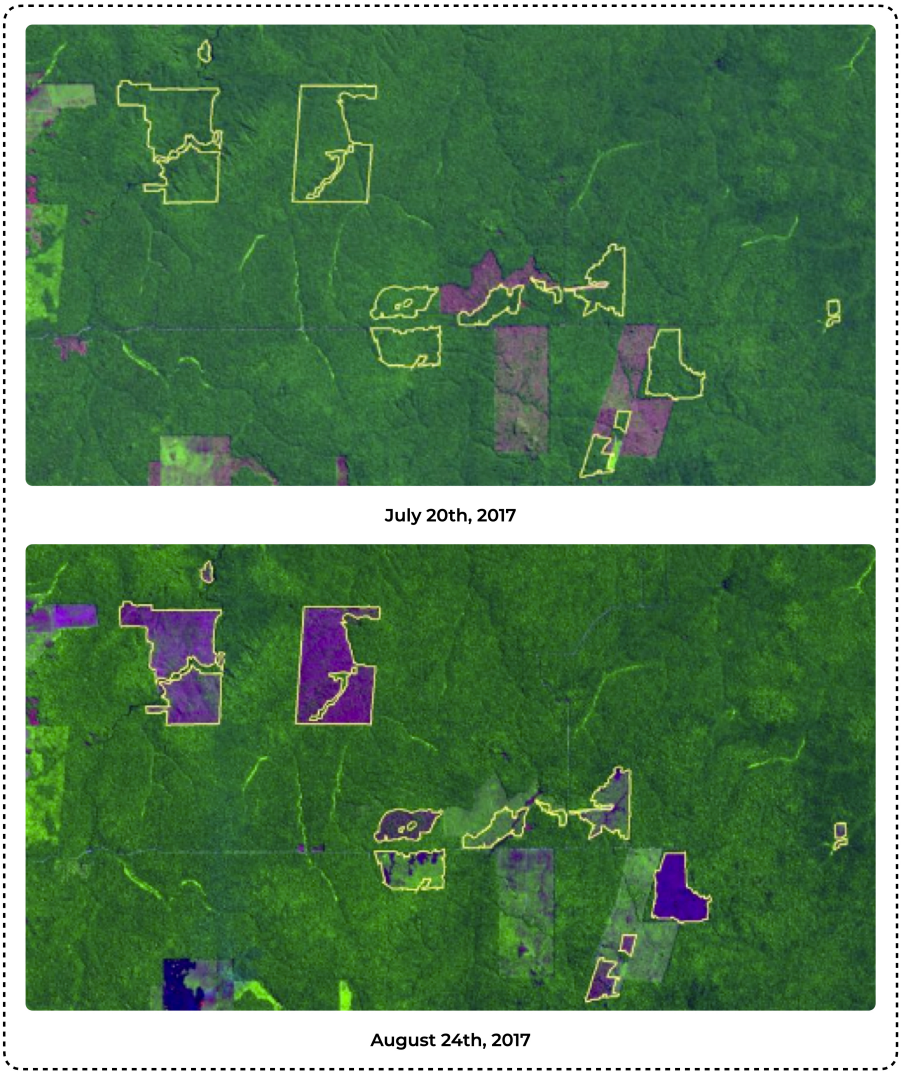
sensing imagery (RSI), analyzed by an expert team who carries visual interpretation out of the images considering attributes such as tonality, shape, texture, and context. Thus, polygons are drawn in the raster, as illustrated in Figure 2.4. In order to avoid those already mapped sites and regions with non-forest, masks are employed. As far as the collection of rasters is concerned, those with the lowest cloud cover are considered, from the month of August on-wards. After selection, the data is highlighted to accentuate the clear-cutting areas. These procedures are performed manually in the geospatial data processing software called TerraAmazom (MAURANO et al., 2019).

DETER is a system developed to detect daily changes in the coverage of forest areas when occurring clear-cutting bigger than three hectares. As with PRODES, the detection is carried out by specialists who carry out visual interpretation. Moreover, it is used the Linear Model of Spectral Mixture (MLME), in order to get the ground and shadow fractions applied to images of 64 meters of spatial resolution (MSR), with a composition of colored bands of the Landsat, ResourceSat and CBERS satellites. Deforestation maps produced by PRODES in the previous year are used to help detect changes in vegetation cover. Such as bare soil, low vegetation, and coverage traces of burns or degradation (INPE, 2019).

In an effort to classify the types of LULC, the TerraClass *Cerrado* was conceived by INPE and the Brazilian Agricultural Research Corporation (EMBRAPA). This project is based on the maps produced by PRODES, which contributes to the monitoring of 15 classes, such as agriculture, building areas, mining, water bodies, and primary and secondary natural vegetation (EMBRAPA, 2021). In general, the methodology adopted consists of pre-processing and segmenting images, visual interpretation, and supervised and unsupervised classification algorithms applied to time series and spectral information data extracted from Landsat 8, Terra (MODIS), and Aqua (MODIS) satellites, whose spatial resolution are 30 and 250 meters respectively (INPE, 2013).

On the other hand, Artificial Neural Network (ANN) algorithms have been employed in several domains of forecasting and classification for RSI data, due to, mostly, learning, adaptation, and generalization knowledge capabilities. The ANN can detect and extract patterns from the data and learn with these features, mainly, generating this knowledge to the unknown data. With regard to the image classification task, the ANN model is fitted based on the training labeled or unlabeled dataset, i.e., supervised or unsupervised learning (PLAZA et al., 2008).

Figure 2.4 - Deforestation polygons mapped overlaid on a satellite image.



Yellow polygons represent the further deforestation in *Boca do Acre*, located in southest of Amazon between 2017 and 2018.

SOURCE: Adapted from INPE (2019).

2.2 Artificial neural networks

ANN, conforming Haykin (2001), is an area of study in the field of Machine Learning (ML). It is inspired by the human brain's capability of data processing in order to pattern detection of high and low levels, and produce an output corresponding to the inputs. In other words, an ANN is a mathematical model composed of non-linear functions, known as neurons, in which every single neuron performs in parallel, however, connected to each other in layers (PACIFICI F., 2008).

A single neuron is composed basically by input data, bias, weight, and activation function, as described by Haykin (2001) in two equations,

$$u_k = \sum_{j=1}^{m} (w_k j x_j) \tag{2.1}$$

and

$$y_k = \varphi(u_k + b_k) \tag{2.2}$$

since, x_j is the input signal to the neuron; w, is the synaptic weight of neuron k for each input; u_k is the output of this operation which, as far as it is concerned, is added to the bias, b_k, which has the effect of an affine transformation to the output; and φ is the activation function, seen in the literature as a threshold for the output y_k of the artificial neuron. Therefore, a neuron can receive one or more input signals and produce an output signal, which feeds neurons in the next layer. Thus, these units could be connected fully or partially in different layers to each other (PACIFICI F., 2008).

Concerning the activation function, there is precisely one for each kind of issue, such as classification and regression. However, in classification tasks usually are employed *ReLu*, for hidden layers in order to change the negative values to 0 and maintain positive values, hence it does not activate every neuron at the same time; *Sigmoid* regularly used in the output layer for binary classification tasks, to produce the parameter ϕ of the Bernoulli distribution, setting values close to 0 assumes 0, and those values close to 1 assumes 1; *Softmax*, used in the last layer of the network, produces a probability distribution over a discrete variable given a number of classes, accordingly, it is used for multi-class classification tasks (GOODFELLOW et al., 2016).

According to Haykin (2001) the learning process takes place by the way the synaptic weights adapt to the input data at each iteration. This learning, in general, can carry out in a supervised, which is introducing to the network a vector (professor) containing label to respective input data and thus assisting it in correcting of learning error rate; or unsupervised, whose learning process does not require labels to the input data, i.e., the model learns patterns and ways to represent every class, such occur in clustering algorithms.

2.2.1 Deep learning

Important advances have contributed to increasing in the learning ability of ANN models, such as deep learning (DL) which brings a paradigm of robustness and efficiency, regarding the network architecture and learning. The DL algorithm stands out for the hidden layers, densely connected to each other, in order to process the signals coming from the input layers and produce outputs in the last layer. The training of these networks has a high cost of using computational resources, however, this cost may be lower when the model is trained. Yet, this is conditioned to the type of architecture, volume, and size of input data. Given that DL models are designed to process a large volume of (GOODFELLOW et al., 2016) data.

In Haykin (2001), one of the aspects that makes these models more efficient and with learning capacity is the backpropagation algorithm. It is employed during the training phase, aiming to obtain the error signal e from the output of neuron j, in the interaction n, which is equivalent to subtracting the desired response d from the output of neuron y, as described in the equation below.

$$e_j(n) = d_j(n) - y_j(n) \tag{2.3}$$

These outputs are produced by the propagation of the input signals through the neuron layers, forward. Afterward, the sum of the total energy of errors is obtained from the sum of the instant value $\frac{1}{2}e_j^2(n)$ of each neuron in the output layer,

$$\mathcal{E}(n) = \frac{1}{2}\sum_{j \in C} .e_j^2(n) \tag{2.4}$$

Therefore, the average of $\mathcal{E}(n)$ produces the average error energy $\mathcal{E}_{med}$. This mean squared error energy function is the result of the sum of all $\mathcal{E}_med$, for each interation n, normalized in relation to the amount of training set standards (N), in the following equation.

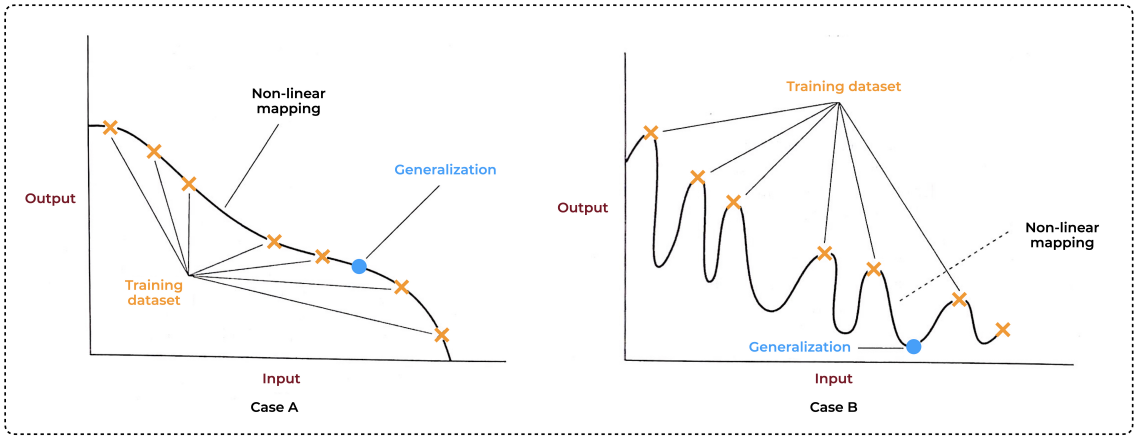
$$\mathcal{E}_{med} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{E}(n). \tag{2.5}$$

Thus, the network's output layer is calculated as the error energy, which is employed for the backpropagation of synaptic weights adjustment and consequently minimizes

the cost of energy. This is an efficient method to increase the learning rate in multilayer networks, such as multilayer perceptron (MLP). This feedforward model is widely used as a basis for many decision-making problems, as the network learns to bring forth results that approximate the inputs, through a non-linear mapping between inputs and outputs (PACIFICI F., 2008).

For Goodfellow et al. (2016) the DNN learning process is to understand patterns over the input data, but also to be able to generalize its knowledge to "unknown" data. In agreement Haykin (2001), this process can be understood as a learning curve adjustment. The neural network operates non-linear interpolations between the input and output data, whose outcomes predicted have the inclination to be closer to the correct labels of the data, regarding supervised tasks. Figure 2.5 exemplifies how the learning process of a neural network occurs.

Figure 2.5 - Nonlinear input and output mapping of a neural network in the learning.



Case A: good generalization based on the training data; Case B: poor generalization, overfitting regularization to training data.

SOURCE: Adapted from Haykin (2001).

Observe Figure 2.5, both cases are concerning the model's generalization ability over the training step. In the first case, the ANN has a good performance, whose generalization follows the training dataset outflow, whereas, in the second case, the model learned a pattern divergent from what was expected. Because of this, the balancing of the adjustment and regularization of the network's parameters is essential, although, excess of control over the results can harm both the ability to learn and the ability to generalize learning.

According Goodfellow et al. (2016), Dropout is a low-cost and efficient regularizer. It works based on affine and non-linear transformations, void the output value of the processing units by multiplying them by zero. It allows the neural network to learn more representative features of the classes, in order to evenly distributed among the neurons. For example, for tree classification, a neuron learned that a tree has leaves, then the dropout omit this information to the next neuron to learn another characteristic, such tree has branches as well.

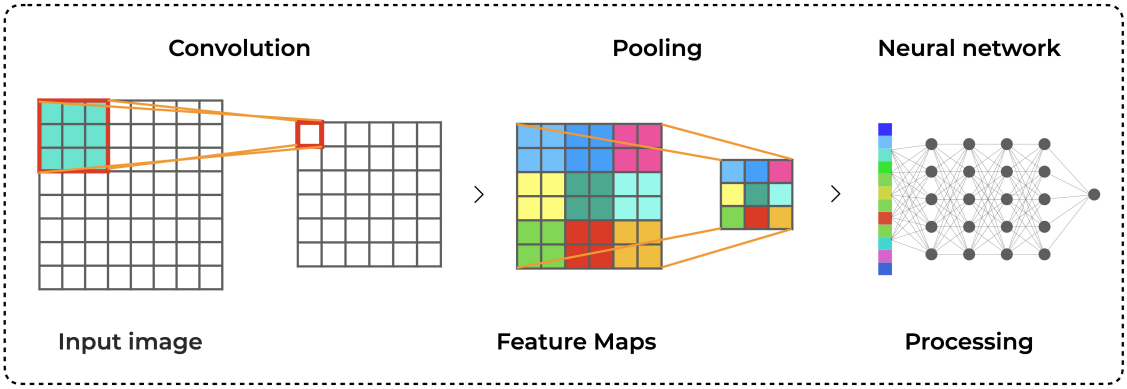
Furthermore, during DNN training, the optimization of the cost function is a point to ensure that the other steps achieve advantage learning rates. Therefore, solutions such as stochastic gradient descent (SGD) and adaptive moments (Adam) are widely used for divergent problems. The SGD works with a fixed learning rate, decreasing it over the training time, and reaching lower true values for the cost function, also allows convergence to occur even with a large dataset. Adam, however, applies first and second-order bias corrections taking into account the points of origin, as well as estimating the first-order gradient by incorporating the momentum rate (GOODFELLOW et al., 2016).

2.2.2 Convolutional neural networks

Convolutional neural networks (CNN) are a type of MLP specialized for processing one-dimensional (1D) time series data with regular time intervals and highly invariant two-dimensional (2D) images (GOODFELLOW et al., 2016). The main feature of CNNs is the convolution operation, which extracts features from the input data automatically. These networks are inspired by the visual cortex of a brain, since the ability to extract patterns from low and high level. Basically, a CNN contains the convolution, pooling, and a MLP layers, as shown in Figure 2.6.

In Figure 2.6, the pixels of the input image are accessed by a window that slides across the entire image, also known as the local receptive field, in which are multiplied by a kernel (matrix) of the same size. The outcome, thus, is a feature map. Moreover, according to Haykin (2001), pooling layers are used after the convolution, which summarize the feature maps information getting the maximum value or calculating the average of the values within a window stride. Consequently, this operation reduces the resolution and the sensitivity of the character map output regarding image distortions, hence another feature map is created. Afterward, those feature maps are flattened, i.e., the feature maps are converted into a single vector, and processed in the dense layers, a fully-connected neural network.

Figure 2.6 - An example of convolution and pooling operation.



The input image is represented by a matrix, whose square means a pixel; the local receptive field has been depicted in cyan blue delimited in red edges; in the pooling step, nine different colors represent the local receptive field in each stride; in the last feature map, these same colors mean the new pixel values coming from pooling operation; the connected gray cycles intend the fully-connected neural network.

SOURCE: Adapted from (GOODFELLOW et al., 2016)

CNN models are widely used for image classification, either in terms of pixel level or just image labeling. The first approach is known as pixel-based classification, according to Soille (2004). This method, for Liu and Xia (2010), means classifying each pixel of the image individually, in contrast to another method that aggregates each pixel of the image, considering spatial and spectral information in objects, segmentation techniques, and then classifying them. Known as object-based classification. With regard to the second approach, a single label is assigned to the entire image.

2.2.2.1 Metrics for performance evaluation

There are several models of CNNs dedicated to CC and PBC, such as DenseNet, VGG16, U-Net and DeepLabv3, for various image classification tasks. However, to evaluate the performance and accuracy of these models, metrics are used. In line to this, the most metrics applied are:

a) Accuracy. The number of correct predictions per the total number of predictions (SCIKIT LEARN, 2023a);

b) F1-Score. This metric is defined as

$$f1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

c) Intersection-Over-Union (IoU). A common evaluation metric for semantic image segmentation (NIU, 2021), defined as

$$iou = \frac{TP}{(TP + FP + FN)}$$

Precision is the ratio $\frac{TP}{TP+FP}$, where TP is the number of true positives and FP is the number of false positives. Recall is the ratio $\frac{TP}{TP+FN}$, where FN is the number of false negatives. All these metrics range from 0 to 1, where 0 is the worst performance and 1 is the best one (SCIKIT LEARN, 2023b). Also, qualitative aspects of the model were taken into accounts, such as understanding the context of the images and the decision-making logic in this classification task. Regarding the IoU metric, the intersection between the true and predicted matrices by the model is calculated, hence are considered the TP, as well as FP and FN values (TENSORFLOW, 2022).

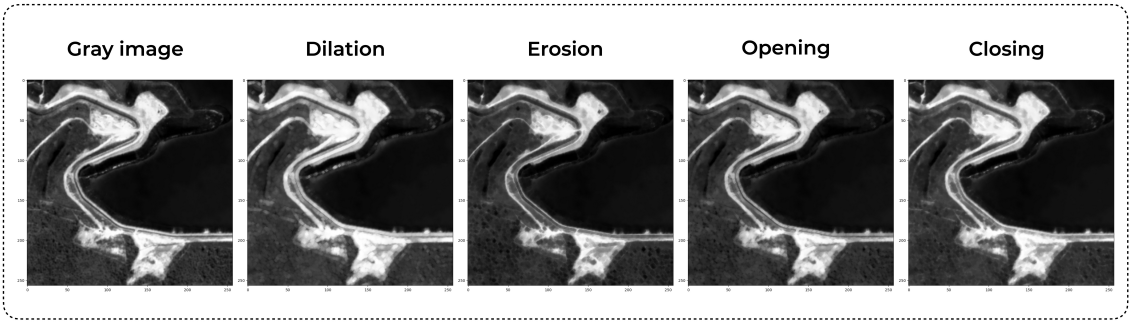
2.2.2.2 Morphological operations

Another way of image processing is the transformations in the image morphology, which is traditionally applied in order to segment, classify, or merely improve the image's quality. Soille (2004), defines morphology as a technique for studying the spatial form of object structures and is characterized by the cross-fertilization of theory, methods, and algorithms. According to Sreedhar and Panlal (2012), the most use morphological operations are dilation, erosion, opening and closing, usually applied to a gray input image, reproducing a new output image.

In Sreedhar and Panlal (2012), dilation stretches the width of those regions where pixels are maximum values, and shrinks regions where pixels are smaller or negative values, i.e., bright regions are dilated relative to dark. Erosion, on the other hand, removes small regions of maximum values and increases the width of areas with pixels of minimum values. The opening operation is obtained by erosion followed by dilation, while the closing is the reverse process, dilation followed by erosion.

In Figure 2.7, note the dilated and closed images, the entire light-colored region of the bare soil has been stretched, and omitting some dark details within the bare soil area. However, in the images that were eroded and opened, darker details became more evident, as seen in as can be seen in the shadows of the savanna vegetation surrounding the exposed soil area, as well as the vegetation. The texture of images is smoothed, especially dilated and closed ones. Gray tones change to light when dilated, and darker when erosion is applied.

Figure 2.7 - Morphological operations.



Four morphological operations applied in a grayscale satellite image are presented: dilation, erosion, opening, and closing.

An application example is to combine these operations with thresholds for image segmentation. Such as Otsu method by Otsu (1979) which, from a gray level histogram, can choose an optimal threshold automatically, based on the integration, furthermore is characterized by its non-parametric and unsupervised nature. Reddy et al. (2022), had used these and other algorithms to create and annotate a dataset in order to train a CNN for semantic segmentation.

2.3 Related work

There are several researches on the LULC classification in the *Cerrado* biome that apply DL and ML techniques for processing RSI data, both concerning contextual, pixel-based, and object-based classification. According to Fonseca et al. (2021), among the ML algorithms, the most used is RF, from high resolution images to low spatial resolution, while for DL, it is U-Net neural network. Those algorithms are efficient, each one for a specific case. RF works well with small amount of data, whereas U-Net and other robust networks, work even better for a large dataset (WANG et al., 2022a).

In Zhang et al. (2021) was considered the RF and SVM algorithms combined with Bayesian optimization parameters, aiming to propose a systematic method to automatically tune the hyperparameters for classification tasks. In this experiment was considered an image set composed with images from the Sentinel-2A/B satellite, which has a spatial resolution of 60 meters. The first step concerns the preparation of the images, pre-processing and creation of labels images manually, optimization of the hyperparameters of the model, and, thereafter, the training of the model. Regard to second, the performance evaluation, consisted of comparing the optimized model

with the traditional SVM and RF method not optimized through metrics. The results, therefore, are optimistic and show a small advantage of using the optimized model over other techniques.

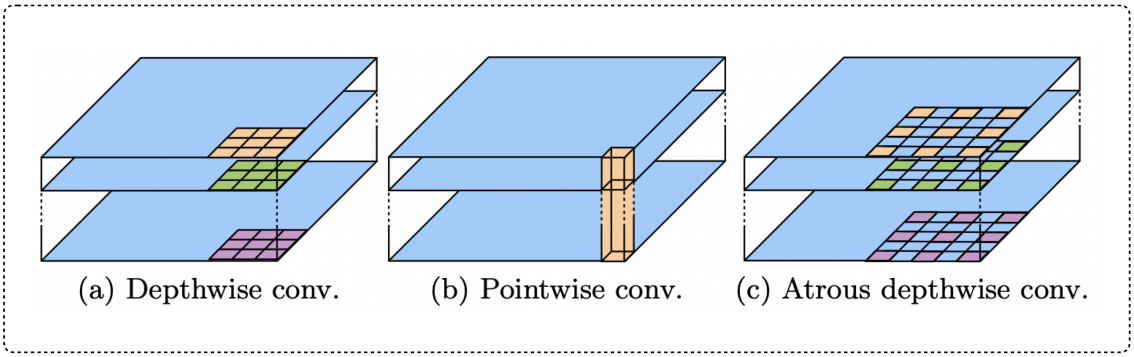
On the other hand, Hänsch and Hellwich (2018) had proposed the use of stacked random forests (SRF) for the classification pixel-based of Polarimetric Synthetic Aperture Radar images. It combines two ensemble learning strategies, average the output of several base the individual trees within the RF, and the estimation of the base of these trees as features for a subsequent model. This experiment was applied to two very different datasets, the first is a fully polarimetric images, with approximately 1.5 meters of resolution, by E-SAR sensor satellite, and the second containing a dual-polarimetric images of 1 meter of resolution by TerraSAR-X satellite. For both datasets, reference masks were created manually. According to the authors, the results were significant and accurate, since they contain considerably less label noise, smoother object boundaries, as well as present a good convergence rate in the first levels, despite convergence to saturate quickly.

As far as semantic segmentation is concerned, DL models such as U-Net and DeepLabv3plus are widely implemented for RSI data. These kinds of networks reduce the resolution of feature maps by the pooling layer, down-sampling, detecting the threshold between each element of the image. After, the network does the opposite process, recovering the spatial information of the image, up sampling, taking into account the boundaries between groups of pixels that share similar features. This structure is known as encode-decode Chen et al. (2018).

The U-Net Network was designed by Ronneberger et al. (2015), extended from Long et al. (2015), as a semantic segmentation model, applied to a set of biomedical images. According to the authors, it is characterized by having many layers of down-sampling and up-sampling and, mostly, by the ability to learn with few training samples, although a diversity of representations of these examples is necessary. DeepLabv3+, is proposed by Chen et al. (2017) and extended by Chen et al. (2018). Unlike U-Net, applies multiple *Atrous* convolutions layers in parallel, presented in Figure 2.8, whose rates of Spatial Pyramid Pooling are diversified throughout the network, in order to learn the context of the input images at many scales.

Neves et al. (2021) propose a hierarchical physiognomies mapping of the Cerrado considering in two levels of vegetation, such illustrated in Figure 2.1. The study site was the *Brasília* National Park, Federal District, Brazil. Thereafter, eight datasets were taken into consideration composed of images from the WorldView-2 satellite

Figure 2.8 - Different Convolution Methods.



a) A standard kernel convolution 3 × 3 is done; b) combines the convolution outputs in depth in the channels; and c) Atrous convolution with the rate of 2, i.e. means a dilated convolution process with a factor of 2.

SOURCE: Chen et al. (2018)

camera, in which are 2 meters of spatial resolution. In addition, for each image of the level one classes, segmented samples were generated that highlight the level two classes, using the simple linear interactive clustering method (SLIC). Both images and labels were used to train a version adapted of the U-Net network, replacing the Softmax activation function with Sigmoid allowing probabilities to be independently assigned to each class and pixel of the image.

Nevertheless, Zheng and Chen (2021) adapted the U-Net in two segmentation approaches, for binary and other for multiclass classification. In this experiment, the "neighborhood voting" method was used to determine the category of uncertain pixels based on the spatial heterogeneity and homogeneity of the samples. For it, a dataset named GaoFen-2 was used, in which brings together 150 red, green, blue and infrared spectral rasters, with an 3.24 meters of spatial resolution, arranged in five classes, built-up, farmland, water, meadow, and forest. The reference masks were drawn manually. In contrast to the work of Neves et al. (2021), the classification results by the multiclass model are better than the binary model, which were slightly high and strengthen the good performance of these algorithms to process RS images.

Pedrayes et al. (2021) have compared U-Net and DeepLabv3plus performances in two datasets, whose samples have different spatial resolutions and sensors. The UOS2 set is composed of 1958 images of 256 × 256 pixels, with 10 MSR, Sentinel-2 satellite;

UOPNOA has 33,699 images of 256 × 256 pixels, 25 MSR. Both two datasets use the same classes and color palette for the reference masks, as well as corresponding to the same study area, the northern part of the Iberian Peninsula plateau in Spain. Although DeepLabv3plus is limited to working with rasters of up to three color channels, the results demonstrate the superiority over U-net in the first dataset. However, for the second dataset both architectures did not achieve a good performance, due to the low spatial resolution of the rasters.

In Du et al. (2021) a detailed study is carried out deploying DeepLabv3plus integrated to the object-based image analysis (OBIA) method to improve pixel-based classification in high spatial resolution images. Two datasets were taken, Vaihingen dataset and Potsdam dataset, with spatial resolution of 9 and 5 meters, respectively. The outputs are combined in order to classify the segments in the scenes, using the Random Forest classifier on hand-crafted features algorithm. Thus, it was estimated the labels of these outputs through conditional random field (CRF) framework. The performance and quality of the classification was assessed, comparing the use of other networks integrated to the framework. According to the authors, the results shown better with DeepLabv3plus integrated to the proposed method.

Note that the cited works applied different semantic segmentation techniques to manually labeled databases, i.e., reference masks presented to the model to teach it how to classify each pixel of the image correctly. However, a dataset designing requires very time-consuming in terms of mask generation and labeling (PORZI et al., 2019; NIU, 2021). There are proposals that consist of labeling a small amount of data and training a model that will do this for the other unlabeled samples, known as few-shot learning, whose models can be trained with fine-tuning or from scratch (SAENKO et al., 2010; LECUN et al., 1998). Even with small dataset, the generation of reference masks with good semantic quality requires a lot of preparation time, such as the INPE (2019) methodology presented in the 2.1 Section.

In Niu (2021), was used a weak supervised learning in order to generate and label the reference mask. According to the author, in many DL approaches that use this method, the time spent to prepare the dataset is shorter and the accuracy scores of the models are high. This method consists of two categories, response area extraction method and pseudo-label based method, which uses CRF or fully convolutional neural network (FCN) models for mask and bounding box generation. However, the author structured the approach in three steps: i) feature extraction from a residual network module; ii) generation of masks from features, pseudo-labels; iii) training

with pseudo-marked masks and their respective images. The mask generation process relies on the Gaussian Mixture Model (GMM) to classify the foreground and background sampling point. Results were slightly better than U-Net, using IoU, Precision, and Recall metrics.

According to the studies presented, the best accuracy classification is achieved with high spatial resolution satellite images, as well as a dataset that presents a great diversity of elements (NEVES et al., 2021; DU et al., 2021). Moreover, the performance of the models proves to be more efficient using the multiclass classification method instead of the binary classification (ZHENG; CHEN, 2021). However, it was noted that there are no many datasets about the *Cerrado* biome from high-resolution satellite images and covering a more diverse area, available and ready for use with DL techniques. Another important aspect is the generation and labeling of reference masks to support the training of semantic segmentation models, it is a step that requires a lot of time. Although approaches, such as Niu (2021), are efficient in some cases, there are still bottlenecks and different methods that can be explored it. Therefore, in terms of *Cerrado* biome Fonseca et al. (2021), Câmara (2020), and Simoes et al. (2021), corroborate that LULC classification in this biome is not a trivial task.

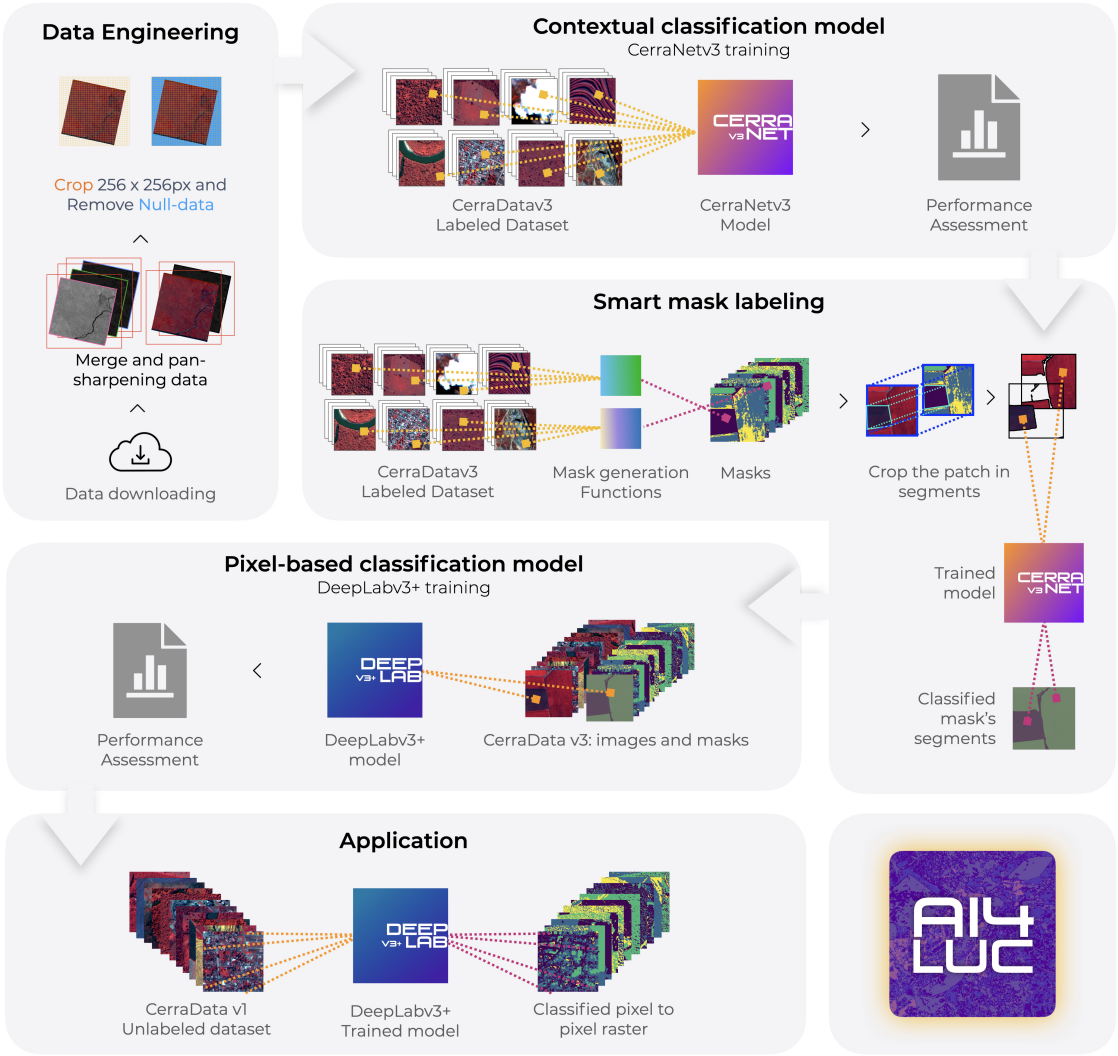
2.4 Final remarks about this chapter

This chapter presented the theoretical background with the main concepts and techniques associated with this master dissertation. Related studies were also presented as well as some of their limitations. In the next Chapter, the AI4LUC method is presented in detail.

3 AI4LUC METHOD

The Artificial Intelligence for Land Use and Land Cover Classification[1] (AI4LUC) is a method based on the methodology of the DETER, TerraClass, and PRODES projects (INPE, 2019), in terms of image interpretation criteria, such as context information and texture, concerning to pixel-based classification. AI4LUC is arranged in three hierarchies: modules, components, and functions, as presented in Figure 3.1.

Figure 3.1 - AI4LUC pipeline.



Every gray widget represents a module of the method; gray arrows indicate the running sequence in the pipeline; the black arrows denote the next running step component within the module; the orange dotted lines indicate input data, whereas in the dark pink the output data.

[1]Project available on the repository: https://github.com/ai4luc/AI4LUC

AI4LUC is a general method developed based on the Python language, the Conda environment, and using geoprocessing packages for remote sensing images, such as GDAL and earthpy. Regarding DL and DNNs, frameworks/APIs such as Tensorflow, Keras, *scikit-learn*, and PyTorch were used. The experiments were run on the SDumont supercomputer, using Bull Sequana X1120 computing nodes where each one has 2 x Intel Xeon Skylake 6252 CPU, 4 x NVIDIA Volta V100 GPUs, and 384 GB of RAM. For experiments that do not require more GPU capacity, a second computer was used with 8GB of RAM, an Apple M1 processor with an 8-core CPU, 7-core GPU, and 16-core Neural Engine.

This chapter presents the procedures adopted in each module of the AI4LUC method, mainly with regard to the novel version of the CerraData as well as the Smart Mask Labeling module, and how the evaluation experiments were carried out.

3.1 Data engineering module

This section introduces the area of interest (AoI) over the *Cerrado* biome; procedures of raster preprocessing; and attributes with regard to Biome *Cerrado* Datasets (CerraData) datasets versions.

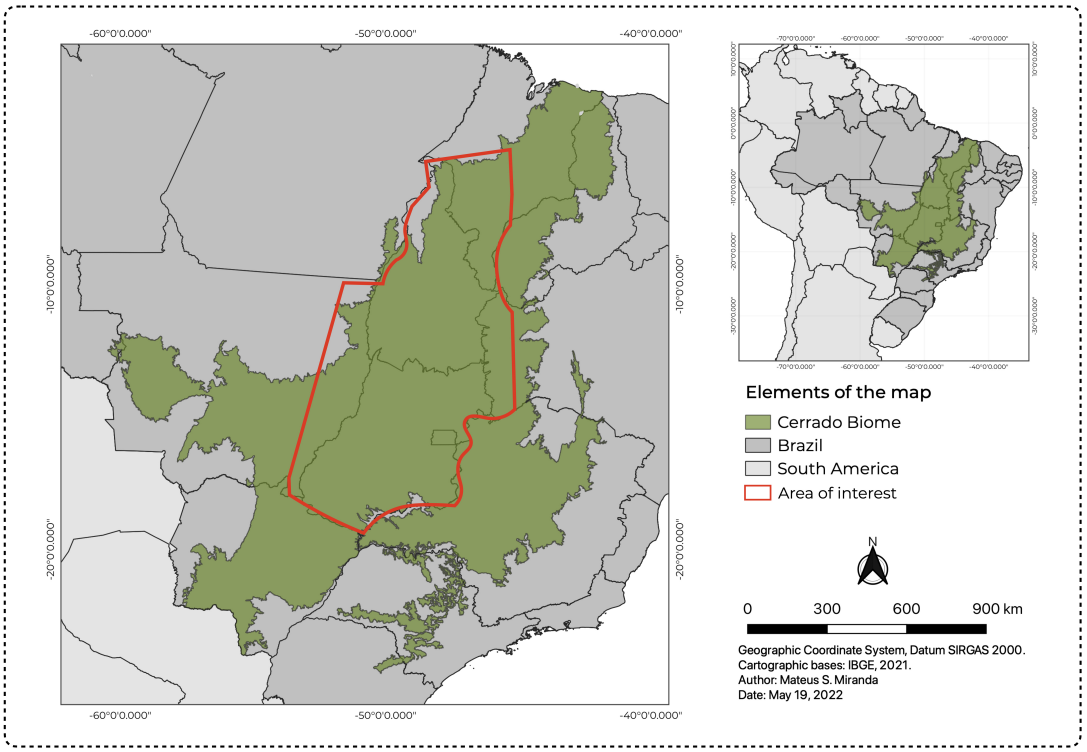
3.1.1 Data collection

A wide AoI over the *Cerrado* biome was defined, as delimited by a red line in Figure 3.2, covering the states *Bahia* (BA), *Goiás* (GO), *Maranhão* (MA), *Mato Grosso* (MT), *Tocantins* (TO), and the unit *Distrito Federal* (DF). This is approximately 44% of the entire biome extension. Each image corresponds to a path and row in the satellite observation grid. Therefore, a heterogeneous region has been selected, i.e., vast diversities LULC patterns. In total 150 rasters were obtained from INPE's Image Catalog platform[2], whose observation period corresponds from February 2020 to February 2022.

The rasters were registered by the wide panchromatic and multi-spectral camera (WPM) of the China-Brazil Earth Resources Satellite (CBERS-4A). This camera was chosen due to the spatial resolution of the near-infrared, green, and blue bands, which have eight meters of spatial resolution, availability of a panchromatic (PAN) band, which can be used to improve the resolution of the four other bands even more. Moreover, WPM provides orthorectified scenes, i.e., images with radiometric and geometric correction of the system refined by the use of control points and a

[2]http://www2.dgi.inpe.br/catalogo/explore

Figure 3.2 - Area of interest.



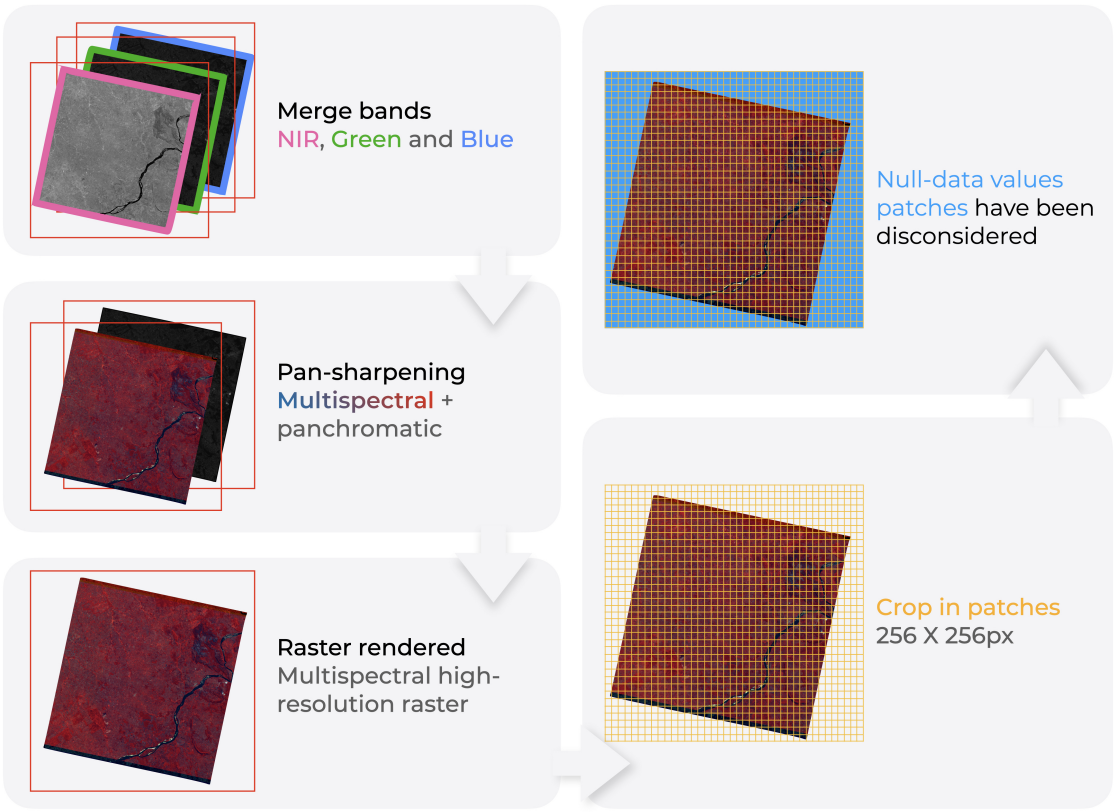
digital elevation model for all spectral bands.

3.1.2 Image preprocessing

Data pre-processing is shown in Figure 3.3. The spectral band of near-infrared (0.77 - 0.89μm; NIR) were associated to the R channel in a false color composition. The green band (0.52 - 0.59μm; G) to G channel, and blue (0.45 - 0.52μm; B) to B channel, from its respective raster. This color composition was chosen because it highlights the vegetation, in shades of red, from the other objects in the scenes, such as water, soil, and fire scars (NOGUEIRA et al., 2016). The stacking of these bands was performed by the *stack* function of the earthpy package.

Subsequently, utilizing the QGis platform, pan-sharpening with the Hue Saturation Value (HSV) method for the panchromatic (0.45 - 0.50μm; PAN) band and the multi-spectral image, producing an image with two meters spatial resolution rendered to RGB, to each raster. Afterward, these rendered images were cropped in patches of 256×256 px, whose geospatial information were kept. Thus, 45,000 samples were created. Though, images patches with null values were dropped out from the dataset,

Figure 3.3 - Procedure of raster pre-processing.



The gray widgets mean preprocessing steps, and the white arrows indicate the sequence and direction of execution.

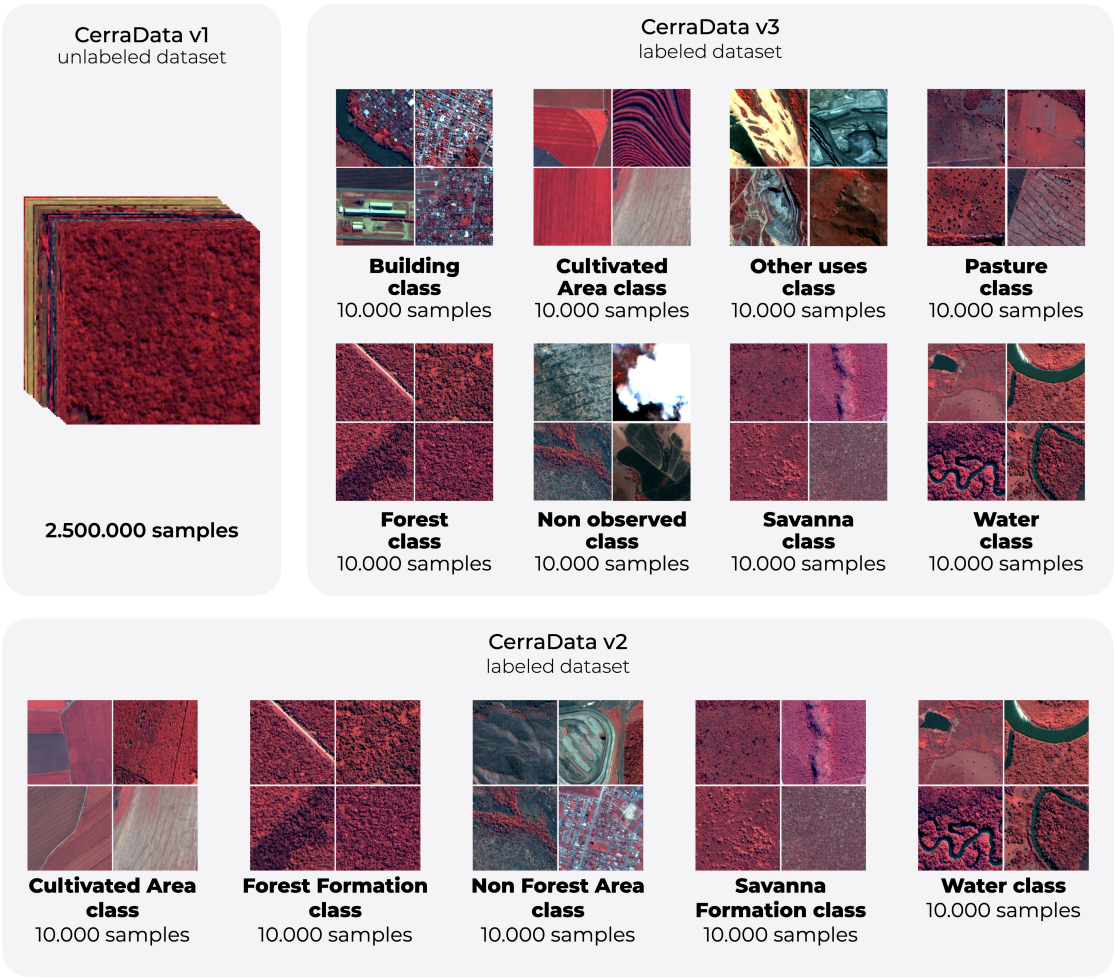
around 20,000 samples. In total, a large dataset with 2.5 million samples was created.

3.1.3 CerraData's datasets

The first version of CerraData has a total of 2.5 million of unlabeled images patches. This unlabeled dataset is organized into five sections corresponding to each origin state, where there are four subsections, i.e., batches of images. This huge version can be valuable for researchers who need a database with a significant amount of patches images. This offers access to researchers and developers to ready-to-use data for tasks of contextual classification and even semantic segmentation on the *Cerrado* biome.

The second version, CerraDatav2 (MIRANDA et al., 2022), has five LULC classes as described in Table 3.1, totaling 50,000 samples. The patches were labeled manually by visual interpretation, supported by some classified images of Cerrado physiognomies published on work of Neves et al. (2021), as well as descriptions of the types

Figure 3.4 - CerraData dataset versions.



Three versions of the CerraData dataset are presented, the earlier version is unlabeled, whereas the two latest are labeled.

of vegetation cataloged by Ribeiro and Walter (2008). Only one label is assigned to the entire patch, considering the class most predominant in the scene. However, there are exceptions. For example, the Water class has no sample in which it prevails. In this case, samples which have “some" water in the scene were labeled as Water. The entire procedure, since pre-processing until image labeling, lasted around three months.

Regarding the third version, CerraDatav3, eight types of LULC were considered, containing 80,000 patches, i.e., there are 10,000 patches per class. The patches were manually labeled via visual interpretation, as done in the second version. Moreover, this dataset was carefully audited by a committee composed of four LULC experts in the *Cerrado* biome, who are research members of TerraClass, DETER,

Table 3.1 - CerraDatav2, classes and their descriptions.

ID	Classes	Description
0	*Cultivated Area*	Pasture, agriculture, and planted trees.
1	*Forest Formation*	Predominance of arboreal formation and riparian forests.
2	*Non-Forest Area*	Urban areas, mining, fire scars, and dunes.
3	*Savanna Formation*	Woodland savanna, typical savanna, rupestrian savanna, shrub savanna, and *vereda*
4	*Water*	River, small lakes, dams, and fish farming ponds.

and PRODES. The class names, as described in Table 3.2, are based on the thematic mapping of the TerraClass project.

CerraDatav3 quality audit consisted of verifying and classifying 250 labeled samples from the dataset. The sample selection criterion consisted of the diversity of elements in the scene, especially when it comes from the Non-Observed Area, Savanna, Other Uses, and Pasture classes, due to the complexity of distinction and identification. These samples were randomly separated into two batches, one with 139 and the other with 111 patches. These batches were organized in a table, in addition to having a column for patches, they also had other columns for coordinates, ID and for assigning labels.

For the first samples batch, the committee experts had to assign up to three labels to each patch, while for the second batch, each sample could be assigned more than three labels. The image classification parameters adopted by the committee were the context and texture of the scene elements. The evaluations were carried out individually by the members who, subsequently, compared and discussed in meetings those divergences of labeling, in which labels mostly assigned by the committee were considered for the patches.

As a result, four new LULC classes were defined in relation to the second version of CerraData, as specified in Table 3.2. The Pasture class was created from the

Table 3.2 - CerraDatav3, classes and their descriptions.

ID	Classes	Description
0	*Building*	Building urban and rural areas.
1	*Cultivated Area*	Agriculture with one, two or more cycles, perennial and semi-perennial.
2	*Forest*	Arboreal formation and riparian forests, galleries, drought, and forestry.
3	*Non-Observed Area*	clouds, cloud shadows, fires and fire scars.
4	*Other uses*	Mining, rocky outcrops, beaches, and dunes.
5	*Pasture*	Grassland formations and herbaceous forage vegetation of cultivated species.
6	*Savanna Formation*	Woodland savanna, typical savanna, rupestrian savanna, shrub savanna, and *vereda*.
7	*Water*	River, small lakes, dams, and fish farming ponds.

Cultivated Area category, while Other Uses, Non-Observed Area, and Building were defined from the Non-Forest Area class. The classes' names and their specification were rearranged based on the LULC thematic classes from the *TerraClass* project, which has eleven categories, carried out by the audit committee.

3.2 A contextual classification model

The CNN for contextual classification play a large role, data management and mainly mask labeling assistance in the third module of the method. Therefore, CerraNetv3 and CerraDatav3 were considered. This dataset was split into two subsets, 79,200 samples for training and 800 samples for testing. Among the test subset samples, 250 of 800 samples were used in the dataset audit. Thus, in this chapter, the CerraNetv3

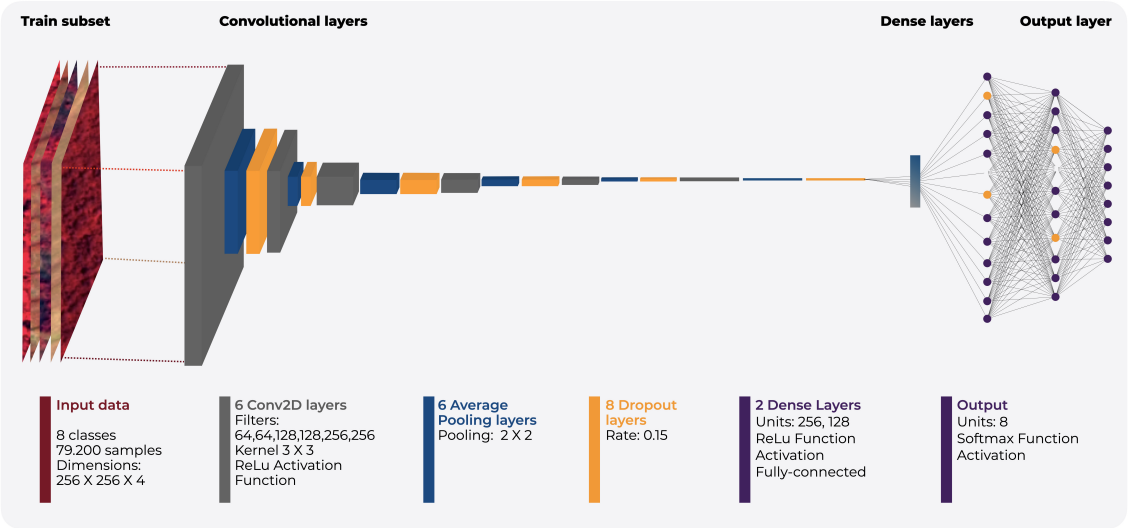
network specifications are documented.

3.2.1 CerraNetv3

The origin of the name CerraNet is a tribute to the *Cerrado*, designed to classify land use and land cover from satellite images. CerraNet is a deep learning neural network designed to initially perform binary classification of preserved and non-preserved areas of the biome (MIRANDA et al., 2021b), capable of working with satellite images with a spatial resolution of eight and ten meters. The network has four convolutional layers, plus two Maxpooling and Dropout layers between each convolution, as well as 3 dense and dropout layers and the output layer with a neuron.

The second version of the network, however, was updated to make multi-class classification regarding the state of water volume as Normal, Low and Critical, in supply dams from the state of *São Paulo*. Therefore, the main architectural changes were summarized in four convolutional layers with 128 filters in the first two and 64 filters in the last two, adding a dense layer and dropout, and the inclusion of two more neurons activated by the Softmax function in the output layer. In addition, works with satellite images with a spatial resolution of two meters (MIRANDA et al., 2021a).

Figure 3.5 - CerraNetv3 network architecture.



The convolutional (gray), Average Pooling (blue), Dense (purple), and Dropout (orange) layers are represented; the red dotted lines indicate input data, and the dark lines represent the connection between neurons in the layers forward.

As well as the previous versions, CerraNetv3 makes contextual classification, aiming

to manage data and create their respective labels for training the segmentation model. The main updates consist of the inclusion of convolutional layers and the reduction of dense layers, and the replacement of the *Maxpooling2D* layers with the *Averagepooling2D* layer, in order to optimize the model and increase accuracy in the contextual classification task. This version was designed for the CerraData v3, whose search for the best hyperparameters was chosen manually.

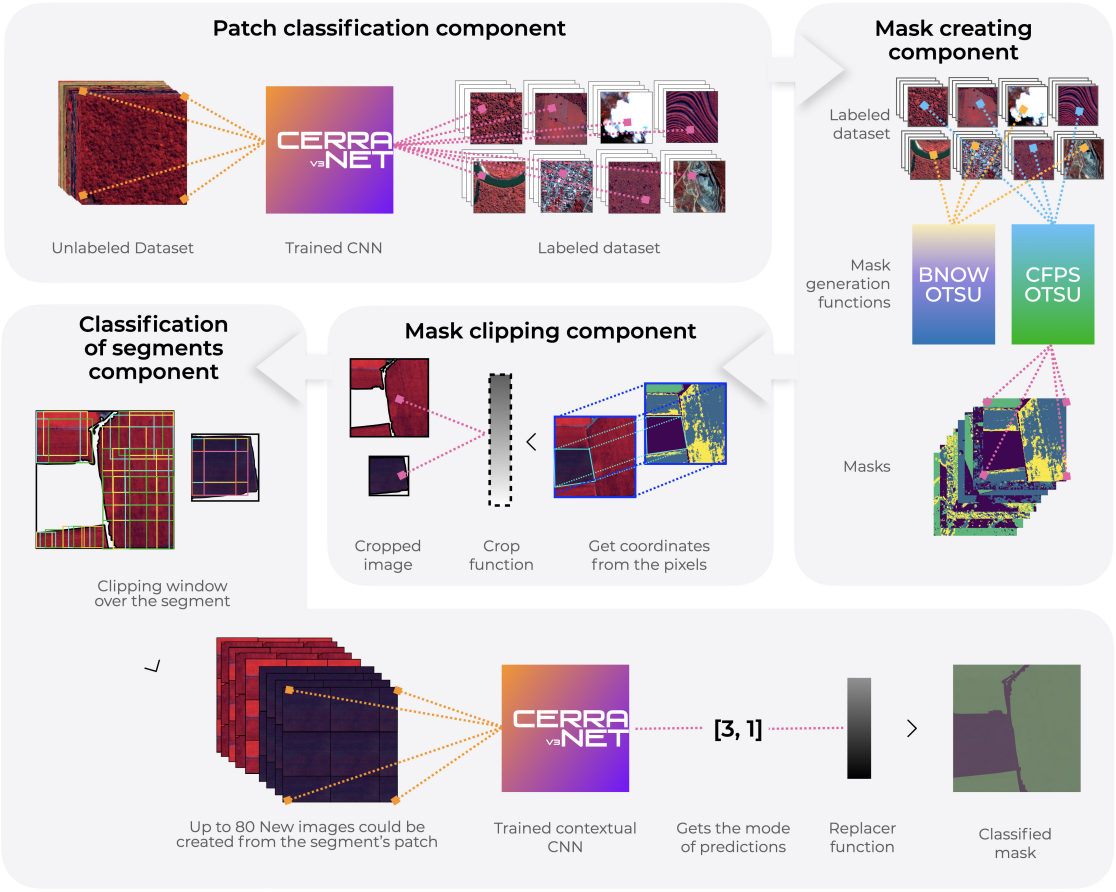
Note in Figure 3.5, the model received an input shape of 256 × 256 × 3, batch size of 128, whose images are filtered in the six deep convolutional layers, with kernel 3 × 3, activated by the *ReLu* function, the first two with 64, the two middle layers with 128 and the last two with 256 filters. Subsequently, *averagepooling2D*, pooling 2×2, and Dropout of probability 0.15 layers were employed. The feature maps were transformed into a one-dimensional vector, by the Flatten layer, and processed in the dense layers activated by the ReLu function. The output layer, however, has eight neurons activated by Softmax function, which corresponds to the number of classes.

The model performance was optimized with *Adam* method, whose parameters are *learning rate* of 0.0001, *beta* 1 of 0.9, *beta* 2 of 0.999, *epsilon* of 1e-07, consisting in a descending stochastic gradient method that is based on adaptive estimation of first and second-order moments. Moreover, to calculate the learning gain and loss rates, the metrics *accuracy* and *categorical crossentropy* were employed. Regarding the training phase, up to 80 epochs are estimated, since the loss rate is monitored at each iteration and ensures that training is interrupted if there is no decrease. This stopping criterion considers up to 8 epochs as waiting time. Consequently, the best model is saved at the end of training.

3.3 Smart mask labeling

Automated generation of the masks to support semantic segmentation is a very helpful direction to alleviate the efforts researchers must employ in real life settings. Within AI4LUC, the smart mask labeling module provides the automated mask generation and labeling. The masks are created via filters, thresholds, and morphological operations algorithms in order to segment the elements of the image. The coordinates of the mask segments are used to access the image pixels, which are input data to CerraNetv3. The output of this classification label is used to replace the pixels of each segment of the mask. Details on this pipeline are presented in Figure 3.6.

Figure 3.6 - Pipeline of the smart mask labeling.



There are four components, identified in gray; the running sequence is indicated by gray arrows; the orange and blue color dotted lines indicate input data, whereas in pink color dotted lines output data; the outputs 3 and 1 of the model concern the Non-observed and Cultivated Area classes, respectively.

3.3.1 Patch classification component

CerraNetv3 has two roles in this component. Firstly, it is used to categorize data into eight classes, if not labeled, for custom settings in the mask generation component parameters. Secondly, it is to support mask labeling, as described in follow. Therefore, the trained model receives the input data, normalizes it between 0 and 1, and labels it in one of the eight classes.

3.3.2 Mask creating component

For the generation of masks, considering CerraDatav3, two groups of classes that share common characteristics were defined. One group, CFPS, has more vegetation, while the other group, BNOW, the classes have samples containing less vegetation

cover. Thus, two functions were defined to create masks that combine different morphological operations, filters, and threshold Otsu (OTSU, 1979).

Function names are given the first letter of each class name and the threshold name, hence, BNOW-Otsu, designed for the category group Building, Non-Observed Areas, Other Uses, and Water bodies. CFPS-Otsu function, comprising the classes Cultivated Area, Forest, Pasture, and Savanna Formation. These functions produce segmented image, i.e., a mask, which each mask's segment is a random value assigned to the corresponding pixel.

As mentioned earlier, the patch classification component is used to classify the input image in the interest of selecting the filter and its custom settings to the mask creation. Afterward, the input data is converted from RGB to grayscale, in which each grayscale pixel is calculated as the weighted sum value of the corresponding red, green and blue pixels as the default of the function in *Scikit-Image* package:

$$Y = 0.2125(R) + 0.7154(G) + 0.0721(B). \tag{3.1}$$

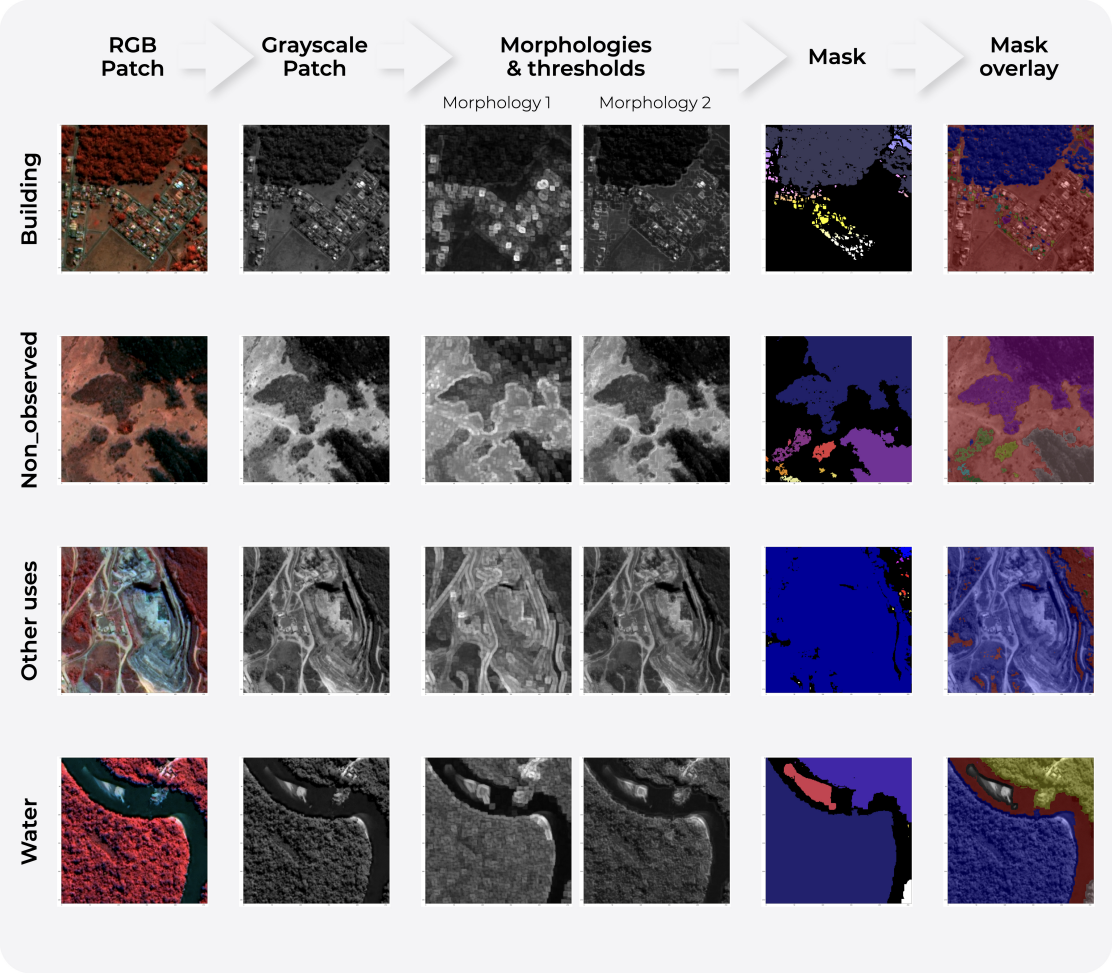
BNOW-Otsu function comprises three morphological transformations, dilation (DIL), erosion (ERO), and diameter closing (DC), and the threshold method Otsu, combined in two different modes aiming to enhance the dark and bright spots in image, respectively, as described in Equation 3.2 below.

$$\begin{aligned} morphology1 &= (DIL(gray\_image) - ERO(gray\_image) + DC(DIL) \\ thresh &= morphology1 > threshold\_otsu(gray\_image) \end{aligned} \tag{3.2}$$

Observe in Figure 3.7, as a result of this first combination, a gray image with bright spots are dilated, creating a subtle light border between the lightest and darkest areas of the image, due to the DIL, but also the addition of the DC. Thus, these brighter regions, when *morphology1* variable bigger than Otsu method, are emphasized. Concerning the darkest regions, a second morphology was considered, as follows in Equation 3.3, combines erosion proceeded by a dilation, which added to the DC increases the diameter of the space between small bright and dark regions.

$$
\begin{aligned}
morphology2 = (ERO(gray\_image) - DIL(ERO) + DC(gray\_image)) \\
thresh = morphology2 < threshold\_otsu(gray\_image)
\end{aligned}
\tag{3.3}
$$

Figure 3.7 - BNOW-Otsu function running steps.



The widgets contain the mask creating sequence indicated by horizontal gray arrows, at vertical the classes names are listed.

Consequently, small bright is removed and the dark spots connected, as shown in Figure 3.7. Therefore, vegetation areas, such as forest, are emphasized and their edges are usually outlined by light pixels, when *morphology2* values are less than Otsu threshold values. Observe that the second morphology preserves some texture of the image while *morphology1* segments are smoothed and thus more homogeneous. Note that there are subtle differences between the generated masks, being the square

of DIL and ERO, threshold diameter and DC connectivity, as well as the threshold adjustment of the Otsu method. These custom adjustments of the parameters were defined by a human, but selected after input image classification by the CerraNetv3 network.

CFPS-Otsu function consists of two morphological transformations, opening (OP) and dilation (DIL) combined with threshold Otsu values. Hence, two masks are obtained in the follow proceed in Equation 3.4,

$$\begin{aligned} morphology &= DIL(OP(gray\_image)) \\ thresh1 &= morphology < threshold\_otsu(gray\_image) \\ thresh2 &= morphology > threshold\_otsu(gray\_image) \end{aligned} \tag{3.4}$$

The morphologies combination creates a gray image whose small dark and bright spots are connected with itself and dilated, whereas the texture of the element is getting smooth and homogeneous, displayed in Figure 3.8. In this way, the filter values those images that mostly contain vegetation, but when there are other classes, the filter can highlight them. The Otsu method, as a threshold, segments similar regions via two logical operators, creating under this circumstance two masks. CerraNetv3 can also be used to classify the input image and thereby select custom settings for the square size parameters for OP and DIL and refine the output of the Otsu method.

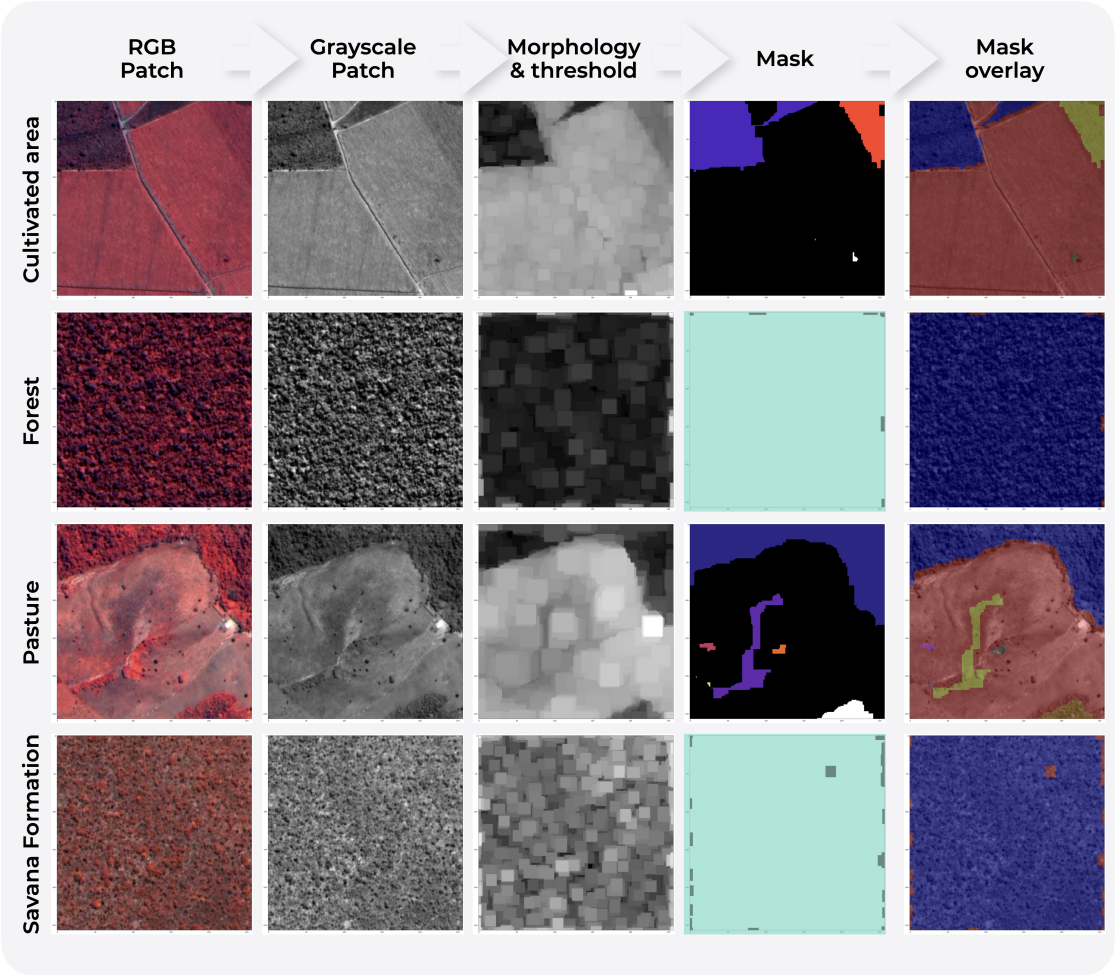
In both functions operations, *skimage measure label* is applied to generate the mask based on thresholds outcomes. This functionality converts the threshold output into integer values whose similar neighboring pixels are connected. For this reason, each segment receives a random integer value, as seen in the Figures 3.7 and 3.8.

3.3.3 Mask clipping component

This component uses the masks as a map to crop the input image into one or more polygons, via its coordinates of each pixel, as presented at the third component in Figure 3.6. However, the area and size of each segment are criteria considered before doing this. Thus, those polygons with an area smaller than 900px are not considered, due to their low contextual information.

Considering that the segments have different sizes, most of the time different from

Figure 3.8 - CFPS-Otsu function running steps.

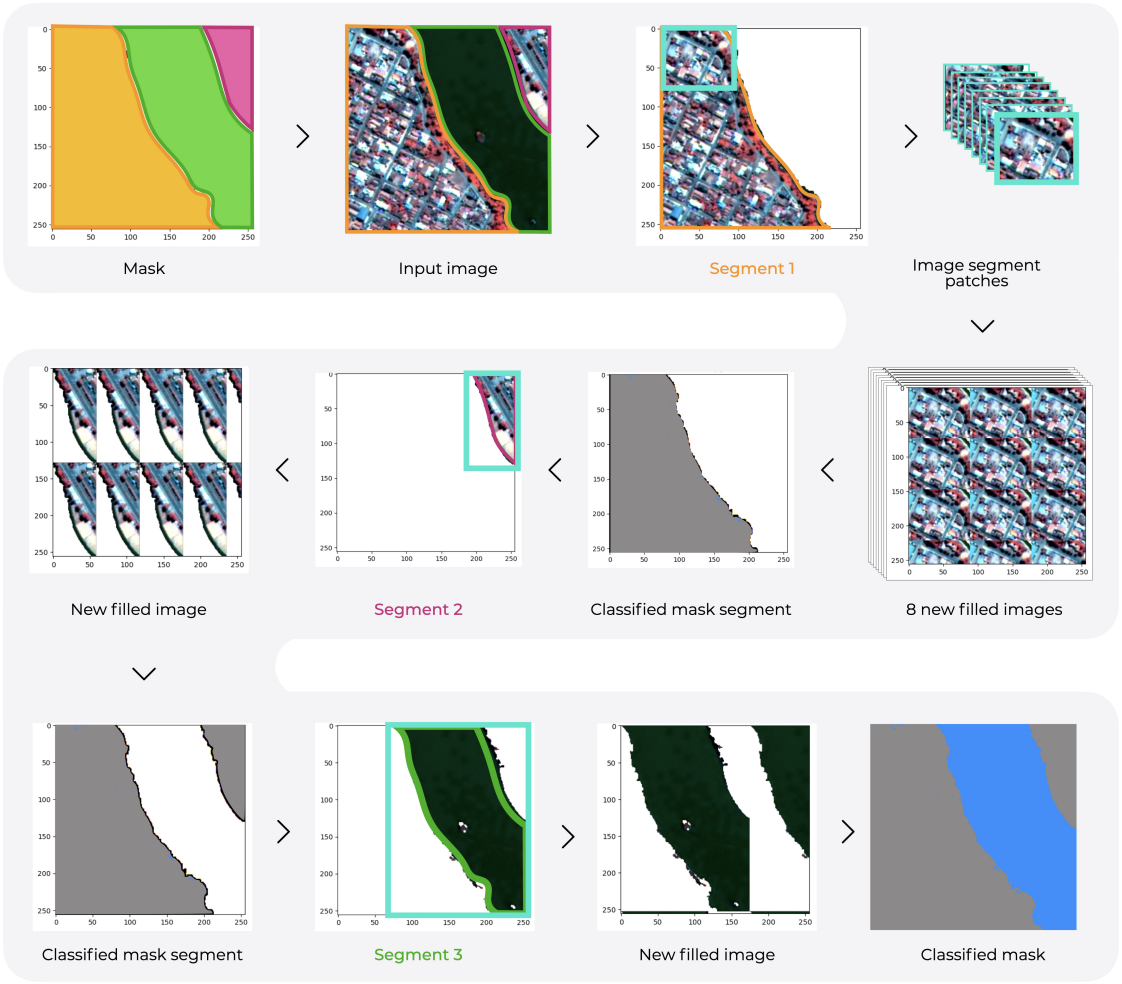


The widgets contain the mask creating sequence indicated by horizontal gray arrows, at vertical the classes names are listed.

the default size of input CerraNetv3, a sliding window was employed in order to go through the entire segment at a step of three pixels, generating up to 80 different windows. But only if the patch holds more than 92% of content, i.e., segments with a smaller area without pixels, in order to obtain homogeneous contextual information. The window is represented as a cyan-blue rectangle in Figure 3.9.

Subsequently, each window is repeated within a new image of 256 $\times$ 256 pixels and RGBA channels, aiming to fill it with each yielded window. For instance, in Figure 3.9, for great segments was considered a sliding window in order to get up to 80 patches in homogeneous regions in the segmented image, at 3px step for the x and y axes. Otherwise, only one window is considered. Therefore, up to 80 filled images can be created per segment of the mask. It is worth emphasizing, these parameters'

Figure 3.9 - A running example of mask labeling.



Three segments, represented by the colors yellow, green and pink; the cyan blue frames is the sliding window that can patch large segments.

values were experimentally selected.

3.3.4 Classification of segments component

Observe the Figure 3.9, CerraNetv3 classifies this filled image with the segment's patch. Concerning cases when up to 80 images from the same segment are presented to the classifier, the predicted label that appears most frequently is chosen. Afterward, the pixels of the mask segment are replaced by the predicted label. The small segments, whose areas are smaller than 900px, are disregarded for the classification by the CerraNetv3, but inherit the label of the subsequent neighbor segment. This procedure is carried out for every segment mask.

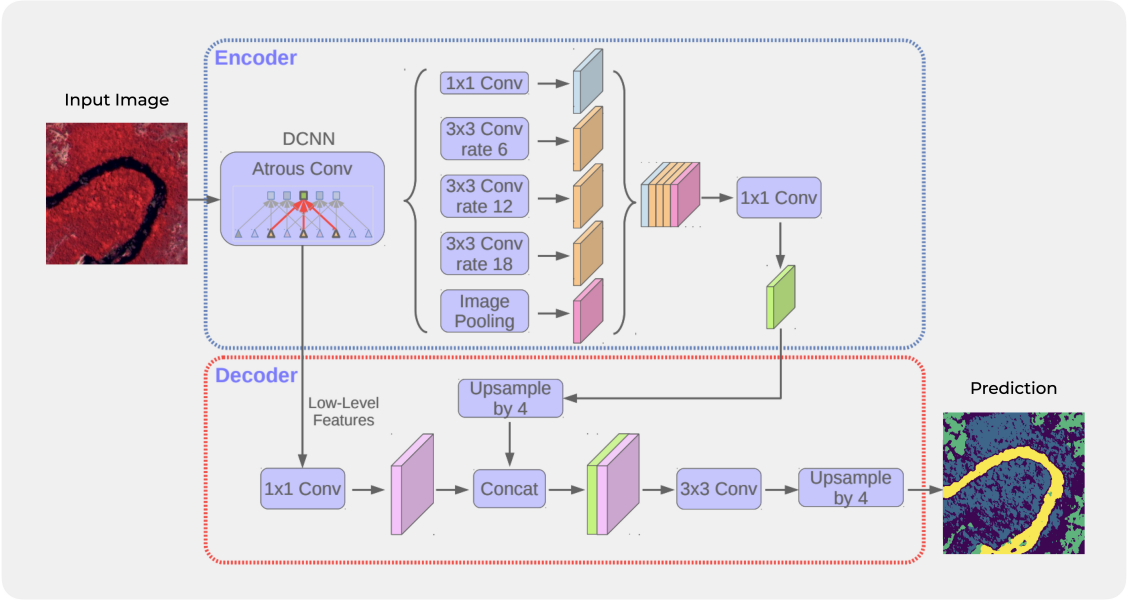
3.4 Pixel-based classification model

This module is based on training a CNN for semantic segmentation. Thus, the DeepLabv3plus network was chosen due to its great performance in terms of classification accuracy as well as segmentation quality, as seen in the work of Du et al. (2021) and Chen et al. (2018). In view of this, this section presents the network. The neural network was imported from the Segmentation Models PyTorch (SMP) package developed by Iakubovskii (2019), which uses the PyTorch framework.

3.4.1 DeepLabv3plus

The DeepLabv3plus was developed for semantic segmentation tasks, proposed by Chen et al. (2018), extended from Chen et al. (2017). This network has encoder module, that contain atrous spatial pyramid pooling layers for feature extraction, and decoder module to segment objects boundaries at input image, as shown in Figure 2.6.

Figure 3.10 - DeepLabv3plus network architecture.



The input image is processed in the encoder module, whose features, in the decoder module, are used in the construction of the segmented image. The output image is a classified mask.

SOURCE: Adapted from Chen et al. (2018)

DeepLabv3plus has a new decodes modules, presented in 3.10, which receive from

encoder module features bi-linearly up-sampled by a factor of four, since are concatenated with their respective low-level feature coming from Atrous Convolutional, before the number of channels is reduced by a 1×1 convolutional layer, after all feature information are refined followed by another simple bi-linear up-sampling by a factor of 4. Thus, the performance was improved, beside all process brings more computation cost (CHEN et al., 2018).

3.5 Performance assessment

Based on the hypotheses, three experiments have been carried out to evaluate the AI4LUC method: i) Performance comparison of CerraNetv3 and ResNet-50 (HE et al., 2015), a state-of-the-art CNN, in terms of contextual classification; ii) Performance analysis of the SML module regarding the generation of masks and labeling, using metrics; iii) DeepLabv3plus prediction performance is compared with U-Net, SML module, and truth mask.

The learning method adopted for training the models used in these experiments was from scratch, e.g., was not utilized in the fine-tuning of pre-trained weights in the models. Regarding performance evaluation metrics, F1-score and Precision were applied for all experiments, while IoU was applied only to analyze the accuracy of semantic segmentation models. For these tests, the CerraDatav3 test subset was considered.

3.5.1 Experiment: contextual classification

3.5.1.1 CerraNetv3 $\times$ ResNet50

In this first experiment, the proposed CNN is compared with the ResNet-50 model, widely employed in tasks of context classification, as well as encoding modules for semantic segmentation networks. Hence, metrics and their image classification were analyzed, highlighting misclassification and good performances for each class. For both architectures were built using Keras and TensorFlow packages, adopting settings of hyperparameters Adam optimization, *categorical crossentropy* for loss calculation, *accuracy* metric for gain calculation, up to 80 epochs, stop control with the patience of 8 epochs, and batch size of 128. In this experiment, the models were trained with the CerraDatav3, split into two subsets, training and testing, as mentioned in Section 3.2. During the training phase, for each network, the best model was saved taking into account the lowest loss value per epoch.

3.5.2 Experiment: smart mask labeling

The CerraNetv3 performance evaluation is defined quantitatively and qualitatively, in terms of accuracy in the labeling of the reference masks and in the contextual classification. However, for the purpose of comparison, masks generated with the functions of the mask generation component, but manually labeled, were used. The manual labeling was based on labels assigned to the samples used in CerraDatav3 auditing. Thus, the masks labeled manually and by CerraNetv3 are presented as input to the assessment metrics.

3.5.3 Experiment: pixel-based classification

In contemplation of the second hypothesis, in this experiment, the accuracy of the DeepLabv3plus model was evaluated. The achieved outcomes were compared with the results produced by U-Net and with the manually labeled masks, regarding the scores achieved through the metrics. The architecture of the U-Net network was imported from the SMP package developed by Iakubovskii (2019), using the PyTorch framework. Both models have been trained with the images from the CerraDatav3 training subset and the reference masks labeled by CerraNetv3.

For the training of both semantic segmentation models, 10,000 CerraDatv3 image patches were considered, as well as 10,000 labeled reference masks, created in the SML module. Regarding the hyperparameters of the network, the *CrossEntropyLoss* function was defined to calculate the cross entropy loss between the input logits and the label; the *Adam* method with a learning rate of 0.01 for model optimization; performance gain with the F1-score, IoU, and accuracy metric; and 40 epochs.

3.6 Final remarks about this chapter

This chapter presented the AI4LUC method. It was detailed how each module, component and function of the pipeline was built and applied. In highlight, the new version of CerraDatav3 was presented and how CerraNetv3 was applied in the SML module. In addition, the method evaluation protocol was introduced.

4 EXPERIMENTAL RESULTS

This chapter presents the experiment evaluations which were accomplished in order to assess the performance of every module of the AI4LUC method, as described in 3.5 Section.

4.1 Contextual classification

The context, texture, and spectral response of objects in the scene are essential features for the contextual classification problem (INPE, 2019; FONSECA et al., 2021). Incoming, the comparison and analysis between CerraNetv3 and ResNet-50 outcomes are presented.

4.1.1 CerraNetv3 × ResNet-50

This section presents the results of the comparison between CerraNetv3 and ResNet-50. Table 4.1 shows the simple average and standard deviation of the three runs performed for the two networks, based on the F1-score and Precision metrics. The highest score is highlighted in **bold** and a * was added in the model name.

Table 4.1 - Performance assessment: CerraNetv3 × ResNet-50.

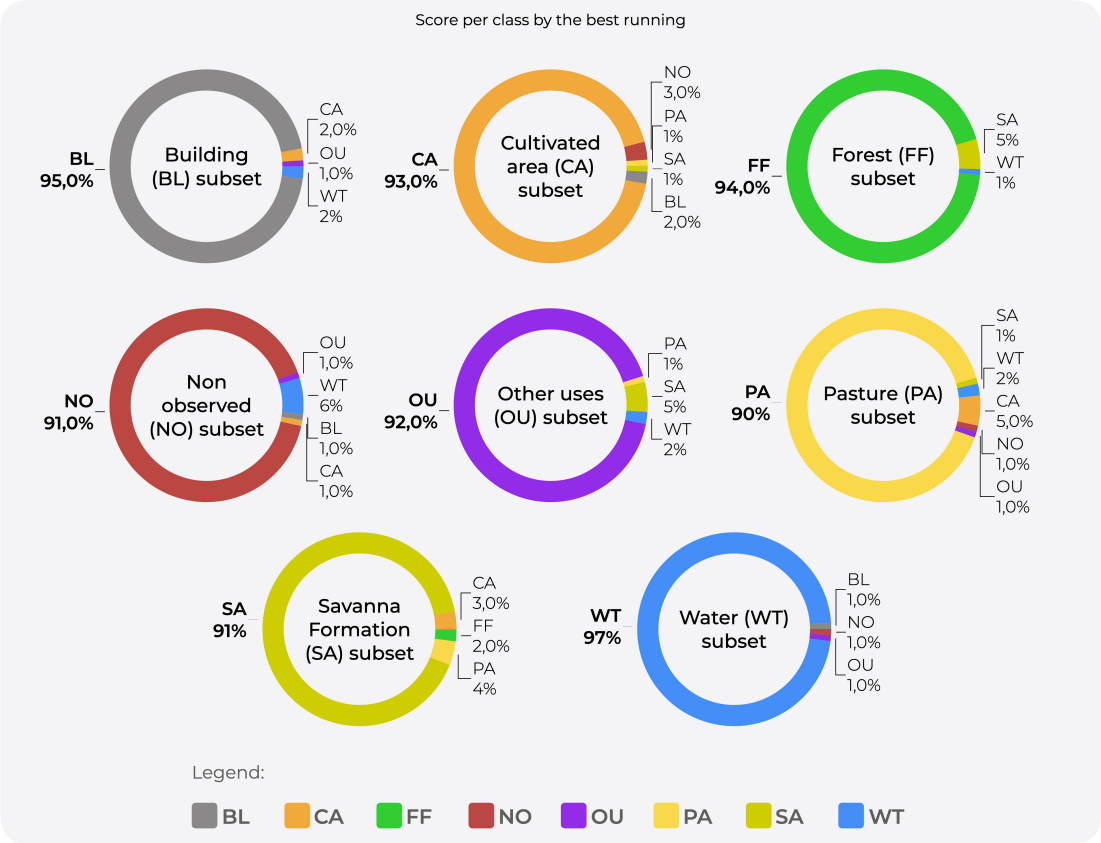
DCNN	F1-score	Precision
CerraNetv3*	**0.9241 ± 0.0013**	**0.9248 ± 0.0012**
ResNet-50	0.9199 ± 0.0015	0.9213 ± 0.0012

The results achieved by both networks are satisfactory, but it is worth underlining the gain of CerraNetv3 in relation to ResNet-50. Bearing in mind that the proposed model has a total of 1,441,736 parameters and ResNet-50 24,114,312 parameters, which usually means more training time and computational cost. CerraNetv3 completed its iteration cycle in an average time of 14 hours and 23 minutes, while ResNet-50 in an average of 70 hours, in the *nvidia long* queue of the SDumont supercomputer. The standard deviation of the two networks denotes good stability regarding in the F1-score achieved in each execution, especially when it comes to training from scratch.

Figures 4.1 and 4.2 detail the total percentage of correct and incorrect classification obtained for each class of the test set, for each best-runned model. The first Figure shows that for the WT class, 97% of the samples were correctly labeled, while 3%

are labeled as BL, NO, and OU. For the NO class, 6% among the 9% of classification errors correspond to WT, so it is speculated that these two classes are confused by the classifier, even though there is not one among the categories in the scene. In subsets FF and OU, 5% of samples were incorrectly labeled as SA. In the case of OU there are images that have savanna, but, simultaneously, rock outcrops. Regard to FF, these errors are associated with tree formation, whose treetops are far from each other, which resembles SA formation.

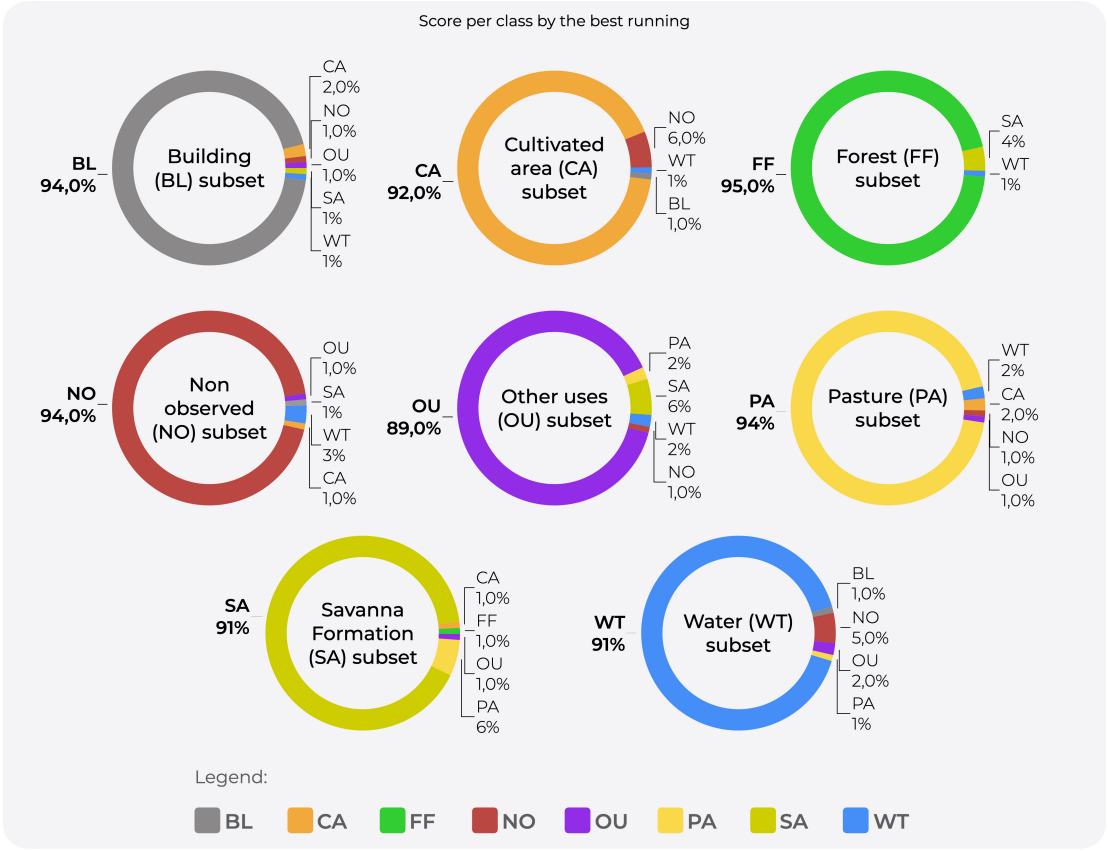
Figure 4.1 - CerraNetv3 classification per class.



The BL subset is the second with the highest percentage of hits by CerraNetv3. This class is incorrectly assigned in at least 1% of the NO, WT, and CA subsets, despite having small built-up areas in these images. The lowest percentage of hits, however, is the PA subgroup, hitting 90%, in which five other classes are mistakenly assigned, mainly the CA label, due to the sharing of certain spectral, texture, and context features. In general, the percentages for each subgroup of the test set are high and indicate low variability in labeling errors, which justifies the stability of

the model.

Figure 4.2 - ResNet-50 classification per class.

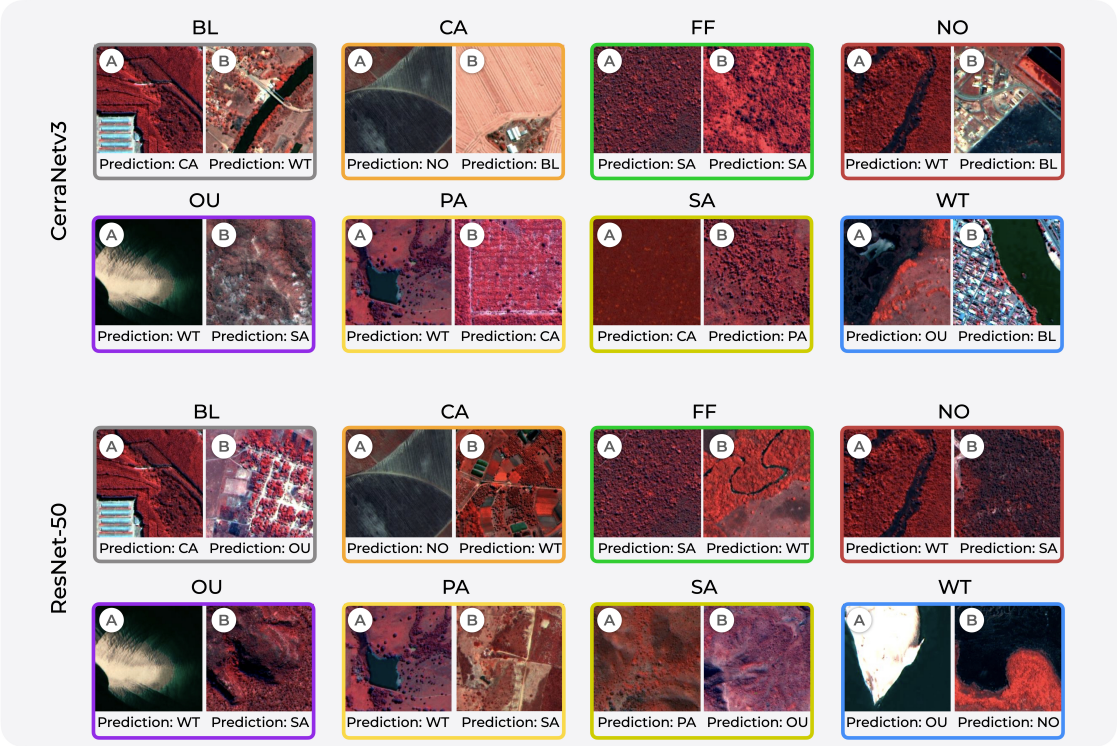


Regarding the percentages of correct classification by ResNet-50, it is noted that some trends of errors remain, as seen between subgroups SA, OU, and FF that, like CerraNetv3, some samples are labeled as SA. There was a significant loss in the percentage of correct answers in the subsets CA and BL, particularly in OU and WT, whose differences with respect to other model are 3% and 6% respectively. Although the SA subset maintained the number of hits, one more mislabeling was registered among the others. However, there were improvements in the percentage of correct classifications, being 1% for FF and 3% for PA and NO.

The misclassification by both models is presented in the Figure 4.3. In a single image patch there are more than one class of LULC, however, only one label is assigned to it. In face of this, it can easily characterize certain interpretation mistakes by the deep learning (DL) model, i.e., for some cases can be considered as false

positives. Although the patch A of the BL class has been labeled as CA, understood a mislabeling, it is correct as well because both classes are in the patch scene, as seen for the A patches of CA, OU, and PA.

Figure 4.3 - CerraNetv3 and ResNet-50 misclassification.



Another interesting mislabeling happens to the NO class, in which the NO object is seen as WT. On the other hand, patch B of the WT class is labeled as WT by ResNet-50, due to similarity with water spectral response. In relation to the PA class, both models had assigned WT to A sample A, because of the water body, which maybe it is a representative feature in the scene. A similar kind of mislabeling is seen in patches B of the BL class by CerraNetv3 and in the B of the FF class. Taking it into account, the models associate the context around the water body to get label it as WT. However, in patch B of the WT class, CerraNetv3 does not consider building features instead of water.

Some phytophysiognomies are confused with each other by the models, such as the A patch of the class FF whose context is similar to savanna formation, yet the trees have a canopy of forest formation. While B of PA class by CerraNetv3, the pasture

is mislabeled as CA, maybe due to the spectral response, as well as the texture. Concerning the SA class samples, the most common mistakes were relating PA, FF, CA, and OU, once some characteristics are shared with each other, in terms of similarity to vegetation composition in the scene.

4.2 Smart mask labeling

Considering hypothesis 1, this experiment was developed which, based on a contextual classification CNN, assisted in the labeling of reference masks for LULC images. Table 4.2 presents the scores achieved for each CerraDatav3 class. To assess accuracy, the metrics IoU, F1-Score, and Precision were taken. In the evaluation, all pixels of the image are computed, thus all classes are considered for metrics calculation. Top scores by class are highlighted in bold. Moreover, the table provides an overall of each metric.

Table 4.2 - Manually labeled versus SML for CerraDatav3's test set.

Classes	**IoU**	**F1-score**	**Precision**
Building	0.1788	0.6760	**0.7384**
Cultivated Area	0.1585	**0.8053**	**0.9397**
Forest	**0.2352**	**0.9803**	**0.9954**
Non-Observed Area	**0.2179**	0.6606	**0.7216**
Other uses	0.1035	0.4977	0.6277
Pasture	**0.1988**	**0.8280**	**0.9353**
Savanna Formation	0.1108	**0.8993**	**0.9946**
Water	0.1783	0.3961	0.5762
BNOW score	0.2939	0.4936	0.5426
CFPS score	0.3897	0.8390	0.8716
Best scores	0.3324	0.8390	0.7838
Overall	0.4621	0.6738	0.7078

Founded on F1-score and Precision metrics, the highest scores are achieved by the vegetation-predominant class group, scoring 0.8390 and 0.8716 subsequently. These classes are more homogeneous than the others, and the generation of reference masks has few segments, which provide more contextual information for the classifier model. For example, among all classes, FF and SA were most accurately labeled. However, with IoU the best evaluations are over FF and PA from the CFPS group, while BL is the best-ranked class in the BNOW group. That is, they are classes that most

intersected between correctly labeled segments.

Regarding the BNOW group, the lowest marking evaluated by F1-score and Precision was the WT class. These misclassifications are correlated with failures in mask generation or due to lack of contextual information, as illustrated in Figure 4.6. In this instance, denotes that other pixels referring to other elements of the scene, whether format, spectral response, or texture, must be presented to the classifier. Consequently, these scores indicate the number of pixels correctly labeled within the Water images subset, considering other classes. Thus, this gap influenced the results of the other classes in the group, which, in general terms, reached an F1-score of 0.4936, Precision of 0.5426 and IoU of 0.2939.

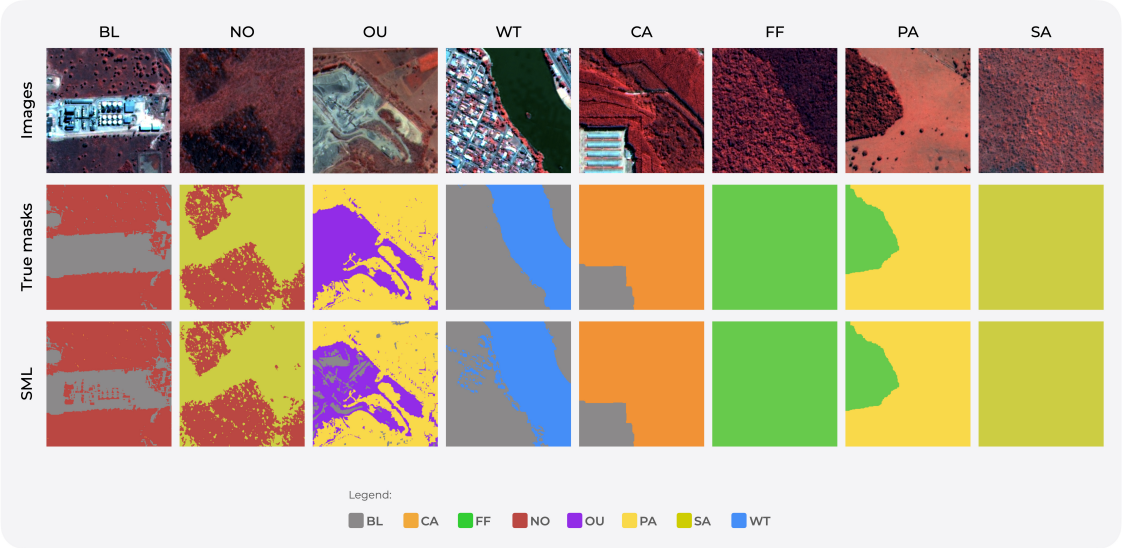
Given classes highlighted in each metric, SML performance was calculated. As a result, it scored 0.3324 with IoU, 0.8390 with F1-score, and 0.7838 with Precision. For this, scores greater than 0.19 were considered for IoU, and 0.70 for the F1-Score as well as Precision metrics. Overall, the classifier in SML performed well, even at the expense of mask generation defeats, recording 0.7078 exactness for all classes in the CerraDatav3 test set.

In addition, another aspect related to performance is the cost of execution time and computational cost. The procedure of generating and labeling the mask on a personal computer takes up to approximately 33 seconds for each mask, assuming up to 80 patches are generated from the segment. However, on the SDumont supercomputer, in this same case, it takes up to approximately 6.4 seconds per mask. Applying this method to produce labeled masks for the CerraDatav3 training set, aiming to use them to train the semantic segmentation models, took about 96 hours per class through the supercomputer, i.e., 9900 image patches.

4.2.1 Predictions analysis

The quality of mask generation implies classification by the CerraNetv3 network. For example, Figure 4.4 shows cases in which one mask from each class in the CerraDatav3 test set were correctly generated and classified. Note that the masks of the CFPS class group are homogeneous and defined, whose elements are properly highlighted, such as seen in the masks of the CA and PA classes' samples. Although there are small elements in the PA and CA's images, SML yielded and assigned to the masks labels belonging to subsequent neighbors. Regarding the FF and SA samples, the classified masks have a single segment that covers the entire image, without noise.

Figure 4.4 - Correctly labeled masks.

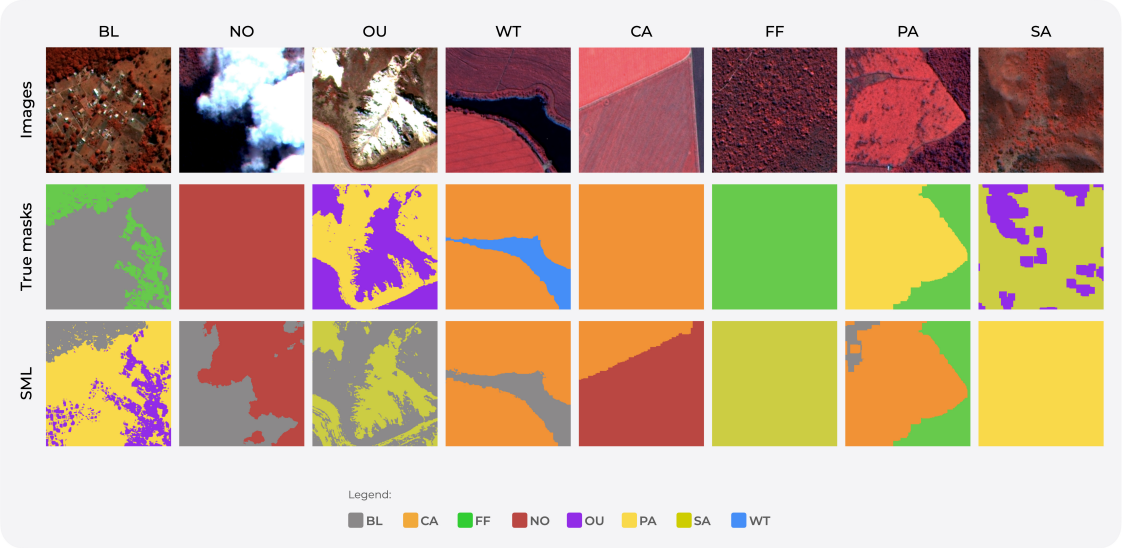


For almost all SML masks segments of the images reproduced results similar to true masks, mainly the examples belonging to the NO and WT classes. However, BNOW mask group has noisier segments that imply labeling accuracy, reported with the metric F1-score and Precision in Table 4.2. For this reason, incorrect labeling may be related to mask generation or classified error, due to little contextual information. In the Figure 4.5 some mislabeled mask cases are displayed.

Mislabeling masks from both class groups was evidently committed by the classifier. Look at the image of FF, its true mask, labeled as FF, and the one classified by SML, labeled as SA. There are spaced tree cover on the scene, as seen in SA, but this vegetation has a forest canopy. In the image of SA, there is the spacing between vegetation and exposed soil, between mountains that are included in the OU class, SML labeled it as PA. Regarding the PA example, the PA segment was mislabeled as CA and BL, although the texture and spectral response resemble CA, there are other types of vegetation in this landscape, which is not characteristic of CA images. In the classification of CA by SML, most of the scene was classified as NO, in which clearly CerraNetv3 considered the NO segment during the classification.

The WT class image has a lake surrounded by a CA, however, in the mask labeled by SML, the lake was classified as BL. Bearing in mind that the segment only emphasizes the water body, it is speculated that the mislabeling by the classifier in the SML needs other features in order to complement the context of the water segment, as it occurs for FF in Figure 4.6. The NO and OU classes, compared to

Figure 4.5 - Mislabelled masks.

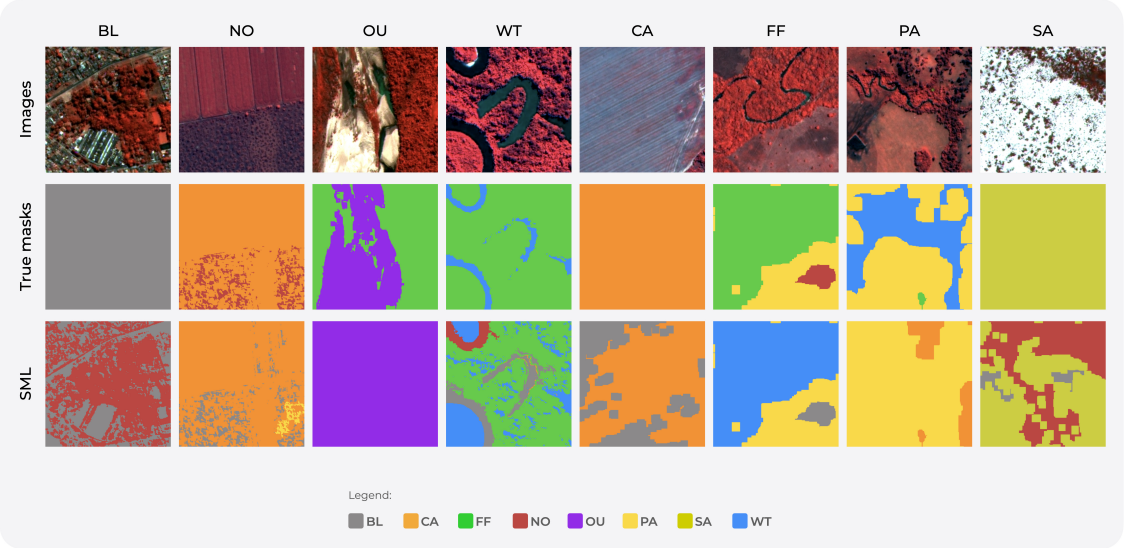


the others, contain more LULC diversity. However, this does not mean a balanced number of samples of each LULC type in each class, therefore the classifier tends to infer incorrectly in some cases, as shown in Figure 4.5. With respect to the BL class masks, in both masks, the small built-up areas were not segmented. Thus, for the true mask, the labels FF and BL were considered, while the classifier in SML assigned PA instead of BL and NO instead of FF.

In addition, image segmentation failures in mask generation can lead to classification errors. Figure 4.6 shows some flaws in terms of the format, size, and quality of the segments. In general, the true and SML masks of the BNOW group are noisy, i.e., small segments, while the segments of the CFPS group have thicker features. However, notice that, in these examples, mislabeling is related to those failed segments. On the other hand, exceptions can be considered, for example, the segment of the FF mask was classified as WT by the SML, since there is a narrow river and forest formation in the scene of the patch.

Based on the metrics in the Table 4.2, as well as these examples in the Figures, it was accomplishable to understand the reasons that made the classifier mess up in the labeling of the WT and OU segments. Look at WT masks in Figure 4.6, their water body segments are smaller than the edges seen in the image, and thus could be mistaken for NO or BL. Hence, the segments of these classes can be dilated to include not only the object of interest, but also information from other elements around it. In general, CerraNetv3 has made a good interpretation for every class,

Figure 4.6 - Failures in mask generation.



even with noisy segments or mislabeling other ones.

4.3 Pixel-based classification

This experiment concerns hypothesis 2, comparing the performance of DeepLabv3plus with U-Net, taking into account the CerraDatav3 test set. It is important to emphasize that both models were trained with 10,000 reference masks generated with SML module from the CerraDatav3 training set. Therefore, the detailed results in Tables 4.3 and 4.4 were calculated from the predictions assembled by the models and the manually labeled masks, accepted as true masks, whose highest scores are highlighted in **bold**.

The masks predicted by DeepLabv3plus correctly match, compared to the true masks, only 0.2805 based on the F1-score metric. And opposite to SML predictions, the lowest score occurred for the FF class, scoring less than 0.0001, as shown in Figures 4.7 and 4.8. While the classes CA, PA, and SA were labeled with more precision than the others, with the metrics F1-score and Precision, described in Table 4.3. The IoU between the predictions and the true masks was substantially low, stressing the BL, NO, and PA classes by marks a little higher than 0.10.

There is a greater difference in Precision for the CA and SA subsets, while the accuracy values for the others are close to the results produced with the F1-score metric. This concerns the properties considered for the calculation of the hits of

each metric. Hence, the Precision metric denotes the accuracy of the percentage of correctly classified samples. Therefore, at least 0.9 of the samples from these subsets are expected to be correctly labeled.

Table 4.3 - DeepLabv3plus predictions for CerraDatav3's test set.

Classes	**IoU**	**F1-score**	**Precision**
Building	**0.1378**	**0.6652**	**0.7250**
Cultivated Area	0.0648	0.3210	**0.9482**
Forest	0.0032	0.0001	7.6348e-05
Non-Observed Area	**0.1169**	**0.6370**	0.6765
Other uses	0.0204	0.0380	0.0629
Pasture	**0.1079**	**0.7970**	**0.8208**
Savanna Formation	0.0667	0.5690	**0.9935**
Water	0.0670	0.1545	0.1588
BNOW score	0.1398	0.2822	0.2470
CFPS score	0.1057	0.2860	0.3294
Best scores	0.2340	0.6570	0.5790
Overall	0.1708	0.2805	0.2822

U-Net scores are detailed in the 4.4 Table. Note the similarity with DeepLabv3plus in the sense of Accuracy for classifying samples from the BL, NO, and PA subsets, whose evaluations outcome are analogous according to F1-score. However, for other subsets, the performance results were less than 0.2, and in an extreme case, all samples of subset SA were mislabeled. Another similar situation to the other network is in regard to assessment results with the IoU metric, of which none exceeds 0.1310.

Given that the models were trained with few samples and consequently the results achieved by both models were low, DeepLabv3plus had a performance gain of nearly 71.35% compared to U-Net. However, considering only the subsets with the best scores and recalculating the accuracy, it scored 0.6570 with F1-score, Precision of 0.5790, and IoU of 0.2340 for the best model. Regarding U-Net, following this idea, 0.6573 was obtained with F1-score, IoU of 0.2302, and Precision of 0.6159. A foreseen reason for the low models' outcomes could be related to the unbalanced amount of segments regarding each class, becoming networks overfitting for some classes. Also, these mislabeling cases can be associated with this unbalancing number of samples.

The accuracy of DeepLabv3plus calculated based on the F1-score for the BNOW

Table 4.4 - U-Net predictions for CerraDatav3's test set.

Classes	IoU	F1-score	Precision
Building	**0.1310**	**0.6693**	**0.6695**
Cultivated Area	0.0330	0.0074	0.0049
Forest	0.0022	6.9006e-05	7.3707e-05
Non-Observed Area	**0.1137**	**0.6431**	**0.6185**
Other uses	0.0085	0.0046	0.0025
Pasture	**0.1025**	**0.7818**	**0.7875**
Savanna Formation	0	0	0
Water	0.0306	0.0776	0.0586
BNOW score	0.1244	0.2596	0.2003
CFPS score	0.043	0.0816	0.0511
Best scores	0.2302	0.6573	0.6159
Overall	0.1195	0.1637	0.1165

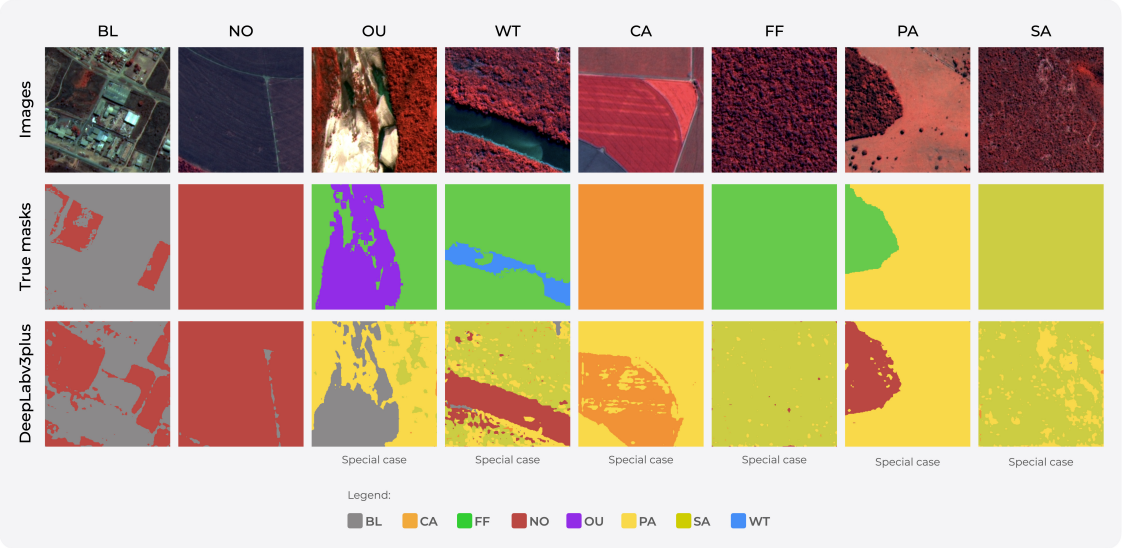
class group, as described in Tables 4.3 and 4.4, was 8.66% higher than U-Net. Comparing them in the CFPS group scenario, DeepLabv3plus had a gain of approximately 250.92% over the other network. Predicated on these analyses, it is speculated that the low scores of each network are related not only to the number of samples used for training, especially in terms of the quality of the reference masks produced from the set of training images, in the SML module. These differences in accuracy and machine learning of these are easily noticeable in the examples illustrated in the following Sub-section.

4.3.1 Models predictions analysis

The examples illustrated in the Figures of this Subsection are evaluated and discussed with regard to the quality of the input image segmentation, as well as the veracity of the segment labeling compared to true masks.

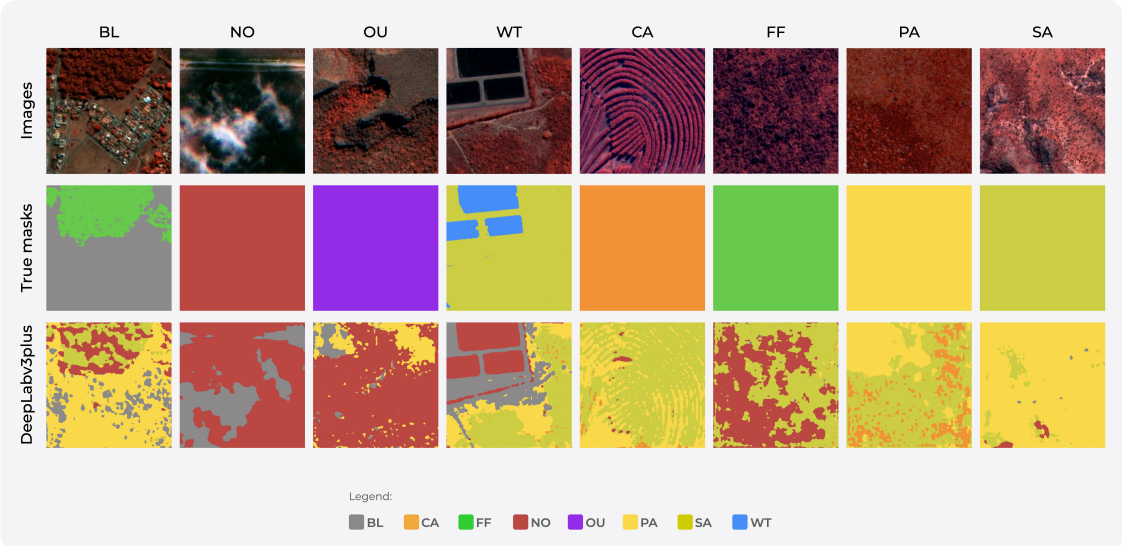
In Figure 4.7 exclusively five samples had their segments partially or completely correctly classified by DeepLabv3plus. The best labeling occurs for the BL and NO samples, which, in addition to the segments being better than those of the true mask, are correctly identified. In the PA sample, the PA segment is correct, but the FF segment is mistakenly classified as NO. While CA and SA the model considered small regions in the images as different segments and assigned them PA labels. The OU and WT class samples had their segments mislabeled, whose performance has already been described in Table 4.3. However, the accomplished segmentation

Figure 4.7 - Correctly labeled masks by DeepLabv3plus.



is more detailed and accurate than the true mask, although there is some noise. Overall, PA was the most often assigned label for special cases, with the exception of the PA sample. In terms of mislabeling, see further details in Figure 4.8.

Figure 4.8 - Mislabeled masks by DeepLabv3plus.

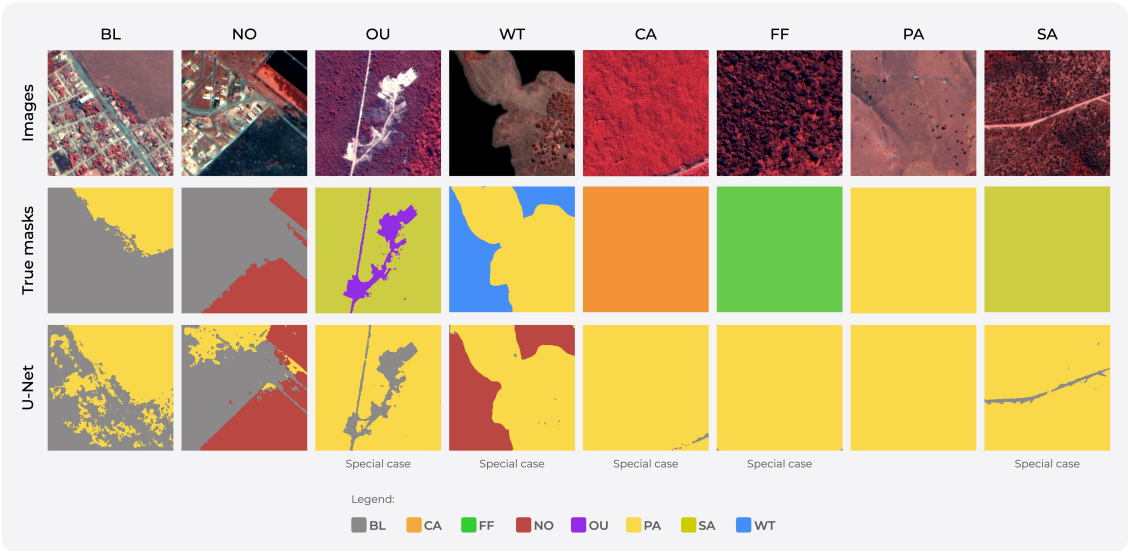


The quality of segmentation is evident in almost all samples by DeepLabv3plus.

Observing the image of the BL class, small polygons are identified as construction as in the image, in contrast to the true mask that considers the entire region as BL, with the exception of the FF segment. The misclassifications of these segments, on the other hand, tend mainly towards SA and PA for the CFPS group, as well as BL and NO for the BNOW class set. These are precisely the classes with the highest F1-scores in Table 4.3. Notice that the model often labels FF and PA as SA, while the CA and SA samples are labeled as PA.

Regarding the U-Net classification successes, only three samples had all their polygons correctly labeled, corresponding to the images of the BL, NO, and PA classes. In the case of the prediction for the WT class image, only one of the segments matches the true mask. However, in all mislabeling, the PA, NO, and BL labels are assigned by the model, as evaluated by the F1-score and Precision metrics, presented in Table 4.4. The PA label is often mistakenly assigned to the CA, FF, and SA classes samples, consequently resulting in low scores on assessments with the metrics. WT segments are in most cases labeled as NO. OU areas are labeled as BL. These errors are also accentuated in the examples displayed in Figure 4.10.

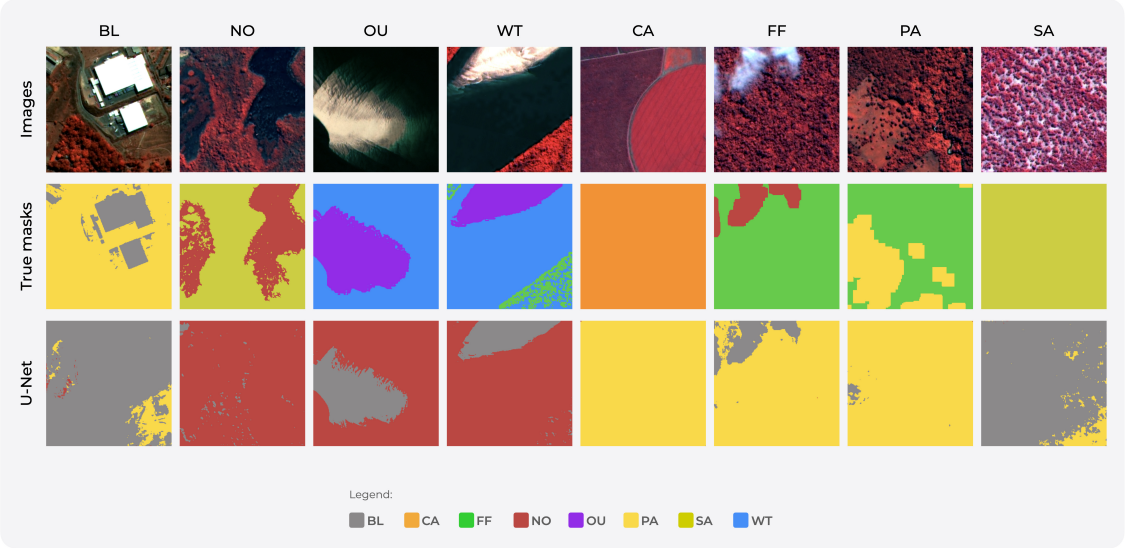
Figure 4.9 - Correctly labeled masks by U-Net.



Unlike DeepLabv3plus, U-Net can not consistently produce good image segmentation, and as a result, mislabeling. Successful cases of segmentation are regularly for those samples that have two elements, as seen in the OU and FF samples. This

aspect can also be observed in Figure 4.9. However, these routine misconceptions or learning hardships by both models are the result of errors that occurred in the mask generation and labeling step in the SML module. In line with this, it is essential to use the entire set of reference images and masks to train these models, the mask generation component for the BNOW group, in the SML module, needs updates in order to make the segments more homogeneous and with less noise.

Figure 4.10 - Mislabeled masks by U-Net.



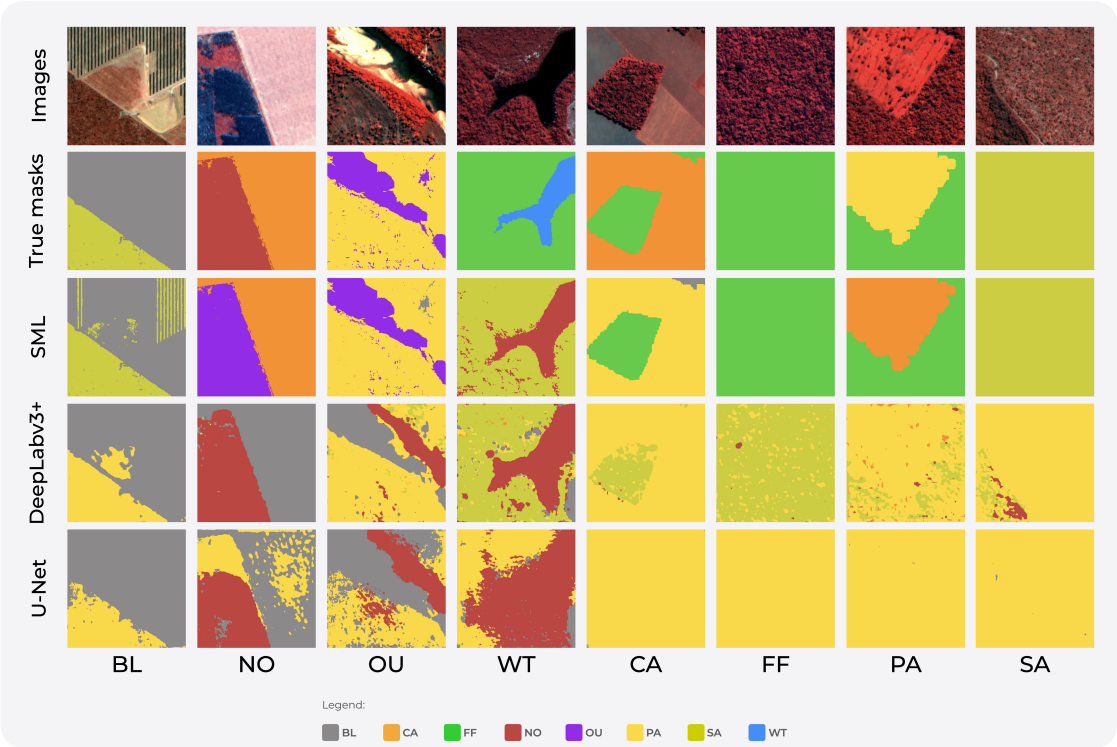
In terms of training speed, U-Net has taken approximately 26 hours to complete 40 iterations, for 32,525,256 trainable parameters. DeepLabv3plus needed two more hours to complete the training, with 26,682,520 trainable parameters. Unlike the other models, these two semantic segmentation CNNs were trained on the basic computer, whose specifications were mentioned in the introduction to Chapter 3. In view of this, DeepLabv3plus is by far the network that requires computational resources but also has good performance compared to U-Net, requiring fewer trainable parameters.

4.4 Overall analysis

This Section associates the results of pixel-based classification performed by Smart Mask Labeling (SML), DeepLabv3plus, and U-Net, in order to review general aspects of segmentation and accuracy in mask labeling, exemplified in Figure 4.11. The context of the segmented scene is very essential for successful classifications.

However, the segment needs to comprise not only the target object, but also other objects that, without contaminating the fragment, contribute to the interpretation of the scene by the model. Among all the segmentation presented in the Figure 4.11, DeepLabv3plus managed to produce good masks. But as far as classifications are concerned, the SML module managed to operate correctly for almost all cases.

Figure 4.11 - Comparison of the outcomes of SML, U-Net, and DeepLabv3plus.



Check out every output of the PA and SA classes images and their true masks. Both SML's outcomes have great segments, despite having made misclassification in PA's mask. However, observing PA's fourth e fifth columns, there are mislabeling and failure in the segmentation, as expected due to the report scores in Tables 4.3 and 4.4. Having regard to SA samples, SML has made them without flaws, meanwhile, DeepLabv3plus and U-Net mislabeled them. As conferred in the previous sections, mislabeling for WT and OU samples is recurrent, which made the semantic segmentation models inherit them. In contrast, the BL class was one of the best identified by everyone in all experiments, as well as PA, mainly by U-Net, and SA, particularly by DeepLabv3plus.

Driven by the significance of lower pixel-based models performance, a question arises is the pixel-based classification model module required at the end of the pipeline to assist in LULC classification? Concerning metrics, the SML module would be sufficiently capable of performing the classification, but the predictions would need to be checked by experts and corrections applied to segment labeling. When it comes to segmentation quality, the DeepLabv3plus network stands out comparing the rest approaches. As for training time, the contextual classification model, CerraNetv3, implemented in the SML module required less training than the segmentation networks. However, regarding the prediction time on a basic computer, the third module takes up to 33 seconds to generate and label a mask, on the other hand, the segmentation network of the fifth module requires less than three seconds to perform the same task.

4.5 Final remarks about this chapter

This chapter presented the results achieved in each experiment, thus responding to the three hypotheses regarding the AI4LUC method. Likewise the particular discussions involving the experiments, an overall analysis was made, comparing the outcomes by SML component, DeepLabv3plus, and U-Net. Next chapter, important aspects of the hypotheses are discussed.

5 CONCLUSIONS

The AI4LUC was developed for the pixel-based classification of LULC types in the Cerrado biome. The module is structured into five modules, internal components, and their functions. All were designed to help with the preparation of the dataset, and training of DCNN models, mainly in the generation and labeling of reference masks to train the semantic segmentation model, in the fourth module. In order to evaluate the efficiency and diligence of the method, three experiments were conducted.

The first experiment, contextual classification, consists in to compare CerraNetv3 and ResNet-50 performances. Based on the results of the first part of the experiment, training from scratch with few samples may not be an adequate approach, since the performance was better with synaptic weights initialized with ImageNet fine-tuning adjustments. And in this application, shallower networks will be more efficient than deep networks, due to the abstraction level of the input data and the saturation of accelerated gain rates.

In the secondary part of the experiment, based on the previous analysis, considered comparing the ResNet-50 with CerraNetv3 performances. In addition to the novel architecture having fewer layers than the other CNN, it does not require as many computational resources as ResNet. However, both achieved satisfactory results, especially because they were trained from scratch. Considering these aspects and the results, the implementation of CerraNetv3 is more sustainable and efficient for the method.

The second experiment aimed to analyze the accuracy of the third module of the method, Smart Mark Labeling (SML). Among all the procedures adopted, the main difference concerns the implementation of CerraNetv3 in mask labeling, which subsequently produced the reference masks used in the third experiment. Compared to the manual labeling procedure in INPE (2019), it is a significant speed and performance gain. Although this experimental module presents mislabeling, mainly for the BNOW group, by virtue of the lack of context in the created and classified segments, better results are achieved for the CFPS class group.

Regarding the third experiment, pixel-based classification, apart from examining the performance of DeepLabv3plus in terms of efficiency with few training samples training, whose reference masks were produced in the SML module, its outcomes were compared with the U-Net model. Both DL algorithms were trained in a basic

computer. However, due to mislabeling in the SML module and consequently the accuracy of the semantic segmentation models, achieving low scores supported by the F1-score, Precision, and IoU metrics.

5.1 Conclusion about the hypotheses

Below, the hypotheses of this master dissertation, shown in Section 1.2 of the Chapter 1, are reiterated:

- Hypothesis 1: The AI4LUC, based on the CerraNetv3 network, can assist in the automated labeling of mask's segments, generated from the satellite scene;
- Hypothesis 2: DeepLabv3+, while trained with CerraDatav3 and the labeled masks obtained with the help of CerraNetv3, will perform better than U-Net, in terms of semantic segmentation;
- Hypothesis 3: The AI4LUC method, based on the integration of two DCNNs, contributes to automate the pixel-based classification of remote sensing images.

Founded on results produced through the experiments, the three hypotheses are considered approved. Regarding the first hypothesis, CerraNetv3 contributed to the two-stage automation of the SML module, i) organization of unlabeled images into categories and thus it allowed the adjustment of customized parameters for each class in the mask generation functions; mostly ii) it was implemented in the classification of mask segments. In terms of Hypothesis 2, DeepLabv3plus obtained better overall scores, whose performance gain of nearly 71.35% compared to U-Net. Hypothesis 3 was accepted because the other two hypotheses were approved. Given as both CerraNetv3 and DeepLabv3plus worked together and ensured good performance, even as a result of the inferior ratings reported in experiment three.

5.2 Contributions and limitations

This research presented scientific and technological contributions but it also has some limitations as described as follows.

5.2.1 Scientific contributions

The AI4LUC was designed based on the INPE (2019) methodology, with the support of members of the TerraClass Cerrado project. The method stands out due

to the SML module, which combines morphological operations and a contextual classification CNN, CerraNetv3, in order to automate mask generating and labeling. According to the first experiment, CerraNetv3 was the best model in terms of accuracy as well as training speed, while ResNet-50 is 17 times more trainable parameters and takes 5 times more to train than the CerraNetv3.

In addition, another contribution was the combination demonstration between two CNNs, CerraNetv3 and DeepLabv3plus, which occurred with the production of reference masks labeled by the contextual classification model used as input data to train DeepLabv3plus. In this way, it is expected to combine other DL models to improve the learning gain and accuracy, consequently.

5.2.2 Technological contributions

As part of the development and improvement of the method, all source codes are open to any developer in the repository, GitHub platform. In addition, for this method, three datasets were designed on the Cerrado biome, called CerraData, containing images recorded by the Brazil-China CBERS-4A satellite, with two meters of spatial resolution. All its versions are available for download and ready to use in DL or ML applications, as well.

5.2.3 Limitations

The BNOW-Otsu function, which is a crucial step in the SML module pipeline, did not yield satisfactory results compared to CFPS-Otsu. This issue has a cascading effect on the third experiment, leading to mislabeling and directly impacting the training and performance of the subsequent semantic segmentation network module. Additionally, the limited number of training samples exacerbates this problem, as only 10,000 images and masks were utilized. To address this, it is advisable to utilize the entire training subset consisting of 79,200 images for training the networks. These factors collectively contribute to the low learning rate of the segmentation models.

Furthermore, it is essential to conduct comparative analyses of performance among other models for both the first and third experiments. This will help identify the strengths and weaknesses of each model and determine which one best aligns with the objectives at hand. Such analyses are crucial for obtaining valuable insights and making informed decisions regarding model selection.

5.3 Future work

Given the aspects previously discussed with respect to the limitations of the method, it is intended to correct the mask generation functions of the SML module. This involves not only mask-generating quality but also the context of the segment for CerraNetv3 and thus improves the classification. In addition, the post-processing of the output images after the generated procedure will be studied. The algorithm optimization with respect to the runtime is an important feature of the module that will be improved.

AI4LUC offers the possibility of adding more tasks into the pipeline, such as a module for generating time series, from which a segment can be selected from the SML module, corresponding to a class, and in this way extract the pixels from the images of a region. However, the possibility of implementing a multi-task learning CNN (WANG et al., 2021) will be analyzed, i.e., a network that has one input and more than one output, to generate segmented images and estimate the height/depth of objects in the scene.

New versions of the CerraData dataset are also being reviewed. At least two versions will be released, one with vegetation indices, and the other version designed with high-resolution Synthetic Aperture Radar synthetic aperture images. Both options are concerning contemplating the types of LULC in the biome. Furthermore a dataset on the Cerrado, it is planned to project for other Brazilian biomes. But for that, specialists will be needed in each case for the development of the dataset.

REFERENCES

AB'SABER, A. N. O domínio dos cerrados: introdução ao conhecimento. **Revista do Serviço Público**, v. 40, n. 4, p. 41–56, jul. 2017. Available from: <https://revista.enap.gov.br/index.php/RSP/article/view/2144>. 8, 9

ADORNO, B. V.; KÖRTING, T. S.; AMARAL, S. Contribution of time-series data cubes to classify urban vegetation types by remote sensing. **Urban Forestry and Urban Greening**, v. 79, p. 127817, 2023. ISSN 1618-8667. Available from: <https://www.sciencedirect.com/science/article/pii/S1618866722003600>. 2

AGÊNCIA NACIONAL DAS ÁGUAS E SANEAMENTO BÁSICO. **As regiões hidrográficas**. Parque Estação Biológica - PqEB, Brasília, 70770-901, 2014. Available from: <https://www.gov.br/ana/pt-br/assuntos/gestao-das-aguas/panorama-das-aguas/regioes-hidrograficas>. 9

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. Available from: <https://doi.org/10.1023/A:1010933404324>. 73

CÂMARA, G. On the semantics of big earth observation data for land classification. **Journal of Spatial Information Science**, 06 2020. 1, 3, 24

CARVALHO, O. L. F. de; CARVALHO JÚNIOR, O. A. de; ALBUQUERQUE, A. O. de; BEM, P. P. de; SILVA, C. R.; FERREIRA, P. H. G.; MOURA, R. S.; GOMES, R. A. T.; GUIMARÃES, R. F.; BORGES, D. L. Instance segmentation for large, multi-channel remote sensing imagery using Mask-RCNN and a mosaicking approach. **Remote Sensing**, v. 13, n. 1, 2021. ISSN 2072-4292. Available from: <https://www.mdpi.com/2072-4292/13/1/39>. 2

CHEN, L.-C.; PAPANDREOU, G.; SCHROFF, F.; ADAM, H. **Rethinking atrous convolution for semantic image segmentation**. arXiv, 2017. Available from: <https://arxiv.org/abs/1706.05587>. 21, 40

CHEN, L.-C.; ZHU, Y.; PAPANDREOU, G.; SCHROFF, F.; ADAM, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: FERRARI, V.; HEBERT, M.; SMINCHISESCU, C.; WEISS, Y. (Ed.). **Computer vision – ECCV 2018**. Cham: Springer International Publishing, 2018. p. 833–851. ISBN 9783030012342. 4, 21, 22, 40, 41

COUTINHO, L. M. **Biomas brasileiros**. São Paulo, SP, Brazil: Oficina de Textos, 2016. 657 p. ISBN 978-85-7975-254-4. 8

DU, S.; DU, S.; LIU, B.; ZHANG, X. Incorporating deeplabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. **International Journal of Digital Earth**, v. 14, n. 3, p. 357–378, 2021. Available from: <https://doi.org/10.1080/17538947.2020.1831087>. 2, 23, 24, 40

EITEN, G. Cerrado: caracterização, ocupação e perspectivas. In: PINTO, M. O. (Ed.). **Vegetação do Cerrado**. Brasília, DF, Brazil: Editora Universidade de Brasília, 1990. 7, 8

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Pantanal é dependente das águas do Cerrado**. Parque Estação Biológica - PqEB, Brasília, 70770-901, 2008. Available from: <https://www.embrapa.br/busca-de-noticias/-/noticia/18023883/pantanal-e-dependente-das-aguas-do-cerrado->. 9

_____. **Bioma Cerrado**. Parque Estação Biológica - PqEB, Brasília, 70770-901, 2014. Available from: <https://www.embrapa.br/cerrados/colecao-entomologica/bioma-cerrado>. 7

_____. **Matopiba: caracterização das áreas com grande produção de culturas anuais**. Parque Estação Biológica - PqEB, Brasília, 70770-901, 2014. Available from: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/105192/1/20140721-NotaTecnica6.pdf>. 11

_____. **GeoPortal TerraClass apresenta dados de uso e cobertura da terra no Cerrado**. Parque Estação Biológica - PqEB, Brasília, 2021. Available from: <https://www.embrapa.br/busca-de-noticias/-/noticia/62900537/geoportal-terraclass-apresenta-dados-de-uso-e-cobertura-da-terra-no-cerrado>. 12

ESRI. **Pixel classification**. Redlands, California, EUA, 2022. Available from: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/image-analyst/pixel-classification.htm>. 2, 3, 4

FONSECA, L. M.; KÖRTING, T. S.; BENDINI, H. do N.; GIROLAMO-NETO, C. D.; NEVES, A. K.; SOARES, A. R.; TAQUARY, E. C.; MARETTO, R. V. Pattern recognition and remote sensing techniques applied to land use and land

cover mapping in the brazilian savannah. **Pattern Recognition Letters**, v. 148, p. 54–60, 2021. ISSN 0167-8655. Available from: <https://www.sciencedirect.com/science/article/pii/S0167865521001677>. 1, 2, 3, 9, 20, 24, 43, 74

FOUREST, S.; BRIOTTET, X.; LIER, P.; VALORGE, C. **Satellite imagery: from acquisition principles to processing of optical images for observing the Earth**. Paris: Cépadureès editions, 2012. ISBN 978-2-36493-036-0. 9

GÉRON, A. **Hands on with machine learning with sckit-learn e TensorFlow: concepts, tools and techniques for building intelligent systems**. [S.l.]: Altas Books; Consultoria Eireli, 2019. 402 p. 2

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge, Massachusetts: Massachusetts Institute of Technology, 2016. 2, 14, 15, 16, 17, 18

HAN, Q.; YIN, Q.; ZHENG, X.; CHEN, Z. Remote sensing image building detection method based on mask R-CNN. **Spring Complex and Intelligent Systems**, n. 8, p. 1847–1855, 2022. 3

HÄNSCH, R.; HELLWICH, O. Classification of polsar images by stacked random forests. **ISPRS International Journal of Geo-Information**, v. 7, n. 2, 2018. ISSN 2220-9964. Available from: <https://www.mdpi.com/2220-9964/7/2/74>. 2, 21

HAYKIN, S. **Redes neurais: princípios e prática**. Porto Alegre: Bookman, 2001. ISBN 0-13-273350-1. 13, 14, 15, 16, 17

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. **CoRR**, abs/1512.03385, 2015. Available from: <http://arxiv.org/abs/1512.03385>. 41

IAKUBOVSKII, P. **Segmentation models PyTorch**. [S.l.]: GitHub, 2019. Available from: https://github.com/qubvel/segmentation_models.pytorch. 40, 42

IANDOLA, F. N.; HAN, S.; MOSKEWICZ, M. W.; ASHRAF, K.; DALLY, W. J.; KEUTZER, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. **arXiv:1602.07360**, 2016. 74

INSITUTO NACIONAL DE PESQUISAS ESPACIAIS. COORDENAçÃO GERAL DE OBSERVAçÃO DA TERRA. **SPRING: integrating remote**

sensingand GIS by object-oriented data modelling. São José dos Campos: INPE, 1996. Available from: <http://www.dpi.inpe.br/spring/portugues/tutorial/segmentacao.html>. 2

_____. **Projeto TerraClass Cerrado mapeamento do uso e cobertura vegetal do Cerrado**. São José dos Campos: INPE, 2013. Available from: <http://www.dpi.inpe.br/tccerrado/index.php?mais=1>. 12

_____. **Incremento anual de área desmatada no Cerrado Brasileiro**. São José dos Campos: INPE, 2023. Available from: <http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/cerrado/increments>. 10, 11

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Brasil em síntese: território**. [S.l.], 2023. Available from: <https://brasilemsintese.ibge.gov.br/territorio.html>. 7

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍíSTICA. **Mapa: biomas e sistema costeiro-marinho do Brasil 1:250000**. [S.l.], 2023. Available from: <https://geoftp.ibge.gov.br/informacoes_ambientais/estudos_ambientais/biomas/mapas/biomas_e_sistema_costeiro_marinho_250mil.pdf>. 7

INSTITUTO DE PESQUISA AMBIENTAL DA AMAZÔNIA. **Matopiba bate recorde histórico de desmatamento no Cerrado**. Brasília, DF, 2022. Available from: <https://ipam.org.br/matopiba-bate-recorde-historico-de-desmatamento-no-cerrado/>. 11

KHAN, A.; SOHAIL, A.; ZAHOORA, U.; QURESHI, A. S. A survey of the recent architectures of deep convolutional neural networks. **Artificial Intelligence Review**, v. 53, n. 8, p. 5455–5516, 2020. Available from: <https://doi.org/10.1007/s10462-020-09825-6>. 73, 74

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998. 23

LIU, D.; XIA, F. Assessing object-based classification: advantages and limitations. **Remote Sensing Letters**, v. 1, n. 4, p. 187–194, 2010. Available from: <https://doi.org/10.1080/01431161003743173>. 18

LIU, Z.; MAO, H.; WU, C.-Y.; FEICHTENHOFER, C.; DARRELL, T.; XIE, S. A convnet for the 2020s. **arXiv preprint arXiv:2201.03545**, 2022. 74

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. **arXiv**, p. 3431–3440, 2015. Available from: <https://arxiv.org/abs/1411.4038>. 21

MA, L.; LI, M.; MA, X.; CHENG, L.; DU, P.; LIU, Y. A review of supervised object-based land-cover image classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 130, p. 277–293, 2017. ISSN 0924-2716. Available from: <https://www.sciencedirect.com/science/article/pii/S092427161630661X>. 2

MAURANO, L.; ALMEIDA, C. A. de; MEIRA, M. Monitoramento do desmatamento do cerrado brasileiro por satélite - prodes cerrado. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 19., 2019. **Anais eletrônicos...** São José dos Campos: INPE, 2019. p. 191–194. Available from: <https://curtlink.com/0XEzBz>. 9, 12

MIRANDA, M. D. S.; SILVA, L. F. A. e; SANTOS, S. F. D.; SANTIAGO JÚNIOR, V. A. de; KÖRTING, T. S.; ALMEIDA, J. A high-spatial resolution dataset and few-shot deep learning benchmark for image classification. In: SIBGRAPI CONFERENCE ON GRAPHICS, PATTERNS AND IMAGES (SIBGRAPI), 35., 2022. **Proceedings...** Natal, Rio Grande do Norte, Brazil, 2022. v. 1, p. 19–24. 2, 28, 73

MIRANDA, M. de S.; MAXIMIANO, R. de S.; SANTIAGO JÚNIOR, V. A.; KÖRTING, T. S.; FONSECA, L. M. G. Classification of the water volume of dams using heterogeneous remote sensing images through a deep convolutional neural network. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA (GEOINFO), 22., 2021. **Proceedings...** São José dos Campos: INPE, 2021. p. 179–188. Available from: <http://mtc-m16c.sid.inpe.br/ibi/8JMKD3MGPDW34P/45U7H3L>. 32

MIRANDA, M. de S.; SANTIAGO JÚNIOR, V. A.; KÖRTING, T. S.; LEONARDI, R.; FREITAS JÚNIOR, M. L. de. Deep convolutional neural network for classifying satellite images with heterogeneous spatial resolutions. In: GERVASI, O.; MURGANTE, B.; MISRA, S.; GARAU, C.; BLEČIĆ, I.; TANIAR, D.; APDUHAN, B. O.; ROCHA, A. M. A. C.; TARANTINO, E.; TORRE, C. M. (Ed.). **Computational science and its applications – ICCSA 2021**. Cham: Springer International Publishing, 2021. p. 519–530. ISBN 978-3-030-87007-2. Available from: <https://doi.org/10.1007/978-3-030-87007-2_37>. 32

NEVES, A. K.; KÖRTING, T. S.; FONSECA, L. M. G.; SOARES, A. R.; GIROLAMO-NETO, C. D.; HEIPKE, C. Hierarchical mapping of brazilian savanna (cerrado) physiognomies based on deep learning. **Journal of Applied Remote Sensing**, v. 15, n. 4, p. 044504, 2021. Available from: <https://doi.org/10.1117/1.JRS.15.044504>. 1, 3, 9, 21, 22, 24, 28

NIU, B. Semantic segmentation of remote sensing image based on convolutional neural network and mask generation. **Mathematical Problems in Engineering**, v. 2021, 2021. 2, 19, 23, 24

NOGUEIRA, K.; SANTOS, J. A. D.; FORNAZARI, T.; SILVA, T. S. F.; MORELLATO, L. P.; TORRES, R. D. S. Towards vegetation species discrimination by using data-driven descriptors. In: IAPR WORKSHOP ON PATTERN RECOGNITON IN REMOTE SENSING (PRRS), 9., 2016. **Proceedings...** Cancun, Mexico, 2016. p. 1–6. 1, 27

OTSU, N. A threshold selection method from gray-level histograms. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 9, n. 1, p. 62–66, 1979. 20, 35

PACIFICI F., D. F. F. S. C. E. W. Neural networks for land cover applications. In: GRAÑA, M.; DURO, R. E. (Ed.). **Computational Intelligence for Remote Sensing**. Berlin: Springer, 2008. p. 267–293. Available from: <10.1007/978-3-540-79353-3_11>. 2, 13, 14, 16

PEDRAYES, O. D.; LEMA, D. G.; GARCÍA, D. F.; USAMENTIAGA, R.; ALONSO, Á. Evaluation of semantic segmentation methods for land use with spectral imaging using sentinel-2 and pnoa imagery. **Remote Sensing**, v. 13, n. 12, 2021. ISSN 2072-4292. Available from: <https://www.mdpi.com/2072-4292/13/12/2292>. 2, 22

PLAZA, J.; PLAZA, A.; PÉREZ, R.; MARTÍNEZ, P. Parallel classification of hyperspectral images using neural networks. In: GRAÑA, M.; DURO, R. J. (Ed.). **Computational intelligence for remote sensing**. Berlin, Heidelberg: Springer, 2008. p. 193–216. ISBN 978-3-540-79353-3. Available from: <https://doi.org/10.1007/978-3-540-79353-3_8>. 12

PORZI, L.; HOFINGER, M.; RUIZ, I.; SERRAT, J.; BULO, S. R.; KONTSCHIEDER, P. Learning multi-object tracking and segmentation from automatic annotations. In: IEEE/CVF CONFERENCE ON COMPUTER

VISION AND PATTERN RECOGNITION (CVPR), 2019. **Proceedings...** Los Alamitos, CA, USA: IEEE Computer Society, 2019. 23

PROGRAMA DE MONITORAMENTO DA AMAZÔNIA E DEMAIS BIOMAS. COORDENAçÃO GERAL DE OBSERVAçÃO DA TERRA. INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. **Metodologia Utilizada nos Projetos PRODES e DETER**. São José dos Campos: INPE, 2019. Available from: <http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes>. 3, 4, 10, 11, 12, 13, 23, 25, 43, 59, 60

REATTO, A.; CORREIA, J. R.; SPERA, S. T.; MARTINS, É. Solos do bioma cerado: aspectos pedológicos. In: SANO, S. M.; ALMEIDA, S. P. d.; RIBEIRO, J. F. E. (Ed.). **Cerrado: ecologia e flora**. Brasília, DF, Brazil: EMBRAPA, 2008. 876 p. 8, 9

REDDY, C.; REDDY, P. A.; KANABUR, V. R.; VIJAYASENAN, D.; DAVID, S. S.; GOVINDAN, S. **Semi-automatic labeling and semantic segmentation of gram-stained microscopic images from DIBaS dataset**. 2022. Available from: <https://doi.org/10.48550/arXiv.2208.10737>. 20

RIBEIRO, J. F.; WALTER, B. M. T. As principais fitofisionomias do bioma Cerrado. In: SANO, S. M.; ALMEIDA, S. P. d.; RIBEIRO, J. F. E. (Ed.). **Cerrado: ecologia e flora**. Brasília, DF, Brazil: EMBRAPA, 2008. 1, 7, 8, 29

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: convolutional networks for biomedical image segmentation. **arXiv**, p. 3431–3440, 2015. Available from: <https://arxiv.org/abs/1505.04597>. 4, 21

SAENKO, K.; KULIS, B.; FRITZ, M.; DARRELL, T. Adapting visual category models to new domains. In: DANIILIDIS, K.; MARAGOS, P.; PARAGIOS, N. (Ed.). **Computer vision – ECCV 2010**. Berlin, Heidelberg: Springer, 2010. p. 213–226. ISBN 978-3-642-15561-1. 23

SANO, S. M.; ALMEIDA, S. P. de; RIBEIRO, J. F. **Cerrado: ecologia e flora**. Brasília, DF: Embrapa Cerrados, 2008. ISBN 978-85-7383-397-3. 7, 8, 9

SANTIAGO JÚNIOR et al., V. A. **Project IDeepS: image classification via Deep neural networks and large databases for aeroSpace applications**. 2022. Available from: <https://github.com/vsantjr/IDeepS>. Access in: 28 Dec. 2022. 5

SANTIAGO JÚNIOR, V. A. A method and experiment to evaluate deep neural networks as test oracles for scientific software. In: IEEE/ACM INTERNATIONAL CONFERENCE ON AUTOMATION OF SOFTWARE TEST (AST). **Proceedings...** Pittsburgh, PA, USA: IEEE, 2022. p. 40–51. 2

SCIKIT LEARN. **Metrics: accuracy score**. [S.l.], 2023. Available from: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score>. 18

_____. **Metrics: f1-score**. [S.l.], 2023. Available from: <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html>. 19

SILVA, B. L. de C. e; SOUZA, F. C. de; FERREIRA, K. R.; QUEIROZ, G. R. de; SANTOS, L. A. dos. Spatiotemporal segmentation of satellite image time series using selforganized map. In: ISPRS PHOTOGRAMMETRY, REMOTE SENSING AND SPATIAL INFORMATION SCIENCES, 2022. **Proceedings...** Nice, France, 2022. p. 255–261. Available from: <https://doi.org/10.5194/isprs-annals-V-3-2022-255-2022>. 3

SILVA, F.; ASSAD, E. D.; EVANGELISTA, B. Caracterização climática do bioma Cerrado. In: SANO, S. M.; ALMEIDA, S. P. d.; RIBEIRO, J. F. E. (Ed.). **Cerrado: ecologia e flora**. Brasília, DF, Brazil: EMBRAPA, 2008. 8

SIMOES, R.; CAMARA, G.; QUEIROZ, G.; SOUZA, F.; ANDRADE, P. R.; SANTOS, L.; CARVALHO, A.; FERREIRA, K. Satellite image time series analysis for big earth observation data. **Remote Sensing**, v. 13, n. 13, 2021. ISSN 2072-4292. Available from: <https://www.mdpi.com/2072-4292/13/13/2428>. 1, 24

SOILLE, P. **Morphological image analysis: principles and applications**. New York: Springer-Verlag, 2nd edition, 2004. ISBN 978-3-642-07696-1. 18, 19

SREEDHAR, K.; PANLAL, B. Enhancement of images using morphological transformations. v. 4, n. 1, p. 33–50, 2012. Available from: <https://arxiv.org/pdf/1203.2514.pdf>. 19

TAN, M.; LE, Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: PMLR INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 36., 2019, Long Beach, California. **Proceedings...** [S.l.], 2019. p. 6105–6114. 74

TENSORFLOW. **Metrics: IoU**. [S.l.], 2022. Available from: <https://www.tensorflow.org/api_docs/python/tf/keras/metrics/IoU>. 19

VAPNIK, V. **The nature of statistical learning theory**. [S.l.]: Springer, 1999. 73

WANG, Y.; DING, W.; ZHANG, R.; LI, H. Boundary-aware multitask learning for remote sensing imagery. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 14, p. 951–963, 2021. 62

WANG, Z.; WANG, J.; YANG, K.; WANG, L.; SU, F.; CHEN, X. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with deeplabv3+. **Computers and Geosciences**, v. 158, p. 104969, 2022. ISSN 0098-3004. Available from: <https://www.sciencedirect.com/science/article/pii/S0098300421002545>. 1, 20

_____. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+. **Computers and Geosciences**, v. 158, p. 104969, 2022. ISSN 0098-3004. Available from: <https://www.sciencedirect.com/science/article/pii/S0098300421002545>. 3

ZHANG, T.; SU, J.; XU, Z.; LUO, Y.; LI, J. Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier. **Applied Sciences**, v. 11, n. 2, 2021. Available from: <https://www.mdpi.com/2076-3417/11/2/543>. 20

ZHENG, X.; CHEN, T. High spatial resolution remote sensing image segmentation based on the multiclassification model and the binary classification model. **Neural Computing and Applications**, p. 1433–3058, 2021. Available from: <https://doi.org/10.1007/s00521-020-05561-8>. 22, 24

APPENDIX A - FEW-SHOT LEARNING

This appendix presents the results of a collaboration between INPE, UNIFESP-São José dos Campos, and UFSCar. This collaboration used version 2 of the CerraData base (i.e. CerraDatav2). The article (MIRANDA et al., 2022) was the main outcome of such a collaboration. Credit author statement are as follows:

- Mateus de Souza Miranda: Dataset creation, Writing - Original Draft;
- Lucas Fernando Alvarenga e Silva: Software adaptation/development, Running experiments, Writing - Original Draft;
- Samuel Felipe dos Santos: Software adaptation/development, Running experiments, Writing - Original Draft;
- Valdivino Alexandre de Santiago Júnior: Conceptualization, Writing - Review and Editing, Supervision;
- Thales Sehn Korting: Writing - Review and Editing, Supervision;
- Jurandy Almeida: Conceptualization, Writing - Review and Editing, Supervision.

A.1 Few-shot learning experiment

An experimental evaluation was conducted considering the CerraDatav2 and a few-shot learning setting. Thus, 11 deep CNNs (KHAN et al., 2020) performances were compared, considering two learning methods: training from scratch and fine-tuning the pre-trained model on ImageNet. Also, the top-performing CNNs as a feature extractor only for two traditional ML algorithms: SVM (VAPNIK, 1999) and RF (BREIMAN, 2001). This investigation was carried out in a collaboration between researchers and post-graduate students at INPE, UNIFESP - *Campus São José dos Campos*, and UFSCar - *Campus Sorocaba*. Detailed information about such a research and experimentation can be read in Miranda et al. (2022).

Using the holdout method and random stratified sampling, the CerraDatav2 was splitted into training, validation, and test sets with 100, 100, and 49,800 tiles, respectively. For a fair comparison, the same splits were used by all the evaluated models. F1-score and accuracy (Acc) were chosen as performance measures. Five replications were performed to ensure statistically sound results. The mean and standard deviation of the performance measures for the test set of all the replications

were reported. Therefore, the CNNs models VGG11, VGG16, ResNet18, ResNet50, SqueezeNet, DenseNet161, InceptionV3, ShuffleNetv2 1.0, ResNeXt50, EfficientNet B4, and ConvNeXt-Tiny (KHAN et al., 2020; IANDOLA et al., 2016; TAN; LE, 2019; LIU et al., 2022) have been selected.

All these models were trained using the following hyperparameters: 100 epochs; early-stopping monitoring of the F1-score of the validation split for 10 epochs with Δ set to 0 (i.e., any amount of improvement reset the early-stopping counter); batches of 32 images; and stochastic gradient descent (SGD) optimizer with a learning rate of 0.001 and a momentum of 0.9. In addition, two different learning strategies were considered: (*i*) randomly initializing the weights, training from scratch; and (*ii*) initializing the weights from the publicly available ImageNet weights, in this case, fine-tuning the pre-trained model. Also, as a preprocessing step, all images were normalized using the Z-score normalization. When trained from scratch, the means and standard deviations of the RGB channels were computed from the training and validation sets. Otherwise, ImageNet statistics were used.

In addition to DNNs, SVM and RF classifiers (FONSECA et al., 2021) were tested. Firstly, for feature extraction, the images through the first layers of the best-performing CNNs in terms of the F1-score were pre-processed. Afterward, the hyperparameters of such classifiers were tuned using a grid search on the validation set. As for SVM, the C hyperparameter has been set between 10^{-1} and 10^{3} and the Gamma between 10^{-4} and 10^{0}, both in steps of powers of 10. Also, three different kernels were considered: linear, polynomial of degree three, and Radial Basis Function (RBF). Whereas RF varied the number of trees in the forest between 10^{0} to 10^{3} in steps of powers of 10, the number of features used to split a node was searched between 100%, 75%, 50%, 25%, square root, and the $\log_2$ from the total amount of features. Both *Gini* impurity and entropy were tested as criteria to measure the quality of the splits. Finally, after the models created by the different settings were evaluated on the validation set, we took the best one according to the F1-score and used it on the test set.

A.2 Results

Tables A.1 and A.2 present, at the top, the results on the test set obtained by every one of 11 DNNs considering two learning strategies, from-scratch, and fine-tuning, respectively. The best DNN, considering the F1-score, has an *, whereas the second best is with **. Moreover, the results for the SVM and RF classifiers with the features extracted by the two best CNNs, are presented at the bottom of the Table.

The best outcome for all 15 ML/DL techniques is highlighted in **bold**.

Table A.1 - Performance assessment: from-scratch approach.

Feature Extraction	**Classifier**	**F1-score**
VGG-11	DNN	45.50 ± 22.1
VGG-16	DNN	51.68 ± 20.1
ResNet-18**	DNN	74.58 ± 3.09
ResNet-50	DNN	59.44 ± 16.9
SqueezeNet	DNN	58.32 ± 7.10
DenseNet-161*	DNN	76.07 ± 1.55
InceptionV3	DNN	64.81 ± 7.26
ShuffleNetv2 1.0	DNN	49.57 ± 12.5
ResNeXt-50	DNN	70.11 ± 3.65
EfficientNet B4	DNN	49.10 ± 15.5
ConvNeXt-Tiny	DNN	54.77 ± 1.50
DenseNet-161	**RF**	$\mathbf{78.18 \pm 1.31}$
DenseNet-161	SVM	77.49 ± 1.87
ResNet-18	RF	77.47 ± 3.20
ResNet-18	SVM	77.22 ± 3.23

Detailing the from-scratch strategy, in which Table A.1 the DNN DenseNet-161 achieved the best F1-score (76.07%), approximately 2% better than the second best, ResNet-18. It is also noted that deeper models (e.g., VGG-16 and ResNet-50) underperformed their shallow versions (e.g., VGG-11 and ResNet-18). Although some approaches performed very well in terms of F1-score, others presented a low mean and a high standard deviation for both measures (e.g., VGG-11, VGG-16, ResNet-50, ShuffleNetv2 1.0, and EfficientNet B4).

However, for ML classifier algorithms, improvements have seemed when CNNs were applied as feature extractors. Enhancements of at least 2% were detected for DenseNet-161 and almost 3% for ResNet-18. Especially, the combination of DenseNet-161 as feature extractor and RF as classifier reached the best result (78.18%). As for the fine-tuning strategy, described in Table A.2, except for ShuffleNetv2 and EfficientNet B4, all other methods presented consistent results with great enhancement compared to the from-scratch strategy. Among all 11 DNNs, VGG-16 has the best F1-score (86.41%) followed by DenseNet-161 (86.38%). On the other hand, the ShuffleNetv2 scored the lowest performance compared to the other architectures.

Table A.2 - Performance assessment: Fine-tuning approach.

Feature Extraction	**Classifier**	**F1-score**
VGG-11	DNN	83.84 ± 2.60
VGG-16*	DNN	86.41 ± 1.22
ResNet-18	DNN	83.87 ± 1.91
ResNet-50	DNN	85.94 ± 2.18
SqueezeNet	DNN	84.49 ± 2.82
DenseNet-161**	DNN	86.38 ± 1.45
InceptionV3	DNN	77.85 ± 2.92
ShuffleNetv2 1.0	DNN	15.16 ± 3.14
ResNeXt-50	DNN	84.85 ± 2.08
EfficientNet B4	DNN	56.42 ± 5.89
ConvNeXt-Tiny	DNN	86.04 ± 2.35
VGG-16	RF	82.51 ± 1.05
VGG-16	SVM	83.59 ± 0.97
DenseNet-161	RF	86.16 ± 0.98
DenseNet-161	**SVM**	$\mathbf{86.57 \pm 1.36}$

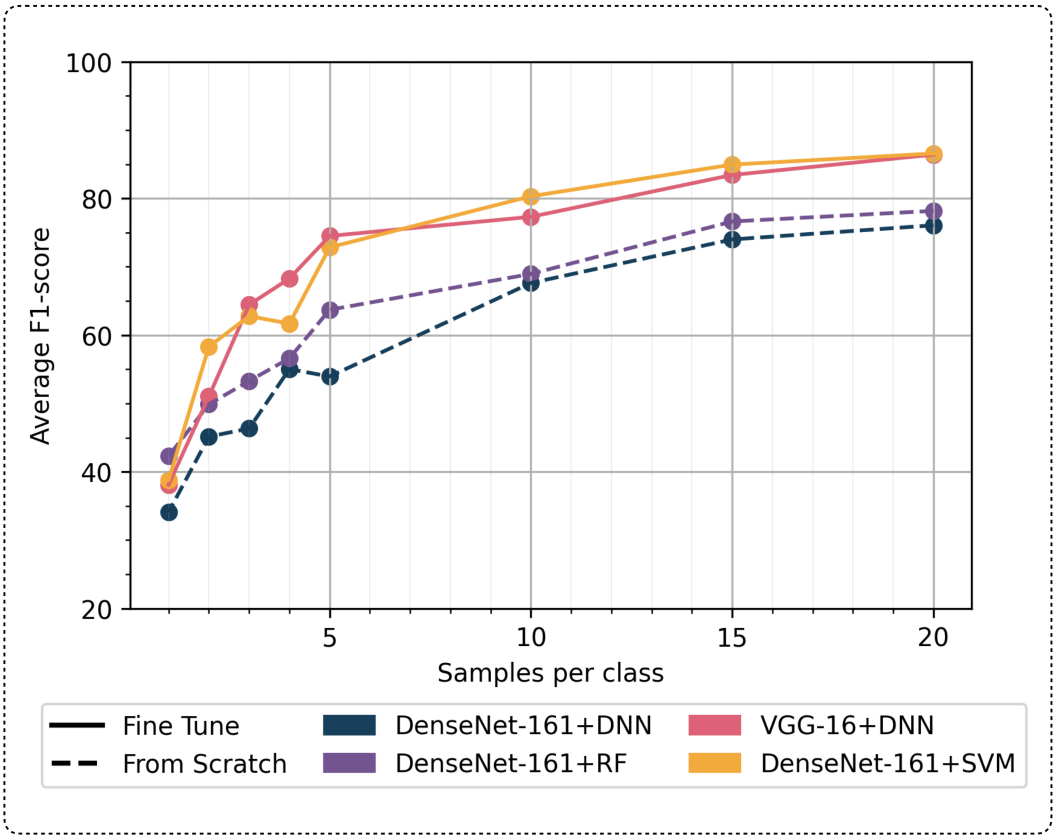
The issue of high standard deviation values observed in the from-scratch strategy is alleviated when pre-trained models are utilized. In addition, notable ML models performances improvements through CNNs as feature extractors show a strategy for training these classifiers and deployment to unknown data. Even though the gaps are small, the best result was achieved when DenseNet-161 extracted the features and SVM performed the classification.

A.2.0.1 Limits of learning from few samples

In order to explore the few-shot learning capabilities of the evaluated models, the number of training samples were decreasing until reaching the minimum, i.e., 1 sample per class. In particular, it was considered the best results network, in terms of training from scratch (DenseNet-161+DNN and DenseNet-161+RF) and fine-tuning based on ImageNet as well (VGG-16+DNN and DenseNet-161+SVM). These models were tested with smaller and smaller training sets, i.e., containing 20, 15, 10, 5, 4, 3, 2 samples, and, finally, only 1 sample per class.

Observed in Figure A.1, the fewer training examples, the worse the model performance. Based on F1-score, these results drop on average from 81%, for 20 samples per class to 38%, for 1 sample per class. However, increasing samples in the training set, e.g., 1 to 4 samples per class, the scores get a gain of 25%. Therefore, more samples, more learning performance gain by the DL and ML models. Regardless the

Figure A.1 - Few-shot stressing of the best-evaluated models.



remarkable characteristics of the DNNs and classical ML algorithms evaluated in this study, it is clear that smarter strategies are important to obtain the maximum benefit according to the few-shot learning philosophy.

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.