Active Learning for Improved Landcover Classification



Miguel Ribeiro Pereira
Master's Degree in Data Science
Department of Computer Science
Faculty of Sciences of the University of Porto
2024



Active Learning for Improved Landcover Classification

Miguel Ribeiro Pereira

Dissertation Report

Department of Computer Science

2024

Supervisor
João Pedro Pedroso, Associate Professor, Faculty of Sciences of the University of Porto

Acknowledgements

The first acknowledgment goes to my supervisor, Prof. João Pedro Pedroso. I want to thank all the support he gave me during the duration of the thesis, his comprehension and availability to guide me through the project, even when the schedule was tight.

My second thank you goes to (soon to be Dr.) Sofia Cardoso Pereira. Your presence and expertise were invaluable during every step of the way. I am so grateful that you have accepted to advise my thesis, and more than deserved to be titled as one of my supervisors.

I would like to also thank my family and friends' incredible support, which allowed me to get to where I am today. A special note goes to my brother Ricardo and my university peer Francisco, for their helpfulness and input during some of the project's most difficult steps.

A huge thanks to RSeT (and Maria in particular) for sharing the initial dataset and accompanying the dissertation's progress. And last but not least, I want to thank the CIBIO/BIOPOLIS research center for funding the project[1], and mention the support from researchers Luke and Luís.













[1]This work was co-funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 854248 and by the project TROPIBIO NORTE-01-0145-FEDER-000046, supported by Norte Portugal Regional Operational Programme (NORTE2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

Resumo

A plantação de caju na Guiné-Bissau constitui a principal atividade económica do país, em termos de volume de exportação e de população empregada no setor. A grande dependência neste fruto seco leva a uma produção descontrolada que está diretamente associada a alguns dos principais problemas desta nação, entre os quais o aumento dos níveis de desflorestação, insegurança alimentar, e a fragilidade da sua estrutura económica. Apesar do governo nacional estar a tomar medidas para controlar o desenvolvimento destas problemáticas, as soluções apresentadas até ao momento não se têm revelado suficientes. A primeira dificuldade passa pela quantificação e monitorização das plantações de caju, algo que é complexo e dispendioso de praticar a nível nacional. Este trabalho apresenta uma contribuição para o desenvolvimento de uma alternativa à monitorização no terreno automática e económica. A abordagem implementada usa imagens de Deteção Remota, juntamente com algoritmos de *Machine Learning*, para produzir um classificador capaz de detetar plantações de caju em toda a Guiné-Bissau, assim como numa região da República da Guiné, até ao rio Kogon. Utilizaram-se técnicas de *Active Learning* para desenvolver um conjunto de treino tão informativo quanto possível usando um número escasso de pontos, o que permitiu reduzir o peso da tarefa de anotação. Concluiu-se que este método é eficaz na obtenção uma descrição adequada e generalizada da cobertura de caju em toda a região, apesar de apresentar algumas limitações principalmente associadas ao processo de anotação remota dos pontos. Os resultados abrem portas para uma continuação usando novas formas de seleção de pontos a anotar no processo de aprendizagem ativa, bem como classificadores mais avançados.

Palavras-chave: Sustentabilidade, Monitorização da Cobertura do Solo, Deteção Remota, Observação Terrestre, *Machine Learning*, Classificação Supervisionada, *Active Learning*, Google Earth Engine, Python

Abstract

Cashew cultivation in Guinea-Bissau corresponds to the country's main economic activity, both in terms of export volume and population employed in the sector. The heavy reliance on this crop leads to uncontrolled production, which is directly associated with some of the country's main problems, including increased deforestation, food insecurity, and the fragility of its economic structure. Although the national government is taking measures to control the development of these problematics, the solutions presented so far have not proven to be sufficient. The first challenge lies in quantifying and monitoring cashew plantations, a complex and expensive task to carry out on a nationwide scale. This work presents a contribution toward developing an alternative for automatic and cost-effective field monitoring. The approach implemented uses Remote Sensing imagery, along with Machine Learning algorithms, to create a classifier capable of detecting cashew plantations throughout Guinea-Bissau, as well as in a region of the Republic of Guinea, up to the Kogon River. Active Learning techniques were employed to develop a training set that is as informative as possible using a scarce number of points, reducing the burden of expert annotation. It was concluded that this method proved effective in obtaining an adequate and generalized description of cashew coverage throughout the region, despite presenting some limitations mainly associated with the remote point annotation process. The results open the door to continued work using new methods of selecting points to annotate in the Active Learning process, as well as more advanced classifiers.

Keywords: Sustainability, Land Cover Monitoring, Remote Sensing, Earth Observation, Machine Learning, Supervised Learning, Active Learning, Google Earth Engine, Python

Contents

Acknowledgements iii

Resumo iv

Abstract v

Contents viii

List of Tables ix

List of Figures xi

List of Abbreviations xii

1 Introduction 1

1.1 Motivation 1

1.2 Objectives 2

1.3 Thesis Outline 3

2 Background 4

2.1 Guinea-Bissau: development plans and the cashew sector 4

2.2 Remote Sensing 8

2.2.1 Sentinel-2 and other alternatives 8

2.2.2 Derived features — vegetation indices, CCDC and GLCM 11

2.3 Machine Learning and Active Learning 13

3 State-of-the-art **17**

3.1 Image Classification with Land Cover Mapping 17

3.2 Cashew Orchard Detection and Monitoring 19

4 Data and Methods **22**

4.1 Overview 22

4.1.1 Tools and Libraries 22

4.2 Data 23

4.2.1 Train-test splits 26

4.3 Feature acquisition 28

4.4 Classification models 30

4.5 Active Learning 32

4.5.1 Query Strategies 32

4.5.2 Labeling Process 33

4.5.3 Initial Train Set and Batch Size 34

5 Developments **35**

5.1 EDA 35

5.1.1 95% density level ellipses 35

5.2 RSeT results 36

5.2.1 Preliminary Classification and Land cover map 37

5.2.2 First AL pipeline 39

5.3 Random Sampling results 41

5.3.1 Feature selection 44

5.3.2 Hyperparameter Optimization and 7-class vs. 2-class results 46

5.3.3 Second AL pipeline 48

5.4 Margin Sampling results 50

5.4.1 Third AL pipeline 52

5.4.2 Post-processing and final 2021 Land cover map 54

6 Conclusions and Future Work **57**

6.1 Conclusions 57

6.2 Future Work 58

Bibliography **64**

Appendix **65**

List of Tables

2.1 The 13 S2 bands. 9

4.1 Table of RSeT and RS nº of points. 26

4.2 Table of RSeT and RS train and test split percentages per class. 27

4.3 Tabular schema of the RSeT dataset. 31

5.1 RSeT test results after hyperparameter tuning. 37

5.2 RSeT land cover map class percentages. 39

5.3 RS test results after hyperparameter tuning. 46

5.4 RS train and test splits using the binary annotation nomenclature. 47

5.5 Absolute and relative Nº of MS samples per class. 52

List of Figures

2.1 2002-2017 Annual Cashew nut production and accumulated deforestation. 5

2.2 Evolution of a forest region transformed into a cashew plantation. 6

2.3 Sustainable Development Goals list. 7

2.4 Multi-spectral image representation. 9

2.5 Atmospheric transmissivity across the light spectrum for S2 and Landsat. 10

2.6 Visual representation of the CCDC algorithm. 12

2.7 Schematic of an Active Learning pipeline. 14

2.8 Margin Sampling visualization. 15

3.1 Cantanhez 2019 land cover map. 20

3.2 Cashew trees detection result. 21

4.1 Tools and Libraries Flowchart. 23

4.2 Visual representation of the Region of Interest. 24

4.3 Geographical location of RSeT data. 25

4.4 Feature acquisition flowchart 29

5.1 95% density level ellipses for each RSeT dataset class. 36

5.2 RSeT best model confusion matrix. 37

5.3 RSeT initial 2021 Land cover map. 38

5.4 BTs two most likely classes' barplots. 40

5.5 First AL BTs and RS results. 41

5.6 Geographical location of RS data. 42

5.7 Barplot of RS patch sizes. 43
5.8 Barplots of RS samples' class counts. 43
5.9 RS feature selection comparison. 44
5.10 Barplots of feature types after Above Average Importance feature selection. 45
5.11 Results for different C and γ hyperparameter values. 48
5.12 RS-AL results with different initial training sets. 49
5.13 Geographical location of MS data. 51
5.14 Comparison between Margin Sampling and Random Sampling. 53
5.15 Percentage of discarded points per MS-AL iteration. 54
5.16 Scores after Sieve filter with different parameter values. 55
5.17 Final 2021 Land cover map. 56
5.18 Final 2021 Land cover map with an overlay of Guinea-Bissau communities. 56

1 Representation of the manually obtained polygons for Savanna. 65
2 95% density level ellipses in the RED-NIR space. 66
3 95% density level ellipses in the SWIR2-NIR space. 67
4 Location of the 9 sub-regions used for the 1^{st} AL configuration. 68

List of Abbreviations

AL Active Learning

ANN Artificial Neural Network

API Application Programming Interface

BOA Bottom of the Atmosphere

BTs Breaking Ties

CCDC Continuous Change Detection and Classification

CDI Cloud Displacement Index

CNN Convolutional Neural Network

CRS Coordinate Reference System

CSV Comma Separated Value

DL Deep Learning

DT Decision Tree

EDA Exploratory Data Analysis

EO Earth Observation

EVI Enhanced Vegetation Index

ESA European Space Agency

FREL Forest Reference Emissions Level

GDP Gross Domestic Product

GEE Google Earth Engine

GLCM Gray-Level Co-Occurrence Matrix

HP Hyperparameter

MCLU Multiclass Level Uncertainty

ML Machine Learning

MS Margin Sampling

NBR Normalized Burn Ratio

NDVI Normalized Difference Vegetation Index

NIR Near-infrared

PCA Principal Component Analysis

REDD+ Reduced Emissions from Deforestation and Degradation

ResNet Residual Network

RF Random Forest

RGB Red-Green-Blue

ROI Region of Interest

RS Random Sampling

S2 Sentinel-2

SDG Sustainable Development Goal

SVC Support Vector Classifier

SVM Support Vector Machine

SWIR Short-wave infrared

TOA Top of the Atmosphere

UN United Nations

UNFCCC United Nations Framework Convention on Climate Change

VCS Verified Carbon Standard

Chapter 1

Introduction

1.1 Motivation

The Republic of Guinea-Bissau is located on the West African coast. Its mainland is bounded by Senegal to the north, Guinea to the east and south, and the Atlantic Ocean to the west. It also includes a group of 88 islands that constitute the Bissagos archipelago, as well as other islands that lie off the coast.

With a total area of $36,120\,\text{km}^2$, most of Guinea-Bissau's land cover corresponds to different types of savanna woodlands, tropical and sub-tropical forests, as well as other types of open or closed-canopy landscapes [1]. This vegetation is marked by the two seasonal behavior that are typical in this region: a rainy season from May to October, and a dry season from November to April. These environments hold a large biodiversity, both in terms of plant and animal variability.

With a population estimate of around 2.1 million by the end of 2023, it holds roughly a quarter ($\sim 500,000$) of its population on the country's capital, Bissau [2]. Characterized as a developing country, the economy revolves around the primary sector (farming and agriculture, forestry and fishing being the most noticeable activities). It is estimated that the Agriculture sector alone covers one-third of the Guinea-Bissau GDP, and employs $68-80\%$ of its entire adult population [3].

There is an overreliance on cashew nut cultivation and commerce in Guinea-Bissau. This product accounts for over 90% of the national exported goods, and with more than 70% of the families living from it, Guinea-Bissau shows multiple characteristics of a single-product economy [3, 4]. This dependence on cashew severely undermines the country's agricultural and economic growth potential. New plantations, associated with slash-and-burn agricultural techniques, are also important driving factors for the country's speeding deforestation rate [1]. Therefore, the established economic structure is fragile, and its roots are major threats to the biodiversity of the region, which can become a deep humanitarian, economic, and environmental crisis.

To address these issues, the government is keen on implementing measures that are aimed towards sustainable development. A few of those involve the implementation of a nationwide

forest monitoring system and a reform of the agriculture sector, where cashew production is inserted, for a better regulated and more controlled production [1]. However, monitoring and quantification of these events on a nationwide scale and across a large time span is an incredibly hard task, especially if done completely on-site.

Thus, an alternative solution involves the remote monitoring of the country's land cover. A viable approach could be based on the imagery of orbiting satellites, which monitor either a specific or an almost global area and obtain a new image of the Region of Interest (ROI) with a certain periodicity. Some corporations provide this as a service, but there are also some freely accessible options. For instance, the Copernicus Sentinel-2 (S2) is a mission developed by ESA, constituted by 2 remote satellites (the first launched in 2015, and the second in 2017) that orbit the entire globe, being able to produce an almost entirely worldwide image across 13 different spectral bands, with a revisiting time of 5 days, up to a 10m spatial resolution [5]. The image collection produced by the S2 mission (or other similar projects and services) creates a large amount of spatial and temporal information that is ready to be used for the implementation of a cost-effective monitoring solution.

The goal of this work is to automate the process of identifying cashew orchards with the help of satellite data following a Machine Learning (ML) perspective. However, the usage of supervised learning techniques requires the acquisition of land cover classes for a sizeable number of samples. The problem is that this is a slow and costly process, which requires a lot of expert time to perform the annotations. This is where an Active Learning (AL) approach may be appropriate — using a heuristic to select only the most valuable and informative points to label and add to the train set. This has the potential of achieving optimal performance with an ML classifier, whilst significantly reducing the labeling effort required to automate the process of detecting cashew.

1.2 Objectives

The main objective of this dissertation is the production of a land cover map that indicates the location of cashew orchards in the region of interest — Guinea-Bissau and the northern part of Guinea, up to the Kogon river —, with the help of remote sensing data and machine learning techniques. More specifically, the goals of this dissertation are:

1. To obtain a dataset that is sufficiently broad and representative of the whole ROI.
2. To explore Active Learning techniques, their capacity to efficiently select points to label, and their subsequent performance.
3. To produce a land cover map based on Machine Learning predictions over the entire ROI for the year where we have ground truth data (2021).

1.3 Thesis Outline

This dissertation is organized in the following manner:

- **Chapter 2** introduces the background and context needed for the understanding of the following sections. This includes the characteristics of the country and Region of Interest (ROI), some important notions of Earth Observation (EO) and Remote Sensing techniques, and an overview of how the field of Active Learning can be applied in the context of spectral image classification.
- **Chapter 3** regards the state-of-the-art methods for satellite imagery classification, following a Machine Learning (ML) approach, and states some advancements and techniques specifically for the problem of remote cashew detection.
- **Chapter 4** describes the data used during the thesis and the characteristics of the acquisition process. It also mentions the methods, tools, and setups needed for the Machine Learning pipeline, including feature selection, hyperparameter optimization, evaluation, and the heuristics used during the Active Learning (AL) process.
- **Chapter 5** states all the relevant developments and results of the work, from all the steps of the implemented pipeline, up to the production of the final ROI land cover map.
- **Chapter 6** concludes the dissertation and overviews future work possibilities.

Chapter 2

Background

2.1 Guinea-Bissau: development plans and the cashew sector

Guinea-Bissau presents multiple characteristics of a developing country: the nation's electrification rate is still low, not exceeding 45% in 2020 [1]; at least 70% of the working population is employed in the agriculture sector [3]; malnutrition rates were reported to have risen from 22.2% to 28% between 2012-2018 [3]; a significant chunk of the population (between 11% and 51%, depending on the year and region) suffers from chronic food insecurity or scarcity [1]. The humanitarian situation is aggravated by high rates of poverty, illiteracy, and overall lack of knowledge of good agricultural and dietary practices [3, 6]. Key sectors such as transport, industry, water resources, health, and energy remain largely underdeveloped [1]. Political instability is clear, with two coup attempts happening in the last two years (February 2022 and November 2023), which is a main risk factor for the execution of any improvement plan made to tackle these issues [3].

The problems do not appear just at a humanitarian level. In recent years, the annual deforestation values have been noticeably higher, and seem to be increasing at a faster pace [7]. The environmental costs of this phenomenon go above just the clear biodiversity and habitat loss. Intensive deforestation is known to indirectly increase the probability of floods [8], with the effects already being visible in Guinea-Bissau specifically: from 2015 onward, the increase in floods in southern and eastern regions has destroyed 8% of the country's total rice production [3], one of their most consumed crops. Climate change, exacerbated by local human activity and exploration, is an increasing factor for food insecurity. All of this fits with Guinea-Bissau being ranked as the fourth country most vulnerable to climate change, according to the 2019 Notre Dame Global Adaptation Initiative country index [3].

We note three main causes that contribute to increasing deforestation: firstly, the expanding electricity sector, as it is estimated that over 80% of the primary energy source in the country is biomass [1]; secondly, illegal logging of wood for commercialization has also been in expansion in recent years [7]; and thirdly, the increase of agricultural practices, mainly the production

of cashew, rice and peanuts. The entire primary sector (agriculture, farming, forestry, and fisheries) accounts for at least 60% of the country's GDP [9], with a special note to cashew, which covers between 18%-55% (depending on the region) of the entire planted land [10]. The relation between cashew cultivation and deforestation rates is visible in Figure 2.1: from 2002 through 2012, researchers noticed a significant correlation, concluding that 89% of the deforestation numbers in those years can be directly attributed to the expansion of cashew production. Between 2013 and 2017, despite the correlation decreasing, there is still a statistically significant association between the two phenomena [7].

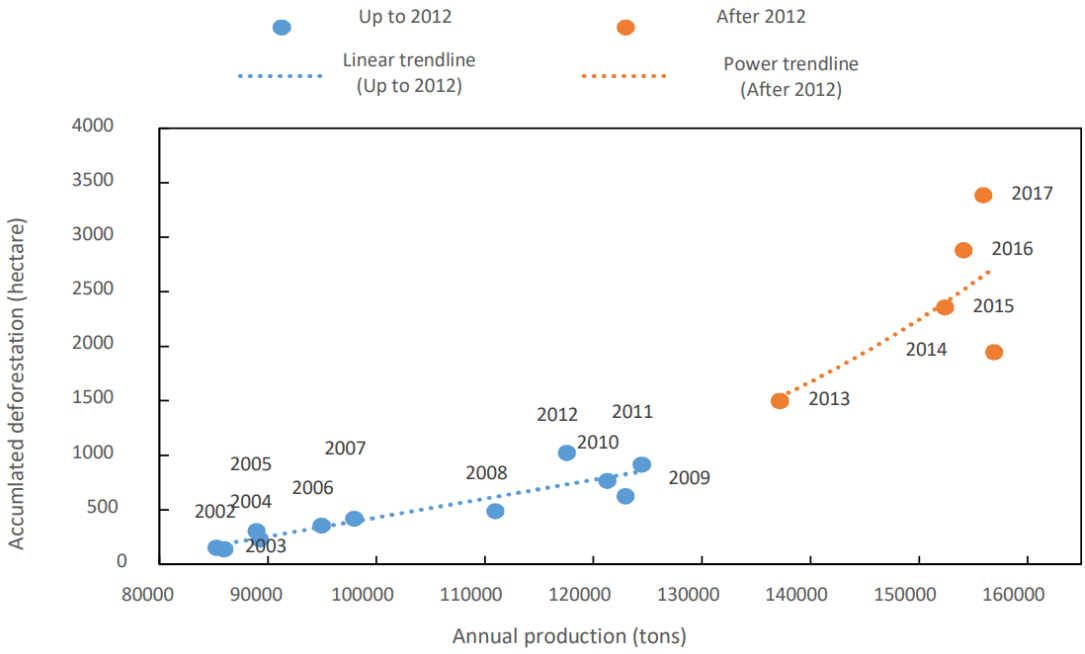


Figure 2.1: Annual Cashew nut production and accumulated deforestation of native forest, from 2002 to 2017. Image taken from [7].

The cashew sector is a main topic that is directly or indirectly associated with a multitude of the aforementioned affairs. In an environmental sense, it is a driving force associated with the transformation of the country's native biome and annual cumulative deforestation increase, with indirect impacts on native species and organisms [11]. An example of a typical cashew orchard emergence and evolution is shown in Figure 2.2. In an economic sense, this crop corresponds to a noticeable percentage of the country's exported goods (90%), covers over 50% of the total arable land, and employs up to 80% of the labor force of the entire country [3]. However, the available data and statistics regarding cashew production and plantation are considered unreliable and far below the real values, due to a lot of cashew plantations and transactions being left unregistered [12, 13].

The fact that the country's economic sector is so centered around this nut also makes it much more fragile. Typically, single-crop economies are very susceptible to pests and climate conditions, factors that heavily dictate an orchard's production [9]. Since this crop is underused for food and is mostly for export, it is also susceptible to market shifts [14]. Besides that, rudimentary

agricultural practices and lack of awareness of an orchard's age threaten long-term production values, as it is estimated that after 26 years, most trees give barely any production [10].

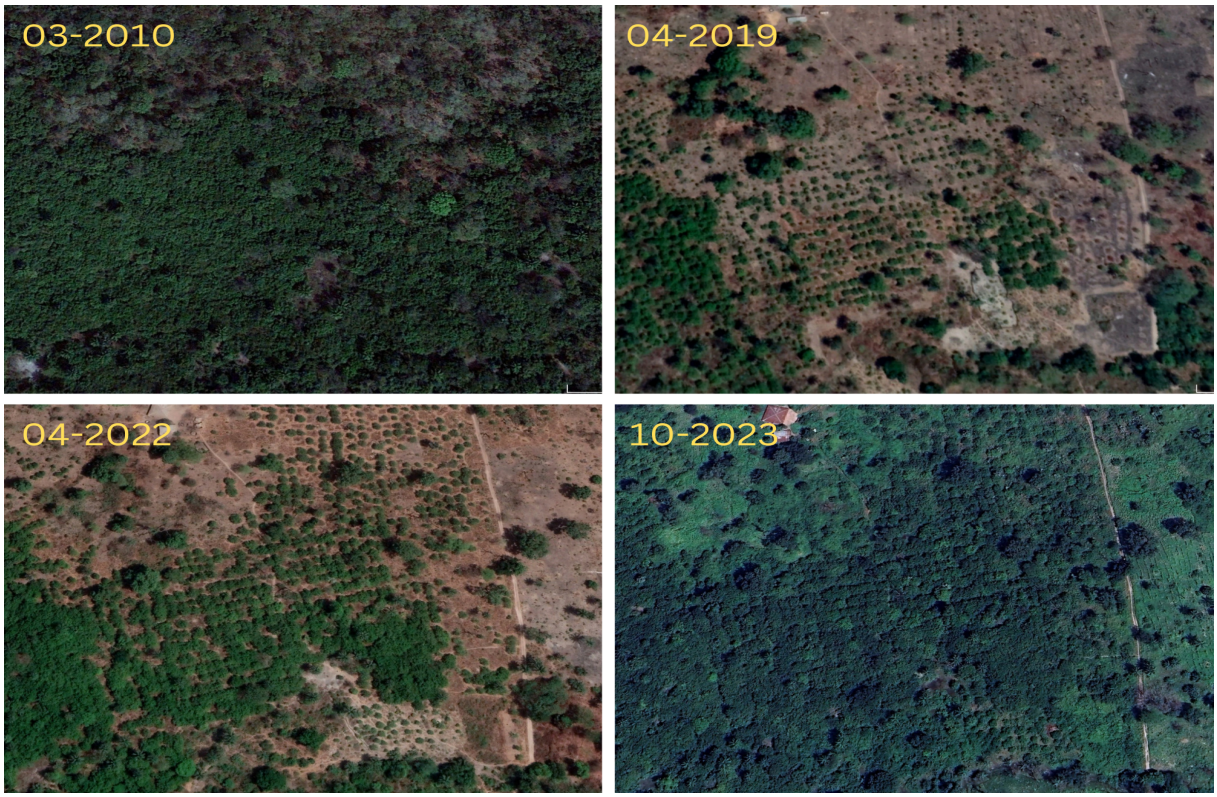


Figure 2.2: Evolution of a forest region transformed into a cashew plantation. In the span of 13 years, the existing vegetation was wiped out, and the land became almost fully covered by cashew trees. Images taken from Google Earth Pro.

The national government is not oblivious to any of these issues and shows interest in finding resolutions that are more aligned with nationwide sustainable development. In 1995, they ratified the United Nations Framework Convention on Climate Change (UNFCCC), a UN near-global initiative focused on preventing dangerous and irreversible human interference in the climate system [15]. In the case of Guinea-Bissau, this commitment involves an economic and technological transition with green initiatives from the ground up, receiving supporting funds for climate change activities, like the reduction of Greenhouse Gas emissions and carbon emissions, and the capacity to adapt and resist the already existing effects of climate change.

Also, as a UN member-state, the country must comply with the more recently imposed (2015) Sustainable Development Goals (SDGs). This initiative was set as a universal call to action to end poverty and protect the planet, being divided into 17 different sections, shown in Figure 2.3. Their areas are very diverse, as they cover humanitarian, social, economic, and environmental measures. However, the 17 SDGs are integrated, following the belief that action in one area is likely to affect the outcomes in others, and that development must balance sustainability across all those different fields.



Figure 2.3: The 17 Sustainable Development Goals.

From their commitment to the 17 goals, there are mainly three of them closely related to the theme of this dissertation. **SDG 13 (Climate Action):** the country is required to find ways for human exploration to occur with minimal Greenhouse Gases and carbon emissions, as well as sustainable land management, without significantly damaging the native and existing forests, as part of the Reduced Emissions from Deforestation and Degradation (REDD+) program. **SDG 2 (Zero Hunger):** the country's agricultural practices are leading to saturated and less productive fields over time [14], following a cash-crop mentality: the main objective of the produced crops is for selling and profiting and not for subsistence, especially cashew. This single-crop structure, with mostly economic intents, compromises the country's long-term sustainability and food management. **SDG 15 (Life on Land):** Habitat loss is threatening the survivability of many native species, such as African Forest elephants, *Loxodonta cyclotis* [11]. The government pledged to support the country's rich biodiversity, taking measures for controlling deforestation and land cover changes, as well as expanding the protected areas spread across the nation [16].

There are other goals that are directly or indirectly related to the purposes of this thesis. This includes, for example, SDG 1 (no poverty), which is the one Guinea-Bissau prioritizes for the 2030 Agenda [3]. But in broad terms, using information that is publicly available [16], the Sustainable Development Report shows that Guinea-Bissau ranks 156 out of the 166 parties. With only 20% of the proposed targets being set to be achieved by the deadline of 2030, this gives them a bottom 10 ranking position, showing there have been clear difficulties associated with the implementation of the necessary measures across most categories.

2.2 Remote Sensing

A more thorough land cover and land use monitoring plan would impact the resolution of some of the aforementioned issues, with special attention to cashew production, considering the central role it takes in the entire nation's operation. We need some basis of information to describe the problem using quantifiable metrics and reliable estimates, not only to have a better grasp of the actual state of the situation and act accordingly, but also because such estimates are required by programs like the SDGs, to decide if there is the need to direct funding for the implementation of initiatives to achieve their proposed results.

Monitoring should be done continuously, with some periodicity, to understand the evolution of the matter. This can require a large amount of resources, funding, and manpower. Hence, arises the need for a cost-effective land cover monitoring approach, that requires small amounts of resources, is easily scalable to the entire country, and can be done during long time periods. Considering the requirements, a Remote Sensing approach might present itself as a very effective solution, and will be explored throughout the dissertation.

2.2.1 Sentinel-2 and other alternatives

For a remote automated approach, satellite or aerial imagery options are available. An example of such an alternative is the European Copernicus Sentinel-2 program [5]. This mission is composed of two orbiting satellites (the first launched in 2015, and the second in 2017) capable of almost global coverage, which are phased at $180°$, allowing for a high revisiting frequency and the production of a new global image in just five days. The data gathered across multiple years is free and open for public use. The S2 satellites collect information about the reflectance values of different regions of the electromagnetic spectrum (that is, the proportion of light in a certain wavelength interval that is reflected by the environment) across a total of 13 discrete bands, as shown in Table 2.1. These bands are mostly scattered across the visible range — RGB bands — and the infrared range — Near-infrared (NIR) and Short-wave infrared (SWIR). The spatial resolution of the bands (the distance represented in a pixel unit) also differs: 4 bands have 10m resolution, 6 have 20m resolution, and 3 have 60m. The images obtained by S2 are called multi-spectral images, due to the number of bands being in the units-tens range, and most bands covering a rather broad wavelength interval. This contrasts with Hyper-spectral imagery, for which the data is captured across hundreds or thousands of bands, each covering a very narrow wavelength range.

Satellites like S2 produce a geographic data type called a raster. The data is stored as a pixel grid, where each pixel is attributed to specific coordinates (according to a pre-stipulated Coordinate Reference System). When there are multiple bands, one can think of a raster file as a stack of these gridded images, one per each satellite band, like it is presented in Figure 2.4. Rasters can also be combined based on an aggregation function if they present any spatial overlap, through a technique called compositing.

Band Number	Band Description	Wavelength Range (nm)	Resolution (m)
B1	Coastal aerosol	433–453	60
B2	Blue	458–523	10
B3	Green	543–578	10
B4	Red	650–680	10
B5	Red-edge 1	698–713	20
B6	Red-edge 2	733–748	20
B7	Red-edge	773–793	20
B8	Near infrared (NIR)	785–900	10
B8A	Near infrared narrow (NIRn)	855–875	20
B9	Water vapour	935–955	60
B10	Shortwave infrared/Cirrus	1360–1390	60
B11	Shortwave infrared 1 (SWIR1)	1565–1655	20
B12	Shortwave infrared 2 (SWIR2)	2100–2280	20

Table 2.1: Spectral characteristics of the 13 bands of Sentinel-2 (S2): wavelength ranges and resolutions for each band. Source: [17]

Because of the near-global coverage, Sentinel collects data in smaller chunks called tiles, which represent a 109.8km $\times$ 109.8km area of the Earth's surface, allowing for more efficient processing. Tiles can still be assembled to form a spatially continuous image through a process called mosaicking.

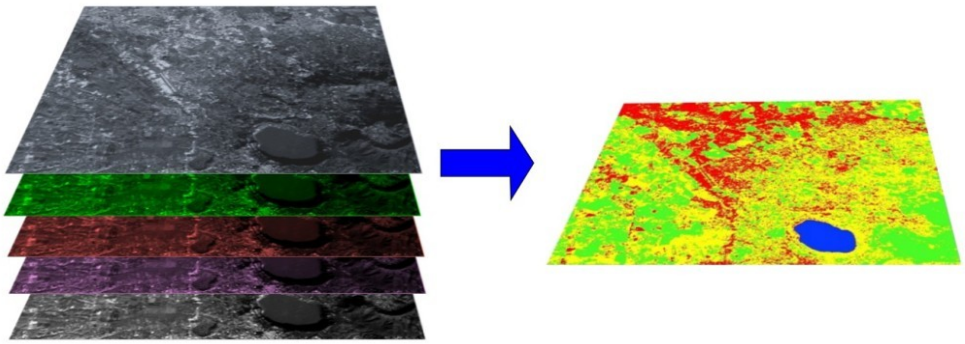


Figure 2.4: Representation of a multi-spectral image (raster data). Image taken from QGIS's Semi-Automatic Classification Plugin docs.

Alternative sources can be used for land cover monitoring or land cover change tasks. The Landsat [18], made jointly by NASA/USGS, is the longest continuous space-based record of Earth's land in existence, starting as far back as 1972. The characteristics of the image collection produced by this mission are very similar to what S2 offers, but there are two major disadvantages to be considered for our project. The first is the revisiting time: the most recent Landsat mission (Landsat-9) produces a new world image every eight days, in comparison to Sentinel's five days. The second (and biggest) reason is the spatial resolution: all but one band of the Landsat satellites have $30+$m spatial resolution, in comparison to S2 mostly having 10m and 20m bands. Higher spatial precision and resolution are important factors for cashew orchard monitoring, due to

specific characteristics of this class (mostly its spatial structure and organization) [12]. As such, for this project, we have decided to work with Sentinel data.

There exists a wide range of options besides Landsat and Sentinel-2. In recent years, satellites with sub-meter spatial resolution have become more and more common. Also, drone or aerial image acquisition techniques are becoming more prevalent [19–21]. However, most of these solutions are not openly accessible. They can also be unreliable in terms of spatial and temporal coverage, as most do not have such a large, near-global coverage (which makes it difficult to find imagery for Guinea-Bissau and West African countries in general), and they typically do not follow a consistent monitoring and image acquisition period. These characteristics make them unsuitable for the intended use in this dissertation.

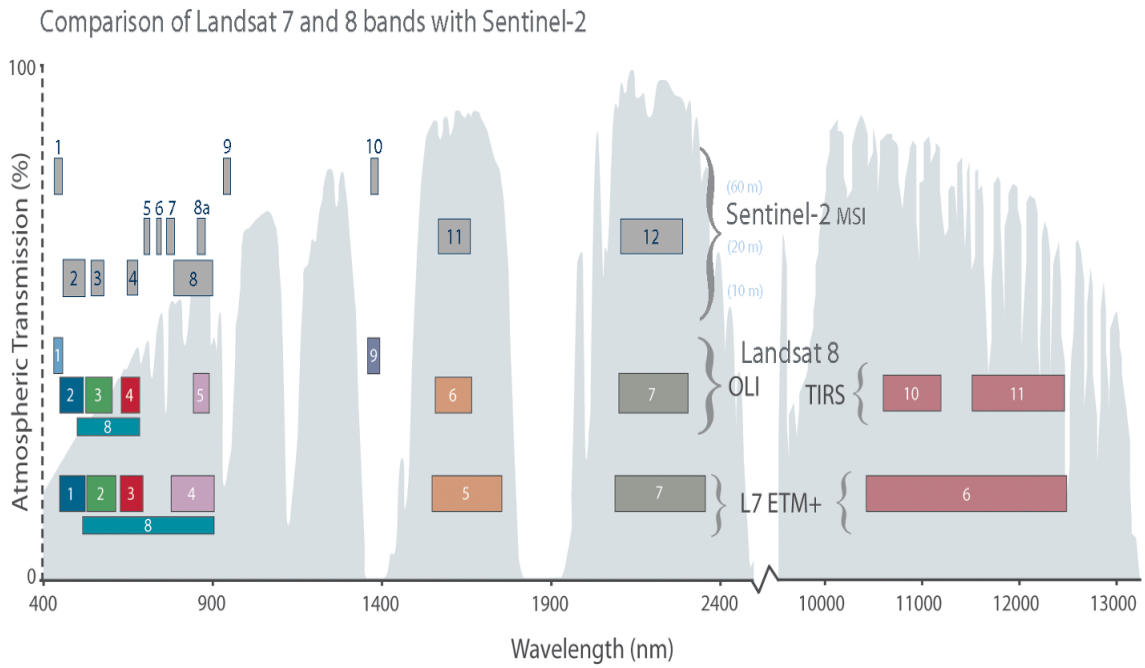


Figure 2.5: Atmospheric transmissivity across the light spectrum. Also presents the comparison of Top of the Atmosphere (TOA) transmission values of Landsat 7, Landsat 8 and Sentinel-2 satellites. Taken from User Guides — Sentinel-2 MSI — Overview

It is important to note that S2 and other orbiting satellite acquisitions typically measure Top of the Atmosphere (TOA) transmission values. The atmospheric transmissivity varies significantly across the light spectrum, as shown in Figure 2.5. For orbiting satellites, the light detected is affected by absorption and refraction phenomena in the different atmospheric layers. For land cover uses, certain corrections can be made to estimate the Bottom of the Atmosphere (BOA) reflectance rates at the Earth's surface level instead, removing the atmospheric effects. Since 2018, BOA images for the Sentinel program have been derived from the initially obtained TOA images, and are also openly distributed. However, for such intended EO purposes, bands like Band 1, 9, and 10 are typically discarded: they are designed to quantify some possible atmospheric influences, such as the diffusion and absorption of light by atmospheric gases and aerosols or disturbances provoked by smoke (from wildfires or pollution), so after the initial atmospheric correction step they can

typically be removed [14].

2.2.2 Derived features — vegetation indices, CCDC and GLCM

One of the main parts of this thesis involves the identification of different types of vegetation. Most plants have a noticeable behavior regarding their transmittance and reflectance, called the red edge: they have lower reflectance values in most of the visible spectrum, with a slightly higher green reflection than red and blue. However, there is a region of rapid reflectance increase in the NIR range of the spectrum, due to chlorophyll being highly transparent to higher wavelength radiation. This distinct vegetation behavior is better captured by metrics like the Normalized Difference Vegetation Index (NDVI) instead of the S2 original reflectance bands. This band can be obtained by taking the Red and NIR bands to create an index normalized in the range $[-1, 1]$ that is easily interpretable: the higher and healthier vegetation density, the higher (closer to 1) will be the NDVI index.

However, NDVI has its limitations, mostly due to being highly sensitive to certain conditions (atmospheric effects, clouds, soil wetness). Other indices, such as the Enhanced Vegetation Index (EVI), are designed for similar purposes without showing these disadvantages. There are two used adaptations of this index: the first, typically called EVI, is based on the reflectance values of three bands, NIR, Red and Blue; the second, called EVI2, is a two band adaptation, defined with just NIR and Red reflectances. Both metrics fill some of the limitations of NDVI, mainly by being less chlorophyll sensitive (thus more responsive to canopy structure alterations), less affected by atmospheric influences, and also less prone to saturation [22].

Aside from vegetation indices, features that regard temporal information might be interesting to explore. The land cover of Guinea-Bissau is composed of geographical categories such as savanna woodlands, for example. This biome has a clear seasonal variability, where the land cover is practically filled with green vegetation during the humid season, but most of it disappears in the dry season, leaving just a few sparse (sometimes deciduous) trees in a vast majority of uncovered land. Another example where temporal information can be useful is for the identification of cashews, the class of interest. Cashew orchards usually follow three different stages of maturity, that are independent of the season: an early development stage with very small trees and practically non-existent canopy coverage; a middle-aged phase where the trees' canopy is still increasing, and overall orchards do not yet reach complete canopy coverage, but now the trees are big enough to be detected by low-medium resolution satellite imagery; a final stage with fully grown cashew trees, associated with a complete or near-complete canopy coverage. From a spectral signature point-of-view, orchards in the early phase have a very similar signature to that of plain uncovered soil, whereas final stage plantations have a similar aspect and signature to some types of dense forest. However, temporal features might detect the changes and growth happening in a cashew orchard, which would be a clear distinctive behavior, compared to typical forests and empty fields.

So, a possible way of extracting temporal features from a time series of satellite raster images

(also called an image collection) is by obtaining the coefficients from the Continuous Change Detection and Classification (CCDC) algorithm [23]. This algorithm computes a time series modeling for each individual pixel, together with a change detection heuristic. For each pixel, a time series of the detected values is defined for each of the bands present in the image collection. But for some of the bands, called breakpoint bands, a change detection heuristic is applied to the entire series. This can split the time series into multiple segments, and for each, a harmonic regression model is fitted (a series of sines and cosines, usually up to 3 different frequencies). The coefficients of the harmonic regression of a certain segment can then be extracted and used as temporal features for a Machine Learning algorithm to perform a classification task. An example of this algorithm in action is presented in Figure 2.6.

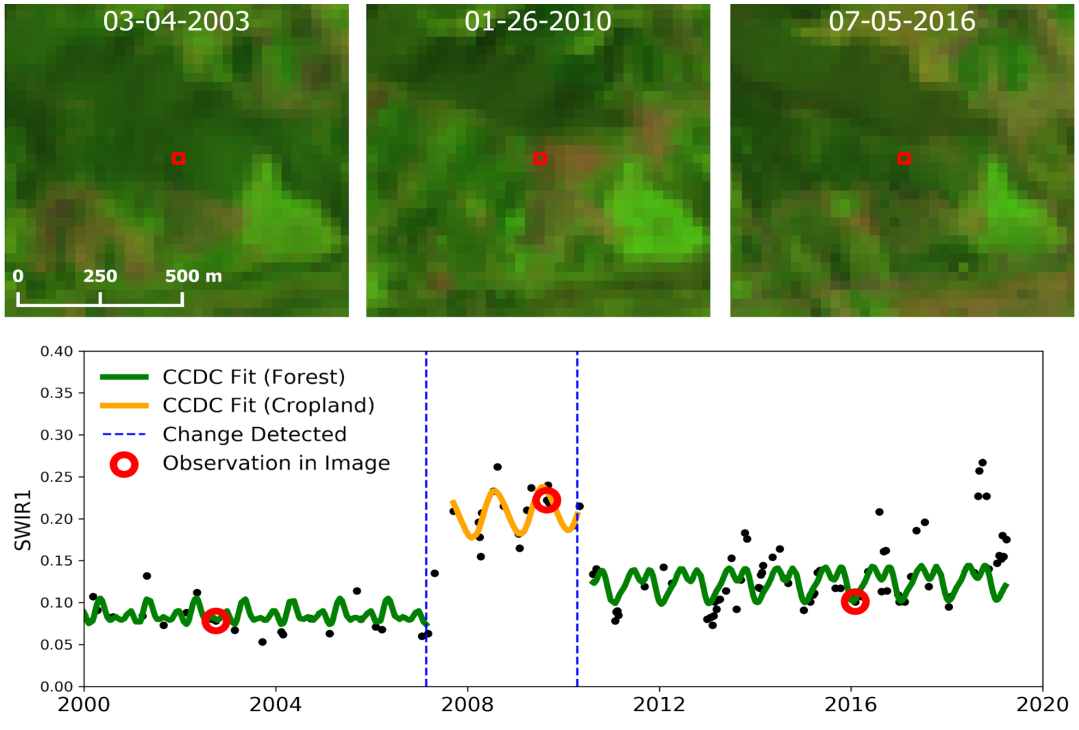


Figure 2.6: Visual representation of the CCDC algorithm working for the pixel highlighted in red. It shows the time series of the only breakpoint band used by the authors, Short-wave infrared 1 (SWIR1). The change detection algorithm finds 2 breaks, cutting every bands' time series for this pixel into 3 segments. The coefficients are then used to train a classifier that predicts the land cover type of the pixel during the time interval of each of the segments. The pixel in question suffers from a land cover change, initially being a forested region that around 2007 was transformed into cropland, going back to forest in 2010. Image obtained from [24].

The usage of CCDC temporal features for cashew detection and land cover mapping in Guinea-Bissau has already been explored, despite not yet commonly appearing in the literature. The most widely used alternative features to the initial bands and vegetation indices are spatial contextual features [25, 26]. Many times (though not always), cashew orchards are planted following a clear structure, as shown in Figure 2.2 — this happens when cashew is planted in a grid-like appearance that is alike the naturally growing aspect of forests or other vegetation types. This

is easy for a human to detect with the naked eye, but for an ML algorithm that only takes the information of one individual pixel, it is much harder to catch.

One method used to calculate spatial contextual features is called Gray-Level Co-Occurrence Matrix (GLCM), which considers the relationship between neighboring pixels to provide texture information for a central pixel. This method works with gray-scale intensities — usually take value from 0 to 255 — and analyses how often they occur together [27]. Statistics describing the intensity levels in a central pixel's $n \times n$ neighborhood can be derived, even in specific directions and distances. Some metrics include the Sum Average, Sum Variance, or Sum Entropy, the local Contrast of that neighborhood, Correlation, and Dissimilarity between pairs of pixels, and other metrics, up to 18 different statistics [28].

Both CCDC and GLCM are valuable algorithms for extracting information of different natures, but their limitations are similar. They both are computationally expensive: GLCM works with a neighborhood of each pixel; CCDC works with a time series for each individual pixel, applying change detection algorithms that can find a varying number of breaks, followed up by a harmonic regression. When used to extract features for a ML classification task, they also contribute to a significant increase in the dimension of the feature space: for each initial band and training feature, up to nine CCDC coefficients may be obtained (e.g., in the typical case of harmonic regression with three frequencies); GLCM obtains up to 18 spatial contextual features for each initial band.

2.3 Machine Learning and Active Learning

The goals of this dissertation include obtaining a dataset that is sufficiently broad and representative of the entire Region of Interest, as well as producing a land cover map by classifying each pixel in the data using a Machine Learning algorithm. As we are in a supervised learning setting, there has to be a previous annotation step as part of the project, which can be done either on-site or remotely. Both on-site and remote annotation have drawbacks: field annotation missions typically involve significant manpower — especially at a nationwide scale —, much above the available project resources; the remote approach is less resource-heavy, as it can be done by one expert alone, but this process is still time-consuming and less reliable. In this scenario, where clear resource or time budgets exist that limit the number of points we can acquire, an approach based on Active Learning (AL) is interesting to explore.

A typical AL pipeline is represented in Figure 2.7. An initial ML classifier is trained on a primary (and usually small) training set. Then, that model evaluates the samples that are still unlabeled, to understand which ones are more worth classifying. The Query Policy regards this process of selecting new points to label: the unlabeled samples are ranked through an acquisition function (a heuristic associated with the most important factor we are looking for and distinguishes the samples based on that criterion), and depending on certain characteristics like the number of points to select or budget constraints, a small set is passed on to the Oracle. This entity (usually

an expert) is tasked with determining their labels and reporting them back to the pipeline. With them, the newly labeled samples can be added to the training process, and with the upgraded training set, the classifier is either retrained or fine-tuned. Active Learning usually works as an iterative process: the Query Policy can be applied again, using the upgraded model, and the steps to expand the training set with other informative samples are repeated. This can be done for as many iterations as possible given the budget constraints, or until we have reached a satisfactory performance. Therefore, Active Learning follows the idea of smart labeling: efficiently selecting points to constitute our training dataset instead of spending resources with unimportant or even prejudicial samples, thus minimizing the burden of the experts that do the annotation step [29].

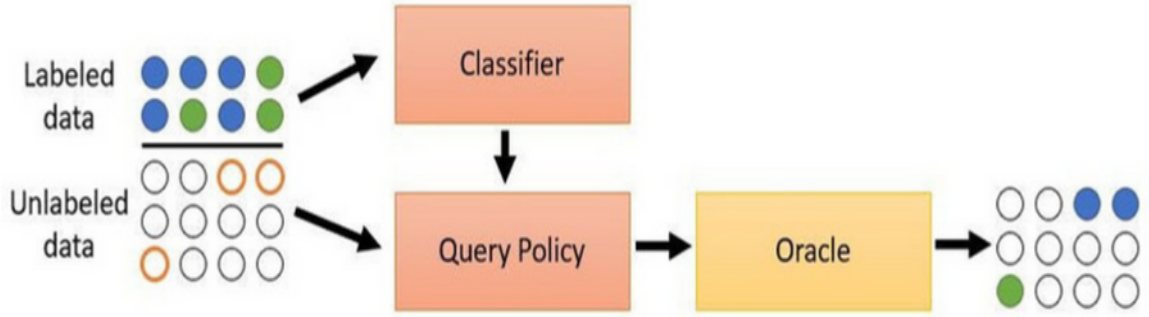


Figure 2.7: Schematic of a typical Active Learning pipeline. An acquisition function (Query Policy) is used to evaluate which samples are worth labeling. These are given to an oracle, the entity that labels them. The newly annotated samples will then be added to the training set, and can then be used to train a now more powerful model.

The previous explanation and the ideas progressed during this chapter follow the pool-based approach to Active Learning: the unlabeled samples constitute a large unlabeled pool U that at any point can be evaluated by the Query Policy and requested points for labeling and expanding the training set. Other approaches could be applied, like membership query synthesis or stream-based sampling [30], but pool-based sampling heuristics seem like the most natural option for this dissertation since we are going to work with a large (and closed) unlabeled pool of S2 data with hundreds of millions of pixels. Usually, heuristics are conceptually defined to iteratively select one new sample at a time — that is, the acquisition function ranks all the unlabeled points, and the one that maximizes or minimizes it (depending on the type of function) is selected for labeling; just that one sample is added at each AL iteration. However, point-by-point selection is very inefficient, especially when the unlabeled pool is quite large. As such, most heuristics are applied in a batch-like fashion: in each AL iteration, the Query Policy selects a batch of unlabeled samples to give to the Oracle and add to the training set, thus reducing the number of iterations and speeding up the process. But the process of choosing the batch size is contentious [31] — it depends on the characteristics of the data, the ML classifier used, and the context where it is applied.

In pool-based approaches, multiple acquisition functions can be defined which focus on different measures of informativeness. The most used heuristics fall into the category of uncertainty sampling [25, 30–35]. This is where the acquisition function evaluates an unlabeled sample's

attractiveness by some metric of the classifier's prediction confidence, which changes depending on the heuristic. For example, Margin Sampling is the simplest one that uses margin values to acquire new data: it selects the unlabeled points that are closest to the decision boundary [36, 37]. A two-class example of this heuristic in action is represented in Figure 2.8. The correspondent acquisition function is defined in Equation 2.1, where $f(\mathbf{x}_i, \omega)$ represents the distance of unlabeled sample $\mathbf{x}_i$ to the hyperplane defined for class ω in a one-against-all setting for multiclass problems [37]. Despite the multiclass formulation, this heuristic is mostly used for binary classification, where there is only one hyperplane [37] — the Query Policy selects the unlabeled sample $\mathbf{x}_i$ closest to the decision boundary, which corresponds to the minimum $|f(\mathbf{x}_i)|$ measure. When $N > 2$ classes are involved, an adaptation called Multiclass Level Uncertainty (MCLU) is preferred.

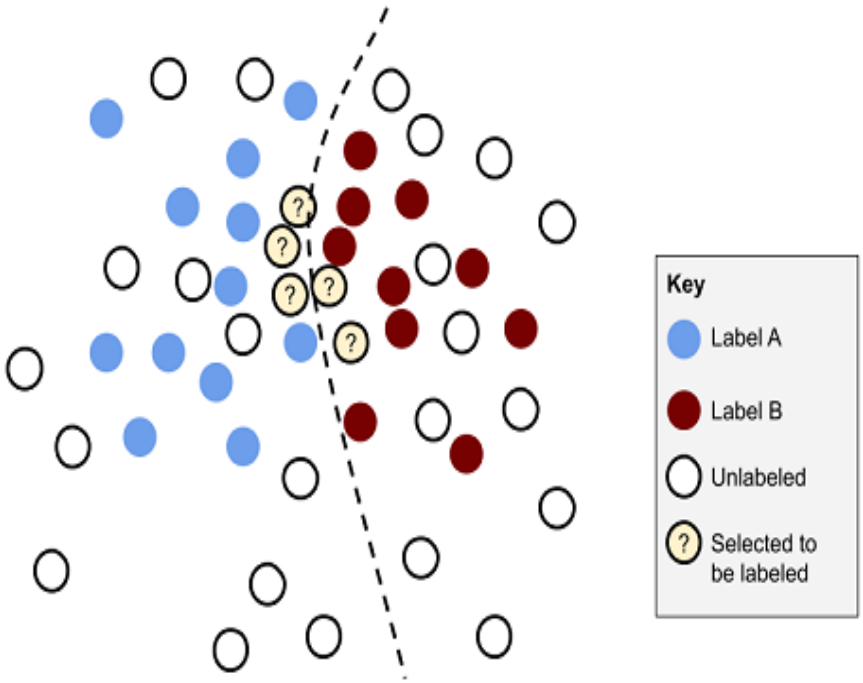


Figure 2.8: Visualization of a Margin Sampling iteration. A classifier (SVM) is trained on the labeled data (blue and red colored dots), stipulating a decision boundary. The unlabeled samples that are closest to the decision boundary are selected for labeling.

$$\hat{\mathbf{x}}^{MS} = \arg\min_{\mathbf{x}_i \in U} \left\{ \min_{\omega} |f(\mathbf{x}_i, \omega)| \right\} \tag{2.1}$$

Margin-based heuristics are adequate to use in conjunction with margin-based classifiers such as SVMs. However, for classifiers of different natures, the main AL alternatives are posterior-probability based heuristics. These are associated with models that can give a probability estimate for each of the classes, which may then be used to define the point's informativeness — model examples are Artificial Neural Networks with a final softmax layer, ensemble-based classifiers (which can also define probability estimates for each class), and even SVMs if using Platt's scaling

to output probabilities [38], though this is less common. Breaking Ties (BTs), which is sometimes also called Best versus Second Best (BvSB), is highlighted in this category: when a classifier outputs a probability distribution $f : \mathcal{X} \rightarrow [0, 1]$, the heuristic identifies the two most likely classes for an unlabeled sample, and calculates the difference of the two classes' probabilities, as shown in Equation 2.2:

$$\hat{\mathbf{x}}^{BTs} = \arg\min_{\mathbf{x}_i \in U} \left\{ \max_{c \in (1,\dots,N)} f(\mathbf{x}_i, c) - \max_{k \in N, k \neq c} f(\mathbf{x}_i, k) \right\} \tag{2.2}$$

where $f(\mathbf{x}_i, c)$ denotes the estimated probability that the unlabeled sample $\mathbf{x}_i$ belongs to class c. Breaking Ties is an almost direct adaptation of the MCLU formulation that deals with probability estimates, with a very similar interpretation: if the difference in Equation 2.2 is large, that means the most likely class for an unlabeled sample $\mathbf{x}_i$ is much more likely than the second possible class, underlying there isn't much prediction confusion. However, if the difference is small, it means there is a high model uncertainty between (at least) two of the possible classes, hence this is a point likely to be very informative and worth labeling.

The formulations in Equations 2.1-2.2 are presented in a point-by-point fashion — they show how to choose just one single unlabeled sample $\mathbf{x}_i$ at each iteration. However, they are easily convertible to a batch-like approach: instead of selecting just the point that, in the current iteration, minimizes the respective acquisition function, the b unlabeled samples with the lowest acquisition function values are selected.

Uncertainty sampling algorithms have some drawbacks: they tend to focus on very specific regions of the input space, for example close to the decision boundary (like MS and MCLU by nature do), leaving other regions of the input space practically unexplored; also, creating a small set selected by a specific metric typically bias the classifier towards some of the classes [39]. Both phenomena threaten a model's robustness and generalization capacity [40]. Hence, the second most common family of AL heuristics in Remote Sensing Image Classification are representativeness-based algorithms [41–43]. They focus on efficiently selecting samples that are highly informative, but also scattered across the feature space, usually by involving some form of aggregation or clustering of the unlabeled data to choose points that are further away from already known regions [41, 42]. A known representativeness approach is Core Set [42], a greedy algorithm based on the k-center problem, which chooses the points that in the current iteration minimize the distance of the unlabeled set to the labeled set. These representativeness heuristics are usually affected by outliers in the unlabeled pool — to mitigate their effect, a budget of points that are the furthest away from the entire unlabeled pool is often defined, with these samples being disregarded as options for labeling. They also tend to be computationally heavy, due to applying a clustering or an optimization algorithm for such a large number of base pixels on the unlabeled pool. To solve this issue, researchers suggest firstly selecting a subset of the unlabeled pool (either randomly or chosen by another AL heuristic), and only then applying the representativeness-based heuristics to just that subset [41, 42].

Chapter 3

State-of-the-art

This section presents a literature review and description of how multiple parts of the Remote Sensing and Image Classification pipeline are typically implemented. This dissertation follows a Machine Learning (ML) approach: we want to know how algorithms can use satellite or aerial raster data in order to classify the elements of a certain Region of Interest.

3.1 Image Classification with Land Cover Mapping

Supervised learning techniques are the most commonly used in the context of Machine Learning for multi-spectral image classification. Different input data types can be considered, depending on the type of ML algorithm used. The simpler option is to work with each pixel individually: the information is defined as a vector with the input bands' values (corresponding to the training features), used as input for a ML model to predict a pixel's class. This tabular data type is suited for classical ML classifiers, which were the initially used alternatives — more noticeably SVMs and ensemble based algorithms (mostly Random Forests, but also XGBoost, Extra Trees, and others) [25, 31, 34, 35, 44]. However, having access to neighboring pixels can be useful to identify certain land types, which can be done by working with patches of pixels: considering a raster image as a (n, m, D) matricial object, where D is the number of features (bands) and n and m represent the number of pixels on the $2D$ dimensions of the Earth's surface, a patch is just a sub-raster with shape (k, l, D), with $k < n$, $l < m$. With the expansion of Deep Learning (DL), algorithms like Convolutional Neural Networks (CNNs) [21, 26, 32, 44–46] and Residual Networks (ResNetS) [21, 32, 46] have been gaining more and more research space, since they work better with matricial inputs such as patches, or raster data in general due to their their grid-like structure. However, the network output can follow two possible approaches: either predict the class of every pixel on the input patch, or use the patch's information and spatial context to predict the class of only the patch's central pixel.

An alternative data type is called a superpixel [25, 35, 44, 46]. Superpixels are calculated for a certain raster, using techniques like Simple Linear Iterative Clustering [44], and typically

correspond to groups of pixels that are close to each other and have roughly similar features. Unlike patches, which have a very structured form (they all have the same size and are usually rectangle or square-shaped), these can have different numbers of integrated pixels, with varying sizes and shapes. Superpixels are typically described by calculated spatial statistics, somewhat similar to what GLCM does. So, each superpixel can be characterized by a vector input and is thus suitable to be interpreted by traditional classifiers, allowing these algorithms to work with groups of pixels at a time. Since this is a significantly larger base unit than one pixel, the image processing and classification tasks become more efficient.

Supervised learning requires the acquisition of labels in the early stages of the pipeline. However, the class of a pixel (or other units such as a superpixel) is an arbitrary concept. Its definition can change depending on the specific project application (for instance, a pixel might be labeled as "forest" in one scenario and as "vegetation" in another). Class and satellite granularity also play a role, as lower-resolution data likely has to be associated with coarser definitions, since there is no spatial capacity to distinguish very specific classes. This process is further complicated when there are pixels that capture characteristics of multiple land types, leading to ambiguity or uncertainty in a pixel's membership to a particular category. In such cases, a fuzzy classification approach may be a more suitable alternative.

Regarding model performance, researchers still debate the advantage of the spatial feature extraction capacity of DL algorithms. CNNs and similar networks can use a stack of convolutional layers to extract spatial features with increasing levels of abstraction by themselves. This capacity might remove the need to add spatial bands to the input features, which has the advantage of reducing the dimensionality of the feature space. However, some authors still see advantages in considering manually added spatial features in Deep Learning inputs anyway [47]. On the other hand, since classical models typically only work with vector inputs instead of matrix-based ones, they are unable to extract latent spatial features by themselves. Hence, their only options to access spatial context are through manually obtained spatial features (such as the GLCM statistics) or through the usage of superpixels. Nevertheless, some researchers suggest that, with certain types of contextual features, traditional classifiers can perform similarly or even slightly better than CNNs, especially at low-medium satellite resolutions [44]. In this scenario, traditional Machine Learning algorithms are preferable, as they are lighter and require noticeably fewer training resources than their DL counterparts.

Cloud filtering is a main concern when dealing with satellite imagery. Since satellites are in orbit, certain regions or even entire images can be covered by clouds at the satellite's monitoring instant. When the main focus is the analysis of land cover, these cloudy pixels should be masked (which means discarding them from further analysis), since the values measured at the Earth's surface are the only ones that matter. For the context of image time-series like the image collection given by S2, there is the possibility of removing just the cloudy pixels, or instead removing an entire image where more than x% of the pixels are cloud masked [14]. Sentinel also has a Cloud Probability band that can be used for this purpose (mask pixels where the probability is above a user-defined threshold). Other metrics were also created with this goal in mind, such as the

Cloud Displacement Index (CDI). This index is based on the optical separation in elevated objects due to sensor parallax [48] — elevated objects like clouds are detected under parallax and can be separated from ground objects. Therefore, stipulating a CDI threshold may also be a cloud filtering approach.

Data augmentation techniques can also be considered during pre-processing. These are especially used in conjunction with DL models, as they are data-hungry algorithms. Considering that most DL techniques work with data patches, the techniques are adapted to this data type: patches are multiplied using rotation or reflection transformations [26, 32, 41, 49]. Variation of brightness or contrast techniques are also used [21]. Despite this caution about the number of unit samples, there isn't much pre-processing done in regards to feature dimensionality. Some authors apply PCA to reduce the number of initial features, which can be quite large if we work with spectral and manually constructed spatial features, for example the ones obtained with GLCM [34, 41, 50], but most authors instead work with all those features directly, possibly applying a feature selection step at a later stage.

3.2 Cashew Orchard Detection and Monitoring

The problem of cashew monitoring in Guinea-Bissau already appeared in the literature. This project follows Sofia Cardoso Pereira's Master's thesis [12, 14], which focused on implementing a remote monitoring system based on supervised learning methods for Cantanhez National Park, a protected area in the Tombali region. It is important to clarify that the current dissertation is not a continuation of that work [14], as her project's goals were achieved during her dissertation. The main problematic has multiple similarities, but the current dissertation is an expansion of her thesis, as we plan on reaching nationwide remote coverage, using Active Learning procedures to efficiently obtain a new dataset that is representative of the entire country. Thus, there are new goals and research fields to be applied.

Sentinel-2 Surface Reflectance data was used on that thesis [14]. In an initial step, 4 composites were made, one for each month between January 2019 and April 2019, with the median used as the criterion. For each of the monthly composites, the training features were comprised of nine of the thirteen S2 spectral band values, as well as extracted GLCM features for those bands. The annotations were offered by RSeT [51]: 24252 pixels obtained in polygon groups and divided into six classes: Forest, Cashew, Sparse Vegetation (including savanna), Mangroves, Water, and Other (a class containing urban areas, roads, bare soil, and farmland). A Support Vector Classifier (SVC) was trained with this ground truth data, and its predictions were utilized in the production of a Cantanhez land cover map, presented in Figure 3.1.

This work has been followed up in recent years, maintaining the focus on cashew detection, but also on deforestation monitoring. One article approached the mapping of smallholder cashew plantations in the West African nation of Benin [52]. They applied a Spatio-Temporal Classification with Attention (STCA) model for the Planet Basemaps multi-temporal image collection, using the

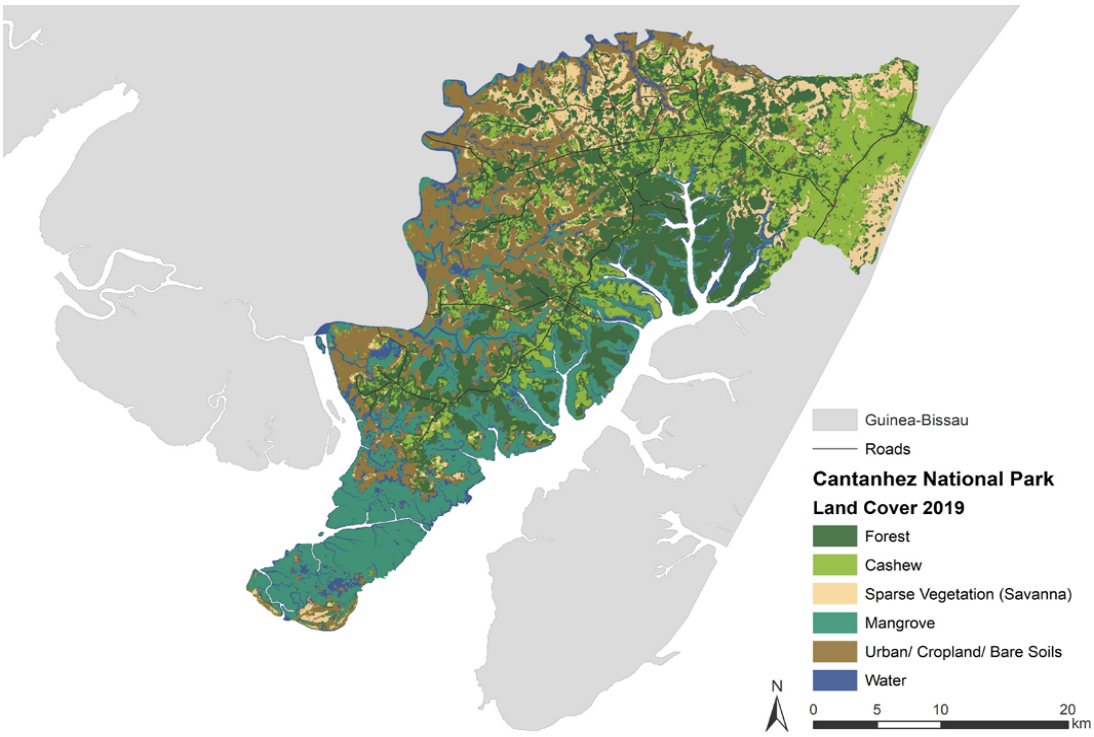


Figure 3.1: 2019 final land cover map obtained with SVM-based classification over Cantanhez National Park. [12]

predictions to create a land cover map of the monitored regions. However, they took this process further: after identifying where the orchards were located, they focused on those regions only. A Clustering Augmented Self-supervised Temporal Classification (CASTC) model leveraged to distinguish two kinds of cashew orchard tree-spacing practices (high vs. low-density plantations). This allowed for more reliable cashew production estimates, since different types of plantations (with more or less trees and in multiple evolutionary stages) output different production values.

The work presented in [12, 14] also served as a basis for problematics other than cashew. A group of researchers used a similar idea to tackle the classification of land cover types with similar spectral signatures in Pará, Brazil [53]. According to the authors, the cocoa agroforest and native forest have very close spectral behavior, which makes their distinction difficult through spectral values alone, at a specific point in time — something that also happens with cashew and forest in Guinea-Bissau. As such, they found that the usage of a time-series of median composites over five months instead of a single composite of the region helped the separation of these classes — which is an expansion of the usage of four median composites from January to April to obtain training features at different times used in [14].

All the previously mentioned articles are focused on land cover mapping, using model predictions as the land type class for the pixels in the Region of Interest. However, some take the mapping of cashew orchards to another level: researchers have tried to determine the specific number and size of every cashew tree on their ROI [20, 21] — this knowledge can be used to perform better and more realistic estimates of that region's cashew production values. They use models like

Faster R-CNNs [20] or FG-Unets [44] for the purpose of object segmentation and detection (in this case, detection of cashew trees), which can produce a more specific mapping, like the one presented in Figure 3.2. We note our project's focus was on land cover mapping techniques, and did not attempt to follow an object detection approach — we did not have enough resources, mainly high-resolution satellite or drone imagery, to obtain the level of detail needed to delimit individual trees in a plantation.

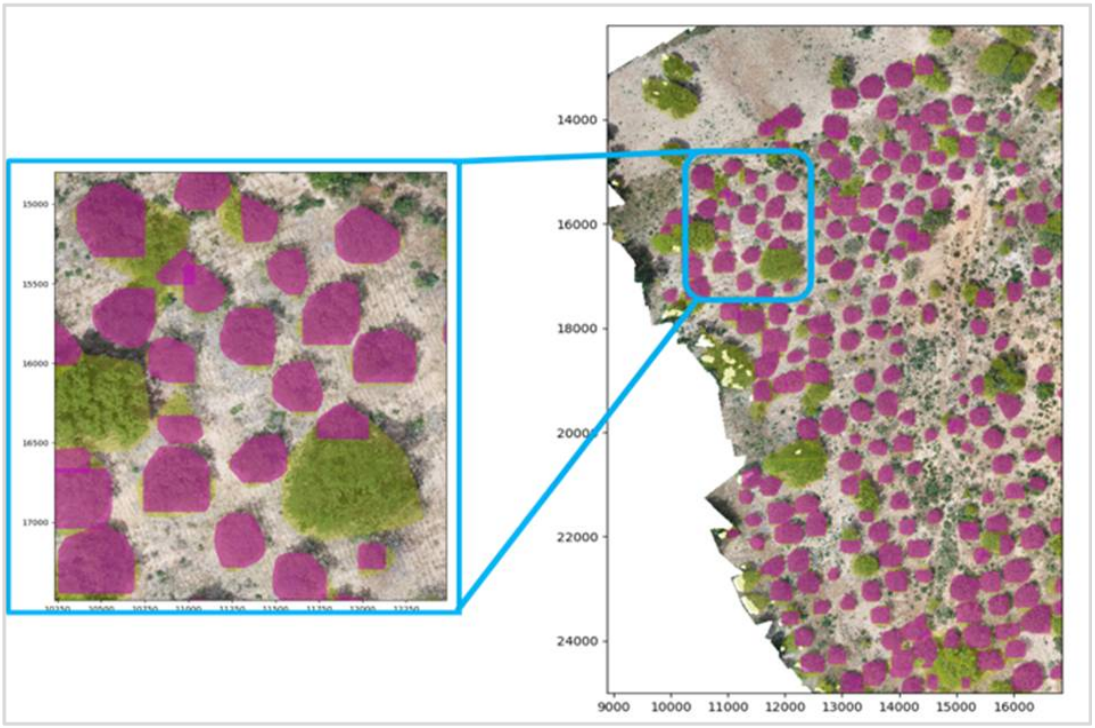


Figure 3.2: The result of cashew trees detection. Yellow patches represent non-cashew trees, whilst the reddish patches correspond to detected cashew. [20]

Chapter 4

Data and Methods

4.1 Overview

This dissertation concerns a Machine Learning approach for characterizing land cover across a defined Region of Interest utilizing a fully remote pipeline, which includes the data acquisition step. In this chapter, we describe the characteristics of the data used, as well as the methods and techniques applied during each ML step — data and feature acquisition, preprocessing, classification models considered, hyperparameter tuning, and Active Learning processes. The main results will be presented in chapter 5.

4.1.1 Tools and Libraries

A flowchart of the tools and libraries is presented in Figure 4.1. Google Earth Engine was used primarily for the data acquisition and pre-processing steps — it has the ability to deal with large amounts of raw data in a distributed and efficient manner, so it was used until a smaller data aggregate that could be worked on locally was obtained. Sentinel-2 data was considered for the period of 2017 (since the second Sentinel satellite was launched and operational) up to 2021 (the year of analysis), forming an image collection with bands corresponding to the surface reflectance values, vegetation indices, CCDC coefficients, and GLCM spatial contextual features. Each band's minimum and maximum values over the ROI were also obtained and stored for later min-max normalization. This tool was also used to perform the cloud masking task of the data cleaning process. GEE has a native Javascript code editor, but is also available through a Python API, mainly through the *geemap* library [54]. Despite containing most of the Javascript API features, the Python package was mainly used for dynamic data visualization and downloading the filtered and processed raster data, partitioned into several tiles (GeoTIFF files).

The remaining stages of the project's Machine Learning pipeline were implemented in Python. The *numpy* and *pandas* libraries were used to process tabular data. The main tools for reading, processing, and visualizing raster data were *rasterio*, *rioxarray* and *geopandas*. The ML section

of the dissertation was based in *sklearn*, considering Support Vector Classifiers and Random Forests with Decision Tree base learners. Hyperparameter search was done with *hyperopt* [55]. This library provides a Bayesian optimization approach for the hyperparameter tuning task that is more efficient than GridSearch or even RandomSearch since it focuses resources on more promising regions of the HP space.

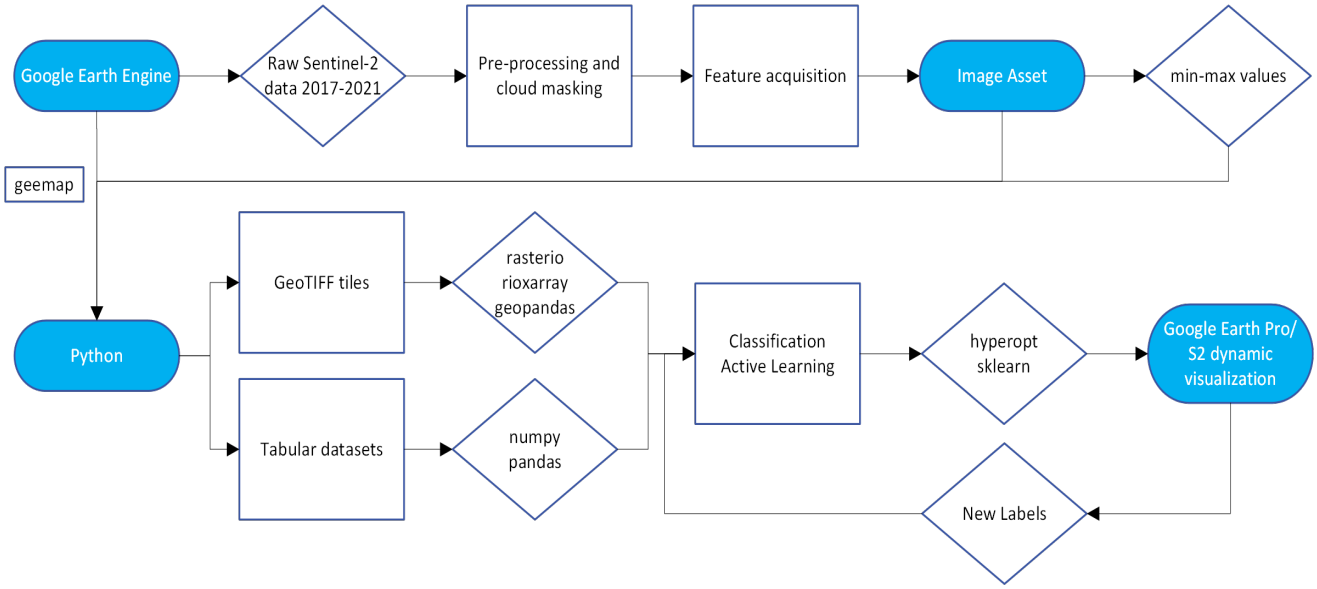


Figure 4.1: High-level flowchart of the tools and libraries used during every step of the dissertation's pipeline.

The remote annotation task — necessary for the Active Learning part of the project — was mainly done in Google Earth Pro, where the Historical Imagery view allows for the visualization of multiple high-resolution images across a large timespan of the region, which supports the labeling process by helping to identify temporal behaviors that are characteristic of certain classes (such as seasonality). However, this view has its limitations, mainly due to some regions in Guinea-Bissau having reduced Historical Imagery coverage. Thus, a dynamic visualization procedure was implemented in Python using monthly median composites from S2 satellite data, which allowed for terrain visualization of every month between 2017 and 2023 (this has noticeably worse resolution however, which is a limitation for the remote labeling task).

4.2 Data

The area of study is shown in Figure 4.2 and is divided into two main parts. The first part corresponds to the entire country of Guinea-Bissau (mainland and islands included). However, a second region in the Republic of Guinea — up to the *Kogon* river — was added to the ROI during a collaboration with researchers from CIBIO/BIOPOLIS [56] for the draft of an African Forest elephant emergency plan [57]. Three main datasets were considered throughout the dissertation to perform the analysis of the entire Region of Interest:

- **RSeT dataset:** The initially used labeled set was offered by RSeT [51], corresponding to 5504 annotated samples only from Guinea-Bissau, as shown in Figure 4.3. The data was originally collected for the development of a 2021 Forest Forest Reference Emissions Level (FREL) connected to the *VCS: Community based avoided deforestation project*, developed by the Institute of Biodiversity and Protected Areas (IBAP) and the Bio-Guinea Foundation. [58]

- **Random Sampling dataset:** The second dataset was created during the Active Learning phase of the project and corresponds to 4498 samples that were randomly selected in the whole Region of Interest and labeled by the author. Aside from other uses, the random selection procedure serves as a baseline comparison for other AL heuristics, in order to assess their efficiency in selecting an informative training set.

- **Margin Sampling dataset:** The third was obtained during the final Active Learning experiments, with a total of 1406 points labeled according to a Margin Sampling (MS) algorithm, which could also select any point in the entire Region of Interest.

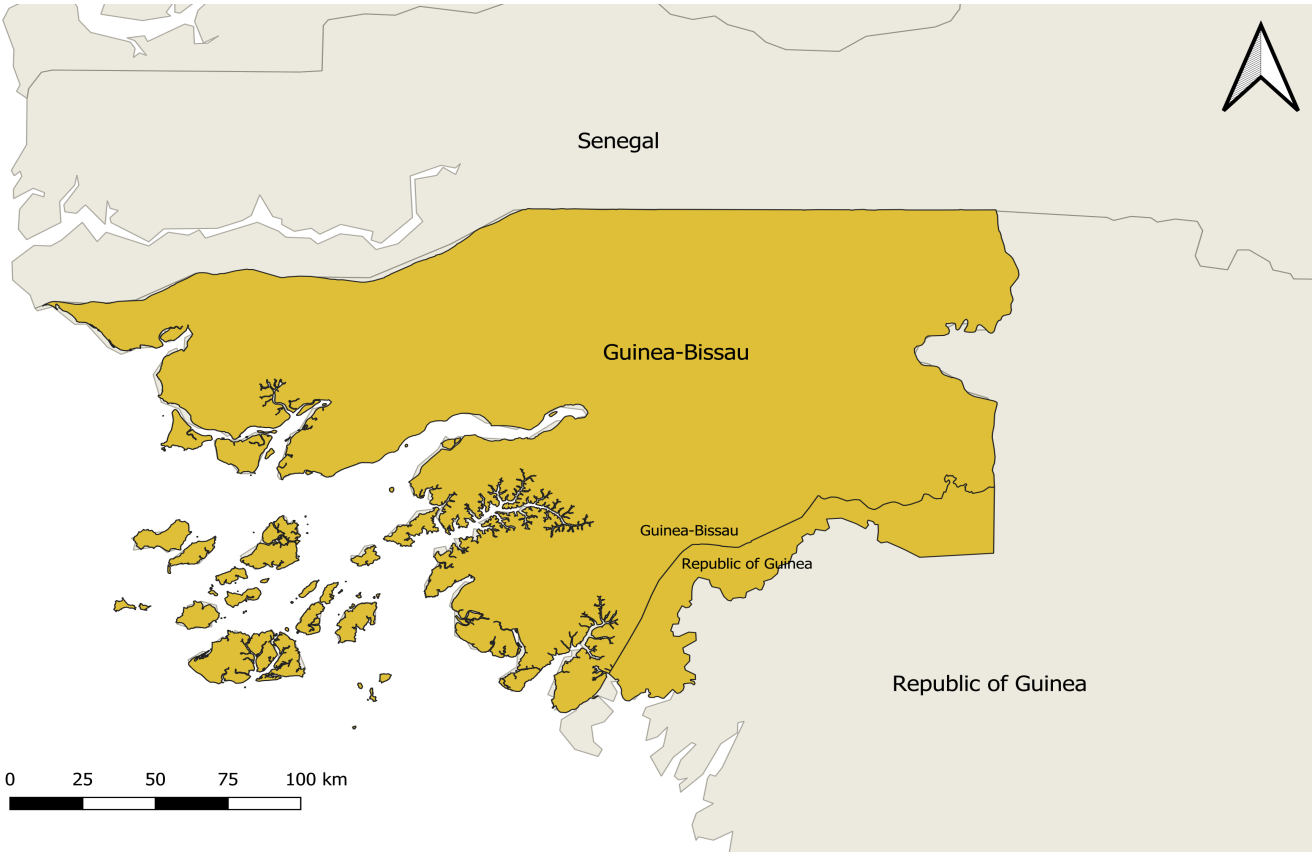


Figure 4.2: Visual representation of the Region of Interest. Image taken in QGIS.

Another smaller dataset with just 602 samples was obtained during the first round of Active Learning experiments (section 5.2.2). It was probed in certain sub-regions of the Region of Interest and served only to withdraw preliminary conclusions about the feasibility of the AL

process. Therefore, it was not used for any major output of the dissertation, and thus not mentioned at the same level as the other three.

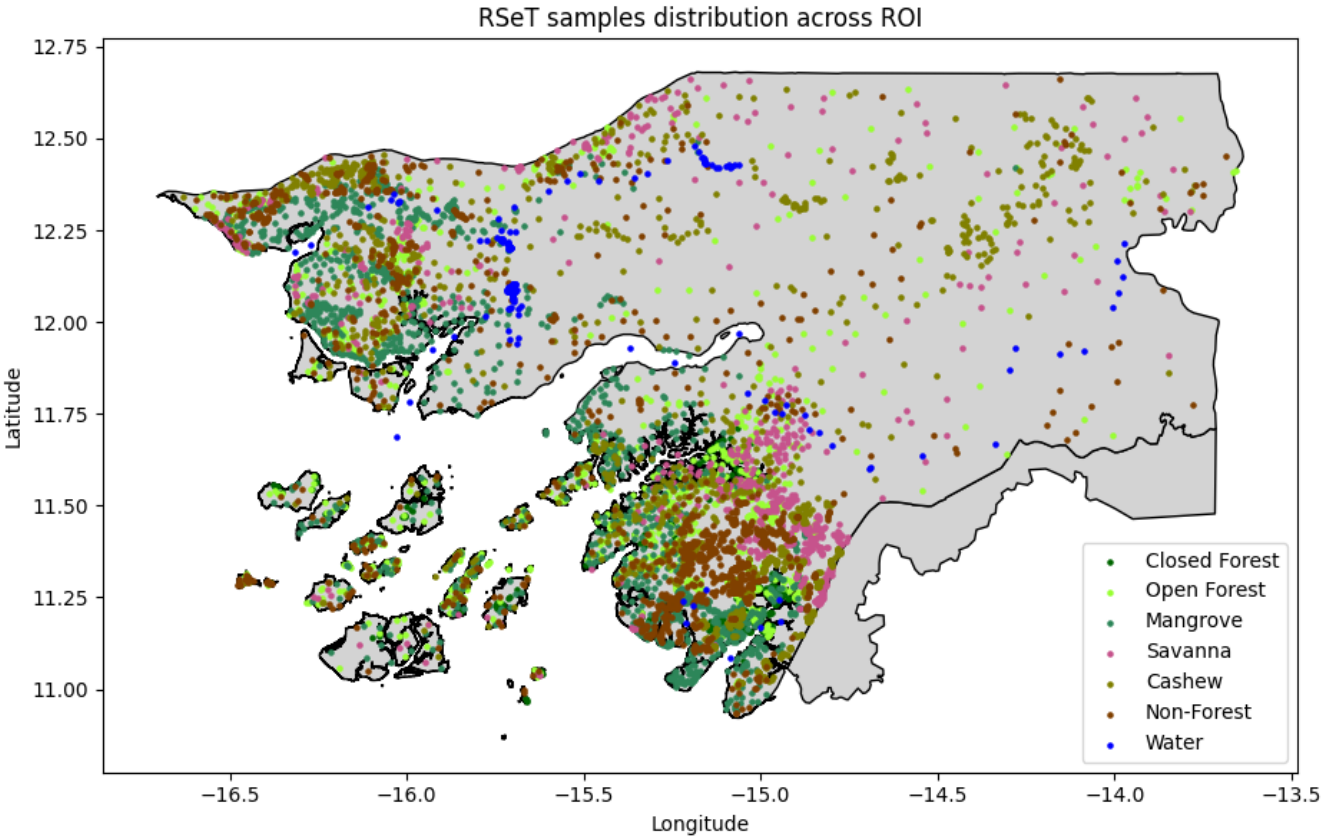


Figure 4.3: Geographical location of the RSeT labeled data points.

RSeT labels were assigned according to a seven-class system: Closed-canopy Forest, Open-canopy Forest, Mangrove, Sparse vegetation/ Savanna woodlands, Cashew orchards, Non-Forest, and Water. Non-Forest is considered an "Others" class — it aggregates pixels corresponding to houses or buildings, roads, farm fields, and empty land. The labels of the other two datasets (Random Sampling and Margin Sampling ones) were determined by the author following RSeT's scheme; however, an eight class was added for pixels that are definitely not cashew, but which cannot be assigned any of the other six RSeT classes — these can still be useful in a Cashew vs. Non-Cashew classification setting (more details about this will be presented in the following sections).

Table 4.1 shows the number and percentage of points per class for the RSeT and Random Sampling datasets (the Margin Sampling dataset specifics will be shown later). Regarding the RSeT dataset, cashew is the class with the most collected points, closely followed by Mangrove and somewhat closely by Non-Forest. A small fraction of samples were taken of Water, as well as Closed Forest. Class proportions change for the randomly collected data, with the most prominent now being Savanna, followed by Non-Forest and closely by Cashew, with the remaining classes (including the added eighth non-cashew class) being much less predominant.

The RSeT annotations were taken in 2021. Previous research suggests that these labels should not be directly transposed into the following years, due to ongoing land cover changes across the

country and spectral signature variations across different periods [14]. Therefore, we assigned Random Sampling and Margin Sampling labels according to the respective land cover in 2021 for consistency and restricted the analysis to 2021 to make optimal use of the available labels.

	RSeT dataset		**RS dataset**	
	Nº points	**% of dataset**	**Nº points**	**% of dataset**
Closed Forest	222	4.0	9	0.2
Open Forest	905	16.4	289	6.4
Mangrove	1217	22.1	417	9.3
Savanna	642	11.7	1303	29.0
Cashew	1222	22.2	993	22.1
Non-Forest	1152	20.9	1024	22.8
Water	142	2.6	111	2.5
Other (Non-cashew)	-	-	352	7.8

Table 4.1: Table with the number and percentage of points per class for the RSeT and Random Sampling datasets, following the seven class scheme plus the eighth added class of non-cashew data.

4.2.1 Train-test splits

From the three major datasets used during the project, both RSeT and RS sets were split into train and test sets. Train sets were used for every model development task — feature selection, hyperparameter optimization, and model selection —, as well as to train the selected classifiers. Meanwhile, the test sets served to evaluate model performance on previously unseen data and to grade the evolution of the Active Learning iterative processes. The Margin Sampling dataset was only used for model training. More details about why two test sets from different origins were utilized will be presented in the results of section 5.

During the splitting process, the samples' spatial correlation was taken into consideration, since points that are very close to each other (especially from the same class) tend to have very similar signatures. A part of a class's variability comes from geographical or regional differences, thus it is usual for neighboring data points to all go to either training or testing together [14] — not doing so can inflate the test set classification results and overestimate the actual model's performance. To apply a grouped split, we used *Scikit-learn*'s built-in *StratifiedGroupKFold* function. The method splits the groups as a whole (instead of individual samples) into K folds through a stratified process (which is usually considered for cases with imbalanced datasets). It can be used for train-test splitting following a Holdout approach instead, simply assigning certain folds for training

and the remaining for testing. The dataset RSeT provided was not collected in groups — each pixel was individually sampled, with no direct neighboring pixels being included —, so to be able to apply to the splitting algorithm, we manually defined polygons for the pixels of each class in GEE, with each polygon being attributed a unique ID — one example is shown in Figure 1 in the Appendix, which represents the drawn polygons for the class Savanna. Conversely, both RS and MS data were collected in patches (blocks of 3×3 pixels), so each patch was attributed with a unique ID and used as the grouping column for the stratified grouped split algorithm.

The *StratifiedGroupKFold* train-test split was made using 6 folds for the RSeT dataset (five folds are used for training and one for testing), therefore 6 different Holdout train-test splits can be defined from this approach. Due to the heterogeneous size of each class's polygons, there was some variability regarding the proportion of points per class depending on the fold. So, we chose the train-test split that more closely followed the entire RSeT dataset class proportions. This resulted in a split where 4559 samples are assigned to training and 943 to testing, an approximate 83%-17% proportion. For Random Sampling data we decided to change the approach to a $50-50$ split — using two folds instead of six —, resulting in a training set with 2181 samples and a test set with 2224. The main reason for this change was the observed class proportion variability between the 6 formed folds, as mentioned for the RSeT dataset — reducing the number of folds (which translates to a proportional increase of the test set and train set decrease) makes it easier to guarantee they have similar characteristics. However, we also wanted to ensure the test set was as diverse and representative of the ROI as possible — following conclusions from sections 5.2.2 and 5.3 —, which is more likely to happen if we choose a bigger test set.

	RSeT dataset		**RS dataset**	
	Train (%)	**Test (%)**	**Train (%)**	**Test (%)**
Closed Forest	94.1	5.9	100.0	0.0
Open Forest	76.4	23.6	52.2	57.8
Mangrove	80.7	19.3	45.3	54.7
Savanna	85.0	15.0	55.3	44.7
Cashew	84.4	15.6	47.0	53.0
Non-Forest	84.4	15.6	48.3	51.7
Water	90.1	9.9	75.7	24.3
Other (Non-cashew)	-	-	35.5	64.5

Table 4.2: Train and test split percentages of each class (that is, the percentage of points from class y that goes to training and the remaining that is left for testing), for the RSeT and Random Sampling datasets.

Table 4.2 details the class percentage distributed into training and testing, for both RSeT and

RS datasets. It is noticed that the minority classes, mainly Closed Forest and Water, are not as well separated, especially so for the Random Sampling case: since RS contains just one Closed Forest polygon (as seen in Table 4.1) there is no other option than to this class only appearing either in training or testing. Meanwhile, since Water is an easily separable class — as concluded by the analysis of the 95% density level ellipses (Figure 5.1) —, it is likely that one could train a model capable of well describing this class or forming an adequate test set for it despite a disproportionate split. However, for the majority classes — mainly Savanna, Cashew, Non-Forest, and Mangrove —, the samples were well distributed between the sets, according to either the $83\% - 17\%$ or $50 - 50$ split proportion respectively given by the 6-fold and 2-fold splits.

4.3 Feature acquisition

A high-level flowchart of the feature acquisition step is presented in Figure 4.4. Sentinel-2 TOA image collection (available in GEE) was used as the basis. Of the 13 original S2 bands, only nine were considered: bands B1 (Coastal Aerosol), B9 (Water Vapour), and B10 (Cirrus) were discarded since they are mostly used for atmospheric corrections and are typically ignored for surface-level applications [14]; band B8A was also dropped due to its spectral similarity and intersection with band B8. However, 4 vegetation index bands were added to the working Image Collection, corresponding to NDVI, NBR, EVI, and EVI2 index bands. Therefore, we still ended up with an Image Collection containing 13 different bands. To obtain training features from this temporal Image Collection, it was already determined that the spectrally similar classes are separated the most by the end of the dry season [14]. Following that conclusion, we created a median composite of the Image Collection for the entire month of April 2021, the final month of the dry season. A 13 band Multi-spectral Image of the entire ROI was obtained, with each band being adequate to use for a Machine Learning process.

GLCM spatial features weren't considered during the dissertation's initial phase — when working with just the RSeT data. This labeled set was originally used by RSeT in conjunction with the CCDC temporal features for the classification task. Thus, at the beginning of the project, the focus was only on temporal features (aside from the 13 S2 and vegetation index bands). We used the CCDC implementation that is already available in GEE or in GEE packages [59]. However, the algorithm and the available implementation were originally designed for Landsat data. In order to be used jointly with Sentinel-2 satellite imagery, the parameters of the algorithm most likely need to be adapted, due to S2's faster revisit time and smaller timespan in comparison to Landsat. But it is extremely complicated to search for the best combination of parameters: GEE does not have any native parameter (or hyperparameter) search methods, and the algorithm is computationally heavy, difficulting the implementation of any search procedure that does not break GEE's usage limits. As an alternative, we found one available guide [60] regarding the usage of CCDC with S2 data, so the parameters used (statistics thresholds for the change detection procedure, minimum number of observations to declare a break, or the λ coefficient for the regression procedure) followed the ones presented on the guide, with some modifications

following RSeT's research: every one of the 13 bands was used for break detection; Green and SWIR2 were considered for cloud masking (*tmaskBands*); the maximum number of segments to search and store per pixel was defined as four. We also utilized the cloud coverage technique available in the CCDC implementation [59] which is based on the CDI metric, but without changing the default 65% threshold.

The output of the CCDC algorithm resulted in an image with 540 bands, corresponding to 135 per segment. Looking only at the Image defined by the temporal segment for each pixel that overlaps 2021, this contained bands regarding the CCDC harmonic regression coefficients, but also some defined by the CCDC pre-processing function (such as the Elevation, Aspect, Water Occurrence or Rainfall in the ROI) [59]. Some bands were ignored due to their clear auxiliary purpose (multiple bands regard the timespan of the respective segment, the number of observations on that segment, or other statistics of the segment, and as such are not suited to be used as training features for a supervised learning algorithm). The SIN3 and COS3 coefficients (the third-degree frequencies of the harmonic regression) were also completely discarded because they were zero for almost all of the bands. 113 bands remained after this primary selection — 104 CCDC coefficients and 9 defined by the pre-processing function. This 113-band image was merged with the April 2021 median composite image, resulting in the final 126-band image, with each band being suited to use as a training feature, and was stored as a GEE Asset. From it, we obtained our first usable dataset for a ML supervised learning task: GEE's *sampleRegions()* function allowed to get the 126 bands' values for the 5504 RSeT data points, which were stored as a CSV file and exported.

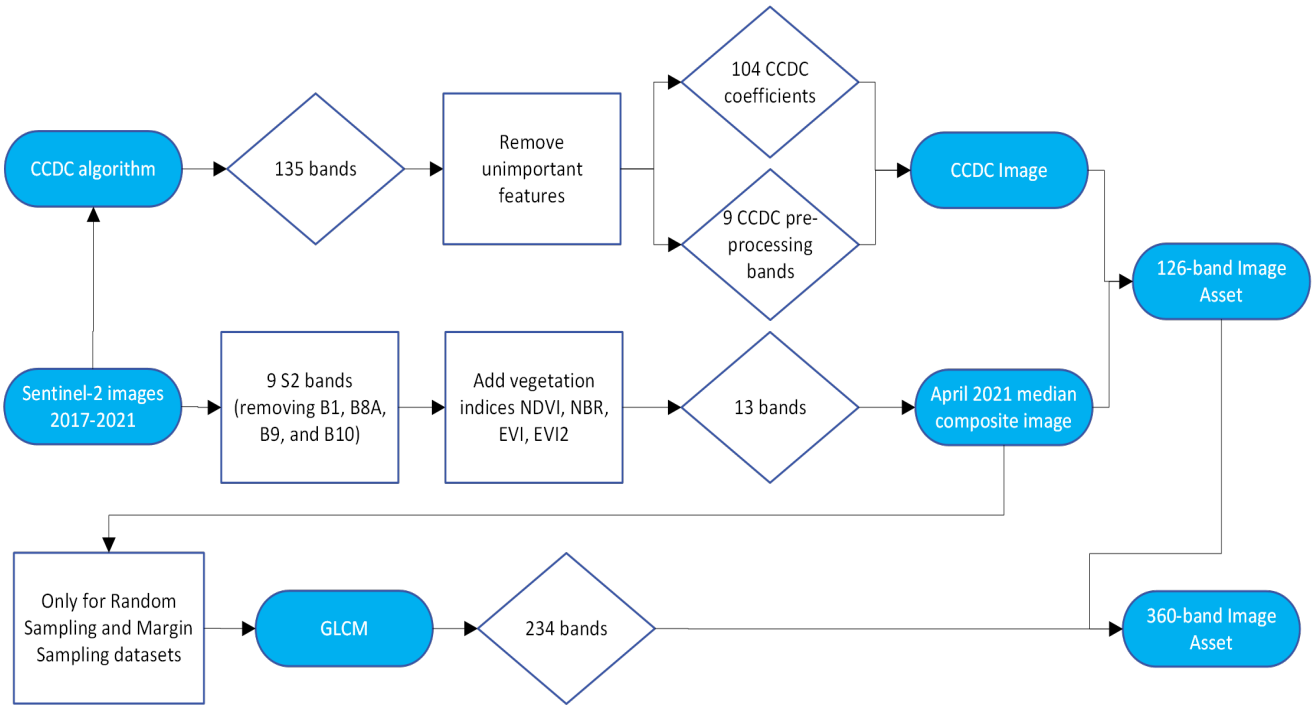


Figure 4.4: Flowchart of the feature acquisition step. Three main types of features are obtained: April 2021 median composite bands, CCDC coefficients and GLCM statistics at a later step.

The statistics from the GLCM algorithm were only added when working with the regionwide Random Sampling and MS-AL datasets. GEE's *glcmTexture* method determines 18 statistics per initial image band, so it was only applied to the 13 spectral plus vegetation index bands, not the CCDC coefficients or the other added features. All 13 bands describe continuous metrics, but GLCM only works with grayscale intensity levels, usually ranging from 0 to 255. Therefore, every band had to be re-scaled and discretized to fit into this range. The algorithm was executed with the spatial kernel for feature extraction selected as a 5×5 square matrix, resulting in an output Image with $13 \times 18 = 234$ spatial bands. A second Image Asset with all the possible training feature bands was formed, merging this 234-band Image from GLCM with the 126-band Image previously obtained, resulting in a total of 360 bands and possible training features. Some of the classifiers we tested during the model selection phase are very sensitive to the scale of the data, so the minimum and maximum values for every band were calculated and stored for a later data normalization step with a min-max scaler.

Due to the size of the feature space after adding the GLCM features, we compared the applicability and performance of different feature selection algorithms. *scikit-learn*'s *RFECV* (Recursive feature elimination with cross-validation) procedure was initially used, optimized for Cashew F1-score. The recursive elimination process selected 250 of 360 features, a number that led us to search for a smaller feature space that would not result in predictive performance decay. The alternative selection procedure was based on the *feature_importances* score present in *scikit-learn*'s RF implementation: the 67 features with above average feature importance (*scikit-learn*'s *feature_importances* score over $1/360$) produced the deepest cut that did not result in any noticeable performance decay. Results will be shown in section 5.3.1: they were measured using the RS training data, following 10 different train-validation splits based on a stratified grouped 10-fold algorithm, with each feature space being used in conjunction with the best classifier — a Random Forest — that was previously obtained using RSeT data.

Thus, when working with the RSeT dataset, there were 126 features considered for the Machine Learning task that came directly from the S2 measurements or the CCDC algorithm. Meanwhile, when working with the RS and MS datasets, we only used the 67 variables from spectral, spatial, and temporal natures that weren't discarded during the feature selection process.

4.4 Classification models

Two main Machine Learning classifiers were considered: Support Vector Machines and Random Forests. SVMs had already shown the best cashew detection results in previous research [12, 14]. RFs were considered as the main alternative due to their integration with Google Earth Engine — the Machine Learning section of this platform is still underdeveloped and does not yet support many different classifier types; however, it has a RF implementation and can also use an imported pre-trained RF to perform classification and other tasks on the platform, which could be beneficial depending on the computational effort of the Active Learning task and other steps

of the project's pipeline. Both algorithms are suited to work with tabular data — they cannot interpret entire patches at once like a CNN does because they are not constructed to deal with matricial inputs. Therefore, the labeled datasets were all stored in a tabular fashion, with the schema being represented in Table 4.3: each row corresponds to an individual pixel; the columns include the training features used for the classification task, the Class attributed to each pixel, as well as the group ID — the identification of that pixel's polygon, in the case of RSeT samples, or its corresponding patch, for RS and MS points. Bayesian hyperparameter optimization was done using *hyperopt*, with the search space defined as below:

	F_1	F_2	F_3	...	F_m	Class	groupID
P_1							
P_2							
P_3							
...							
P_n							

Table 4.3: Schematic representation of the RSeT dataset. A tabular schema is obtained where each row represents a specific pixel P, and each column is its value for every training feature F, the Class, and the corresponding unique group identifier groupID.

- **Random Forests**:
  - Number of estimators: uniform int distribution between 50 and 450, with steps of 10.
  - Maximum base DT depth: uniform int between 1 and 12.
  - Maximum features per DT split: uniform int between 3 and 70, with steps of 3.
- **Support Vector Machines**:
  - Cost: Log uniform search space between -3 and 1 $\rightarrow [10^{-3}, 10^{1}]$
  - Gamma: Log uniform search between -10 and 0 $\rightarrow [10^{-10}, 10^{0}]$
  - Kernel: Choice between linear, polynomial, or *rbf*.
  - Degree: Choice between 2, 3 or 4, conditional for a polynomial kernel.

The metric for the optimization search procedure was the mean Cashew F1-score, determined over a 5-fold grouped stratified sampling split on the RSeT and Random Sampling training sets. This takes into consideration the results on different validation sets for the selection of the best classifiers during HP search. After that, the best performing RF and SVM could be evaluated on the RSeT and RS test sets, which were completely unaccessed up to this point. This avoided any form of leakage that could influence the test results during the Active Learning evaluation process.

Despite RSeT data being natively comprised of seven classes (and RS labels following the eight-class scheme), we also tested the approach of working with just two classes, Cashew and Non-Cashew. This follows the hypothesis that the ML classifier does not have to invest as many resources in distinguishing Non-Cashew data types, and can thus increase the sensitivity of cashew detection. Of course, this would have the tradeoff of losing the specificity of the seven classes, as the land cover description is instead done in broader terms. Therefore, during hyperparameter tuning, the search procedure was run using both options, and the binary class approach would only be considered if the results showed improved cashew scores — which happened for the RS data, as will be presented in section 5.3.2.

4.5 Active Learning

4.5.1 Query Strategies

For the Active Learning step of the project, we wanted to obtain new data from at least two selection processes. The first would be through Random Sampling: this is typically considered the baseline for the selection of new points, since it can be interpreted as a brute force way of acquiring new samples, taking no considerations from the data or model characteristics. In the literature, every AL heuristic is compared with this approach to see if the associated selection process is more efficient at building an informative training set than a simple random train set expansion.

The second form of selection is going to be through an actual AL heuristic, but the choice depends on the classification model. We have considered working with RFs and SVMs during the dissertation, and have actually used each model during different stages of the dissertation. RFs were used at first with RSeT data, whilst SVMs showed better performance with the Random Sampling data onward — and especially so when working with the Cashew vs. Non-Cashew binary classification task, as will be presented in section 5. So, for RFs, we decided to use Breaking Ties: this is an uncertainty sampling algorithm based on posterior probability values, which is suited for ensemble-based classifiers, and was the most seen heuristic for AL pipelines in multi-spectral image classification during the AL Background section (2.3). Meanwhile for Support Vector Classifiers, since they are margin-based algorithms, Margin Sampling (MS) was picked instead, which consists of selecting the unlabeled data that is closest to the current decision boundary. We opted for the two pool-based sampling heuristics mentioned above instead of more complex alternatives due to simpler AL algorithms typically showing more consistent results across different models and datasets with different characteristics [40].

4.5.2 Labeling Process

In the Active Learning stage, the labels were acquired according to each pixel's class in the 2021 land cover, following the seven-class scheme introduced by RSeT, even if the classification task was being done with just the binary labels — this allows for the possibility of drawing back to the seven-class classification task if desired in the future. However, aside from the first round of Active Learning experiments (section 5.2.2), points were also annotated using the aforementioned eighth class: Other (Non-cashew). This way, certain pixels where there is annotation uncertainty between two or more possible classes (for example, Savanna and Non-Forest), with neither of them being Cashew, could still be used for Cashew vs. Non-Cashew binary classification tasks — for seven-class land cover classification procedures, points from class eight were discarded.

The label acquisition process proceeded as follows: the first AL experiment (section 5.2.2) was made in a smaller setting to have a quicker assessment of the performance and the limitations of the application. For that, the algorithm could only probe points from nine sub-regions (shown in Figure 4 in Appendix) spread throughout the Region of Interest. During this process, points were selected individually using Breaking Ties, resulting in a sparse dataset just like the one RSeT offered. But after the first round of Active Learning, the selection started being made in patches, as we noticed that labeling individual pixels was a rather slow process. In each iteration, a smaller number of pixels across the ROI was selected, either through Random Sampling (section 5.3) or Margin Sampling (section 5.4); then, the coordinates of the immediate neighbors were jointly stored and given to the oracle, which annotated the nine points in the 3×3 immediate neighboring patch, instead of just the single central pixel.

This change is capable of speeding up the labeling process up to nine times, since when looking at one pixel, the neighboring ones are also observed by the oracle and are likely to have similar characteristics and belong to the same class. It also leads to a dataset that is more appropriate to work with Deep Learning algorithms in the future like patch-CNNs, since they can take advantage of this patch data type for spatial feature extraction. However, there is the drawback of having an effective training set dimension that is not accurately represented by the number of labeled samples in each dataset, due to the spatial autocorrelation between neighboring pixels — that is, pixels from the same patch might have such similar behaviors that the inclusion of the nine points in the 3×3 patch does not add much more information to the labeled set than just considering the central pixel. Despite this note, the advantages of the patch-based label acquisition procedure were considered greater than this disadvantage, which is also mitigated by the grouped train-test splitting algorithm we considered to prevent inflated test scores.

Not all samples requested by the Active Learning heuristics were labeled by the Oracle (the author). Often, cases were too difficult or near impossible to classify with confidence, because of: the fuzzy nature of land cover classification (multiple sections of the ROI show two or more classes intertwined, resulting in pixels with mixed behaviors that are not suited to be attributed to just one of the seven considered RSeT classes); reduced Historical Imagery coverage (the classification of land cover types such as Savanna or Cashew is extremely dependent on multiple

and temporally spread high-resolution satellite images, in order to identify their seasonality or growth, respectively, and distinguish them from more stable classes such Non-Forest or Open Forest); noticeable land cover changes happening without sufficient contextual imagery to understand whether they took place before 2021 or only after. Samples with not enough annotation certainty were still noted and stored in a separate file, only to stop the Query Policy from requesting the classes of those points to the Oracle again.

4.5.3 Initial Train Set and Batch Size

The initial training set has to be stipulated before beginning new data acquisition, as this is an iterative process where the data acquired by the Query Policy depends on the previously obtained classifier, which is directly influenced by its training set — thus, the initial training set is the first step of the entire process, and changing it would result in a different AL iterative selection. During the first AL experiments (section 5.2.2), the entire RSeT training set was used as the initial one. However, at the end of section 5.3.3, two conclusions are going to be shown: there is no noticeable advantage in using RSeT data to train a regionwide classifier, and there is no significant disadvantage in starting with a very small training set (the first AL iterations might show higher variability, but the final performance is not degraded). Therefore, after the first round of experiments, a small initial training set was defined from Random Sampling acquired data. We separated a total of 11 polygons, corresponding to 93 samples from the RS dataset that were separated from the rest for just this purpose. This set has at least one polygon of every class besides Closed Forest to ensure the initial classifiers already have access to different types of behaviors from the get-go, having two polygons of Savanna, Cashew, Non-Forest, and Other (Non-Cashew).

The AL results were mainly measured in terms of Cashew F1-score, the class of interest, but also according to Balanced Accuracy. The choice of AL batch size (number of newly acquired samples per iteration) depends on the size of the ROI and unlabeled pool, and corresponds to a trade-off between computational efficiency and more informative samples: bigger batch sizes will acquire more points in fewer iterative steps, but this might allow the process to also select samples that are less informative or valuable. For each BTs iteration in the first AL experiment (section 5.2.2), the batch size goal was set at around 100 points. However, every batch was actually formed by 140 individual points to account for discarded samples not labeled by the Oracle. Meanwhile, for each MS-AL iteration (section 5.4.1), we selected the 34 pixels with the highest uncertainty (smallest margin values) and obtained their corresponding 3×3 patches, resulting in a total of $34 \times 9 = 306$ samples per iteration — again, not all of them were classified by the Oracle. The increase in batch size after the first round of experiments was done to speed up the AL process, reducing the number of iterations needed, but also because the selected samples were still highly uncertain.

Chapter 5

Developments

After describing the data and methods in chapter 4, this chapter now presents the main results obtained during the span of the dissertation. The chapter starts with a brief Exploratory Data Analysis (EDA) section that gives insight into certain characteristics of the data and the classification task. Then, the developments obtained with each of the three used datasets (the RSeT, Random Sampling, and Margin Sampling datasets) are presented, as well as the results for every Active Learning stage of the project.

5.1 EDA

5.1.1 95% density level ellipses

Figure 5.1 presents a bi-dimensional representation of the 95% density level ellipses for every RSeT sample for each class, which allows for an initial overview of the characteristics, separability, and variability of the seven considered classes. The representation is usually done in the Red-NIR or SWIR2-NIR bidimensional spaces [14], which is especially informative for vegetation data types, as it can show the red edge behavior in a low dimensionality. From all of the classes, Non-Forest presents the highest variability, which is expected for an "umbrella" class that contains multiple subclasses (farmland, road, buildings) that do not fit with the other defined ones. The Savanna 95% ellipses show a big overlap with Non-Forest in both representations — this is likely due to a similar spectral signature and a seasonal behavior that is close to some types of farmland (where the green coverage also varies across seasons). Meanwhile, Water is a usually well-separated class, due to its spectral signature being very distinct from that of vegetation or any type of soil. The most significant overlap with Water occurs for areas classified as Mangrove, a group of trees that typically exist along coastlines or tidal rivers — noticeable in the SWIR2-NIR space. Finally, we notice the clear overlap between Cashew and the Forest types, most noticeably Open-canopy Forest. Even though the intersection might not be as visible in other band spaces, these classes are known for having very similar spectral signatures [14], which complicates the classification task.

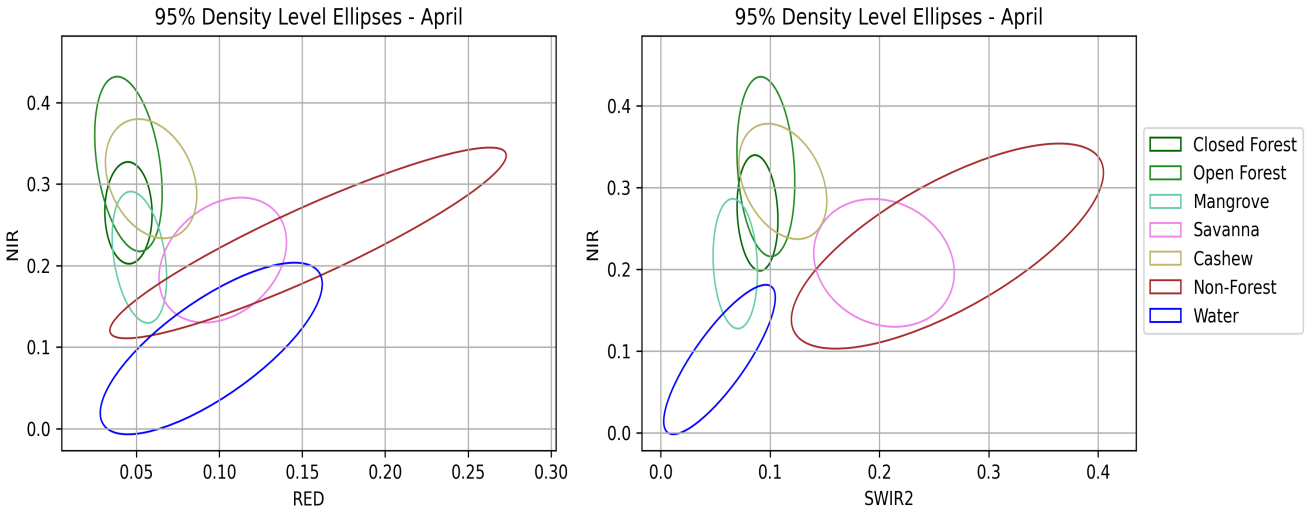


Figure 5.1: 95% density level ellipses for each RSeT dataset class, in the RED-NIR (left) and SWIR2-NIR (right) bidimensional representation spaces, for the April 2021 monthly median composite.

In order to ensure the end of the dry season is the best time for the identification of cashew orchards, we also obtained monthly median composites in GEE for the 12 months of 2021, and sampled the band values for the RSeT data in order to plot 95% density ellipses for every month, as shown in Figures 2-3 in Appendix. This sanity check led to the same conclusion made in [14]: the overlap between Cashew and Open-canopy Forest is more significant during the humid season, decreasing more and more up until the end of the dry season. This observation reinforced the decision to continue to work with spectral information from April 2021.

5.2 RSeT results

The first ML and AL results were all obtained working with the RSeT dataset of 5504 labeled samples. The characteristics of the dataset and train-test split were already discussed in the Data and Methods section (respectively, sections 4.2 and 4.2.1). Therefore, the results will be shown starting from the hyperparameter tuning part. The score defined for the Bayesian optimization search procedure done by *hyperopt* was the mean F1-score for Cashew, the class of interest, determined over a 5-fold grouped stratified sampling split of the training data. The best performing RF and SVM models during HP search were then evaluated on the test set, with the results being shown in table 5.1. Based on the findings, we have decided to work with the Random Forest algorithm: balanced accuracy is slightly lower than that obtained with the SVM, indicating that the RF likely has less overall sensitivity especially for minority classes (as the Overall Accuracy ends up being over 3.5% above the SVM results); however, the F1-score for Cashew is also roughly 3% higher for RF, and we have considered this the main criterion for model selection, as it will allow for a better description and identification of cashew orchards, without losing much overall descriptive power. The chosen RF is comprised of 98 base learners (DTs), with a maximum depth of 8, and 19 maximum features considered per split.

	Accuracy	Balanced Accuracy	F1-score Cashew
RF	**0.916**	0.822	**0.884**
SVM	0.88	**0.84**	0.856

Table 5.1: Test set results for the best Random Forest and Support Vector Machine after hyperparameter tuning.

5.2.1 Preliminary Classification and Land cover map

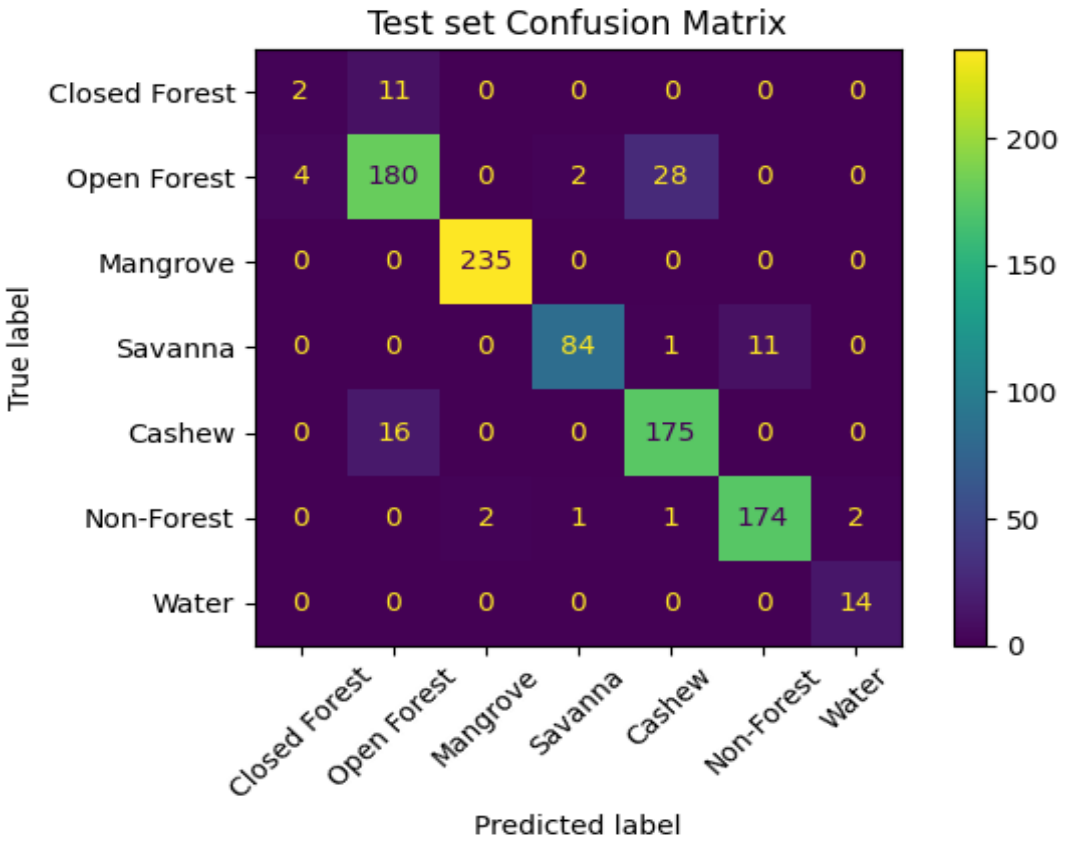


Figure 5.2: Confusion matrix of the best model after Hyperparameter tuning.

The test confusion matrix for the selected model is presented in Figure 5.2, and there are three noticeable things to be taken away from it. The first conclusion is that Closed Forest is being neglected by the classifier, as most of its test cases are being confused with Open-canopy Forest. This likely happens considering the spectral similarity between the two (as seen in Figure 5.1) and that Closed Forest is the second least prominent RSeT class (Table 4.1), which fits the previous conclusion that the Random Forest might be less sensitive to certain minority classes. The second observation regards the slight confusion between Savanna (or Sparse vegetation) examples and Non-Forest. This can happen since some examples of Non-Forest present a seasonal variability that is similar to that of Savanna, where the region becomes almost fully covered with green vegetation in the humid season; the confusion can be amplified if the Savanna examples

have a very small remaining vegetation fraction during dry season months, which can be hard for a medium-resolution satellite to pick up. Thirdly, the most evident confusion — despite still being small — happens between Open Forest and Cashew. These are two classes with a very similar spectral signature, as was already hinted in the 95% density level ellipses; hence, it is natural for the two classes to have some level of intrusion, resulting in a few practically indistinguishable cases.

With the best performing RF model, we then obtained a preliminary Land cover map for the year 2021. The Random Forest can use the GEE Image Asset which contains the training features as bands for all the pixels in our ROI (as this was all prior to the addition of GLCM statistics, we were still working with 126 bands) in order to make class predictions, which can then be considered as the land cover type of that pixel. The first land cover map obtained this way is shown in Figure 5.3. River courses and sea stretches are visible across the region. The extensions of the water courses typically appear to be covered by either Mangrove or Non-Forest. Meanwhile, Open Forest mainly appears in the southern half of Guinea-Bissau (mainland and islands), with virtually no coverage in the Northern half or the Republic of Guinea's fraction. Sparse vegetation covers a big chunk of central and Northeastern land, with Cashew appearing mostly in the Northwestern part of Guinea-Bissau. Of course, this is only a broad description, as the land cover distribution across the ROI is much more complex, with a lot of fuzzy regions with intertwined classes — as it appears, for example, on the Southeastern fraction, with a lot of mixtures between predicted Cashew, Savanna, and Non-Forest regions.

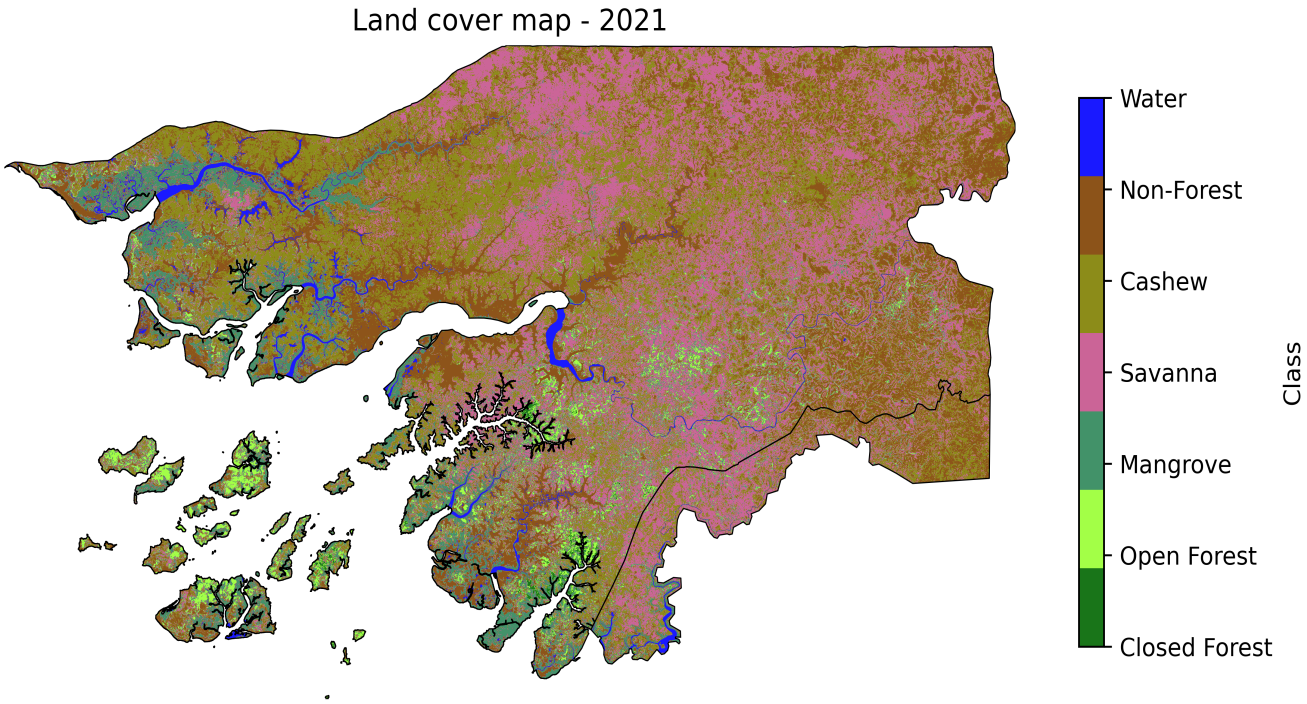


Figure 5.3: Initial 2021 Land cover map, obtained with the predictions of the best RF classifier.

The percentage of predicted pixels per class across the ROI is presented in Table 5.2. Some proportion dissimilarities between the land cover map and the RSeT data labels (Table 4.1) are

noticeable. Across the region, much more Savanna appears than what the existing percentage in the initial dataset would suggest. Meanwhile, the forested region (Closed and Open canopy Forest) is estimated to be just barely over 10% of the total country's land cover. This information also allows for an initial estimate of the amount of land cover overtaken by Cashew, as the model predicts that roughly 19.4% of the area belongs to this class.

Closed Forest	Open Forest	Mangrove	Savanna	Cashew	Non-Forest	Water
0.3	9.9	8.5	37.3	19.4	21.6	3.0

Table 5.2: Percentage of each class on the land cover map obtained with RF predictions.

5.2.2 First AL pipeline

The first AL experiment was made in a smaller setting — only across 9 sub-regions spread throughout the ROI, shown in Figure 4 in Appendix. New samples were acquired using Breaking Ties, since we are working with Random Forests, as well as Random Sampling for comparison, but both acquisition heuristics were only done for three iterations, resulting in a selection of 302 new labeled samples added to the training set for each procedure. A total of 420 pixels were selected by Breaking Ties and evaluated by the Oracle, which constitutes a $118/420 \approx 28.1\%$ rejection rate (selected points that were not classified).

Figure 5.4 shows the absolute values of the two most likely classes for the selected pixels (which are the two classes used to define the BTs metric of uncertainty). Closed Forest, Mangrove, and Water were all classes noticeably less involved with the most uncertain selected samples. For both Mangrove and Water, this is likely due to the classes being well separable from everything else: they have distinct spectral signatures and are thus likely to have a higher associated prediction confidence. Meanwhile, Closed Forest shows similarities with Open Forest, but it is also a minority class on the initial training set, so the RFs are most likely disregarding this class and biased to predict others instead, also resulting in less associated prediction uncertainty. Savanna and Cashew are the classes that mainly appear in the two most likely classes of the selected samples, followed by Open Forest and then Non-Forest. During the classification process, we noticed that the confusion of the newly labeled Open Forest points was mainly happening with Cashew — mature orchards have an almost identical aspect and signature to a stable forest. There was also a noticeable difficulty regarding Savanna and Non-Forest, as certain Non-Forest samples include farmland or empty soil that happens to have similar seasonal behavior to that of sparse vegetation regions. However, some confusion between Savanna and Cashew ends up happening as well: despite cashew trees being evergreen, there are plantations where the interstices are used for growing other crops, possibly leading to a seasonality similar to that of Savanna woodlands.

The first AL results measured in terms of Cashew F1-score are shown in Figure 5.5. Comparing BTs (sometimes also called Best versus Second Best) with RS, the first plot point (the zero on the x-axis) is identical because it represents the results with the initial training set (before the

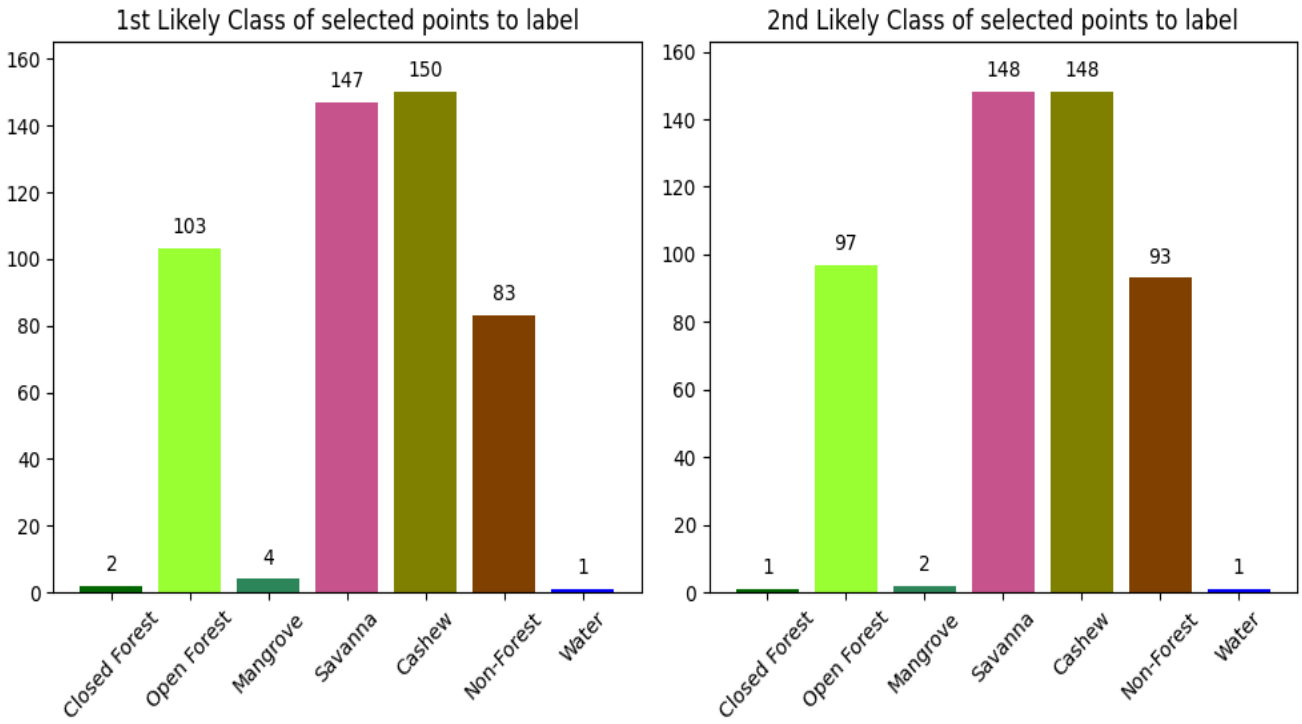


Figure 5.4: Barplots stating the number of occurrences of the most likely and second most likely classes for the BTs selected points.

addition of new samples), which is the same for the two heuristics. Then, BTs starts performing slightly better for the first two iterations, but on the last iteration, the results are almost the same again. The overall conclusion is that the updated RFs are not gaining significant predictive boosts with the addition of newly labeled samples: aside from the first AL iteration, the results decay and end up worse than with the initial training set. This sub-optimal performance led us to stop the current AL process after the third iteration until we could determine the cause of the poor results. One possibility is that we had already reached the performance peak with the RSeT dataset — in this scenario, there would be no advantage in expanding the training set, and the idea of AL would have little applicability. There is also the chance that the labeled samples of the two origins (RSeT data or labeled during the dissertation) do not coincide: as they are labeled by different Oracles, the followed classification criteria might not be the same, which could result in a mix of behaviors, increased class heterogeneity, and a performance decay, which we do see happening.

A final option could be that the evaluation of the Active Learning iterative performance is not adequate or capable of showing the classifier's descriptive improvements when it is equipped with the newly added samples. This hypothesis comes from the fact that the RSeT dataset (which served to define both the initial training set and the test set used so far) is concentrated on the Northwestern and Southwestern fractions of the country (Figure 4.3). Few samples are spread across the central and eastern parts of Guinea-Bissau, and none appear in the added Republic of Guinea region. Naturally, these regions have different behaviors from the RSeT concentrated ones: new class behaviors might become visible in different geographical areas; also, the satellite acquisition process suffers from changes in sunlight exposition, climate conditions, or atmospheric

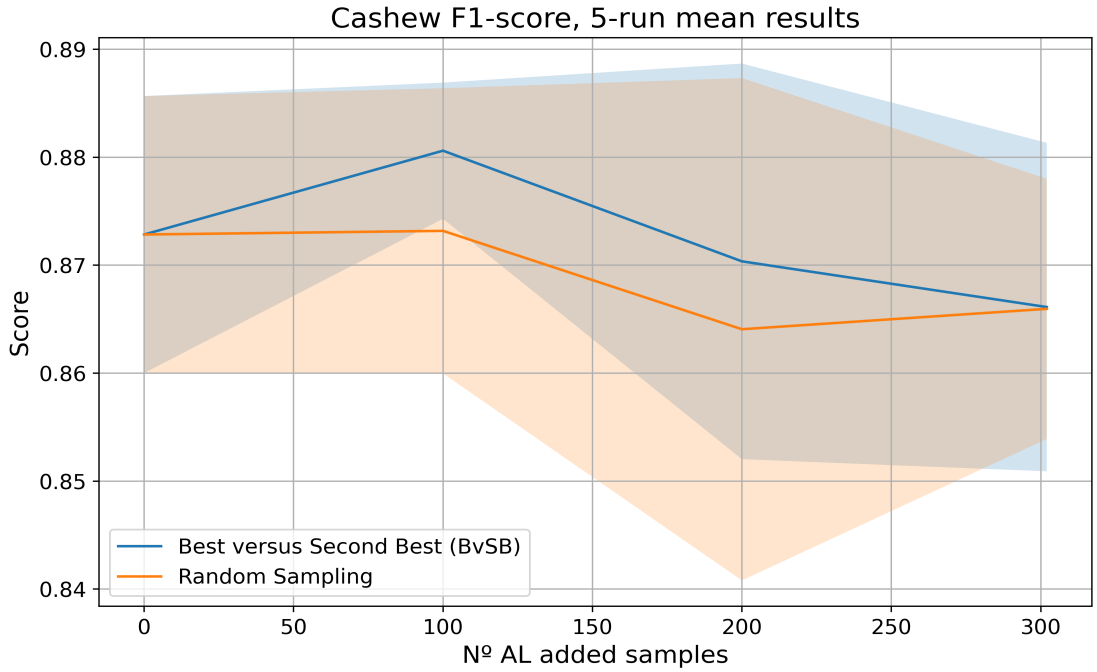


Figure 5.5: First AL results. Comparison of Random Sampling with the Breaking Ties selection heuristic. Results were averaged for 5 RFs with different initial random states.

effects that can interfere with the measured spectral signature. Therefore, the stipulated RSeT test set (created exclusively with points in limited areas, based on a fold resulting from the stratified grouped 6-fold algorithm applied to the RSeT dataset) may be evaluating a specific set of behaviors that is not representative of the entire Region of Interest, and the upgrades made with the new informative and diverse samples end up conflicting with the previous data. But this would not mean the classifiers are worsening in any way, as they would be more generalizable and less specific to certain geographical behaviors.

This last scenario is the only one of the three hypotheses where the area of AL, done by the author, would help improve the classification capability with the selection of a smaller training set. But it imposes the task of stipulating a new test set and the question of finding one that is representative enough of the whole Region of Interest. This is why we also worked with a test set from regionwide Random Sampling data, as will be shown in the next section.

5.3 Random Sampling results

After the poor initial AL results, we decided to define a new test set. Considering the hypothesis of the inadequacy of using RSeT data to form a test set, and with the expansion of the analysis to the entire ROI instead of just the nine sub-regions, we decided to create a new test set from the randomly acquired nationwide samples, mainly because this selection process is already going to be used as a baseline for any other selected AL heuristic and is likely to produce a roughly uniformly scattered dataset.

In this process, we initially obtained 4498 randomly selected samples distributed in 3×3 patches from a total of 6000 evaluated ones — an approximate 25.0% rejection/not-classified rate. The distribution of the patches across the Region of Interest is shown in Figure 5.6. Comparing with RSeT data in Figure 4.3, it is clear that the RS labeled samples are much more uniformly scattered, with a better coverage along Guinea-Bissau's mainland, but also some along the Republic of Guinea's added region. Figure 5.7 shows the distribution of patch sizes (that is, the amount of patches that have a certain number of labeled pixels per patch). From the 527 patches with at least 1 labeled point, 455 or roughly 86% are fully labeled (that is, the entire 3×3 region is classified). The remaining 14% have at least one point that could not be annotated and had to be discarded. The latter cases cannot be used for DL or patch-based algorithms, but they can be leveraged by classical models such as RFs and SVMs, which were the ones considered during the dissertation.

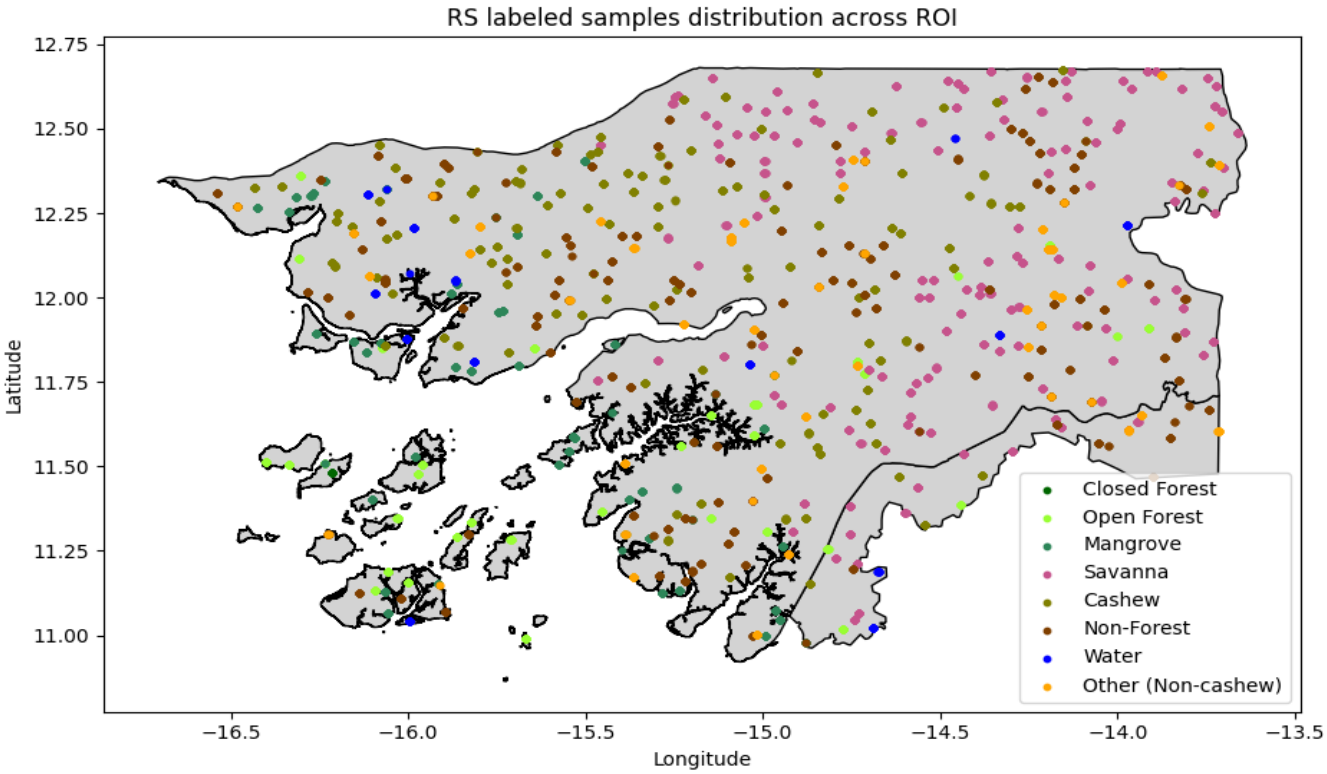


Figure 5.6: Geographical location of the RS selected and labeled data.

Figure 5.8 presents two barplots: the class counts including discarded samples (left), and the class percentages in the clean and usable dataset (right). It is interesting to compare the class proportions with the ones present on the RSeT dataset (Table 4.1). Cashew and Non-Forest are still two major classes in the new RS data, with nearly identical percentages and representation as before. Closed Forest and Water are again the two less prominent classes, with Closed Forest particularly only having one patch represented on the randomly selected set. The main differences between the datasets appear with Mangrove — which was more strongly represented in the RSeT dataset than now —, and Savanna — proportionally more predominant in RS data.

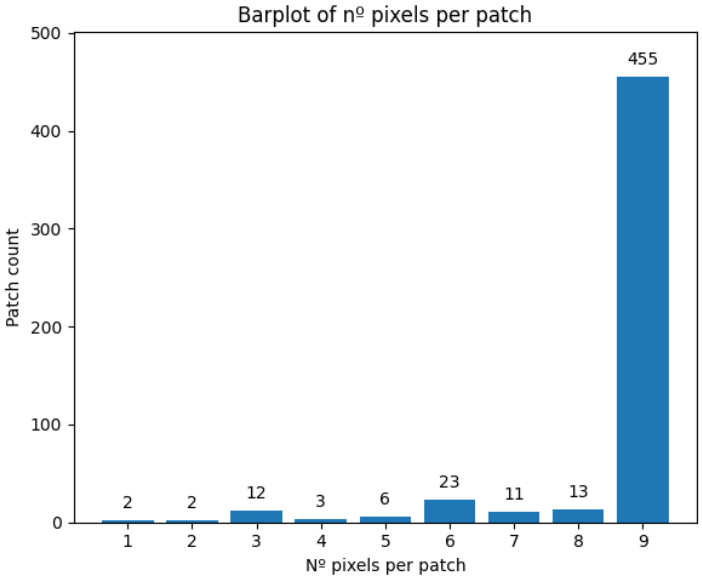


Figure 5.7: Barplot of the number of RS patches by patch size (number of labeled pixels that the patch contains).

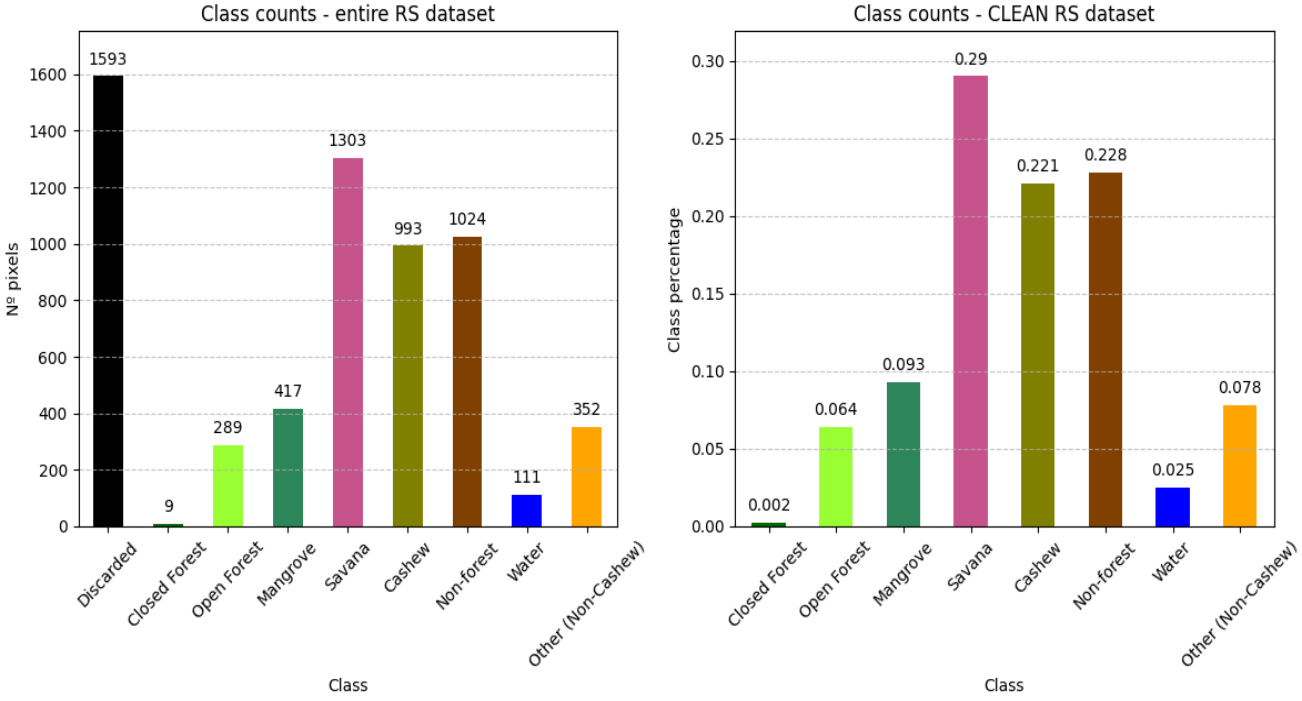


Figure 5.8: Barplots of RS samples' class counts, including discarded samples (left). Percentage of annotated points per class in the clean RS dataset (right).

The Random Sampling dataset shows class proportions that are much closer to what was observed across the region during the classification step — especially for the majority classes —, and also to what was estimated in the initial 2021 land cover map (Table 5.2). Considering its roughly uniformly spread across the region, it is likely a more suitable font of data to create a test set than the RSeT dataset — one that more properly evaluates a model's generalizability and predictive

capacity of diverse behaviors. This led to the train-test split mentioned in section 4.2.1 of the Data and Methods chapter, presented in Table 4.2.

5.3.1 Feature selection

During the Feature Acquisition section (4.3), we mentioned the addition of GLCM statistics resulted in an Image Asset composed of 360 bands, and discussed the need to apply a feature selection technique. This process was done with the RS dataset since the GLCM statistics were only added to the pipeline after working with the initial RSeT dataset, and also because the feature space — just like the classifier and the initial training set — had to be stipulated before advancing to the Margin Sampling procedure, as it directly affects the progression of the AL iterative process. The feature selection step occurred before the RS hyperparameter tuning, so it was done with the best selected RF architecture from RSeT data.

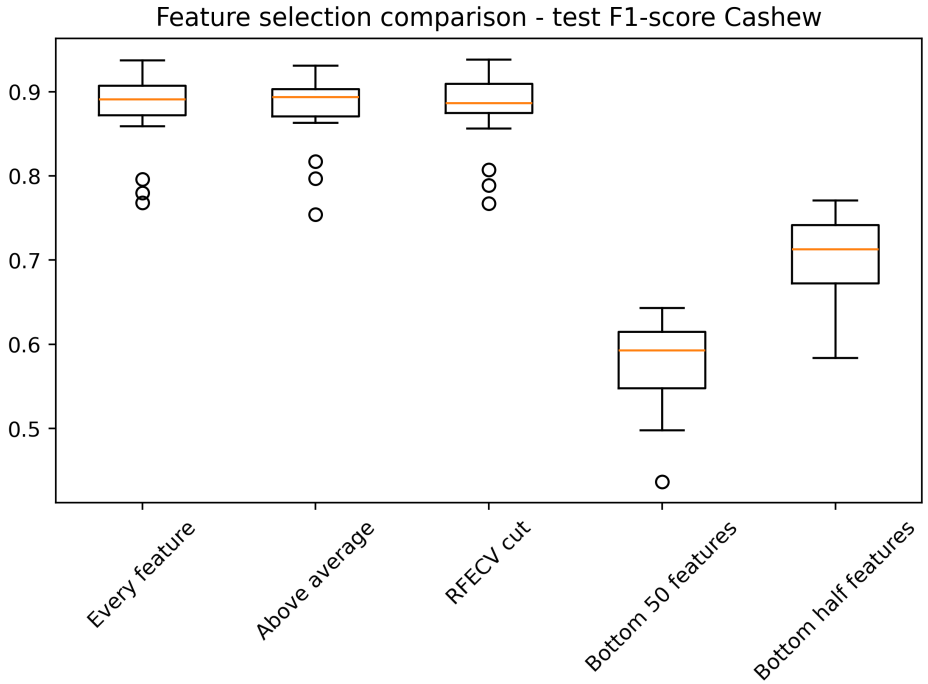


Figure 5.9: Feature selection comparison, with Stratified Grouped 10-fold on the training set. Five different feature sets were compared, presented from left to right: every feature (360), features with above average importance values (67), the result of recursive feature elimination RFECV (250), and the bottom 50 and bottom half (180) of features in terms of importance scores.

Figure 5.9 presents a performance comparison between different feature selection procedures across 10 different RS train-validation splits. Boxplots were drawn to contrast using every feature versus the 250 not eliminated by RFECV and the 67 features with an above-average importance score. Results are also compared to the cases where the RF classifier only had access to the bottom 50 features or bottom half (180) features in terms of feature importance score. The most important conclusion taken from the boxplots is that there does not seem to be any major

advantage in using every feature or the ones selected by RFECV instead of just the above-average importance ones. The top whisker of the above-average importance boxplot is slightly below the other two, but it also shows the highest median, slightly above 90% Cashew F1-score. With such similar results, we decided to only use the 67 above average features from now on, as it performs the deepest cut of the feature space, without visibly sacrificing predictive power. We notice the bottom variables are not inherently worthless: Cashew F1-score (especially using the bottom half features) is above 0.5, meaning the classifier is able to extract information for the classification task; but there are still no noticeable advantages in using them, as it would only increase the dimension of the feature space.

Figure 5.10 presents the nature of the 67 selected features. More than half of them correspond to spatial contextual statistics, indicating that the addition of GLCM was beneficial. The CCDC coefficients only account for roughly a quarter of the selected features. It is possible they could be more valuable if the CCDC algorithm was parametrized more properly, but as previously mentioned, this is extremely difficult to do in Google Earth Engine. At last, the remaining roughly one-quarter of features correspond to either the original S2 bands, vegetation indices, or added bands during CCDC pre-processing, indicating other forms of pixel-wise satellite information are also useful.

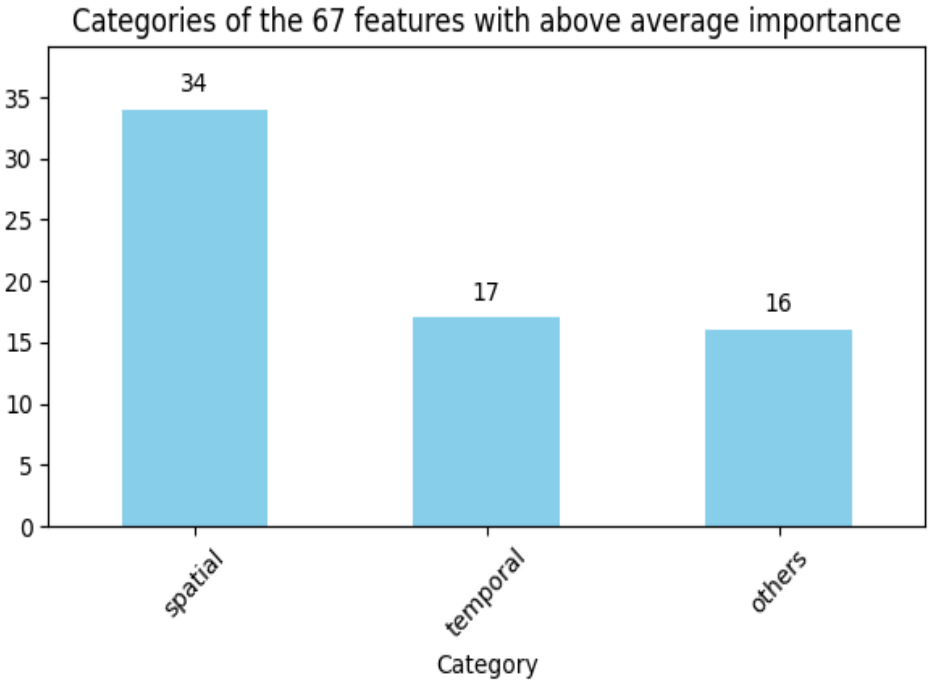


Figure 5.10: Barplots of feature types after Above Average Importance feature selection. Spatial refers to GLCM-obtained features, Temporal are CCDC coefficients, and Others corresponds to the original S2 bands, vegetation indices, or other features obtained during the pre-processing.

5.3.2 Hyperparameter Optimization and 7-class vs. 2-class results

The hyperparameter tuning step was made similarly as before, and as mentioned in section 4.4. The Bayesian optimization search procedure from *hyperopt* was used, with the selected metric being the mean Cashew F1-score, averaged over a stratified grouped 5-fold split on the Random Sampling train dataset. The search was made for the seven classes usage case (discarding the points annotated in class eight), as well as for the binary Cashew Non-Cashew classification task (inserting class eight into the general Non-Cashew binary class). The best seven-class model was also a Random Forest, this time with 73 estimators, using just 2 features per DT split, which could have a maximum depth of 10. However, for the binary classification problem, the best algorithm was an SVM, with a 4^{th} degree polynomial kernel, a Cost parameter equal to 0.0117 and gamma equal to 0.6439.

The RS test set results of the best models are presented in Table 5.3, with the performance of the best RSeT Random Forest also being shown for reference (that is, the best architecture during the RSeT dataset hyperparameter tuning phase, but now trained with RS data). It is clear there isn't a significant performance difference between the two seven-class Random Forests. This is likely due to the architecture of the best RSeT and best RS seven-class RFs being very similar, with the biggest difference happening in terms of the maximum number of features per split. Besides this, we notice the increased Cashew sensitivity when reducing the number the classes to the binary problem. There is a boost in Balanced Accuracy — likely because with just two classes there are no minority classes blatantly discarded by the classifier, as it happened for Closed Forest and the best RSeT RF for example. However, the F1-score for Cashew increases by approximately $3-4\%$ over the seven-class classifiers. In this scenario, despite the information loss associated with the change to a coarser annotation scheme (from seven classes to two), we decided to follow the binary classification approach in order to increase cashew detection and allow for better land cover mapping and estimate-making regarding this crop. This means that from this point onward — for the rest of the Random Sampling section but also the Margin Sampling section — we are going to work with the best performing two-class SVC.

	Balanced Accuracy	F1-score Cashew
Best RSeT model	0.811	0.802
Best RS model - 7 classes	0.803	0.814
Best RS model - 2 classes	**0.905**	**0.843**

Table 5.3: Test set results for the best RF and SVM after hyperparameter tuning.

The binary classification problem is still slightly imbalanced, due to the $28.3\% - 71.7\%$ Cashew/Non-Cashew class proportion. A simple approach to deal with the imbalance is by setting the *class_weight* parameter of *scikit-learn*'s Support Vector Classifier implementation to "balanced", which sets class importance weights during the SVM optimization process to accommodate for the difference in class proportion, increasing sensitivity to the minority class (which in this case is

Cashew). The division of Cashew and Non-Cashew points using the previous train-test split (with eight classes, the seven defined by RSeT plus the added Other Non-cashew class) is very evenly distributed, as indicated in Table 5.4. We are going to keep this eight-class train-test split for the binary case. The fact that it is well distributed not only regarding Cashew and Non-Cashew labels but across the eight classes is an advantage — having similar Non-Cashew proportions in training and testing, but where one has a majority of examples from a specific class (p.e. Savanna) and the other from a different one (p.e. Non-Forest) would likely lead to a discrepancy between train and test characteristics.

Class Names	**Cashew**	**Non-Cashew**
Train percentages	0.47	0.506
Test percentages	0.53	0.494
Total counts	993	3505

Table 5.4: RS total number of points and class percentages for train and test splits, using the binary annotation nomenclature.

Figure 5.11 shows a performance comparison using different Cost and gamma hyperparameter values between three SVM configurations: the one resulting from the hyperparameter search procedure, with a fourth degree kernel and balanced *class_weight* parameter; a second degree polynomial kernel SVM with balanced *class_weight*; and a fourth degree one that does not use the *class_weight* parameter. This comparison allows us to determine the influence of certain configuration changes on the descriptive capacity of the classifier. The figure shows that there are optimal performance peaks for both Cost and gamma hyperparameters for all 3 kernel configurations. The peaks appear on the same hyperparameter spots for both fourth degree configurations, but the second degree classifier benefits from slightly larger Cost and gamma values. The overall maximum performance is also very similar between the three architectures. However, both of the non-optimal configurations can only match the peak performance of the optimal HP configuration in one metric: the second degree reaches an equivalent Balanced Accuracy score as the optimal SVM, but slightly lacks in terms of Cashew F1-score. Meanwhile, the opposite happens with the fourth degree and no *class_weight* configuration.

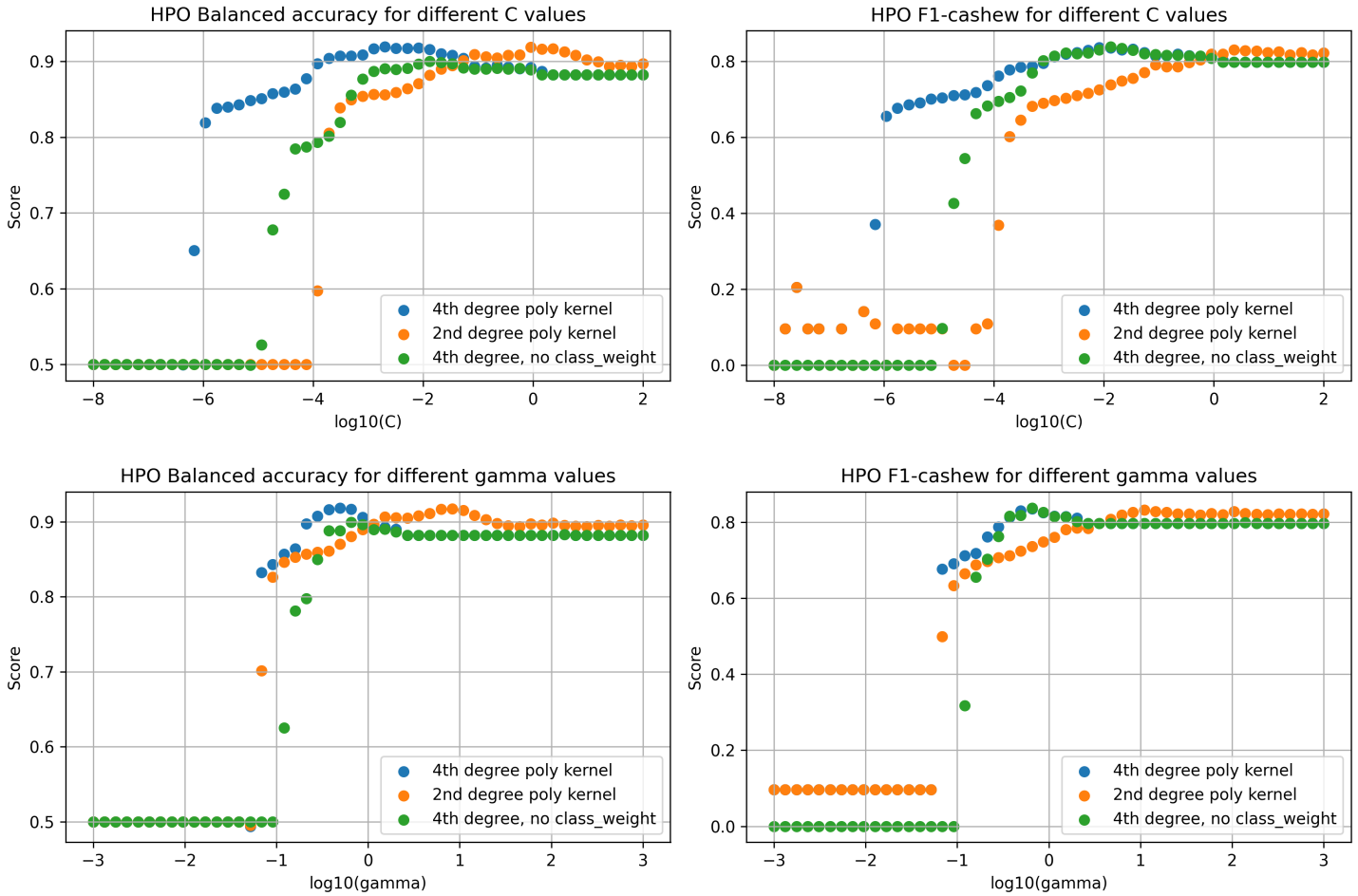


Figure 5.11: Balanced accuracy and F1-Cashew for different C (top) and γ (bottom) hyperparameter values, for different SVM parameter configurations: a quartic polynomial kernel with balanced *class_weight* (the selected configuration), a quadratic kernel with balanced *class_weight*, and a quartic with no *class_weight*.

5.3.3 Second AL pipeline

The RS-AL results will be shown with more detail during the following section (5.4), where they will be directly compared to the Margin Sampling ones in terms of performance. Chronologically, at this point, we still had to understand what to do with the RSeT dataset. It will not be used for testing — the test set considered from this point onward is exclusively formed with Random Sampling data points —, but it could still serve as an initial training set, from which we could expand by adding RS or AL-based samples. However, there are a total of 5504 RSeT points, and only 2181 RS ones available for training, since the rest is used for testing. With a noticeable size disparity, the following AL results might be shadowed by the larger initial training set — meaning we would need more RS added data to start noticing clear model improvements. Also, not every RSeT sample may be informative and necessary (for example, there is a high Mangrove proportion in this dataset which could likely be reduced, considering this class is

easily separable and identifiable). Thus, the first RS-AL results were obtained with different initial training sets, using various percentages of the RSeT dataset. Figure 5.12 shows the Active Learning curves obtained by adding RS training samples to different initial RSeT training sets — the only procedural difference between the curves is the initial RSeT training set. Each AL iteration in every curve is constituted by the random selection of 34 selected pixels and their respective patches. Since not all patches end up with nine labeled pixels, as for some cases we could not classify the entire patch, there can be a slight variability in terms of iteration batch size; so, the curves are represented in terms of mean batch sizes per iteration.

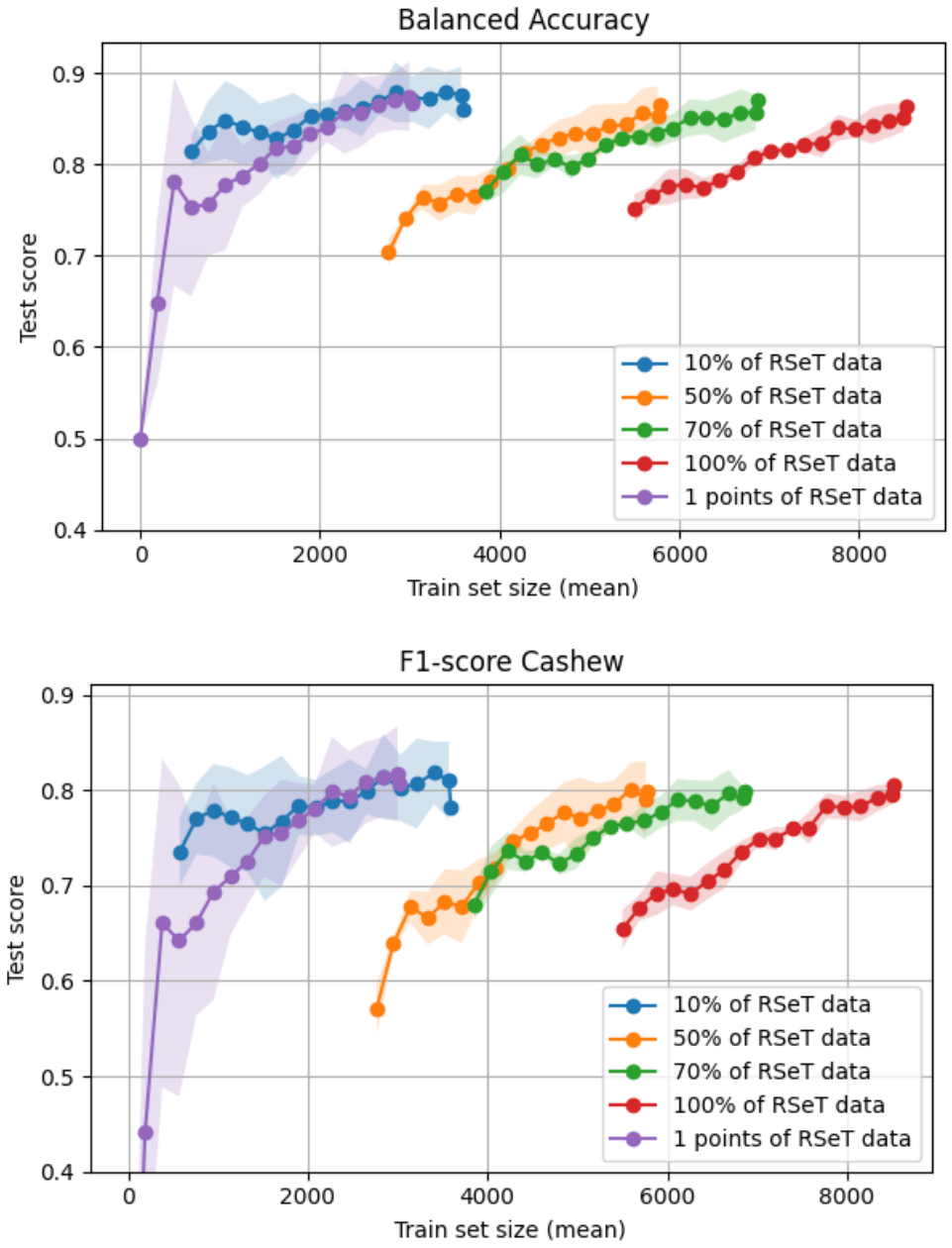


Figure 5.12: RS-AL iterative Balanced Accuracy (top) and Cashew F1-score (bottom) results, with different initial training sets. Each curve represents 5 runs of initial training set selection and the order for adding RS labeled data.

Opposite to what happened during the first AL experiment in section 5.2.2, we can see the gains of the Active Learning procedure — independently of the initial training set. Considering that the main change between the procedure presented here and the one followed during the first AL experiment is the change in the test set, the difference in results leads to the likely conclusion that RSeT data is not well suited to evaluate the descriptive gains obtained throughout the Active Learning process and does not make for an adequate test set. In terms of training, a classifier trained on RSeT data is capable of extracting useful information to predict on Random Sampling data, considering the new test set only has RS samples. This is evident, for example, by the first point of the 100% RSeT training data curve (the red one), which represents the case of a model trained with the entire RSeT dataset and evaluated on the RS test set — the classifier does not behave like a random guesser. Despite this, the worse performance of the RSeT training set, evidenced by that plot point, states a degree of incompatibility between the two datasets: the RSeT dataset is likely not generalizable enough, and thus not fitted to describe the behavior of a more comprehensive and diverse set, likely the case of the RS test set.

In terms of differences between different initial RSeT train set sizes, we notice that using more samples leads to more stable results. Since we are not working with sizeable datasets, there is likely a degree of variability associated with the fraction of data selected, which is naturally reduced when the training set is larger. However, larger initial sets are also defined by slower performance increases, which had been previously theorized: it takes a higher number of added samples to significantly increase the train set size in order to provoke a noticeable decision boundary shift. In fact, the final performance appears to be lower for the bigger RSeT percentage curves than for the 10% or just 1 initial RSeT point curves — especially in terms of Balanced Accuracy.

At this point in the dissertation, we concluded there seems to be no need for the RSeT dataset anymore. It is inadequate for testing the regionwide AL processes and introduces very few benefits when used as part of the training set. Therefore, during the following section for the MS and RS comparison, RSeT data was not used in any step.

5.4 Margin Sampling results

The reason for using Margin Sampling as the selection heuristic was made clear with the results from section 5.3.2 — during the Random Sampling hyperparameter tuning and model selection process, we switched to a binary approach, with the best model being a Support Vector Classifier. As such, Margin Sampling is a more adequate heuristic to work with margin-based ML classifiers than the previously used Breaking Ties heuristic — discussed in section 4.5.

The Random Sampling test set was also used to evaluate the performance of the Margin Sampling process — using the same test set ensures there can be direct comparability between the Margin Sampling and Random Sampling iterative procedures. The initial train set was built with 11 polygons from the RS train set, as presented in section 4.5. After section 5.3.3 it becomes clear

why we decided to use a small initial train set — there is a naturally higher result variability, but the final performance is not degraded by the small initial set (in fact, the opposite can happen), and the AL iterative gains are not shadowed by the larger initial set. For each MS-AL iteration, we selected the 34 pixels with the highest uncertainty (smallest margin values), and their corresponding 3×3 patches were obtained, resulting in a total of $34 \times 9 = 306$ samples per iteration. Since not every sample is classified by the Oracle, the number of points added to the training set per iteration is naturally smaller. The locations of the probed patches are presented in Figure 5.13. In contrast to the RSeT dataset (Figure 4.3), the MS acquired data shows a much smaller point density in the Western coastal area as well as the islands.

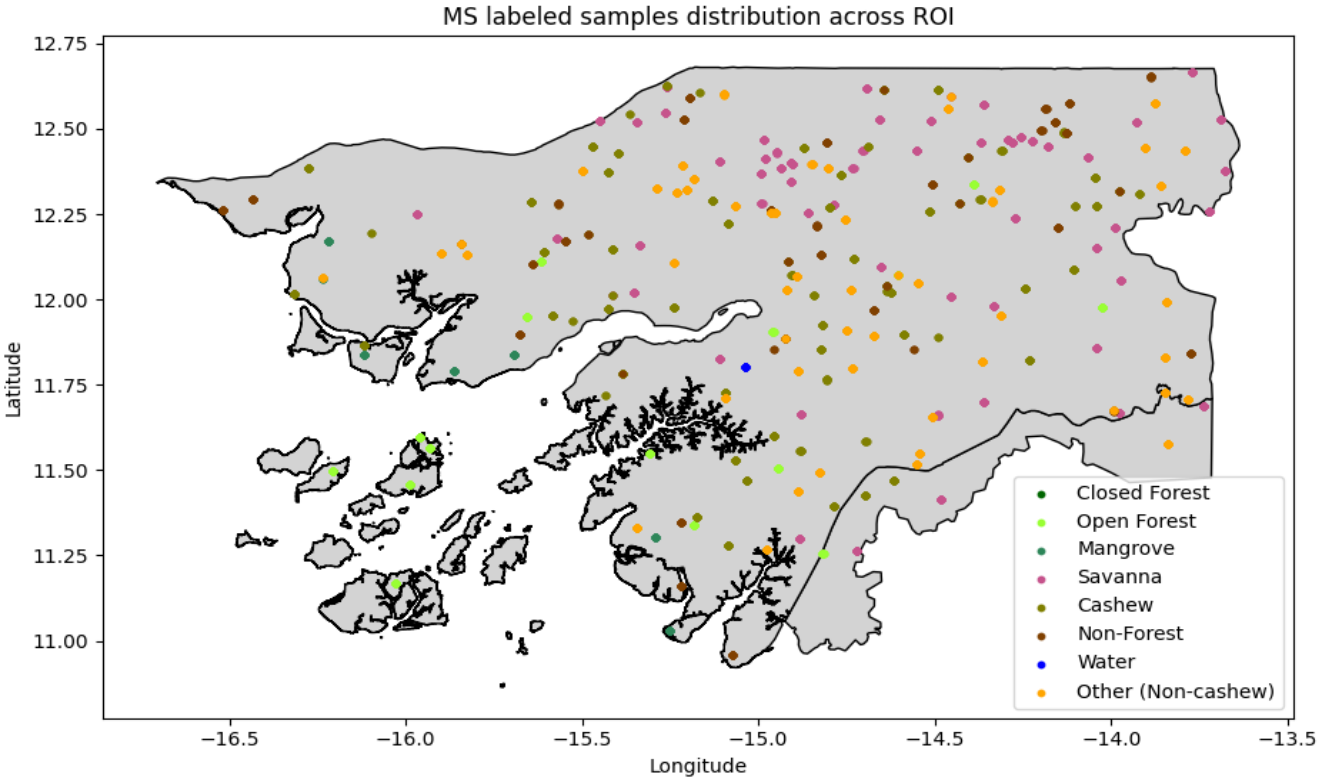


Figure 5.13: Geographical location of the MS selected and labeled data.

The Margin Sampling dataset resulted in a total of 1909 labeled samples, where 503 correspond to Cashew and 1406 to the union of the Non-Cashew classes: a $26.3\% - 73.7\%$ Cashew/Non-Cashew class proportion. The total number of points per class and the respective percentages are presented in Table 5.5. Notice the clear majority of Savanna and Cashew data. It is natural for Cashew to appear as one of the most probed classes — since it is one of the two binary classes, the most uncertain pixels selected by the Query Policy are uncertain with Cashew. The labeling task was also more challenging than during the Random Sampling acquisition — more fuzzy data or borderline cases were required by the AL process, which justifies the proportion increase in class-8 (Other Non-cashew) labeled data. Meanwhile, Water (and Mangrove) samples were rarely sampled by the MS heuristic, corroborating the idea that this class is highly separable and associated with a higher prediction certainty. Despite Closed Forest having some degree of spectral similarity with Cashew [14], no samples were probed. This might be the result of

this class not appearing in the initial training set and having to be discovered by the model, but it might also not have appeared during Margin Sampling, which is an uncertainty sampling procedure, due to other classes having stronger similarities to Cashew (in the case of forest types, the resemblance of mature cashew orchards to Open-canopy Forest might be stronger, as hinted by the 95% density level ellipses in Figure 5.1). Besides, this class represents a very small fraction of the region's land cover (as stated by RSeT researchers), so the likelihood of a Closed Forest point being selected for labeling is naturally smaller.

Class Names	**Total Nº points**	**% of points**
Closed Forest	0	0.0
Open Forest	132	0.069
Mangrove	57	0.03
Savanna	511	0.268
Cashew	503	0.263
Non-Forest	261	0.137
Water	9	0.005
Other (Non-Cashew)	436	0.228

Table 5.5: Absolute and relative Nº of MS samples per class.

5.4.1 Third AL pipeline

The AL evolution for Random Sampling and Margin Sampling are jointly seen in Figure 5.14, with the performance being defined in terms of Balanced Accuracy and Cashew F1-score. Both selection procedures translate to improvements in the test set description. However, Margin Sampling does so more efficiently, with steeper increases during early iterations: just after two AL steps, the classifier obtained an F1-score of 81.2% using Margin Sampling, in comparison to the 71.2% score using randomly selected data. Model peak performance is also higher with MS labeled training data, reaching 87.2% Cashew F1-score with 1254 training samples — in comparison, RS's peak happens at 85.4% F1-score with 1599 samples.

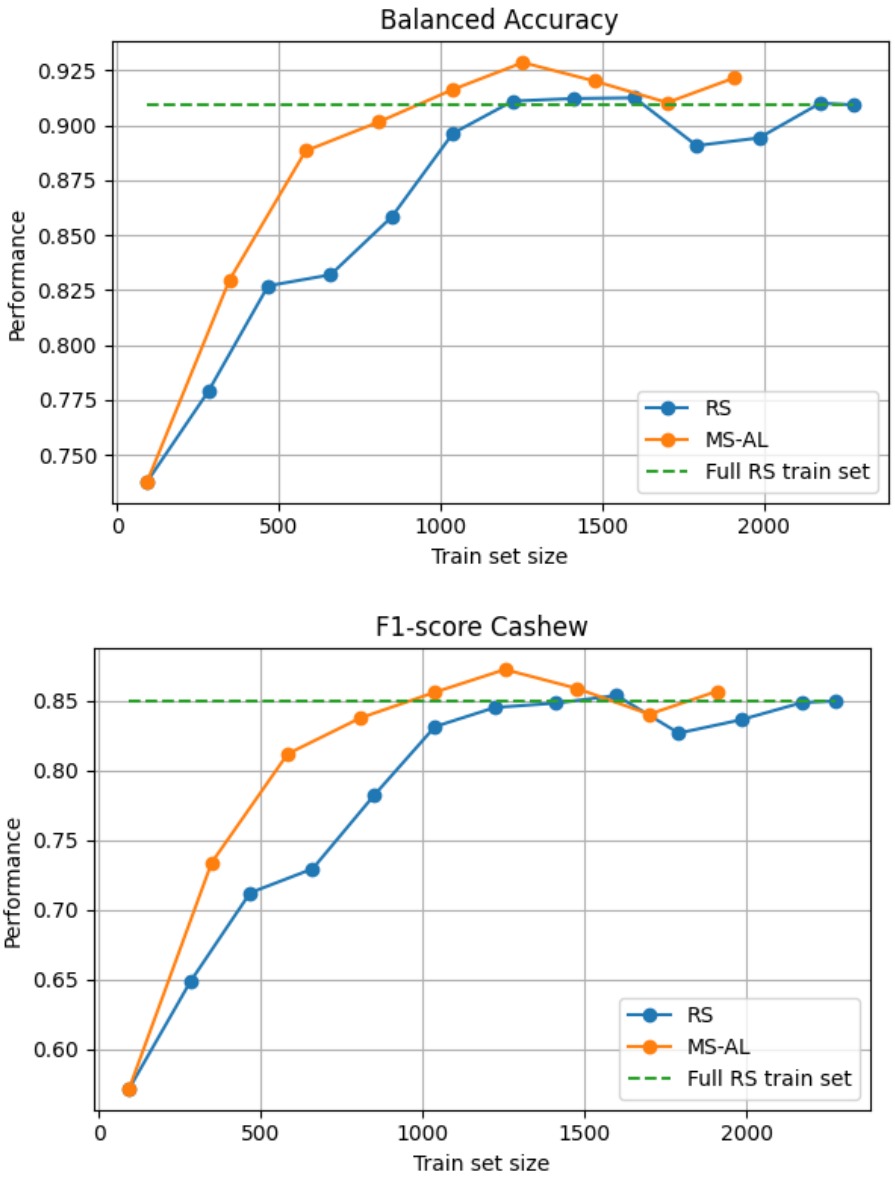


Figure 5.14: AL iterations comparison between Margin Sampling and Random Sampling. Results are presented in terms of Balanced Accuracy (top) and Cashew F1-score (bottom).

Both MS and RS final results present some degree of oscillation. This can be a sign of fluctuation around the maximum performance possible, due to either limitations of the algorithm used, or fuzziness and intrusion in the data. If the latter is the reason for the peak, it would be unlikely to keep having gains if we continued the data acquisition process. Figure 5.15 shows another reason we decided to stop the acquisition process: the percentage of discarded points (requested by the AL algorithm but that could not be classified) is higher in the final steps, with the last iteration having more than a third of requested points discarded. During the annotation task in Google Earth Pro, this was accompanied by a noticeable increase in labeling difficulty, as the Query Policy started to request the class of more complicated samples, with increased fuzziness between Cashew and other classes (or suspicion about it).

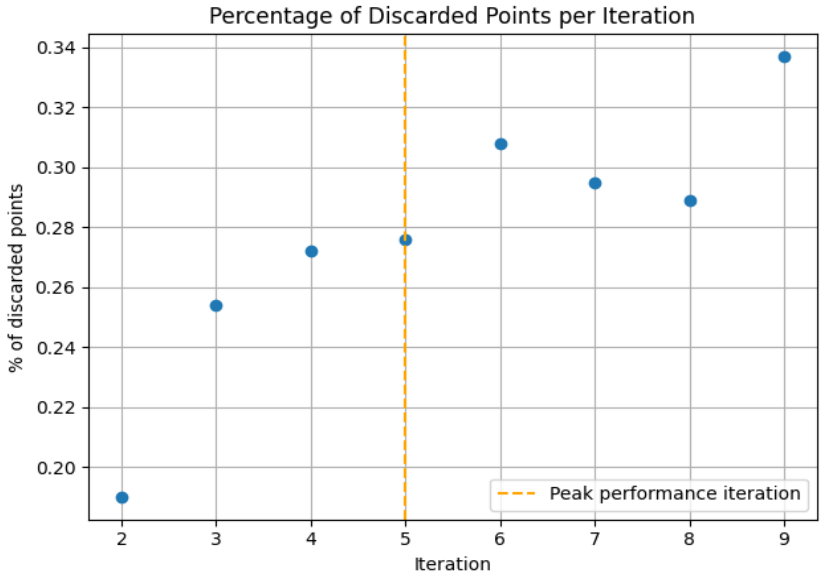


Figure 5.15: Percentage of points for each MS-AL iteration that are discarded by the Oracle during the annotation stage.

Therefore, the MS experiments finished with a Balanced Accuracy of 92.2% and F1-score of 85.7% after 9 iterations, with 1909 samples. Meanwhile, RS finalizes with 90.9% and 85.0% Balanced Accuracy and F1-score, respectively, using 2274 training samples. This finally shows the power of Active Learning in the context of multi-spectral image classification: it allows for a more efficient training sample selection, which can not only improve a classifier's maximum predictive performance but also lead to better results with less labeled data. This can reduce the amount of resources needed for the labeling task and optimize this section of supervised learning. However, an important conclusion of the experiments made during the dissertation is about the difficulty of finding an adequate test set to translate the Active Learning gains, as if we had kept evaluating the AL experiments on the RSeT test set, the verdict about the applicability of this field of research would likely be the complete opposite.

5.4.2 Post-processing and final 2021 Land cover map

With the final classifier from the MS results, we obtained a new land cover map, similar to the one made with the initial RSeT model (Figure 5.3), only now as a two-class Cashew and Non-Cashew version. The mapping with the predictions for every pixel in the ROI was done and stored in a GeoTIFF file. Most post-processing options have to do with smoothing the predictions by removing noise: it is natural for many dispersed pixels or small raster polygons to correspond to a different class than that of their neighboring regions, but this likely corresponds to noise rather than true regions of that predicted land cover type. In this scenario, smoothing the model land type predictions helps increase the land cover map's accuracy. In this dissertation, we used a Sieve filter, which is implemented in the Python *rasterio.features* module. This algorithm finds contiguous polygons of the same class and removes the raster polygons that are smaller

than a user-defined threshold p, which are then replaced by the class of the largest neighboring polygon [14]. As this is a user-defined threshold, we greedily searched for the optimal value for our use case. Different parameter values were evaluated, from 0 to 2000 with steps of 50. For each of them, a Sieve filter was applied, and the classes of the RS test cases were extracted to evaluate performance. Balanced Accuracy and Cashew F1 results are shown in Figure 5.16. For both metrics, the best value appears at $p = 350$: it corresponds to a Balanced Accuracy of 94.0%, and F1-score equal to 89.5%. This is a clear improvement over the unfiltered results obtained right at the end of section 5.4.1.

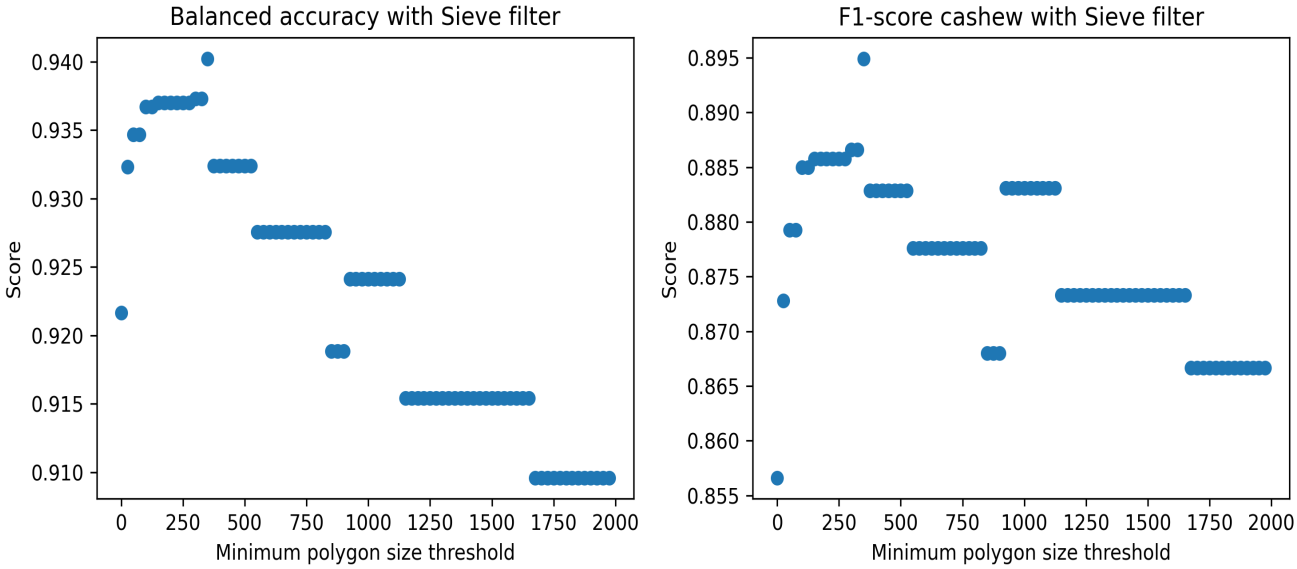


Figure 5.16: Test scores obtained after applying a Sieve filter with different polygon threshold size p parameter values.

In Figure 5.17, we present the final 2021 Land cover map for cashew monitoring, based on SVM predictions and processed with a $350p$ Sieve filter. Looking at pixel percentages, there is a total of 25.6% classified as Cashew. The Northwestern quadrant of Guinea-Bissau is the region with the most detected cashew land area, followed by the Southwestern ROI quadrant and the Northeastern one. On the lower right part of the region, cashew plantations are still scarce. Figure 5.18 shows this final filtered map with an overlay of Guinea-Bissau communities' locations, obtained from the 2017 census data (kindly shared by RSeT researcher Maria Vasconcelos). It is important to note the association between the cashew sector and human activity: all the larger plantations are noticeably overlaid with a higher community density, especially in the Northern part of the country. The lower right fraction does not present many cashew orchards and coincides with a region with fewer registered human communities. The information that can be extracted from the two maps can be used to tackle the overreliance on the cashew sector. It serves as a basis to determine the areas with a more noticeable predominance of cashew orchards (useful to estimate production values), and zones with a stronger cashew-human interaction, which can hopefully assist in the implementation of new measures that protect the region's native land life and strengthen Guinea-Bissau's economic structure.

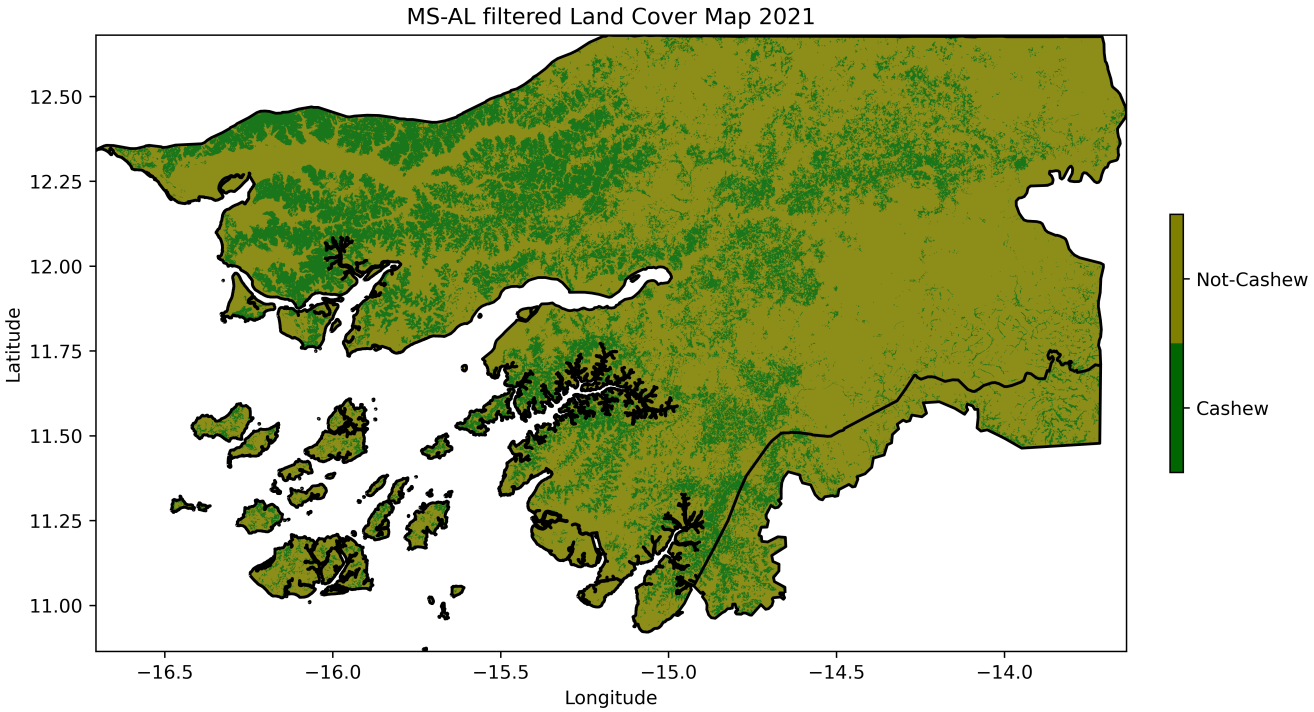


Figure 5.17: Final 2021 Land cover map. Obtained with a post-processing $350p$ Sieve Filter applied to the SVM predictions map.

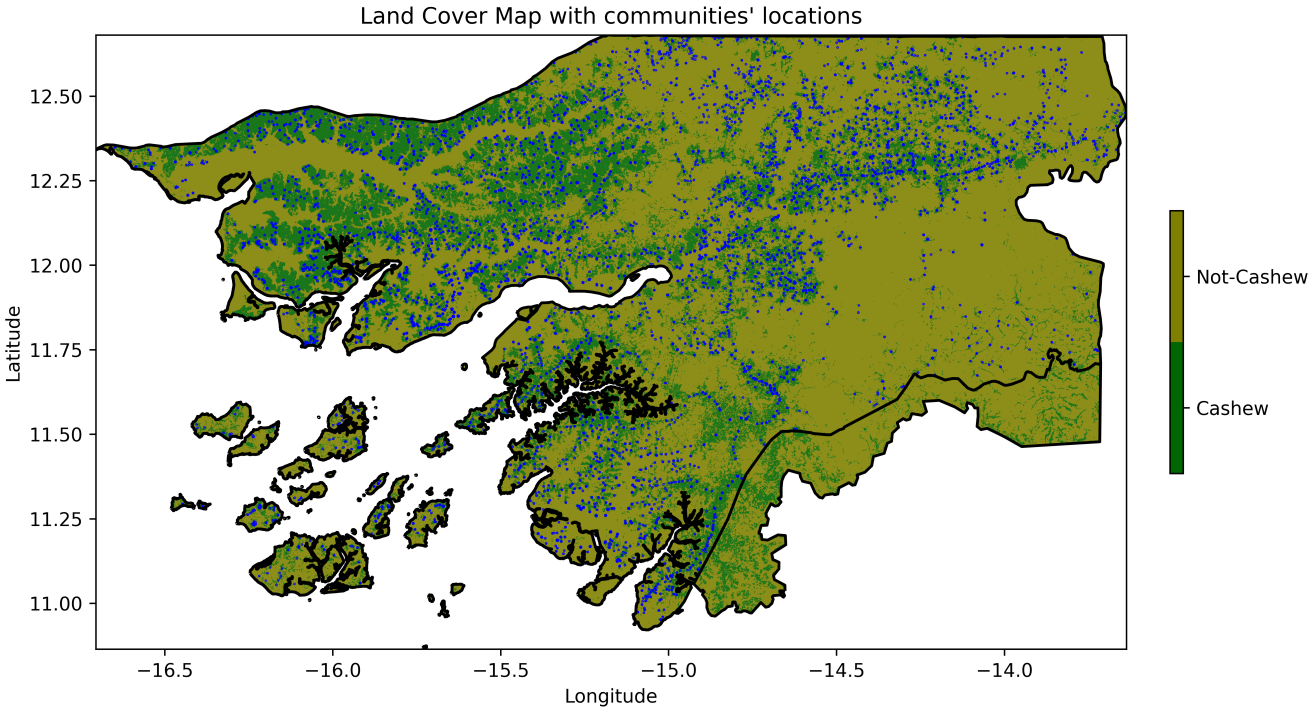


Figure 5.18: Final 2021 Land cover map (Figure 5.17) with an overlay of Guinea-Bissau communities according to the 2017 census, shared by RSeT [51].

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this dissertation, we proposed a remote and cost-effective solution for the task of monitoring cashew orchards, based on Machine Learning classifiers trained with Remote Sensing data. We have also explored how the field of Active Learning can help build a training set that is as informative as possible with little data acquisition effort, following labeling budget constraints. The approach was implemented for images of 2021, with the study being performed over the entire country of Guinea-Bissau and a part of the Republic of Guinea, up to the Kogon river.

A raster image covering the Region of Interest (ROI) was produced, containing bands from Sentinel-2's surface reflectance median composite of April 2021, temporal coefficients from the Continuous Change Detection and Classification (CCDC) algorithm, as well as Gray-Level Co-Occurrence Matrix (GLCM)'s spatial contextual statistics (the latter were only used for the Random Sampling and Margin Sampling datasets). This resulted in an image with 360 bands. They were reduced to 67 during a feature selection analysis, after not noticing any performance losses relative to maintaining every feature or applying a different (and broader) selection procedure.

An initial dataset shared by RSeT showed the spectral overlap between Cashew and other classes, mainly Open Forest. It was observed that this overlap is the smallest in April, at the end of the dry season. We ended up concluding that this dataset was not able to build a classifier capable of predicting across the entirety of the Region of Interest (ROI): it does not provide an adequate test set since it is not representative of the entire region's behaviors, and showed little benefits when used as the initial training dataset of a regionwide Active Learning pipeline.

On the topic of our first proposed goal — to create a dataset that is sufficiently broad and representative of the whole ROI —, a dataset composed of 4498 labeled points was obtained through Random Sampling. The labeled samples were roughly evenly distributed across the entire region and present class proportions that are likely closer to the real land cover percentages. This led us to form a test set exclusively with randomly sampled data that was used for the

remainder of the project, including the final Active Learning experiments.

A Support Vector Classifier was built for the binary classification task of distinguishing Cashew from Non-Cashew. Margin Sampling was used as the selection heuristic, following the second goal of the dissertation: to explore Active Learning techniques, their capacity to efficiently select points to label, and their subsequent performance. Comparing MS and RS results, it was observed that MS allows for quicker Balanced Accuracy and Cashew F1-score improvements using fewer samples, which is important when there are tight labeling restrictions. It also resulted in the highest predictive performance on the test set: 92.9% Balanced Accuracy and 87.2% Cashew F1-score, using 1254 training samples. The results showed the usefulness of this field for the multi-spectral image classification task, integrated into a fully remote methodology. They showed how it can reduce the amount of resources needed for the labeling task and optimize this section of the supervised learning pipeline.

With the final MS results, we used a $350p$ Sieve filter to smooth the predictions' land cover map from noise, following the third proposed objective: to produce a land cover map based on Machine Learning predictions for the year with ground truth data (2021). This post-processing improved the classification results, obtaining a final Balanced Accuracy of 94.0% and an F1-score of 89.5%. Using this, we estimated that approximately 25% of land in 2021 is likely to be cashew. We can also detect the regions where orchards are more prominent, and by overlaying the map with Guinea-Bissau communities' data from the 2017 census, we directly observed the relation between human activity and the cashew sector.

This research proves the potential of a fully remote and automated cashew detection methodology. Contributions can be extracted to figure out the state of this sector in Guinea-Bissau and make plans accordingly, following the country's intended environmentally sustainable development.

6.2 Future Work

It is possible to expand the approach followed during this dissertation in many areas, as part of future research. Firstly, the usage of base satellite data with higher spatial and temporal resolution (for example, the PlanetScope imagery [61]) could improve the detection of such a specific category as Cashew.

Other ML algorithms — particularly patch-based classifiers — can be considered now that 3×3 patch-based data from the RS and MS datasets is available. Also, different AL heuristics are worth exploring: a representativeness-based AL algorithm like Core Set [40] might be able to select an informative training set without focusing on points that are by nature not worth labeling, for example due to fuzziness — something that happened with Margin Sampling.

Meanwhile, updating the land cover map to more recent years would help understand how the evolution of cashew coverage is progressing. This could be done by expanding the training set with samples from different years, either following Domain Adaptation techniques [62], or just

collecting labels from every class in multiple years, which would likely result in a classifier that is even more generalizable and has accurate predictions during a larger timespan.

Despite following a fully remote approach during the dissertation, the addition of an on-site step could improve the capabilities of the whole process by tackling the remote approach's limitations. Similar to the orienteering problem [63], one could use the classifier's current most uncertain regions to define a score for each point in the ROI and plan a field mission to visit a set of the most informative vertices that are difficult to assess remotely.

Lastly, a deeper analysis of the cashew sector and human interaction could be made, as well as a topographic analysis to understand the climate and terrain conditions of the regions most predominantly covered by cashew orchards.

Bibliography

[1] Republic of Guinea-Bissau. *Updated Nationally Determined Contribution In The Framework Of The Paris Climate Agreement*. Oct. 2021. URL: https://www.undp.org/guinea-bissau/publications/updated-nationally-determined-contribution-framework-paris-climate-agreement (cit. on pp. 1–2, 4).

[2] *Guinea-Bissau Worldometer*. 2024. URL: https://www.worldometers.info/world-population/guinea-bissau-population/ (cit. on p. 1).

[3] World Food Programme Executive Board. *Guinea-Bissau country strategic plan (2023–2027)*. Oct. 2022. URL: https://executiveboard.wfp.org/document_download/WFP-0000142930 (cit. on pp. 1, 4–5, 7).

[4] Yekeminan Kone. *Improving cashew cultivation: FAO supports research activities in Guinea-Bissau*. May 2022. URL: https://www.fao.org/africa/news-stories/news-detail/Improving-cashew-cultivation-FAO-supports-research-activities-in-Guinea-Bissau/en (cit. on p. 1).

[5] European Space Agency. *Sentinel-2, Colour vision for Copernicus*. URL: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2 (cit. on pp. 2, 8).

[6] Renata Micha. *2021 Global Nutrition Report*. 2021. URL: https://globalnutritionreport.org/reports/2021-global-nutrition-report/foreword/ (cit. on p. 4).

[7] Abdulai Ismail Seca, Henrique dos Santos Pereira, and Suzy Cristina Predroza da Silva. "Expansion of Cashew Cultivation and its Environmental and Economic Impacts on Guinea-Bissau". In: *Brazilian Journal of Environmental Sciences (RBCIAMB)* 56.3 (2021), pp. 385–397 (cit. on pp. 4–5).

[8] Ronaldo Seroa da Motta. *Estimativa do Custo Econômico do desmatamento na amazônia*. Sept. 2002. URL: https://repositorio.ipea.gov.br/handle/11058/2800 (cit. on p. 4).

[9] Philip Havik et al. "Agro-Economic Transitions in Guinea-Bissau (West Africa): Historical Trends and Current Insights". In: *Sustainability* 10 (Sept. 2018), p. 3408. DOI: 10.3390/su10103408 (cit. on p. 5).

[10] Djassi Ensa. "Cashew Nuts Production in Guinea-Bissau". In: (2022). DOI: 10.52270/27132447_2022_12_3 (cit. on pp. 5–6).

[11] Luís Palma et al. "African Forest elephants persist in Guinea-Bissau but require an emergency conservation plan". In: *Oryx* 58 (2024) (cit. on pp. 5, 7).

[12] Sofia C. Pereira, Catarina Lopes, and João Pedro Pedroso. “Mapping Cashew Orchards in Cantanhez National Park (Guinea-Bissau)”. In: *Remote Sensing Applications: Society and Environment* 26 (2022), p. 100746. ISSN: 2352-9385. DOI: https://doi.org/10.1016/j.rsase.2022.100746. URL: https://www.sciencedirect.com/science/article/pii/S2352938522000544 (cit. on pp. 5, 10, 19–20, 30).

[13] Luís Catarino, Yusufo Menezes, and Raul Sardinha. *Cashew cultivation in Guinea-Bissau – risks and challenges of the success of a cash crop*. Oct. 2015. URL: https://www.scielo.br/j/sa/a/VmksNnMRVxYLGVSWd47C9HP/?lang=en# (cit. on p. 5).

[14] Sofia Perestrelo de Vasconcelos Cardoso Pereira. “Remote Sensing and Machine Learning Tools for Vegetation Monitoring”. In: *Repositório aberto da Universidade do Porto* (2020) (cit. on pp. 5, 7, 11, 18–20, 26, 28, 30, 35–36, 51, 55).

[15] UNFCC. *What is the United Nations Framework Convention on Climate Change?* URL: https://unfccc.int/process-and-meetings/what-is-the-united-nations-framework-convention-on-climate-change (cit. on p. 6).

[16] SGD Transformation Center. *Status of SDG targets for Guinea-Bissau*. URL: https://dashboards.sdgindex.org/profiles/guinea-bissau (cit. on p. 7).

[17] Paweł Hawryło and Piotr Wezyk. “Predicting Growing Stock Volume of Scots Pine Stands Using Sentinel-2 Satellite Imagery and Airborne Image-Derived Point Clouds”. In: *Forests* 9 (May 2018), p. 274. DOI: 10.3390/f9050274 (cit. on p. 9).

[18] NASA. *Landsat Science*. URL: https://landsat.gsfc.nasa.gov/ (cit. on p. 9).

[19] Nitidae. *A pilot study of remote sensing by drone and satellite to strengthen the cashew sector in Mozambique*. 2022. URL: https://www.nitidae.org/en/actualites/une-etude-pilote-de-teledetection-par-drone-et-satellite-pour-renforcer-la-filiere-anacardier-au-mozambique (visited on 04/22/2024) (cit. on p. 10).

[20] Sadouanouan Malo et al. “Cashew Trees Detection And Yield Analysis Using UAV-Based Map”. In: *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE. 2021, pp. 1–5 (cit. on pp. 10, 20–21).

[21] Thierry Roger Bayala, Sadouanouan Malo, and Zakaria Kinda. “DL4Cashew: Deep Learning Model For Cashew Trees Counting”. In: *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*. Vol. 1. IEEE. 2023, pp. 1–7 (cit. on pp. 10, 17, 19–20).

[22] Huete Alfredo, Didan Kamel, and Yuan Yin. In: *Terrestrial Biophysics and Remote Sensing Lab (TBRS)* (). URL: https://web.archive.org/web/20060902144430/http://tbrs.arizona.edu/cdrom/VI_Intro/VI_MOD_VI.html (cit. on p. 11).

[23] Zhe Zhu and Curtis E Woodcock. “Continuous change detection and classification of land cover using all available Landsat data”. In: *Remote sensing of Environment* 144 (2014), pp. 152–171 (cit. on p. 12).

[24] *Continuous Change Detection and Classification (CCDC)*. URL: https://www.openmrv.org/web/guest/w/modules/mrv/modules_2/continuous-change-detection-and-classification-ccdc (cit. on p. 12).

[25] Jiechen Tang et al. "Exploiting Superpixel-Based Contextual Information on Active Learning for High Spatial Resolution Remote Sensing Image Classification". In: *Remote Sensing* 15.3 (2023), p. 715 (cit. on pp. 12, 14, 17).

[26] Xiangyong Cao et al. "Hyperspectral image classification with convolutional neural network and active learning". In: *IEEE Transactions on Geoscience and Remote Sensing* 58.7 (2020), pp. 4604–4616 (cit. on pp. 12, 17, 19).

[27] Girish Ajmera. *Feature extraction of images using GLCM (gray level cooccurrence matrix)*. Feb. 2024. URL: https://medium.com/@girishajmera/feature-extraction-of-images-using-glcm-gray-level-cooccurrence-matrix-e4bda8729498 (cit. on p. 13).

[28] *ee.Image.glcmTexture*. URL: https://developers.google.com/earth-engine/apidocs/ee-image-glcmtexture (cit. on p. 13).

[29] Claudio Persello and Lorenzo Bruzzone. "Active learning for domain adaptation in the supervised classification of remote sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* 50.11 (2012), pp. 4468–4483 (cit. on p. 14).

[30] Janez Kranjc et al. "Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform". In: *Information Processing & Management* 51.2 (2015), pp. 187–203 (cit. on p. 14).

[31] Kashyap Raiyani, Teresa Gonçalves, and Luís Rato. "Abbreviating Labelling Cost for Sentinel-2 Image Scene Classification Through Active Learning". In: *Iberian Conference on Pattern Recognition and Image Analysis*. Springer. 2022, pp. 295–308 (cit. on pp. 14, 17).

[32] Zhao Lei et al. "Active deep learning for hyperspectral image classification with uncertainty learning". In: *IEEE geoscience and remote sensing letters* 19 (2021), pp. 1–5 (cit. on pp. 14, 17, 19).

[33] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning". In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 14).

[34] Huifang Huang et al. "Hyperspectral image classification via active learning and broad learning system". In: *Applied Intelligence* 53.12 (2023), pp. 15683–15694 (cit. on pp. 14, 17, 19).

[35] Chenying Liu, Jun Li, and Lin He. "Superpixel-Based Semisupervised Active Learning for Hyperspectral Image Classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.1 (2019), pp. 357–370. DOI: 10.1109/JSTARS.2018.2880562 (cit. on pp. 14, 17).

[36] Jan Kremer, Kim Steenstrup Pedersen, and Christian Igel. "Active learning with support vector machines". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4.4 (2014), pp. 313–326 (cit. on p. 15).

[37] Devis Tuia et al. “A survey of active learning algorithms for supervised remote sensing image classification”. In: *IEEE Journal of Selected Topics in Signal Processing* 5.3 (2011), pp. 606–617 (cit. on p. 15).

[38] Chandra Prakash Bathula. *Machine Learning Concept 68: Platt’s Scaling*. July 2024. URL: https://medium.com/@chandu.bathula16/machine-learning-concept-68-platts-scaling-b8245421739e (cit. on p. 16).

[39] Shasvat Desai and Debasmita Ghose. “Active learning for improved semi-supervised semantic segmentation in satellite images”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2022, pp. 553–563 (cit. on p. 16).

[40] Romain Thoreau et al. “Active learning for hyperspectral image classification: A comparative review”. In: *IEEE geoscience and remote sensing magazine* 10.3 (2022), pp. 256–278 (cit. on pp. 16, 32, 58).

[41] Usha Patel, Hardik Dave, and Vibha Patel. “Hyperspectral image classification using uncertainty and diversity based active learning”. In: *Scalable Computing: Practice and Experience* 22.3 (2021), pp. 283–293 (cit. on pp. 16, 19).

[42] Chen Ding et al. “Hyperspectral image classification promotion using clustering inspired active learning”. In: *Remote Sensing* 14.3 (2022), p. 596 (cit. on p. 16).

[43] Ozan Sener and Silvio Savarese. “Active learning for convolutional neural networks: A core-set approach”. In: *arXiv preprint arXiv:1708.00489* (2017) (cit. on p. 16).

[44] Dawa Derksen, Jordi Inglada, and Julien Michel. “Geometry aware evaluation of handcrafted superpixel-based features and convolutional neural networks for land cover mapping using satellite imagery”. In: *Remote Sensing* 12.3 (2020), p. 513 (cit. on pp. 17–18, 21).

[45] Renlong Hang et al. “Hyperspectral image classification with attention-aided CNNs”. In: *IEEE Transactions on Geoscience and Remote Sensing* 59.3 (2020), pp. 2281–2293 (cit. on p. 17).

[46] Zhenshen Qu et al. “Deep active learning for remote sensing object detection”. In: *arXiv preprint arXiv:2003.08793* (2020) (cit. on p. 17).

[47] João E Batista. “Evolving robust and readable land cover models for improved forest monitoring”. In: (2024) (cit. on p. 18).

[48] Google for Developers. *ee.Algorithms.Sentinel2.CDI*. URL: https://developers.google.com/earth-engine/apidocs/ee-algorithms-sentinel2-cdi (cit. on p. 19).

[49] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. “Variational adversarial active learning”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 5972–5981 (cit. on p. 19).

[50] Andrea Tassi and Marco Vizzari. “Object-oriented lulc classification in google earth engine combining snic, glcm, and machine learning algorithms”. In: *Remote Sensing* 12.22 (2020), p. 3776 (cit. on p. 19).

[51] RSet. *NGO for international cooperation in renewable natural resource management and adaptation to climate change*. URL: https://www.rset.eu/ (cit. on pp. 19, 24, 56).

[52] Leikun Yin et al. "Mapping smallholder cashew plantations to inform sustainable tree crop expansion in Benin". In: *Remote Sensing of Environment* 295 (2023), p. 113695 (cit. on p. 19).

[53] João E Batista et al. "Optical time series for the separation of land cover types with similar spectral signatures: cocoa agroforest and forest". In: *International Journal of Remote Sensing* 43.9 (2022), pp. 3298–3319 (cit. on p. 20).

[54] Qiusheng Wu. *geemap*. URL: https://geemap.org/ (cit. on p. 22).

[55] *Hyperopt: Distributed Asynchronous Hyper-parameter Optimization*. URL: https://hyperopt.github.io/hyperopt/ (cit. on p. 23).

[56] *CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos*. URL: https://www.biopolis.pt/en/cibio/ (cit. on p. 23).

[57] Luís Palma, Raquel Godinho, and Pedro Beja. *Updating the conservation status of the forest elephant Loxodonta cyclotis in Guinea-Bissau*. 2023 (cit. on p. 23).

[58] RSeT (ONGD) and Instituto Superior de Agronomia da Universidade de Lisboa. *FREL methodology report for scaling up CBADP to national level*. Oct. 2022 (cit. on p. 24).

[59] Paulo Arevalo. *gee-CCDC-tools: Tools and Earth engine apps to interact with the outputs from the CCDC algorithm*. URL: https://github.com/parevalo/gee-ccdc-tools (cit. on pp. 28–29).

[60] Daniel Moraes. *Detecting change with CCDC on Google Earth engine*. Mar. 2022. URL: https://medium.com/@moraesd90/detecting-change-with-ccdc-on-google-earth-engine-7125bfc1a6a (cit. on p. 28).

[61] ESA. *PlanetScope Full Archive*. 2024. URL: https://earth.esa.int/eogateway/catalog/planetscope-full-archive (cit. on p. 58).

[62] Jiangtao Peng et al. "Domain adaptation in remote sensing image classification: A survey". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), pp. 9842–9859 (cit. on p. 58).

[63] João Pedro Pedroso, Alpar Vajk Kramer, and Ke Zhang. "The sea exploration problem: Data-driven orienteering on a continuous surface". In: *arXiv preprint arXiv:1802.01482* (2018) (cit. on p. 59).

Appendix

Supplementary Figures

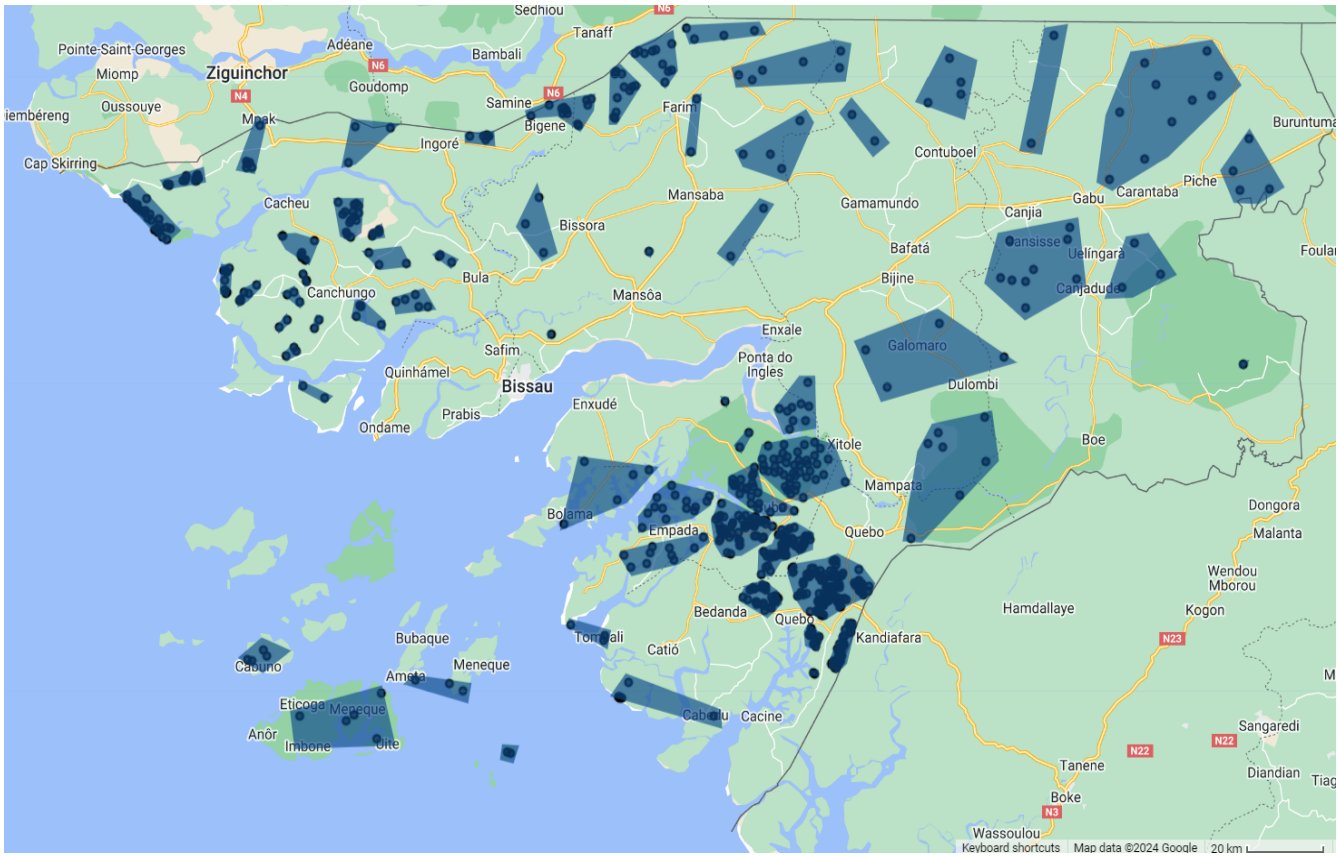


Figure 1: Representation of the manually obtained polygons, for the class - Savanna. Image taken using Google Earth Engine.

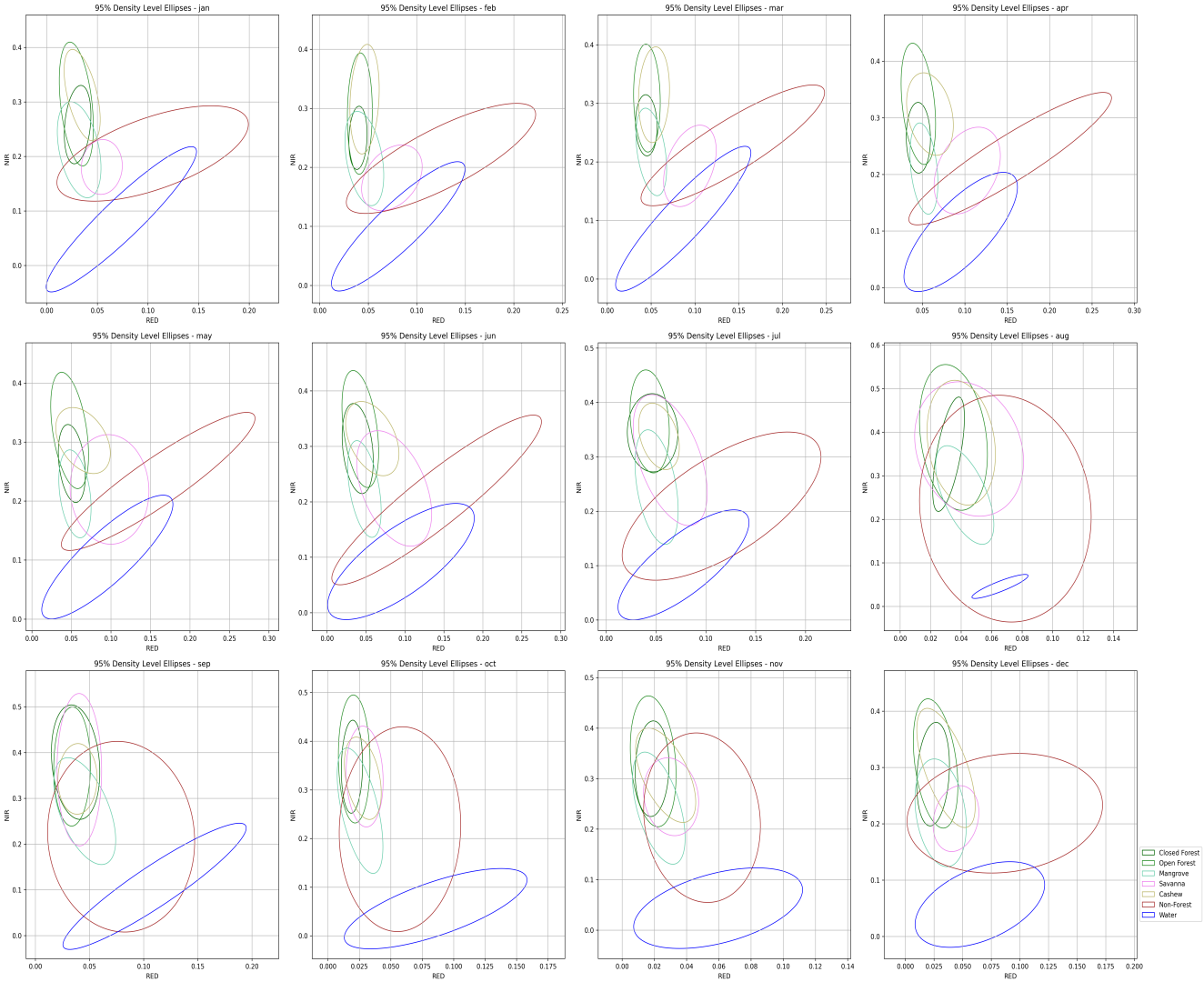


Figure 2: 95% density level ellipses for each RSeT dataset class in the RED-NIR bidimensional representation space, for every 2021 monthly median composite.

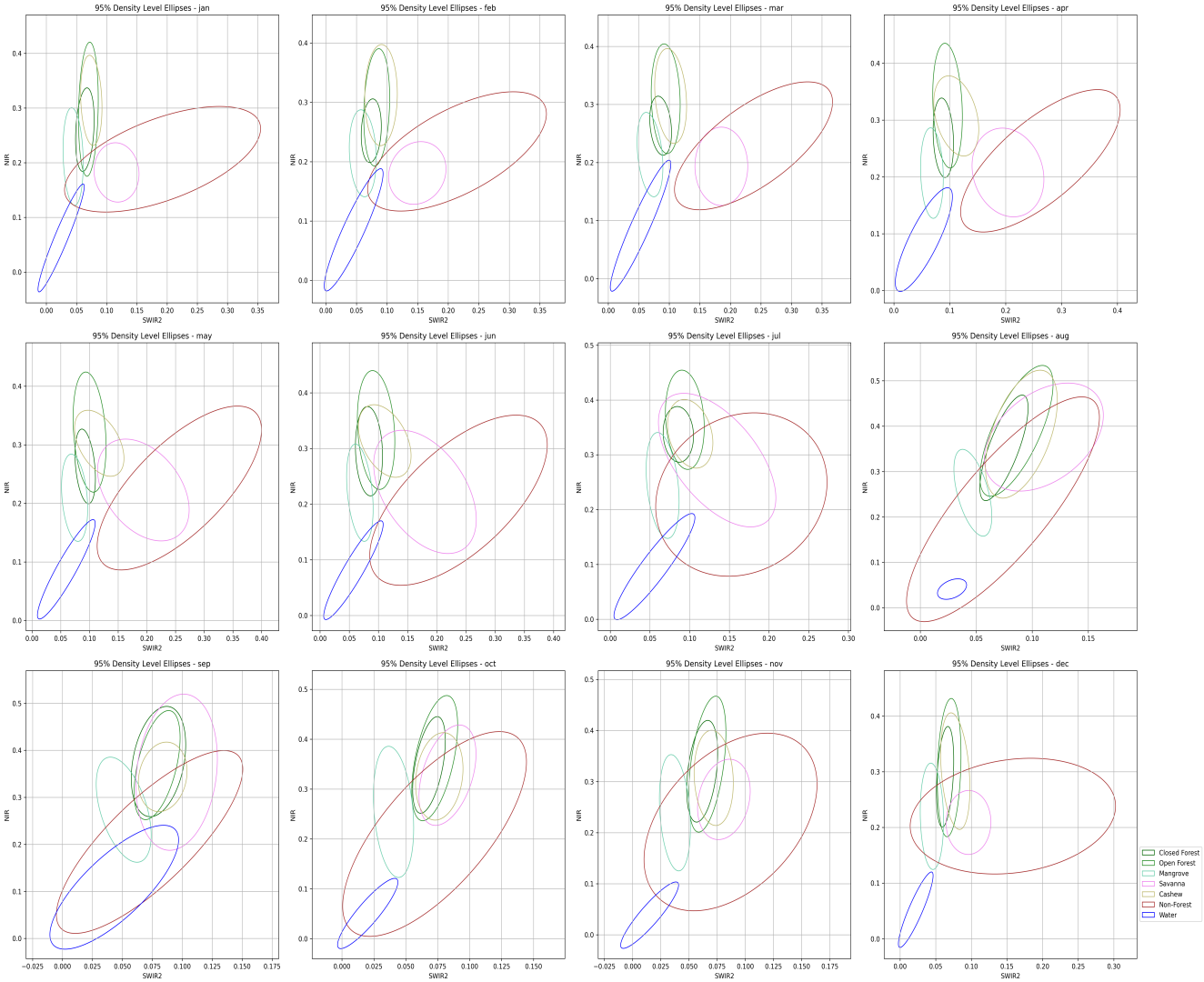


Figure 3: 95% density level ellipses for each RSeT dataset class in the SWIR2-NIR bidimensional representation space, for every 2021 monthly median composite.

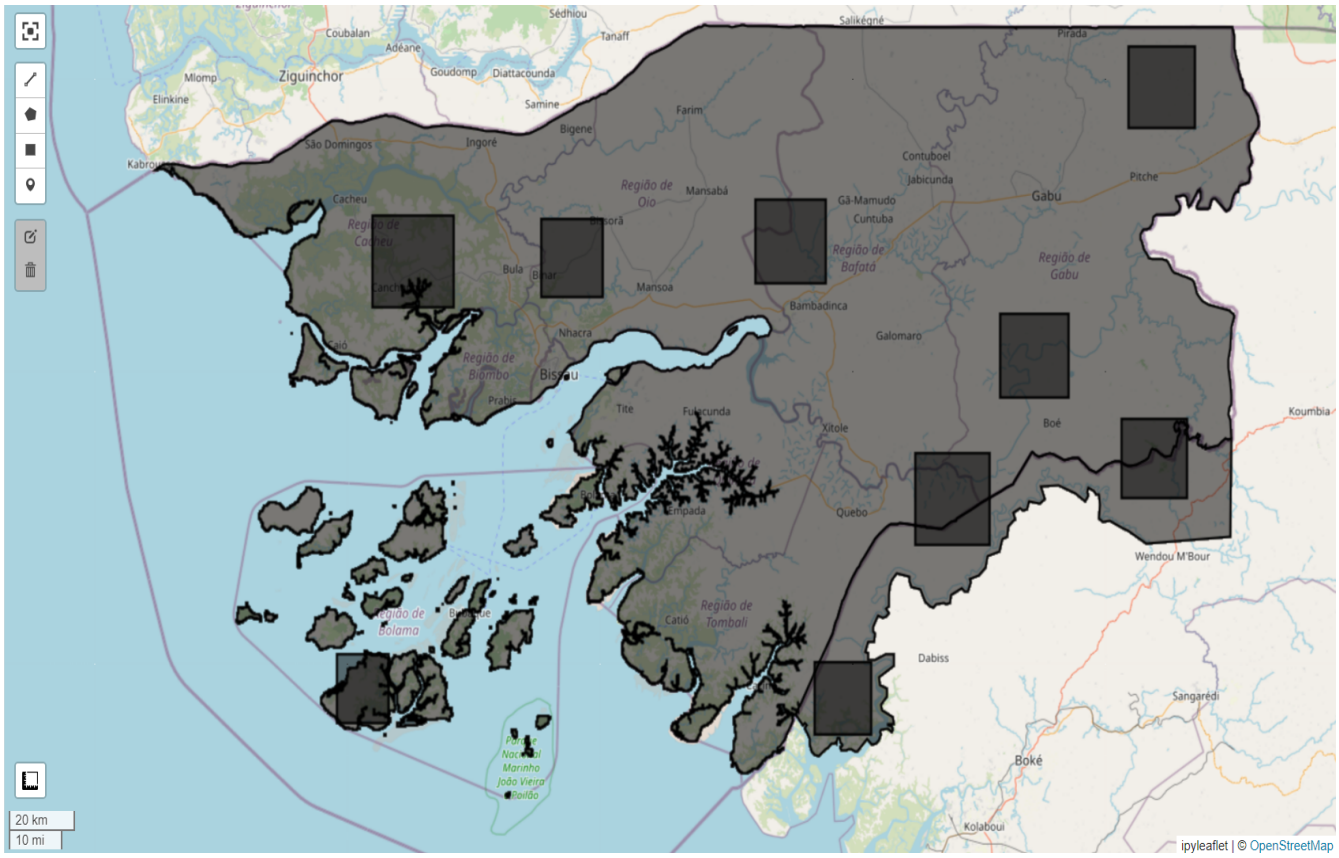


Figure 4: Location of the 9 sub-regions (rectangular areas) used for the 1^{st} AL configuration. Image taken using Google Earth Engine.