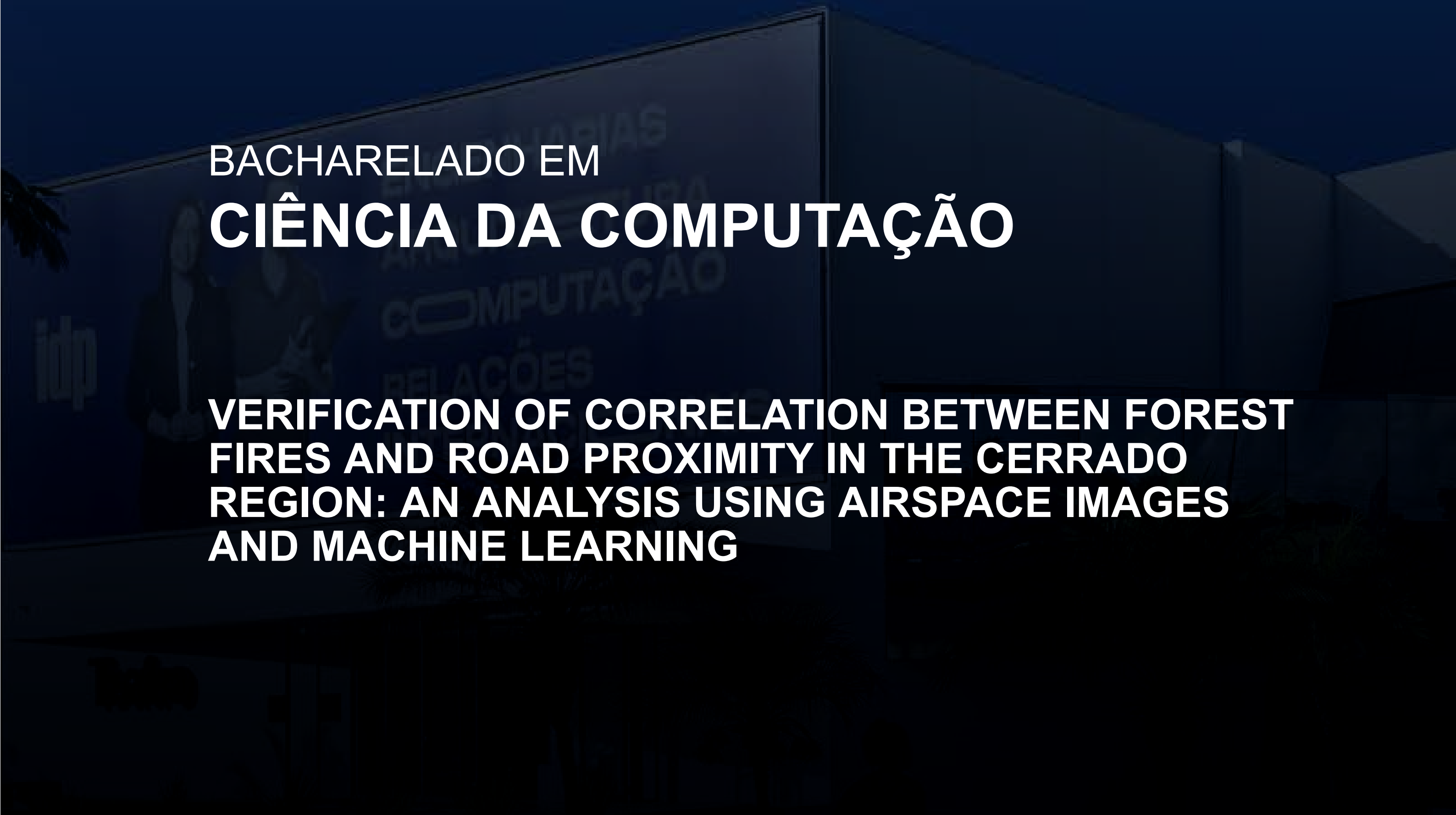


BACHARELADO EM

CIÊNCIA DA COMPUTAÇÃO

VERIFICATION OF CORRELATION BETWEEN FOREST FIRES AND ROAD PROXIMITY IN THE CERRADO REGION: AN ANALYSIS USING AIRSPACE IMAGES AND MACHINE LEARNING

KELVIN GOMES PIMENTEL RODRIGO DOS SANTOS

Brasília - DF, 2025

KELVIN GOMES PIMENTEL RODRIGO DOS SANTOS

VERIFICATION OF CORRELATION BETWEEN FOREST FIRES AND ROAD PROXIMITY IN THE CERRADO REGION: AN ANALYSIS USING AIRSPACE IMAGES AND MACHINE LEARNING

Undergraduate Thesis submitted as a partial requirement for the award of the Bachelor’s degree in Ciência da Computação, at the Brazilian Institute of Teaching, Development and Research (IDP).

Advisor
Rafael Lemos Paes

Brasília - DF, 2025

Verification of correlation between forest fires and road proximity in the cerrado region: an analysis using aerospace images and machine learning

Código de catalogação na publicação – CIP

S237v Santos, Kelvin Gomes Pimentel Rodrigo dos

Verification of correlation between forest fires and road proximity in the cerrado region: an analysis using aerospace images and machine learning/ Kelvin Gomes Pimentel Rodrigo dos Santos. — Brasília: Instituto Brasileiro Ensino, Desenvolvimento e Pesquisa, 2025.

94 f. : il.

Orientador: Prof. Dr. Rafael Lemos Paes

Monografia (Graduação em Ciência da Computação) - Instituto Brasileiro Ensino, Desenvolvimento e Pesquisa – IDP, 2025.

1. Incêndio florestal. 2. Cerrado. 3. Acesso rodoviário. 4. Monitoramento por satélite. I. Título

CDD 006.42

Elaborada por Biblioteca Ministro Moreira Alves

KELVIN GOMES PIMENTEL RODRIGO DOS SANTOS

VERIFICATION OF CORRELATION BETWEEN FOREST FIRES AND ROAD PROXIMITY IN THE CERRADO REGION: AN ANALYSIS USING AIRSPACE IMAGES AND MACHINE LEARNING

Undergraduate Thesis submitted as a partial requirement for the award of the Bachelor's degree in Ciência da Computação, at the Brazilian Institute of Teaching, Development and Research (IDP).

Accepted 08/05/2025

Examination Committee

Documento assinado digitalmente
gov.br
RAFAEL LEMOS PAES
Data: 19/12/2025 21:28:54-0300
Verifique em https://validar.iti.gov.br

Rafael Lemos Paes- Advisor

Documento assinado digitalmente
gov.br
ULISSES SILVA GUIMARAES
Data: 19/12/2025 21:39:28-0300
Verifique em https://validar.iti.gov.br

Ulisses da Silva Guimarães- External Examiner

Documento assinado digitalmente
gov.br
MARINA JORGE DE MIRANDA
Data: 22/12/2025 09:20:50-0300
Verifique em https://validar.iti.gov.br

Marina Jorge de Miranda- External Examiner

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all the people and institutions who, in different ways, contributed to the realization of this work.

First and foremost, to my advisor, Rafael Lemos Paes, for his valuable guidance, patience, and constant support throughout the development of this research. His expertise and dedication were fundamental to the enhancement of this work.

I am grateful to CENSIPAM for the time made available, especially to Ulisses da Silva Guimarães, and for access to meetings that significantly enriched my understanding of the topic.

Special thanks to Pierre Duquesne and Daniel Santos, who kindly provided their time, trust, and access to the OroraTech platform, an essential instrument for data collection and analysis in this research.

To my parents, for their love, unconditional support, and encouragement throughout my entire academic and personal trajectory. Without the foundation and values they provided me, this work would not have been possible.

To my entire family, for their unwavering support, understanding, and encouragement during all the challenging moments of this academic journey. Their belief in my abilities and constant presence were crucial motivators that helped me overcome obstacles and maintain focus on achieving this goal.

To my girlfriend, Nayara Couto, for her unconditional encouragement, understanding during moments of exclusive dedication to this work, and emotional support throughout this journey.

Finally, to IDP Faculty, for the academic structure offered and for the environment conducive to intellectual and professional development.

To all of you, my deepest gratitude for your contribution to this academic journey.

ABSTRACT

The Cerrado biome, rich in biodiversity and naturally fire-dependent, has experienced a significant increase in wildfires driven by human activities, negatively impacting its ecosystems. This study investigates the relationship between wildfire occurrence and road proximity in the central region of the Cerrado, focusing on the states of Goiás and Tocantins, using Sentinel satellite imagery and fire hotspot data from OroraTech. Fire data were analyzed for the months of October and November, a transitional period between the dry season and the onset of rainfall, allowing the assessment of fire events even under climatic conditions less favorable to natural combustion.

The methodology included Principal Component Analysis (PCA) and ensemble learning techniques combining Random Forest, Decision Tree, Multilayer Perceptron (MLP), and K-Nearest Neighbors (KNN). The ensemble model achieved an accuracy of 84.41%, precision of 66.15%, recall of 45.26%, and an F1-score of 53.75%. The segmented analysis indicates that road infrastructure may be associated with different fire dynamics: on dry days, it is more closely related to fire spread, whereas on humid days it shows a stronger association with ignition patterns. The results suggest that, even under less favorable climatic conditions, fire hotspots remain concentrated near roads and within forested areas, indicating anthropogenic influence and potential relationships between road presence and land use and land cover changes. This study contributes to strategic monitoring and wildfire prevention efforts in the Cerrado region.

Keywords: Cerrado biome; Wildfires; Road infrastructure; Anthropogenic influence; Machine learning; Remote sensing.

RESUMO

O bioma Cerrado, rico em biodiversidade e naturalmente dependente do fogo, enfrenta um aumento significativo de incêndios florestais impulsionado por atividades humanas, causando impactos negativos sobre seus ecossistemas. Este estudo investiga a relação entre a ocorrência de incêndios florestais e a proximidade de estradas na região central do Cerrado, com foco nos estados de Goiás e Tocantins, utilizando imagens dos satélites Sentinel e dados de focos de calor da OroraTech. Os dados de incêndio foram analisados para os meses de outubro e novembro, período de transição entre a estação seca e o início das chuvas, permitindo a avaliação de eventos de fogo mesmo sob condições climáticas menos favoráveis à combustão natural.

A metodologia incluiu a Análise de Componentes Principais (PCA) e técnicas de aprendizado de máquina baseadas em ensemble learning, combinando Random Forest, Árvore de Decisão, Multilayer Perceptron (MLP) e K-Nearest Neighbors (KNN). O modelo ensemble apresentou desempenho de 84,41% em acurácia, 66,15% em precisão, 45,26% em recall e 53,75% em F1-score. A análise segmentada indica que a infraestrutura viária pode estar associada a diferentes dinâmicas dos incêndios: em dias secos, observa-se maior relação com a propagação do fogo, enquanto em dias úmidos há associação mais forte com padrões de ignição. Os resultados sugerem que, mesmo sob condições climáticas menos favoráveis, os focos de incêndio permanecem concentrados em áreas próximas às estradas e em regiões florestais, evidenciando a influência antrópica e possíveis relações entre a presença de estradas e alterações no uso e cobertura do solo. Este estudo contribui para o monitoramento estratégico e para ações de prevenção de incêndios florestais na região do Cerrado.

Palavras-chave: Satellite imagery, machine learning, forest fires, road infrastructure, Cerrado biome, remote sensing.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1

1 INTRODUCTION

Recognized as one of the most biodiverse savannas on the planet, the Cerrado is a biome (Figure 1) whose ecology has been shaped by the presence of fire for at least 4 million years. The native vegetation presents remarkable morphological and physiological adaptations that confer resilience to natural fires. However, the contemporary scenario is marked by a new dynamic: fires of anthropogenic origin. These events represent a severe negative impact on local ecosystems, as they differ from the natural fire regime by occurring at unfavorable times of the year, with greater intensity and duration, putting biodiversity at risk [23].

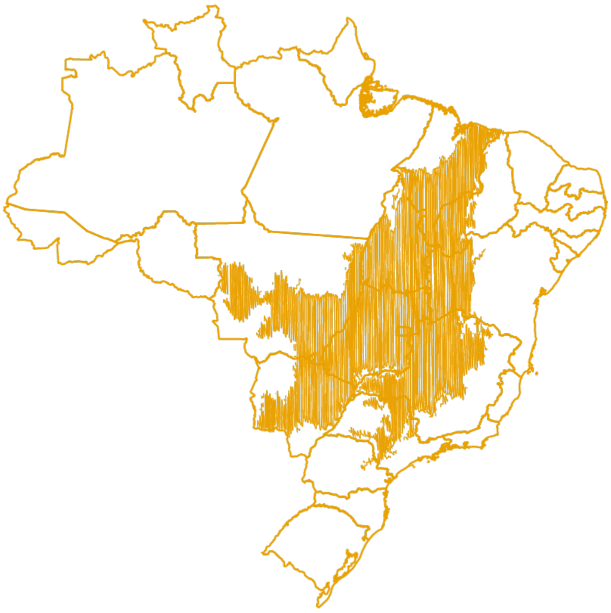


Figure 1: Location map of the Cerrado biome in Brazil. Source: [1]

Table 1: Approximate area of Brazilian biomes

Biome	**Approximate Area (km^2)**	**Proportion of National Territory**
Amazon	4,196,943 km^2	About 49.29%
Cerrado	2,036,448 km^2	About 22% to 24%
Atlantic Forest	1,107,419 km^2	About 29% of the original cover
Caatinga	844,453 km^2	About 11%
Pampa	178,243 km^2	2.07%
Pantanal	150,900 km^2	1.76%

Source: Adapted from [24, 25].

The analysis of Table 1, which outlines the Brazilian biomes, reveals a counterintuitive scenario regarding the impact of fire. Although the Amazon is the largest national biome, with more than twice the area of the Cerrado, it is the latter that leads in total burned area. The disparity is notable: data indicate that between 2003 and 2017, the Cerrado had 2,685,596 km² affected by fire, almost double the 1,382,264 km² recorded in the Amazon during the same period [25]. This seasonal dynamic, which reaches its peak in September [25], coincides with the dry months, when low humidity increases the frequency and intensity of fire in the biome [26].

The expansion of the agricultural frontier intensifies this pressure on the biome. The Cerrado, which covers approximately 23% of the national territory, has already had nearly half of its original vegetation converted into pastures, monocultures, and other anthropogenic uses, severely compromising its biodiversity [27]. Since the 1970s, the biome has become the main hub of food and commodity production in Brazil, driven by factors such as favorable climate, flat terrain, and low land costs, which have attracted farmers from other regions, especially from the South [28]. Vast areas were converted for soybean, corn, cotton, and sugarcane cultivation—soybean being particularly notable [29]—a practice that requires complete removal of native vegetation and intensive use of soil-correction inputs and pest control [30].

In this context of landscape transformation, fire emerges as a fundamental tool for land-use and land-cover change. It is systematically used in agricultural management, mainly in the burning of felled biomass to prepare land for agriculture and livestock [24]. This cultural practice of using fire to clear pastures and cropland may, however, result in uncontrolled fires that spread devastatingly [31]. Burning to eliminate woody material after deforestation and for pasture renewal is common and contributes to soil degradation [32].

The environmental impacts resulting from deforestation and fires trigger a cascade of damages. Habitat loss, fragmentation, and degradation profoundly alter population dynamics and the distribution of native species [33]. Fire not only consumes organic matter on the surface but also affects soil structure and vegetation, potentially leading to total destruction or compromising future development [26]. Combustion releases heat, nutrients, and various chemical by-products into the environment [34].

In addition to ecological damage, the smoke generated represents an increasing threat to global public health. Its emissions contain a complex mixture of pollutants, such as particulate matter, carbon monoxide, and nitrogen oxides. Inhaling these substances can cause oxidative stress, inflammatory responses, reduced lung capacity, and immune system suppression. Smoke plumes, transported over long distances, deteriorate air quality in regions far from the original source. The effects transcend physical health, with reports of mental disorders such as depression and post-traumatic stress disorder in affected populations. Reduced visibility caused by smoke also significantly increases the risk of traffic accidents [35].

Road infrastructure, particularly highways, plays a catalytic role in this cycle of deforestation and fire by facilitating the expansion of the agricultural frontier through access to new areas [30]. The presence of a dense road network is associated with a higher risk of ignitions caused by human activity [32]. The construction of new roads in forested areas frequently indicates ongoing logging activity and signals the intention to convert that land to other uses [33].

Investigating this spatial association between roads and fires can provide crucial support for planning prevention and mitigation strategies. By mapping ignition points near the road network, it is possible to delineate and visualize the most vulnerable segments [32]. The knowledge generated by this correlation allows for the efficient and prioritized allocation of monitoring and inspection efforts, concentrating resources in high-risk areas. Such an approach enhances fire management, since agility in detection and localization is a key factor for reducing damage and optimizing operational costs [32].

Agile and effective fire detection and monitoring are therefore essential to contain fire spread [23]. Continuous monitoring is indispensable for mitigating impacts on ecosystems and for generating data to support public policies for prevention and firefighting. In addition, mapping the location and extent of burned areas enables the creation of thematic maps that help identify high-risk zones and plan strategic actions [25].

Since the 1970s, remote sensing has consolidated itself as the main technology for monitoring natural resources. Currently, satellite-based thermal hotspot detection is the most widespread method. The National Institute for Space Research (INPE), for example, automatically processes more than 200 daily images from ten different satellites, using sensors in the mid-infrared band to identify burning vegetation points across Brazil [32]. Automation in fire detection, driven by modern technologies, emerges as an alternative to the limitations of traditional methods, such as watchtowers, which entail high costs, low efficiency, and risks to operators [25].

The effectiveness of these automated systems lies in their ability to intelligently interpret remote-sensing data. Through sophisticated algorithms, it is possible to go beyond simple heat detection, broadly analyzing the landscape to differentiate types of land cover and identify with high precision the transformations caused by fires or deforestation [23, 24, 33].

Given the clear connection between human expansion, roads, and fire in the Cerrado, it becomes essential to go beyond observation and provide concrete data to support decision-making. This study arises from the need to transform this perception into practical knowledge. The goal is to use data analysis technologies to map and quantify this influence, offering an accurate diagnosis to support more effective prevention policies and the intelligent management of one of the world's most threatened biomes.

1.1 OBJECTIVES

1.1.1 General Objective

To evaluate the influence of proximity to roads as a risk factor for the occurrence of wildfires in the Cerrado, using a method based on remote sensing data and machine learning to quantify this relationship.

1.1.2 Specific Objectives

- **OE(1)** Gather and preprocess geospatial databases related to hotspots or fires and the road network of the Cerrado biome for the study period;

- **OE(2)** Categorize fire events based on the Angström Index, segregating the data into different climatic risk scenarios to assess fire behavior under conditions of greater and lesser natural favorability to propagation;

- **OE(3)** Statistically analyze the correlation between road density and hotspot intensity, comparing the strength of this association under different meteorological conditions;

- **OE(4)** Develop a machine-learning classification model to evaluate the importance of the variable "proximity to roads" in predicting fire risk;

- **OE(5)** Investigate the distribution of burned areas through the association between road infrastructure and different types of land use and land cover (forest, agriculture, or pasture).

2

2 LITERATURE REVIEW

2.1 THE ROLE OF REMOTE SENSING AND ITS CHALLENGES

Remote sensing is a fundamental tool for wildfire monitoring and land-use change detection, enabling periodic surveys of large territorial extensions and rapid assessment of the damage caused by events such as forest fires [36]. It contributes to resource optimization and more effective information management [37].

In the context of wildfires, remote sensing allows the monitoring of fire dynamics and the observation of burned areas shortly after the event [38]. Rapid and accurate damage assessment is essential for developing appropriate mitigation measures [36]. At the same time, remote sensing is crucial for mapping road networks, which are essential infrastructure elements [39].

Moreover, it employs different types of data, such as optical (visible, near-infrared, and shortwave infrared) and thermal imagery [37]. However, this technology faces significant limitations, including cloud cover, which may hinder systematic image acquisition — although new satellite systems in orbit seek to mitigate this issue in tropical regions [39].

Detecting small objects or those with few attributes remains a challenge [40], especially under conditions with distortions or noise [41]. Therefore, preprocessing steps are necessary to enhance subtle features [40]. Complexity further increases when working with historical data, where detailed atmospheric information is rarely available for sophisticated correction models [42].

The mapping of roads and the identification of objects, especially small ones, present significant challenges due to the environmental complexity. Achieving accurate detection requires the continuous application of advanced techniques and the improvement of existing models [40]. Several factors contribute to the difficulty of this task:

The automatic detection of roads in satellite images faces various technical challenges that directly impact algorithm accuracy. One major obstacle lies in the **spectral and spatial similarity** among different anthropogenic structures, where the characteristics of buildings and urban roads often overlap, making their discrimination difficult [40].

Additionally, **occlusion and visual noise** compromise data integrity: vegetation, shadows cast by buildings, and moving elements such as vehicles generate false positives and interrupt the continuity of linear features [40]. These artifacts become particularly critical in dense urban areas, where scenario complexity demands robust filtering methods.

Variable environmental conditions also represent a significant limiting factor. Solar illumination—dependent on time of day, season, and atmospheric conditions—drastically alters the radiometric properties of surfaces [43]. This variability makes color-based attributes unreliable, requiring the use of illumination-invariant characteristics.

The **intrinsic quality of images** also imposes important constraints. Limitations in spatial resolution, geometric distortions, motion blur, and digital compression are exacerbated in dynamic acquisition conditions or with low-cost equipment [43]. These factors combine to produce data with considerable noise, demanding prior correction steps.

To overcome these obstacles—especially for extracting road networks from aerial images—image processing techniques are essential [44]. Image analysis and processing, particularly in domains such as remote sensing, rely on a series of enhancement methods. Resizing, normalization, and noise removal are crucial steps to optimize visual data quality for subsequent analyses, such as change detection or pattern recognition.

2.2 IMAGE PROCESSING TECHNIQUES

2.2.1 Normalization

Min-Max normalization is a technique used to rescale reflectance values of a band to a standard range, such as 0–255. It is commonly used as a preprocessing step in remote sensing imagery. The purpose of this normalization is to adjust the distribution of pixel values to a uniform range, facilitating comparison between different bands or images [37].

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \quad [37] \tag{1}$$

Where:

- x is the original pixel value;
- x_{min} is the smallest value in the band dataset;
- x_{max} is the largest value in the band dataset;
- x_{norm} is the resulting normalized value.

In summary, normalization is a set of techniques with the common goal of standardizing or correcting image data. The approaches differ in mathematical complexity and the factors they aim to correct, whether radiometric, illumination-related, or merely value scaling.

Beyond pixel-value standardization, another crucial preprocessing step is noise removal. This process aims to enhance the visual quality of the image and prepare it for subsequent analyses, such as feature detection or object recognition [45].

2.2.2 Gaussian Filter

Noise may arise from several factors, including capture conditions (illumination, motion) or sensor failures, manifesting as undesired variations that obscure important details [46]. Noise removal techniques seek to attenuate these variations while preserving essential image characteristics as much as possible. Among various approaches, the Gaussian filter is one of the most fundamental and widely used methods for this purpose [45].

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad \text{[45]} \tag{2}$$

Where:

- (x, y) represent the pixel coordinates relative to the center of the filter;
- σ is the standard deviation, a key parameter that controls the degree of smoothing. Larger σ values result in stronger blurring and more aggressive noise removal [47].

Essentially, the Gaussian filter acts as a low-pass filter, smoothing fine details and high-frequency variations typically associated with image noise.

2.2.3 Histogram

In the field of image processing, the histogram of an image is a fundamental graphical representation of the frequency distribution of pixel intensity values (gray levels or colors). For a discrete digital image, the histogram shows how often each intensity level occurs. This representation is crucial to understanding visual characteristics such as contrast, brightness, and the predominance of dark or bright tones [44].

A well-distributed histogram covering the entire dynamic intensity range generally indicates an image with good contrast. On the other hand, histograms concentrated in narrow regions—whether low values (dark images) or high values (bright images)—signal low contrast and potential loss of detail [44].

In the context of burned-area detection, images of burned forests exhibit low reflectance values in the near-infrared and high values in the shortwave infrared, while intact forests show low red-band values. Histogram analysis can reveal these differences [37].

2.2.4 Sobel Filter

The Sobel Filter is a fundamental tool for edge detection in images, estimating spatial gradient magnitude through changes in amplitude along borders [48]. The gradient magnitude S is defined by:

$$S = \sqrt{S_x^2 + S_y^2} \quad [48] \tag{3}$$

where S_x and S_y represent the gradients in the horizontal and vertical directions, respectively [48].

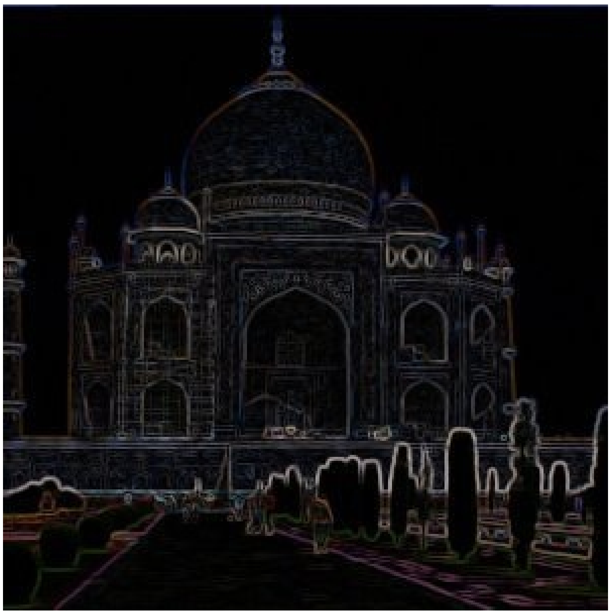


Figure 2: Application of the Sobel filter, highlighting image edges [2].

2.3 COLOR SPACES FOR IMAGE ANALYSIS

2.3.1 RGB

The RGB color model, which represents the additive primary colors Red, Green, and Blue, is widely used and fundamental for conversion into other color spaces [49].

Methods relying on visible bands (RGB) may struggle to distinguish burned areas from other landscape features, especially in urban areas, where pixel-value differences are subtle, and burned forests may not be clearly distinguishable [36].

The RGB model is highly sensitive to noise at low intensities due to its nonlinear transformation and has poor correlation with human color perception, as it does not separate luminance information. This can negatively impact detection in environments with illumination variations—a common challenge in satellite imagery [49].

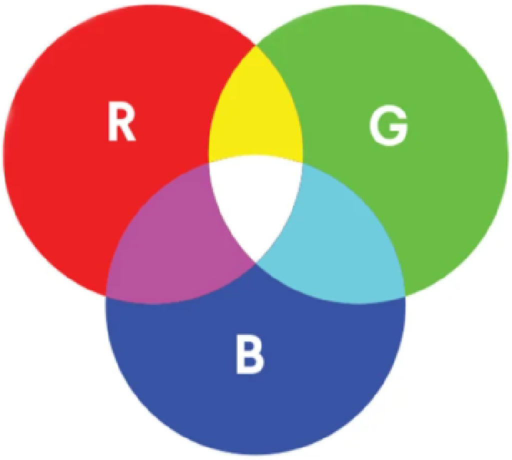


Figure 3: Representation of the RGB color space [3].

2.3.2 HSV

The HSV (Hue, Saturation, Value) color space is a nonlinear transformation of RGB and is described in a cylindrical coordinate system.

HSV is easily interpreted by humans because it aligns with how we perceive colors. It separates chromatic components (Hue and Saturation) from the achromatic component (Value), allowing color information to be treated independently from brightness [49].

The HSI/HSV transformation is popular in detection tasks, and one of its main advantages is having only two components (hue and saturation) closely related to human perception and more resistant to illumination changes [43].

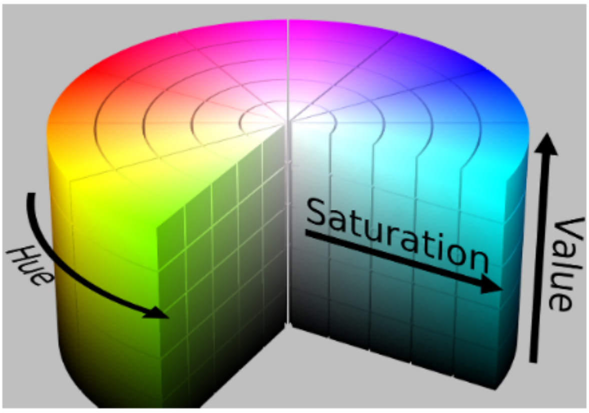


Figure 4: Representation of the HSV color space [4].

2.3.3 LAB

The Lab color space is a uniform color space, meaning numerical distances between colors correspond directly to perceived color differences. It represents luminance with values from 0 (black) to 100 (white). The components a* (red/green) and b* (blue/yellow) represent chromaticity [49].

Chromaticity differences can be computed using Euclidean distance, as hue variations are linear in this space. This makes it ideal for detecting small color differences. Moreover, Lab handles shadows and illumination variations more effectively [49], which is essential for real-world image analysis. The HSV space also provides good segmentation results, with higher accuracy than RGB [49].

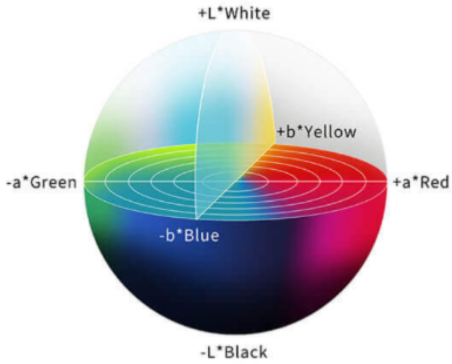


Figure 5: Representation of the Lab color space [5].

The formula for the three-dimensional Euclidean distance, used to measure color differences, is expressed in Equation 4:

$$d(p,q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2 + (z_p - z_q)^2} \quad [49] \tag{4}$$

2.3.4 NDVI — Normalized Difference Vegetation Index

The Normalized Difference Vegetation Index (NDVI) is one of the most widely used indicators in remote sensing for assessing vegetation health and vigor [50]. It is calculated from the reflectance of the Near-Infrared (NIR) and Red (RED) bands, producing values ranging from -1 to +1. Positive values close to 1 indicate dense and healthy vegetation, while values near zero correspond to bare soil or sparse vegetation. Negative values generally represent non-vegetated surfaces such as water or urban areas. Due to its simplicity and efficiency, NDVI is widely used in environmental studies, including vegetation damage assessment. Burned areas typically show a significant decrease in NIR and Red reflectance [51].

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad [51] \tag{5}$$

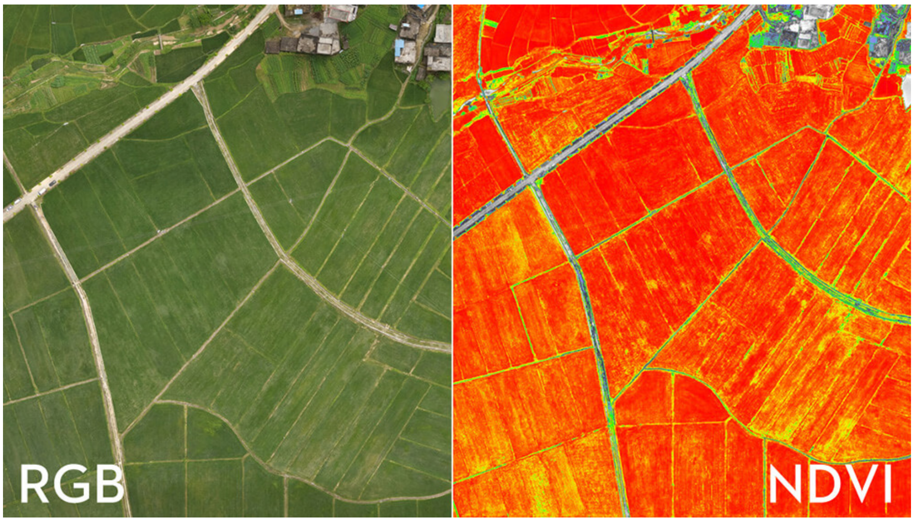


Figure 6: Comparison between the RGB image and the NDVI index [6].

Where *NIR* represents the near-infrared band and *Red* the red band. Its applications include forest and soil moisture monitoring, land-cover change detection, burned-area identification, and natural disaster analysis [52].

2.3.5 NGRDI

The NGRDI is obtained from the values of the green (G) and red (R) bands, and is calculated using the following expression:

$$\text{NGRDI} = \frac{G - R}{G + R} \quad [53] \tag{6}$$

The NGRDI is frequently employed in the evaluation of vegetation characteristics, such as vegetation coverage, leaf area index, and chlorophyll concentration present in plants [53].

Vegetation indices are widely used as explanatory variables in models for estimating agricultural productivity, as they synthesize information related to vigor, density, and physiological state of vegetation. Among them, NGRDI stands out, as it can be applied even when only images from the visible spectrum are available. Its use in agricultural models and in vegetation evaluation studies is reported in [54].

NGRDI is calculated exclusively from RGB bands, making it a relevant alternative in situations where access to near-infrared bands is unavailable, which are necessary for calculating indices such as NDVI. Several studies highlight its potential to enhance vegetated areas and contribute to the identification of vegetation coverage in different contexts. However, because it depends only on the visible spectrum, NGRDI tends to present lower sensitivity when compared to indices that incorporate additional bands, especially in environments with high spectral variability [50, 53].

2.3.6 Machine Learning Techniques

2.3.7 KNN

The K-Nearest Neighbor (KNN) algorithm, or Nearest Neighbor, is a supervised learning technique widely used for classification tasks [55]. Its basic principle consists of assuming that samples with similar attributes tend to belong to the same class. To classify a new point, the method identifies the k nearest neighbors in the training set using a distance metric, usually Euclidean distance. The class assigned to the new point is the one that occurs most frequently among these neighbors [56].

Despite its simplicity, KNN presents some limitations. It is sensitive to the choice of the parameter k, which directly influences model performance, and does not handle categorical variables or missing data well [55]. In practical applications, appropriate selection of k and prior data treatment are essential to ensure good accuracy [57].

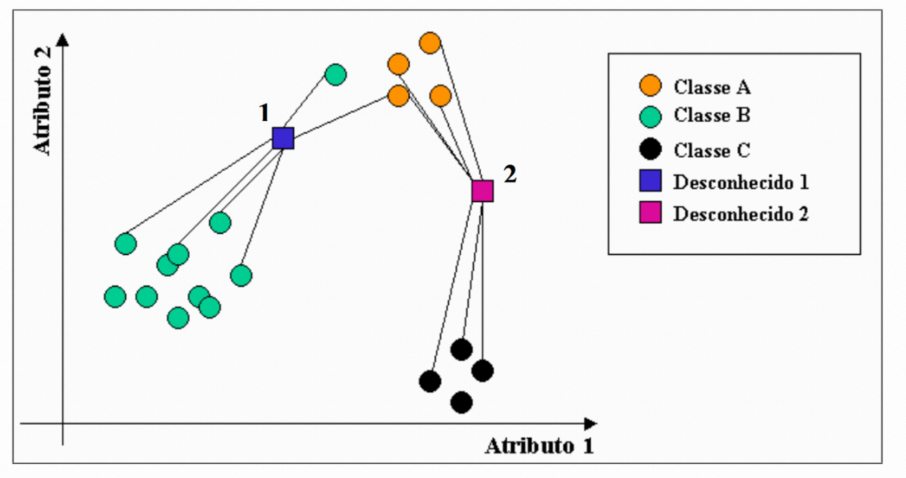


Figure 7: Representation of data grouping into classes using the KNN algorithm [7].

2.3.8 Decision Trees

Decision trees are classification algorithms that operate by creating decision rules based on data characteristics [45]. Each node of the tree corresponds to a characteristic and a decision threshold, and the path from the root to a leaf node represents a series of decisions based on the values of these characteristics [58]. Nodes are split based on impurity criteria to create more homogeneous data subsets. The most common impurity functions include Entropy and the Gini Index. The goal is to maximize information gain at each split [59].

The Gini Index is an impurity measure widely used to evaluate splits in Decision Trees [57]. It represents the probability of a sample being classified incorrectly if labeled randomly. During training, the selected split is the one that produces the smallest impurity value, resulting in purer nodes [60].

The Gini value ranges from 0 to 1, where values close to 0 indicate total purity and values close to 1 indicate great disorder. The Gini Index for a set S is defined by:

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2 \quad [60] \tag{7}$$

where p_i is the proportion of samples belonging to class i. This measure guides the splitting process and contributes to the construction of more interpretable and effective trees.

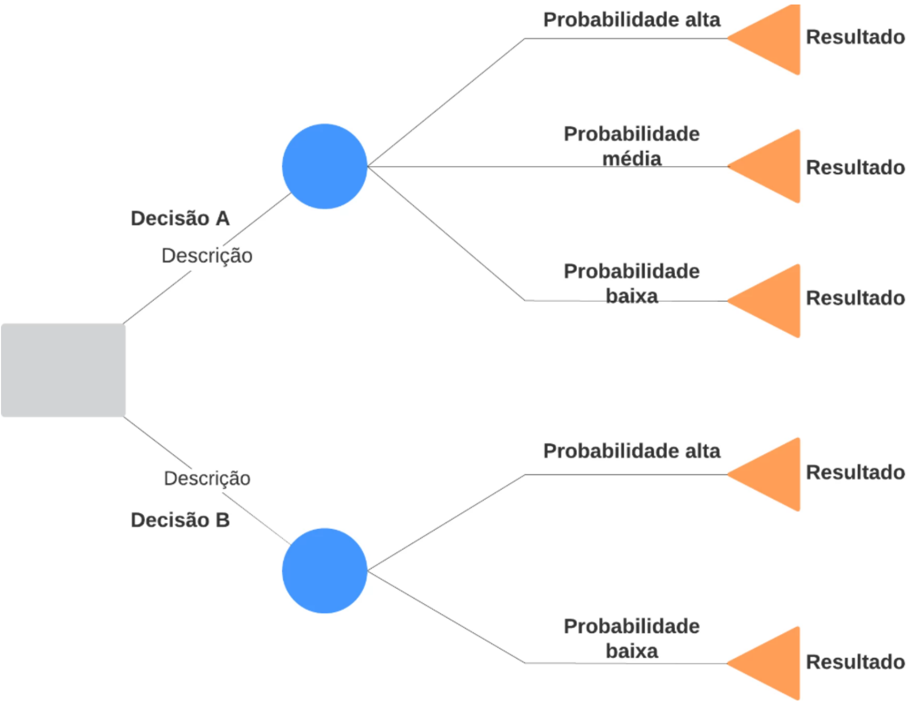


Figure 8: How a decision tree works, where each node represents an evaluated condition [8].

Overfitting occurs when a model learns excessively the details and noise of the training set. Instead of capturing general patterns, the model ends up incorporating anomalies or random fluctuations that are not representative of the real distribution of the data [59]. This results in a model that performs excellently on the training set but fails when applied to new data, generating low accuracy or high error rates in tests [61].

2.3.9 Random Forest

The Random Forest (RF) is a non-parametric supervised learning algorithm that belongs to the family of ensemble methods. It works by combining multiple Decision Trees, so that the final result is more robust and generalizable than that of a single tree [57]. This approach seeks to reduce overfitting and improve predictive performance in complex problems.

During the training process, RF uses two fundamental strategies. The first is bootstrap sampling, in which each tree is built from data subsets generated with replacement. The second strategy consists of aggregating predictions: for classification tasks, the final class is assigned by majority vote of the trees; for regression, the average of the generated predictions is used.

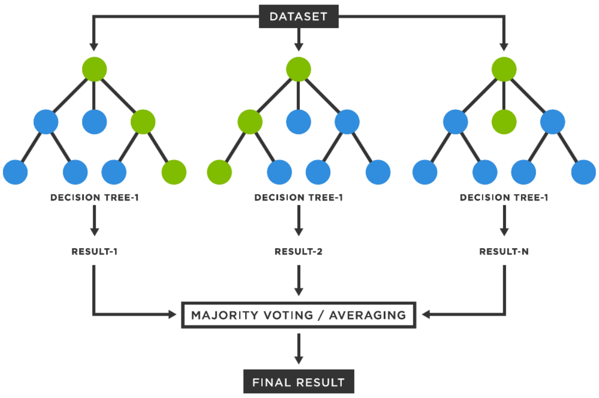


Figure 9: Representation of how the Random Forest algorithm works as a combination of several decision trees [9].

Despite these advantages, RF presents important limitations. Among them are the higher computational cost compared to simpler models, the difficulty of interpretation due to the large number of trees involved, and reduced performance in applications requiring real-time processing [62]. These models can memorize specific details of the training set (Overfitting) [59]. To mitigate this, trees are initially expanded excessively and then pruned to a smaller size, minimizing an estimate of the classification error [63]. Decision trees are also sensitive to outliers and can be easily influenced by noisy data, which leads to inaccurate predictions [64].

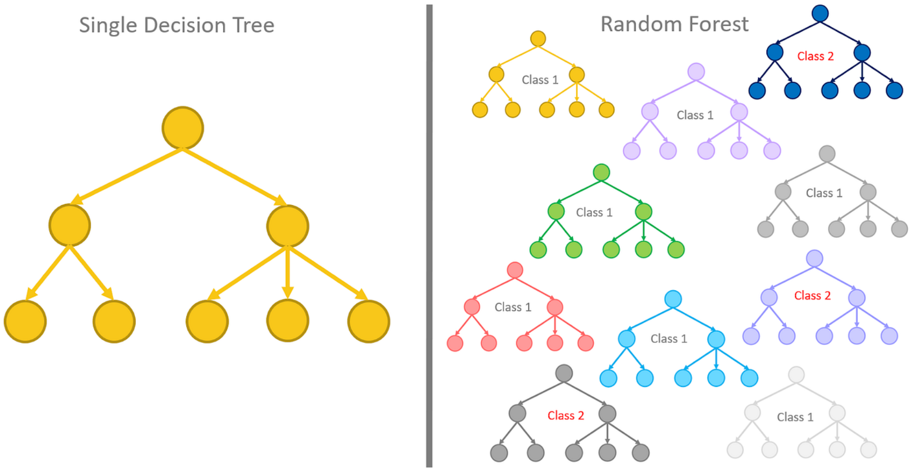


Figure 10: Comparison between a Decision Tree model and a Random Forest model [10].

Still, it remains one of the most widely used approaches in supervised problems due to the combination of conceptual simplicity and high practical performance [60].

2.3.10 Deep Learning Architectures

The Multilayer Perceptron Neural Network (MLP - Multilayer Perceptron) is an important architecture within the field of artificial neural networks (ANNs) and deep learning [65].

An MLP is a type of feedforward neural network. This means that numerical input signals enter through the input layer, propagate through the intermediate layers to the right, and exit through the output layer. Information flows in a single direction, without cycles [66].

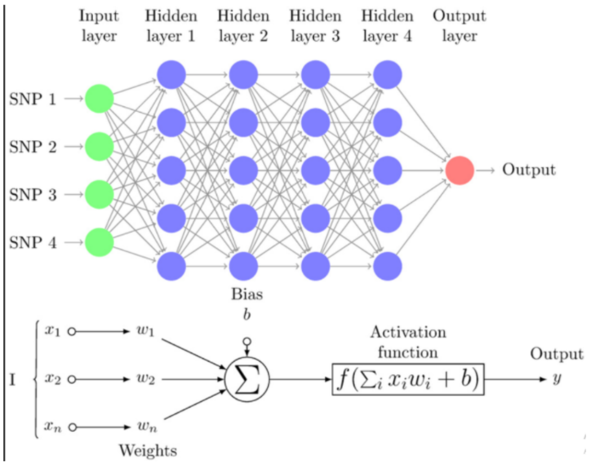


Figure 11: Representation of how an Artificial Neural Network works [11].

- **Input Layer:** It is the first layer of the network and is responsible for receiving raw input data [43].
- **Hidden Layers:** Located between the input and output; MLPs have one or more of them and process information hierarchically [43].
- **Output Layer:** It is the final layer of the network and produces the result of the processing, such as data classification or value prediction [43].

The functioning of each neuron in an artificial neural network follows a structured mathematical process. Initially, the neuron receives multiple numerical signals from the input layer or from neurons in the previous layer. Each of these signals is multiplied by a specific synaptic weight, which determines the strength and type of influence (excitatory or inhibitory) of the connection.

Subsequently, the weighted sum of all inputs is calculated, where each signal is multiplied by its respective weight before aggregation. This operation, called linear combination, produces the net input value (z) of the neuron, as expressed by the equation:

$$z = \sum_{i=1}^{n} w_i x_i + b \quad \text{[67]} \tag{8}$$

Where:

- x_i represents the neuron's inputs;
- w_i are the weights associated with each input x_i;
- b is the bias term, a constant learned along with the weights;
- Σ denotes the sum of all products $(w_i \cdot x_i)$.

The result z is then passed through an activation function, represented by ϕ or f. This function is fundamental because it introduces nonlinearity to the model, allowing the MLP to learn and represent complex patterns that a linear model could not. The final output of the neuron, $a = \phi(z)$, becomes the input to the next layer [67].

- **ReLU (Rectified Linear Unit):** Defined as $f(x) = \max(0, x)$, acting as an identity function for positive values and nullifying negative inputs. It is widely used due to its computational simplicity and effectiveness in mitigating the vanishing gradient problem [68].

2.4 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a statistical and machine learning technique widely used for dimensionality reduction. Its main objective is to transform a dataset with many variables into a new set with fewer variables, known as principal components, preserving as much as possible of the original data variance [43].

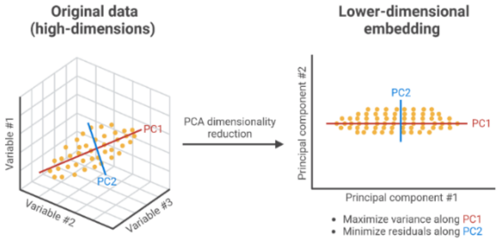


Figure 12: Representation of Principal Components (PCA) and the dimensionality reduction process [12].

PCA (Principal Component Analysis) transforms the original characteristics of a dataset into new uncorrelated variables, called principal components. These components are ordered so that the first ones capture most of the variance, preserving the most relevant information. In high-dimensional data, where correlated features can impair model performance, PCA efficiently reduces dimensionality, compressing the dataset with minimal loss of significant information and improving computational efficiency [43].

In the domain of image recognition, Principal Component Analysis (PCA) is widely employed as a preprocessing technique for high-dimensional datasets, allowing the compression of visual information without significant loss of relevant content [69]. A specialized variant, Local PCA, has proven particularly effective in processing colored images, adapting to the regional characteristics of the scene.

The study by [43] demonstrates that dimensionality reduction through PCA, when integrated with feature selection methods, establishes a synergy that significantly enhances the performance of MLP (Multilayer Perceptron) networks. This combined approach not only mitigates problems related to the curse of dimensionality but also contributes to accelerating the training process and improving the generalization capacity of the models.

2.5 COMPLEMENTARY TECHNIQUES

2.5.1 Ensemble Learning

Ensemble Learning is a powerful and widely used approach in machine learning. Its fundamental principle consists of combining the predictions of multiple models to obtain overall performance superior to that of any individual model. By uniting different hypotheses, the ensemble exploits the particular strengths of each classifier, reducing errors, mitigating biases, and increasing system robustness [59].

Improving performance and generalization: the combination of diverse models tends to reduce prediction error, resulting in higher accuracy and better generalization capacity, especially in complex scenarios or those with noise [59]. Models such as neural networks present excellent capacity to model nonlinear relationships, while modern optimizers such as Adam accelerate and stabilize the training process [45]. Tree-based methods provide interpretability and robustness [58]. Simple algorithms such as k-NN work effectively in similarity-based classifications [70]. The combination of these techniques forms more resilient systems adapted to different forms of data [58].

Bagging is an ensemble technique that trains multiple models (often decision trees) on different subsets of the dataset, generated by sampling with replacement. Each model is trained independently, and the final prediction is obtained by averaging (regression) or by majority vote (classification) [71]. **Boosting**, on the other hand, trains models sequentially, so that each new classifier corrects the errors of the previous one, resulting in a strong and stable classifier [59].

In the figure below we can see the comparison of the two techniques:

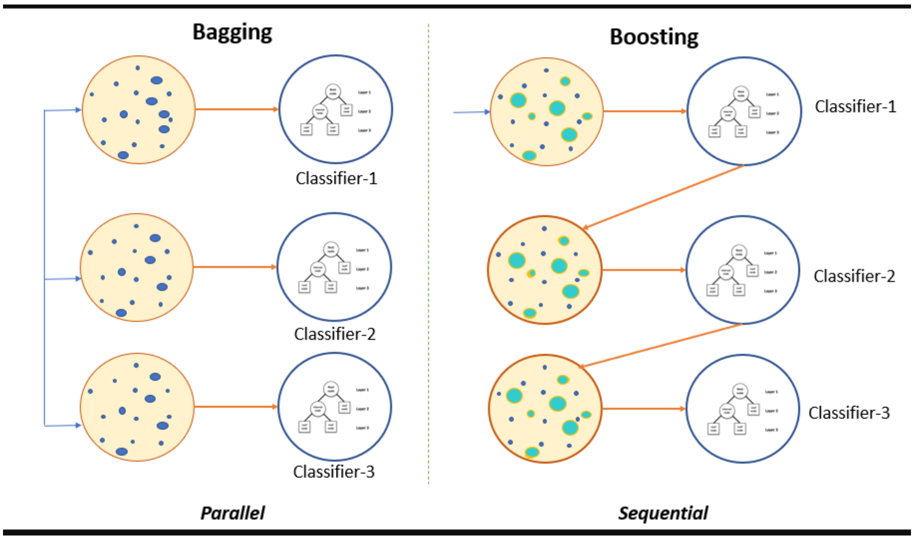


Figure 13: Example of Ensemble Learning combining different classification techniques [13].

2.5.2 AdaBoost Algorithm

The Adaptive Boosting (AdaBoost) is one of the most influential algorithms in the Boosting family. Its central objective is to transform weak classifiers into a strong classifier through an iterative and adaptive process [48, 72].

The functioning of AdaBoost can be summarized in two main ideas:

- **Adaptive weighting of examples:** Initially, all training examples receive the same weight. At each iteration, the algorithm increases the weights of incorrectly classified samples and reduces the weights of correctly classified samples. This mechanism forces the next classifier to concentrate its learning on the most difficult examples [59].

- **Weighted combination of classifiers:** Each weak classifier trained produces a prediction and receives a weight proportional to its performance. Classifiers with lower error rates receive greater weight in the final prediction. Thus, the final classifier is a weighted combination of individual decisions [48].

Ensemble Learning is a powerful and widely used approach in machine learning. Its fundamental principle consists of combining the predictions of multiple models in order to obtain overall performance superior to that of any individual model. By uniting different hypotheses, the ensemble exploits the particular strengths of each classifier, reducing errors, mitigating biases, and increasing system robustness [59].

Improving performance and generalization: the combination of diverse models tends to reduce prediction error, resulting in higher accuracy and better generalization capacity, especially in complex scenarios or those with noise [59]. Models such as neural networks present excellent capacity to model nonlinear relationships [45], while modern optimizers such as Adam [73] accelerate and stabilize the training process. Tree-based methods provide interpretability and robustness [58], and simple algorithms such as k-NN work effectively in similarity-based classifications [70]. The combination of these techniques forms more resilient systems adapted to different forms of data [58].

2.5.3 Otsu's Method for Thresholding

Otsu's Method is a classical automatic thresholding technique used to convert grayscale images into binary images. It is a non-parametric and unsupervised method because it determines the ideal threshold directly from the image histogram without the need for manual adjustments [74].

Its objective is to find the threshold value that best separates pixels into two classes: *background* and *objects*. To do this, the algorithm tests all possible thresholds and selects the one that maximizes class separation (interclass variance), indicating the best distinction between light and dark regions of the image [75].

After the threshold is selected, the image is binarized: pixels with intensity greater than or equal to the threshold become white, while the others become black. The method works best on images with bimodal histogram, although it may present limitations in low-contrast scenarios or complex distributions [76].

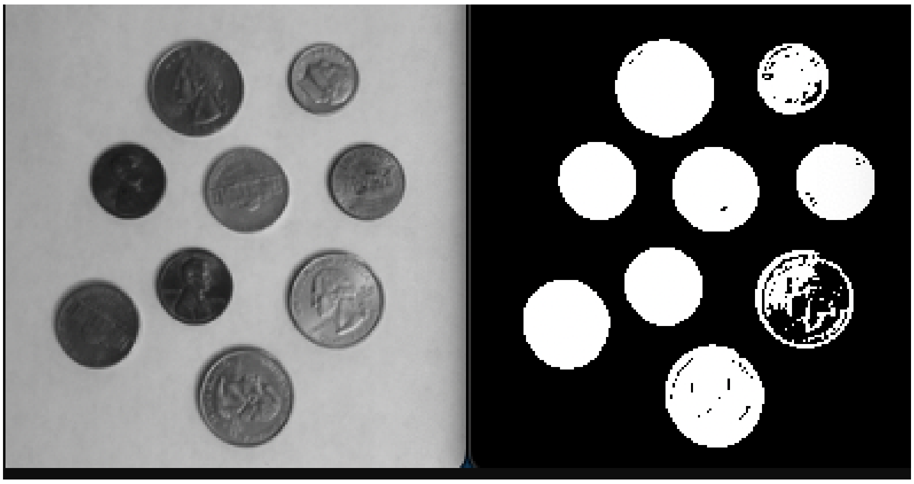


Figure 14: Application of Otsu's method for automatic threshold definition [14].

2.5.4 Image Skeletonization

Skeletonization is an image processing technique that operates on binary images, having been implemented in the work of [77]. This method consists of a thinning process that progressively reduces the lines of a binary image until they reach the thickness of a single pixel, resulting in the extraction of the skeleton of the original image.

The main advantage of the skeletonization process, as demonstrated in the mobile solution for wound analysis, lies in its ability to reduce the complexity of the image representation while maintaining structural information relevant for subsequent morphometric analyses [77].

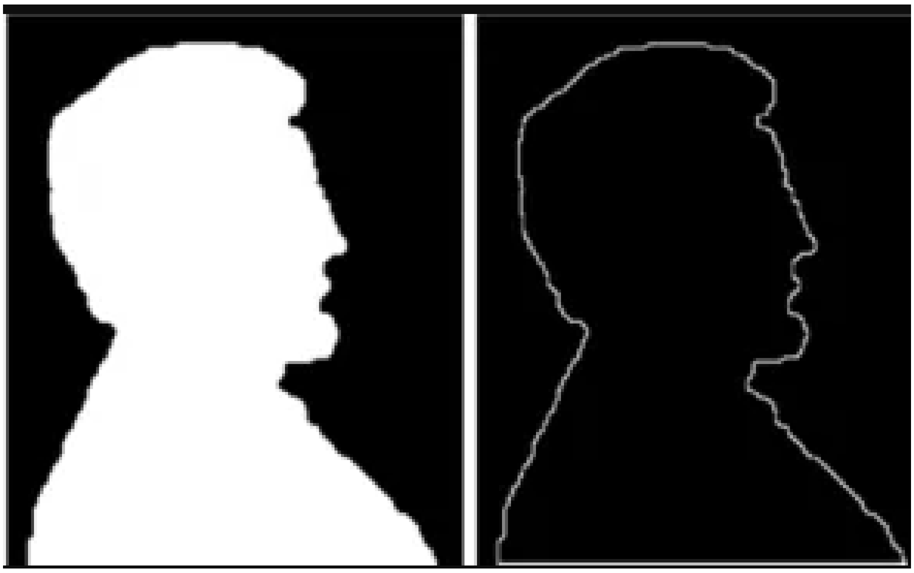


Figure 15: Skeletonization produced from the morphological transformation applied to the image [15].

2.5.5 Pearson Correlation Coefficient

The Pearson Correlation Coefficient (r) is a measure of linear association between quantitative variables that measures the direction and degree of the linear relationship between two variables [78]. The coefficient is based on the standardization of observations and on the sum of the cross product of standardized values of variables X and Y.

The formula of the coefficient is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad [78] \tag{9}$$

where:

- x_i and y_i are the individual observations of variables X and Y
- $\bar{x}$ and $\bar{y}$ are the means of variables X and Y, respectively
- s_x and s_y are the sample standard deviations of variables X and Y
- n is the number of observations

The Coefficient of Determination (R^2), derived from r, represents the proportion of variance shared between the variables.

The value of r varies in the interval $[-1, 1]$:

- **Direction**: The sign indicates positive ($r > 0$) or negative ($r < 0$) relationship
- **Magnitude**: The strength of the linear relationship follows the Dancey and Reidy classification [78]:
  - $0.10 \leq |r| \leq 0.30$: weak correlation
  - $0.40 \leq |r| \leq 0.60$: moderate correlation
  - $0.70 \leq |r| \leq 1.00$: strong correlation
- $r = 0$: indicates absence of linear relationship
- $|r| = 1$: perfect linear correlation

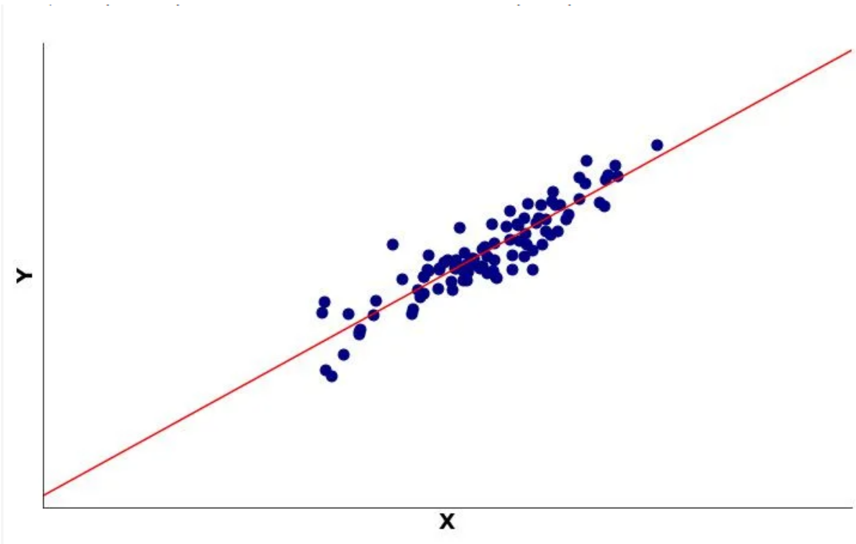


Figure 16: Scatter plot showing positive correlation between two variables [16].

2.5.6 Spearman Correlation Coefficient

The Spearman Coefficient (ρ) is a statistical tool that assesses the extent to which as one variable increases, the other tends to follow in a specific direction. Its main difference is that, instead of calculating using exact values, it uses the position or *ranking* of each data point, eliminating the need for data to follow a perfect distribution [79]. Due to this flexibility, it is the ideal method for analyzing data that possess a general growth trend, even if this relationship does not form an exact straight line [80].

The calculation of the coefficient is based on the difference between the ordered positions of each pair of observations. The formula is defined by Equation 10:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \quad [80] \tag{10}$$

where:

- d_i represents the difference between the ranks of the corresponding variables;
- n is the number of observation pairs (sample size).

The interpretation of the result varies from -1 to +1, where positive values indicate a direct correlation and negative values an inverse correlation. Magnitude between 0.75 and 1.0 suggests a strong to excellent correlation [81].

2.5.7 Calculation and Interpretation of the Ångström Index

The Ångström Index is an empirical fire danger indicator widely used in forest fire risk assessment studies [82]. It combines atmospheric temperature and relative humidity to estimate the favorability of environmental conditions for fire ignition and spread [83].

$$B = \frac{U}{20} + \frac{27 - T}{10} \quad [82] \tag{11}$$

where:

- T: Air temperature in degrees Celsius;
- U: Relative humidity of the air in percentage.

Physical interpretation:

The Ångström Index reflects the combined influence of temperature and atmospheric moisture on fire risk:

- The term $\frac{27-T}{10}$ represents the effect of temperature. As air temperature increases, this term decreases, indicating more favorable conditions for fire occurrence due to enhanced fuel drying [82].
- The term $\frac{U}{20}$ represents the effect of atmospheric moisture. Higher relative humidity increases this term, reflecting less favorable conditions for fire ignition, since moist air reduces fuel flammability [82].

Lower values of the Ångström Index indicate drier and warmer atmospheric conditions, which are more conducive to fire ignition and propagation. Conversely, higher index values correspond to cooler and more humid conditions, reducing fire risk [83].

According to the classification proposed in the literature [82]:

- $B < 2.5$: Very high fire risk;
- $2.0 \leq B < 4.0$: Moderate fire risk;
- $B \geq 4.0$: Low or negligible fire risk.

Due to its simplicity and reliance on readily available meteorological variables, the Ångström Index is particularly useful for operational fire monitoring and early warning systems, especially in regions with limited data availability [83].

2.5.8 Algorithm Comparison Metrics (Benchmarking)

Metrics are used to quantify the performance of architectures. They also assist in the evaluation and optimization of hyperparameters of classifiers [59, 66].

Precision is the proportion of true positives (TP) over all predicted positives (TP + FP):

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall or *Sensitivity*, is the proportion of true positives (TP) over all actually positive cases (TP + FN):

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score is the harmonic mean between Precision and Recall:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy is the ratio of correct predictions to the total number of predictions made:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2.5.9 Related Work

2.5.10 Road Network Mapping with Multispectral Bands

In the work by Hollendonner et al. [36], the authors investigated the automated extraction of road networks using WorldView-3 images from the SpaceNet dataset. The methodology employed a U-Net architecture with DenseNet backbone, utilizing innovative combinations of multispectral bands (Green, Red Edge, and Near-IR2) in false colors.

The results demonstrated significant gains: an increase of 5.4% in F1-Score and 6.5% in IoU compared to the conventional RGB model. This study validates the potential of non-visible bands to improve the segmentation of linear features such as roads, being directly relevant for the exploration of spectral characteristics in this work.

2.5.11 Forest Fire Detection with Ensemble Learning

Xu et al. [37] proposed a fire detection system based on ensemble learning, combining YOLOv5, EfficientDet, and EfficientNet. The strategy used the global classifier (EfficientNet) to validate local detections, reducing false positives to just 0.3%.

Although the domain is different, the ensemble approach and the combination of information at multiple levels (local and global) offer valuable insights for detection architectures in remote sensing.

2.5.12 Mapping with Data Cube and Temporal Segmentation

Chaves et al. [84] explored land use mapping in Mato Grosso by integrating MODIS data cubes with GEOBIA segmentation. Using multiresolutional segmentation and SVM classification, they achieved an accuracy of 0.95 in identifying natural vegetation and crops.

The methodology demonstrates the effectiveness of combining temporal and spatial analysis, reinforcing the importance of image processing and classification techniques for information extraction in complex landscapes.

2.5.13 Comparative Table of Related Works

The table below synthesizes and compares the three analyzed works:

Table 2: Comparative Table of Related Works

Criterion	**Hollendonner et al. [85]**	**Xu et al. [86]**	**Chaves et al. [84]**
Main Objective	Improve automated road network extraction with Deep Learning and multispectral bands.	Develop a real-time fire detection system with low false positive rate using ensemble learning.	Evaluate the combination between data cube and GEOBIA to map natural vegetation and double cropping.
Main Methodology	Semantic segmentation with U-Net and DenseNet using false color bands (Green, Red Edge, Near-Infrared 2).	Ensemble with three models: YOLOv5, Efficient-Det, and a global classifier (Efficient-Net).	Data cube architecture (MODIS ARD) combined with object-oriented segmentation (GEOBIA via MRS) and classification via SVM.
Main Results	Improvement of 5.4% in F1-Score and 6.5% in IoU compared to using only RGB.	Reduction of false positive rate to 0.3%, surpassing individual models.	Overall accuracy of 0.95 in mapping heterogeneous landscapes.
Challenges / Limitations	Occlusions (trees, buildings), unpaved roads, and disconnected segments are still challenging.	Isolated detectors generate many false positives; no model is effective in all scenarios.	Confusion between Cerrado and Pasture; MODIS resolution limits the creation of homogeneous geo-objects.
Relevance for this Thesis	Validates the use of CNNs for infrastructure extraction (roads) from satellite imagery.	Demonstrates the effectiveness of ensemble models for complex environmental tasks.	Demonstrates the potential of combining remote sensing and SVM for accurate LULC mapping in heterogeneous regions.

3

3 METHODOLOGY

3.1 RESEARCH CHARACTERIZATION AND DATA USED

This research was structured to develop and evaluate a method for automatic detection of roads and access routes in areas associated with the first focuses of forest fires. The central objective was to analyze how the presence and density of roads near initial ignition points can aid in understanding occurrence patterns and support prevention and rapid response actions. To this end, 40 fire events recorded by the OroraTech platform in the Cerrado biome during the months of October and November were collected, processed, and analyzed. The research followed the steps outlined in Figure 17 below.

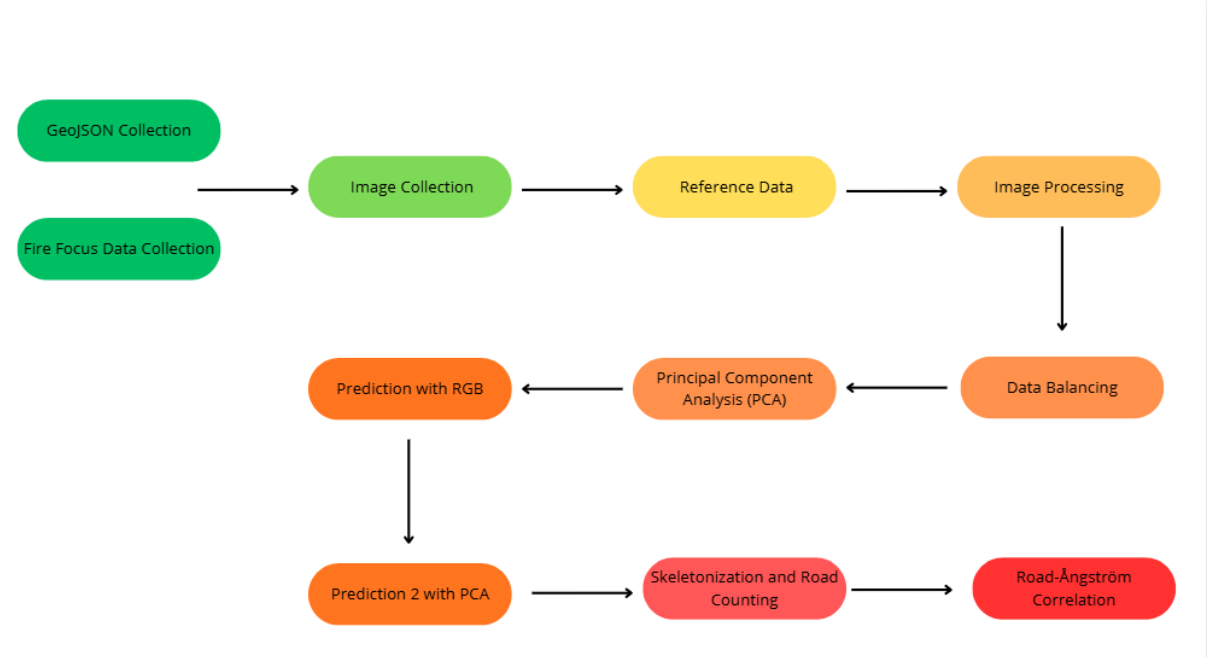

GeoJSON Collection
Fire Focus Data Collection
Image Collection
Reference Data
Image Processing
Data Balancing
Principal Component Analysis (PCA)
Prediction with RGB
Prediction 2 with PCA
Skeletonization and Road Counting
Road-Ångström Correlation

Figure 17: Roadmap of the work developed. Source: Own authorship.

3.2 RESEARCH CHARACTERIZATION AND DATA USED

For each fire event, information was collected through the OroraTech platform, a German company specializing in thermal satellite monitoring. The data included the geographical coordinates of the first focus (in WGS84), detection date and time, temperature of the first focus in degrees Celsius, relative air humidity in percentage, total burned area in square kilometers, and a file in GeoJSON format containing the complete spatial geometry of the event.

Using the GeoJSON file as a spatial reference, Sentinel-2 images were obtained through the Sentinel-Hub platform. Although Sentinel-2 provides several spectral bands (B01;B12), ranging from visible to infrared, only RGB compositions were downloaded.

This choice is due to the fact that the other color spaces used in the research (HSV, CIELAB, derived indices, and principal components) would be generated later in processing from the RGB image itself. In this way, the download was restricted to the essential, avoiding redundancy and reducing the volume of data to be processed.

3.3 ROAD MANUAL VECTORIZATION STAGE

Among the 40 images obtained, 12 were selected for a manual vectorization process of road structures. Using QGIS software, each image was loaded as a raster layer and a new vector layer of linestring type was created with the same spatial reference system. The procedure consisted of manually tracing lines over all visible road structures in the image, including asphalt-paved highways, unpaved secondary roads in earth, rural access routes, and well-defined trails used by vehicles.

During this vectorization process, linear structures that did not correspond to roads were deliberately excluded, such as property fences, forests, shrubs, power transmission lines, and land boundaries. After tracing the lines, each shapefile was reviewed visually at multiple zoom scales to ensure completeness, geometric accuracy, absence of duplications, and correct alignment with the center of the roads. The 12 vectorized shapefiles were then exported in ESRI Shapefile format, preserving geometric and spatial reference system information.

3.4 CONVERSION TO BINARY MASKS

The 12 shapefile files containing the vectorized roads were converted into rasterized binary masks through a script developed in Python using the Rasterio library. Each shapefile was rasterized to the same spatial grid as the original Sentinel-2 RGB images. In the resulting binary encoding, pixels that intercepted any part of a vector line received the value 255 on an 8-bit scale, representing the road class, while all other pixels received the value 0, representing the non-road class. According to Figures 18 and 19 below.

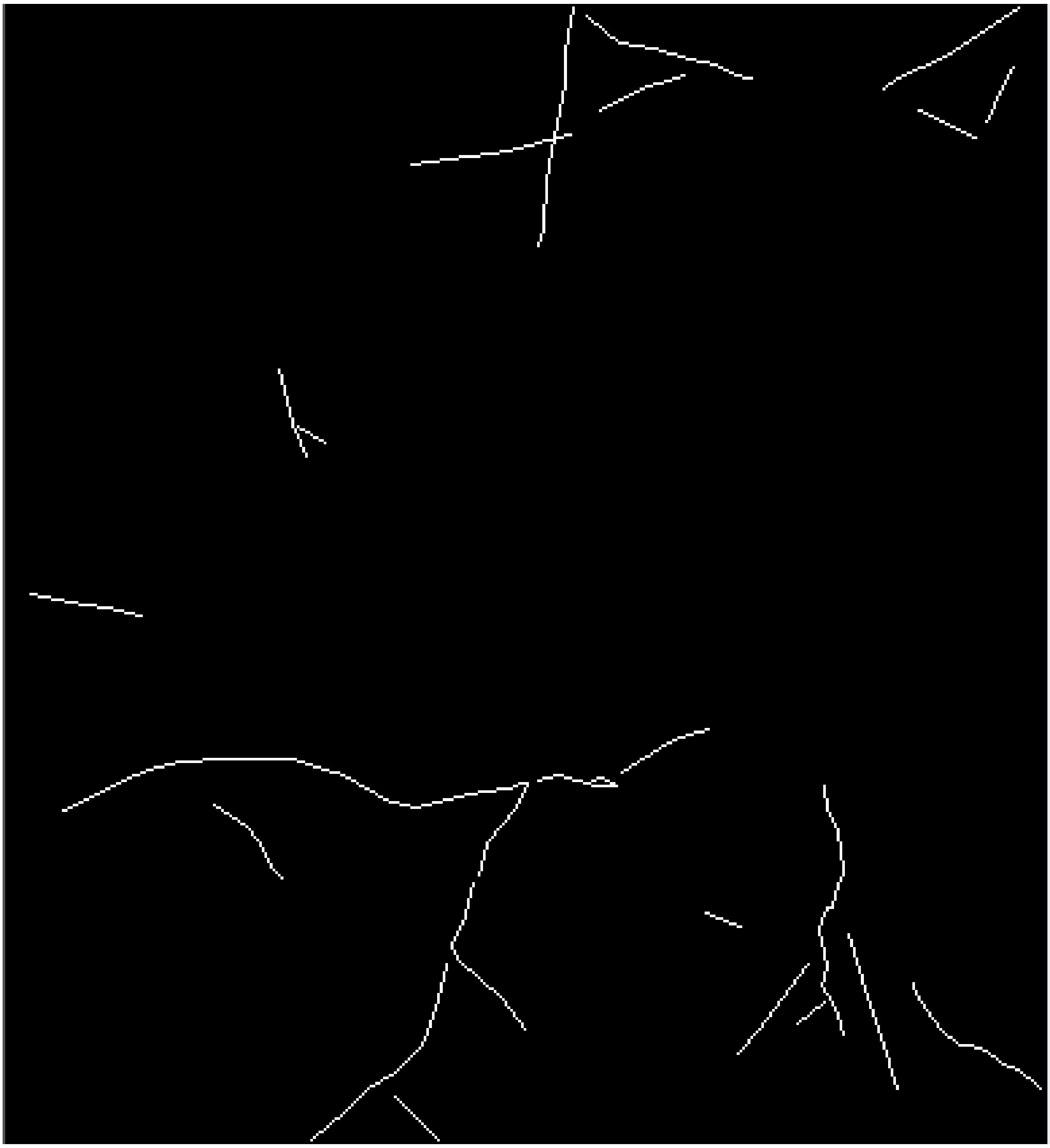


Figure 18: Binary mask developed from the template produced in QGIS. Source: Own authorship.

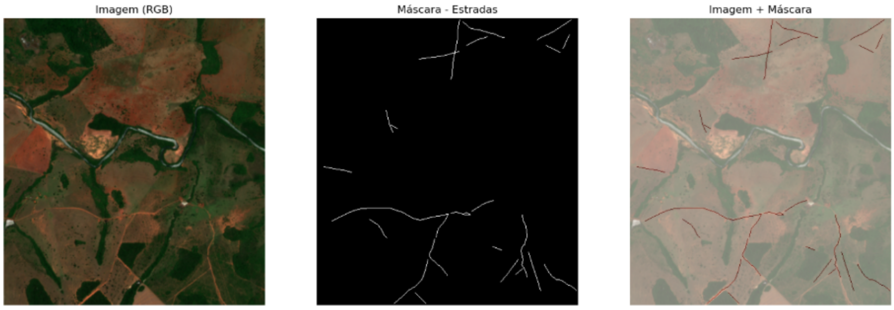


Figure 19: Overlay from the binarized masks and the original image. Source: Own authorship.

It was verified that each binary mask had the same alignment as the original image. To do this, three fundamental aspects were checked: the same geotransformation matrix, the same spatial reference system (EPSG:4326), and the same dimensions in number of pixels. As a confirmation step, each mask was overlaid on the respective RGB image in QGIS, allowing visual observation of whether the contours were correctly matched. In the end, 12 binary masks in GeoTIFF format (8 bits) were obtained, all perfectly corresponding to the vectorized images.

3.5 IMAGE PROCESSING

All 40 RGB images obtained were submitted to two preprocessing operations before feature extraction. The first operation was radiometric normalization, in which each pixel was divided by 255, converting 8-bit values (scale 0-255) to a continuous scale of 0 to 1. This normalization is essential to ensure numerical stability during subsequent feature extraction and model training stages.

The second stage of processing consisted of applying a two-dimensional Gaussian filter, using a 5×5 pixel kernel and standard deviation $\sigma = 1.5$. This operation performs spatial smoothing of the image through its convolution with a Gaussian function, which reduces radiometric noise from various sources, such as thermal sensor noise and atmospheric interference [45]. By smoothing fine details and high-frequency textures, the filter homogenized the overall appearance of the scene, which contributed to a better definition of contours and smooth linear structures characteristic of roads, facilitating their subsequent identification. The procedure was applied independently to each of the three RGB channels, ensuring the preservation of the original relationships between spectral bands.

Figure 20 represents the change in the image.

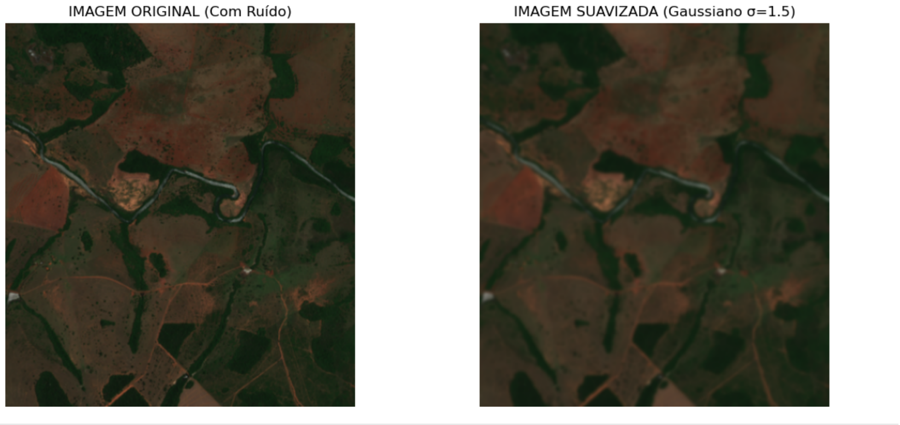


IMAGEM ORIGINAL (Com Ruído)
IMAGEM SUAVIZADA (Gaussiano σ=1.5)

Figure 20: Image smoothed by the Gaussian filter. Source: Own authorship.

3.5.1 Class Imbalance Correction

Analysis of the labeled masks revealed a marked imbalance between classes, since roads appeared as very thin lines. Of the 2,388,146 pixels considered, only 28,594 (1.20%) were positive, resulting in an approximate ratio of 1:82 between road and non-road. To make the set trainable, two correction strategies were applied [87].

First, morphological dilation with 8-connectivity was performed, increasing the thickness of roads, connecting fragmented segments, and reducing the sparsity of the positive class. Then, random undersampling of the negative class was performed, keeping only four background pixels for each road pixel. This approach resulted in a final ratio of 1:4, preserving all 28,594 positive pixels and reducing negative ones to 114,376.

Thus, the balanced dataset used in training contained 142,970 pixels, ensuring greater stability and avoiding learning problems arising from the initial extreme imbalance.

3.5.2 Feature Extraction

Before defining the final set of attributes, an exploratory analysis of the variables extracted from the images was performed. This step had only the function of preliminarily verifying how different descriptors behaved in regions containing roads. It was not a performance evaluation, but an initial procedure to guide the selection of the most appropriate characteristics.

The inspection indicated that perceptual color spaces, such as HSV and CIELAB, reveal relevant differences between pavement, vegetation, and exposed soil [36, 49]. Components associated with luminance, brightness, grayscale, and chromatic axes showed consistent variations along the roads [88]. These results motivated the adoption of multiple colorimetric and structural representations in the composition of the feature set.

Based on these initial observations, two complementary strategies for attribute extraction were developed, described below.

3.5.3 Strategy Based Only on Color Spaces

In the first approach, called Complete Strategy, the preprocessed RGB images were used directly for the generation of a comprehensive set of features. Smoothed RGB channels were extracted, conversions to HSV and CIELAB color spaces, grayscale image, a simple brightness measure, and the magnitude of the Sobel operator for edge detection. In addition to these attributes, an adapted NDVI index was also incorporated, previously calculated from the original bands, employed as an attempt to partially compensate for the absence of the near-infrared band in RGB images [52]. The final set of this approach resulted in 12 features per pixel.

NGRDI, calculated directly from the visible bands of RGB images, presents relevant capacity to highlight vegetated regions even when multispectral sensors are not available, as discussed in [50]. In the present study, the unavailability of the bands necessary for adequate NDVI calculation motivated the adoption of NGRDI as a methodological alternative, since this index can be derived exclusively from information contained in the visible spectrum and still demonstrate good effectiveness in distinguishing areas covered by vegetation. Previous works reinforce this potential, highlighting the use of NGRDI both for vegetation coverage classification and for application in agricultural models and pasture analysis [53, 54]. Thus, its use in this study seeks to supply, at least partially, the absence of NDVI, preserving the ability to perform a consistent spatial analysis of vegetation from available spectral data.

3.6 PREPARING IMAGE SETTINGS FOR COMPARISON

3.6.1 Strategy Based Principal Component Analysis

To understand the data structure and identify possible redundancies between spectral features, Principal Component Analysis (PCA) was applied to the extracted feature set [43].

The analysis revealed that the first components explained most of the variance in the data. The first component (PC1) explained 70.09% of the total variance, the second component (PC2) added 11.92%, the third component (PC3) contributed 7.07%. Thus, only the first three principal components retained approximately 89.08% of the total variance of the data. This finding motivated the development of an alternative processing strategy using images reconstructed from these first three components, exploring dimensional reduction as a regularization mechanism.

In the second approach, called PCA Strategy, the images were initially transformed through Principal Component Analysis (PCA). From the first three components, new images were reconstructed that preserve the main statistical variations of the scene. On these transformed images, the same procedures from the previous strategy were applied: conversions to HSV and CIELAB, brightness calculation, grayscale, and Sobel operator. The objective of this strategy was to evaluate whether the reorganization of spectral information promoted by PCA could facilitate the separation between the classes of interest during model training [43].

The two strategies were designed to be complementary. The first directly explores the original spectral behavior of RGB images, while the second investigates how PCA transformation can reorganize and highlight useful information. Both were grounded by the initial exploratory analysis and by established studies in image processing, ensuring that the classifier receives a rich and representative set of features.

3.6.2 Attribute Extraction, Processing Pipelines, and Modeling

This study adopted a supervised pixel-by-pixel classification approach for road detection in aerial images. The methodological flow was structured in three main stages: (i) extraction of spectral attributes, (ii) data preparation with class imbalance correction, and (iii) training of machine learning models, including independent ensemble strategies.

3.6.3 Attribute Extraction and Color Spaces

The first step consisted of defining descriptors capable of representing each pixel in a manner sensitive to photometric differences between pavement, vegetation, and exposed soil. Multiple color spaces were used due to the complementarity of their spectral responses, allowing specific features of the scene to be emphasized.

Two independent pipelines for attribute extraction were developed, which share the same spectral transformation techniques but differ in how the initial bands are defined.

Pipeline Without PCA In the first pipeline, the image is processed directly from the original **RGB** channels. Based on these channels, spectral attributes were generated in the following spaces:

- **CIELAB (L, A, B)**: separates luminosity from chromaticity [49], approaching human perception;
- **HSV (H, S, V)**: highlights hue and saturation, useful for differentiating materials [43];
- **Grayscale**: representation of light intensity [36];
- **Adapted NDVI (NGRDI)**: NGRDI applied to RGB images to highlight vegetated regions [50].

This pipeline fully preserves the spectral information of the original image, resulting in a richer feature vector, but also more susceptible to noise and redundancy.

Pipeline With PCA In the second pipeline, the initial image is compressed by Principal Component Analysis (PCA). The first three components (PC1, PC2, and PC3), responsible for capturing most of the variance [69]. From these reduced bands, the same transformations are applied:

- conversion to **CIELAB**;
- conversion to **HSV**;
- calculation of **grayscale**;
- calculation of **adapted NDVI**.

3.7 MACHINE LEARNING MODELS

Four classical supervised algorithms were employed in the classification task:

- **K-Nearest Neighbors (KNN)**: suitable for nonlinear decision boundaries and dependent on space geometry [55];
- **Decision Tree**: interpretable model, organized as hierarchical divisions of attribute space [58];
- **Random Forest**: ensemble of multiple independent trees, reducing variance and increasing robustness [57];
- **Multi-Layer Perceptron (MLP)**: neural network capable of capturing complex nonlinear relationships [65].

Each algorithm was evaluated in two distinct hyperparameter configurations, resulting in eight base models per pipeline.

3.7.1 Ensemble Architecture

For each pipeline (with and without PCA), an independent ensemble architecture was constructed. In each case, the eight base models generate preliminary predictions, subsequently combined by an **AdaBoost** meta-classifier with 50 estimators.

Therefore, two ensembles were trained:

1. **Ensemble Without PCA**: operates on the set of attributes derived directly from the original RGB bands;
2. **Ensemble With PCA**: operates on the reduced set of PCA bands.

This separation allows evaluation of the impact of dimensionality reduction on system accuracy and stability.

3.7.2 Comparative Performance

The experimental results revealed important differences between the two ensembles:

- The **ensemble without PCA** presented the best overall performance, obtaining superior metrics of precision, recall, and F1-score. However, the model showed greater sensitivity to noise and higher risk of overfitting, a characteristic associated with the high dimensionality of the feature vector.
- The **ensemble with PCA** presented inferior metrics but demonstrated greater stability, lower variance, and lower tendency to overfitting. Dimensional reduction provided a more compact and less redundant set of attributes.
- In both cases, the use with AdaBoost improved performance relative to individual models, reducing specific errors of each classifier.

In summary, the ensemble based on the **without PCA** pipeline achieved the best absolute performance, while the ensemble **with PCA** stood out for greater robustness and lower risk of overfitting, offering a more conservative alternative for scenarios with high spectral variability.

The result of the mask predicted by the machine learning model using PCA follows the figure below 21.

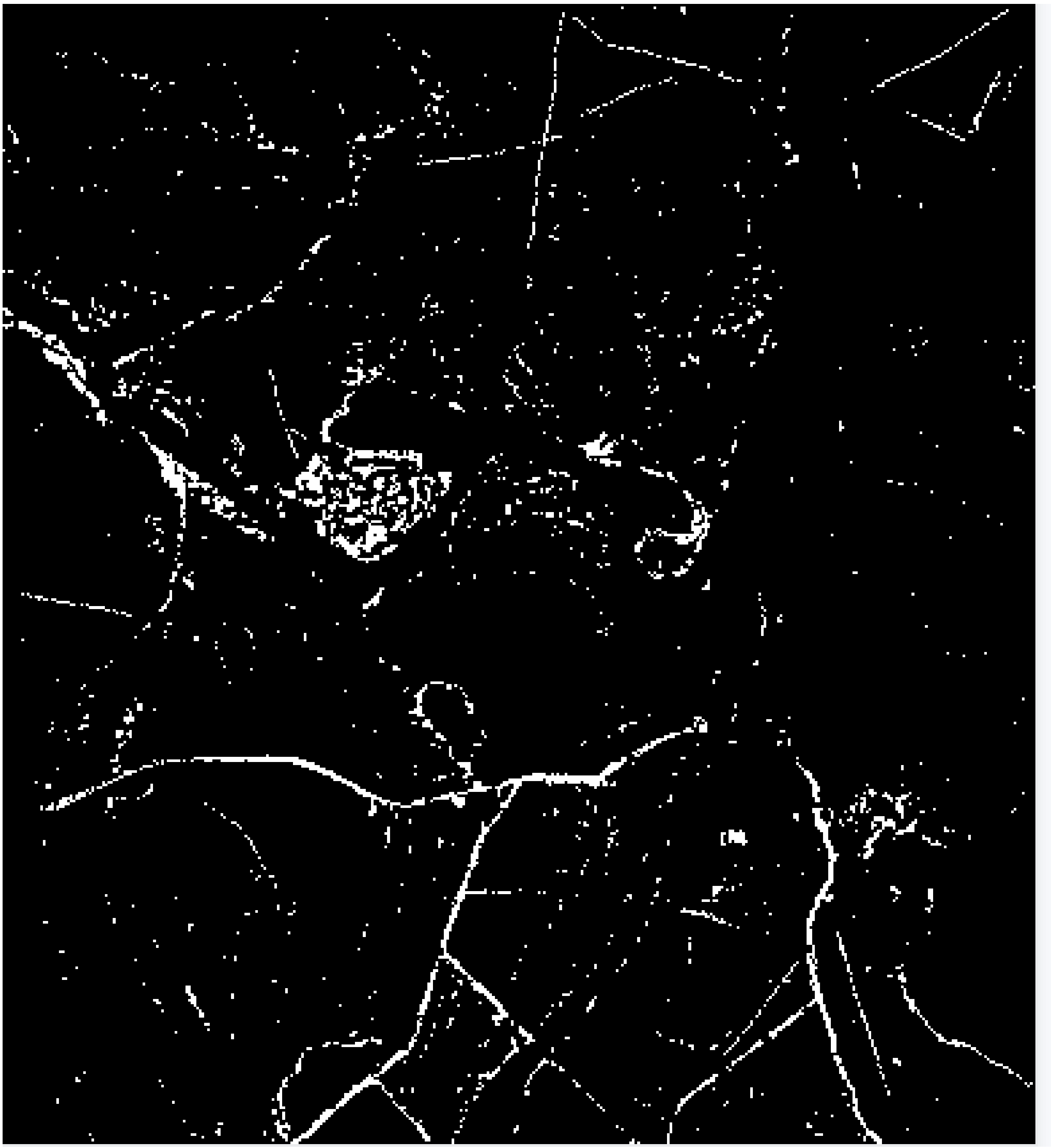


Figure 21: Mask produced by the chosen predictive model. Source: Own authorship.

3.7.3 Post-processing: Binarization, Skeletonization, and Size Filtering

After the generation of prediction masks by the model, a post-processing pipeline was applied to identify, separate, and quantify only the most consistent road segments. The objective of this stage was to reduce noise, remove disconnected fragments, and retain exclusively segments sufficiently extensive for quantitative analysis.

Binarization via Otsu Threshold The raw masks produced by the model may contain intensity variations and ambiguous regions. To standardize this representation, Otsu's method was initially applied, which automatically selects an optimal threshold to separate the road class from the background. The result is a binary mask where pixels belonging to roads are marked with value 1 [74].

Skeletonization of the Binary Mask The binarized image was then subjected to the skeletonization process, whose function is to reduce each connected segment to a central line of unit thickness, preserving its topological structure. This step is fundamental to prevent variations in line width from influencing pixel counting and, consequently, size grouping.

Grouping by Connected Components From the binary skeleton, connected component labeling was applied. Each set of adjacent pixels was treated as an independent road segment. This step transforms the skeletonized mask into well-defined geometric units and allows operation on each stretch in isolation.

Classification of Segments by Size Each identified component had its size measured in number of pixels. Based on this measurement, segments were classified into three categories:

- **small**: up to 200 pixels;
- **medium**: from 201 to 800 pixels;
- **large**: above 800 pixels.

These thresholds were defined empirically, based on experimentation and visual analysis. Small segments tend to correspond to residual noise, spurious connections, or fragmented predictions, while large segments represent continuous and reliable road stretches.

Colorization and Filtering of Segments After classification, each component received a specific color according to its category. In particular, segments classified as **large** were highlighted in red, as they represent more complete and structurally consistent roads. Medium and small segments were considered less relevant for subsequent analysis. Below, Figure 22 shows this colorization of segments.

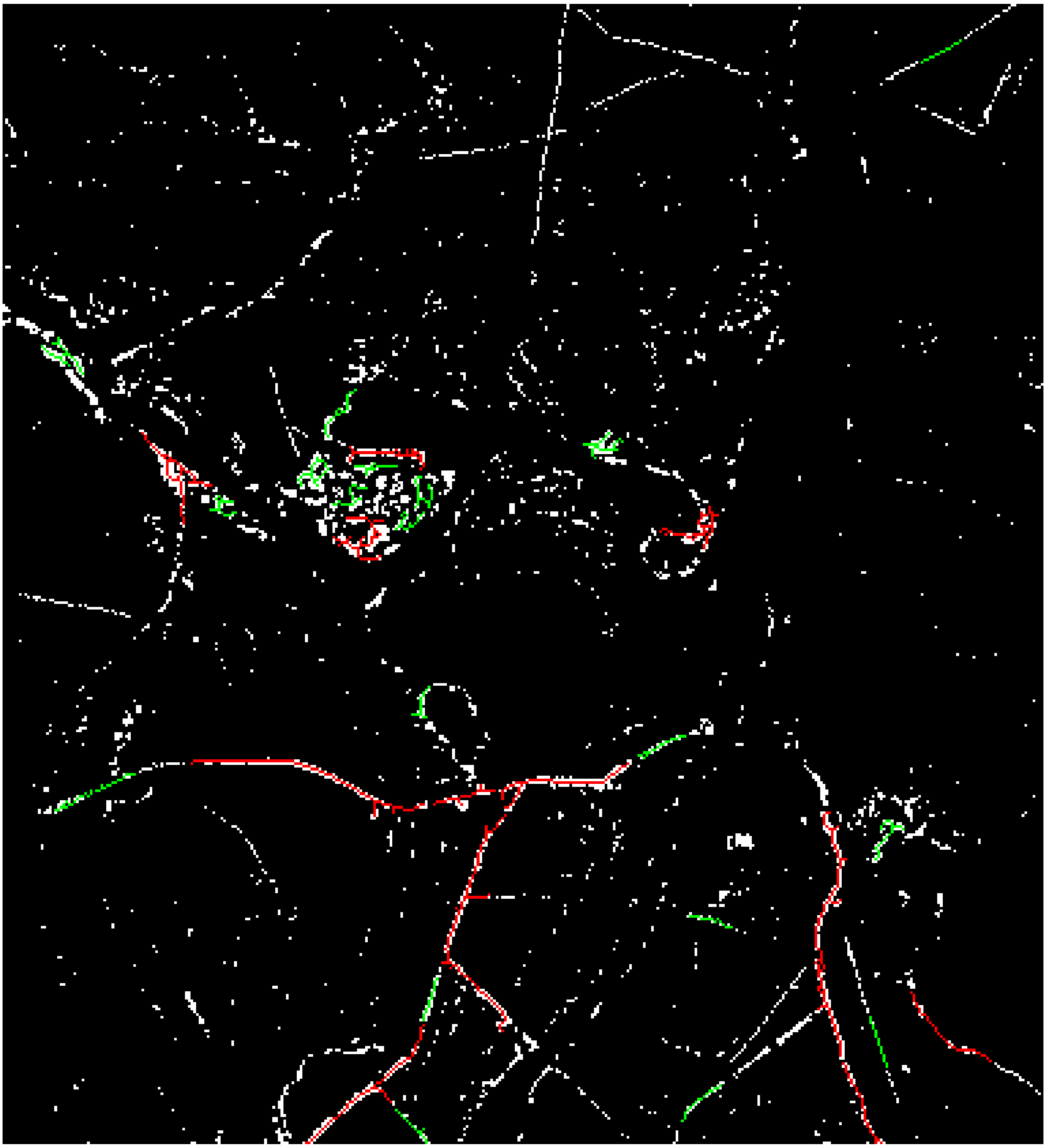


Figure 22: Mask produced from the Machine Learning model, involving PCA and colorization in relation to the size of pixels followed. Source: Own authorship.

Next, all components that did not belong to the large class were removed, resulting in an image containing exclusively red segments corresponding to extensive roads. Figure 23 below shows only the red segments.

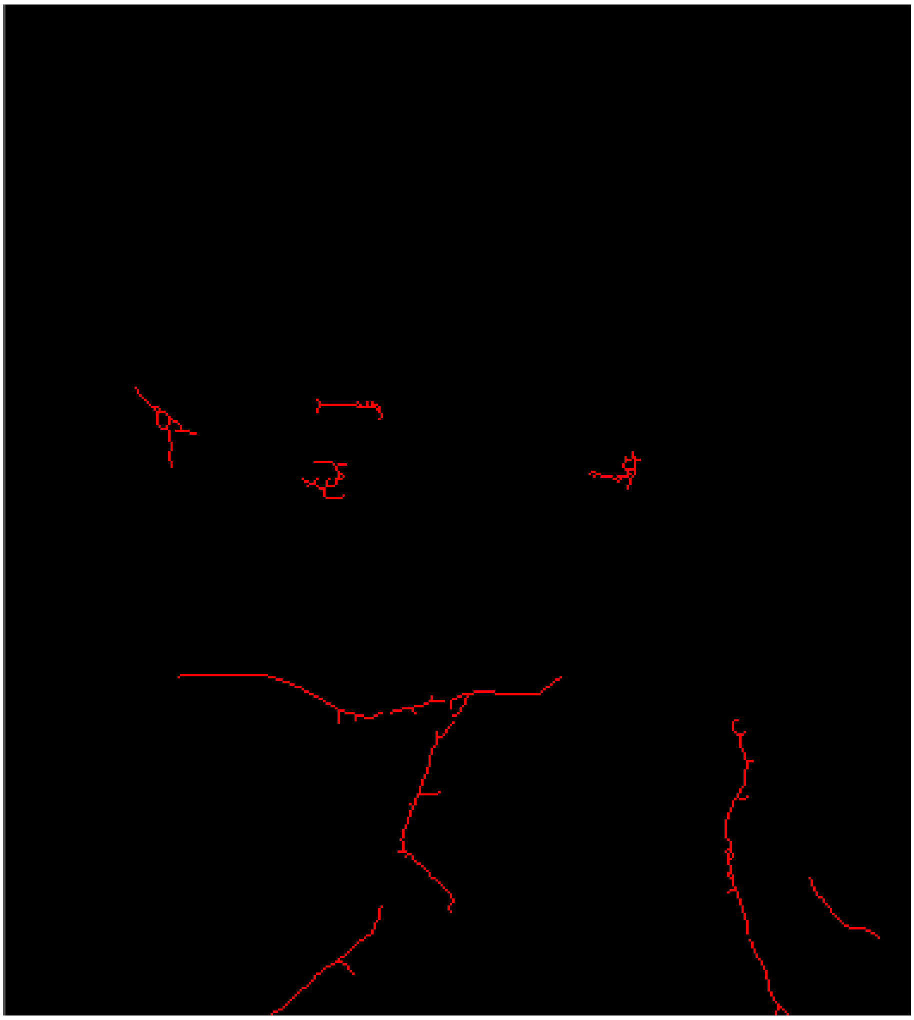


Figure 23: Selection of the largest continuous segments of identified roads. Source: Own authorship.

3.8 COUNTING AND STATISTICS GENERATION

Finally, the mask containing only the red segments was submitted to a process of component and pixel counting. This step produces consolidated metrics on:

These statistics were recorded in CSV files, allowing quantitative comparison between images and between different model configurations.

3.9 STATISTICAL ANALYSIS AND RISK DEFINITION

To investigate the hypothesis of anthropogenic influence on fires, a segmented statistical approach was adopted. The climatic control variable used was the Ångström Index (B).

The cutoff value $B > 2.5$ was used to filter days of "Climatic Safety", that is, periods where humidity and temperature did not favor natural fire propagation [82].

To quantify the relationship between road infrastructure and fire occurrence on these specific days, the **Spearman Correlation Coefficient** (ρ) was chosen.

3.10 STATISTICAL ANALYSIS: THE ANTHROPOGENIC FACTOR AND ROAD INFRASTRUCTURE

To understand the influence of road infrastructure on fire occurrence, a segmented statistical analysis was performed, using the Ångström Index (B) as a climatic control filter. The objective was to isolate fire events where meteorological conditions were not favorable to natural propagation ($B > 2.5$), thus highlighting the human factor. The correlated variables were:

- **Independent Variable:** Quantity of Roads (Total pixels of roads detected in the area).
- **Dependent Variable:** Fire Intensity (Total red pixels identified in the burned area).
- **Control Variable:** Land Use (Forest, Pasture, Agricultural) to identify the motivation for burning.

Results of Anthropogenic Analysis Unlike the hypothesis of spontaneous ignition by dry climate, the application of the Spearman correlation coefficient (ρ) suitable for identifying monotonic trends in nonlinear data revealed a strong positive association between the presence of roads and the magnitude of fires, even under conditions of high humidity.

- **Low Climatic Risk Scenario ($B > 2.5$)**: An extremely high Spearman correlation of $\rho = 0.9088$ was observed between the quantity of roads and the intensity of the burned area.
- **Motivation by Land Use**: When segmented by vegetation type, the **Forest** category presented the highest positive correlation ($\rho \approx 0.50$), while Pasture areas presented null or negative correlation, indicating that road infrastructure acts mainly as a vector for suppression of native vegetation (deforestation) and not just pasture management.

This result suggests that in areas with higher density of road infrastructure, the fire risk is determined predominantly by human accessibility (anthropogenic factor), overcoming the barriers imposed by meteorological variables.

4

4 RESULTS

4.1 IDENTIFICATION AND CORRECTION OF CLASS IMBALANCE

The initial analysis of the 12 binary masks revealed an extreme imbalance between classes. Since roads were represented by very thin lines, there was a very limited quantity of pixels belonging to the positive class. When considering exclusively the labeled images, 2,388,146 pixels were counted, of which only 28,594 corresponded to the “road” class. Thus, the minority class represented approximately 1.20% of the total, establishing a ratio of approximately 1:82 between positive and negative pixels. This initial distribution already indicated the need for specific interventions to enable training and avoid overthinking [87].

- **2,388,146 total pixels**;
- **28,594 positive** (roads);
- **2,359,552 negative**.

To mitigate this problem, two complementary correction strategies were adopted. The first consisted of the application of morphological dilation to the road masks. This operation expanded each positive pixel to its immediate neighbors, connecting discontinuous stretches and giving greater thickness to road segments. Dilation reduced the sparsity of the minority class and produced a more faithful representation of the spatial area occupied by roads.

Despite the gains obtained with dilation, the negative class remained largely dominant. Thus, a second correction strategy was implemented: random undersampling of non-road pixels. For each positive pixel, only four negative pixels were retained, obtaining a final ratio of 1:4 between classes. This approach proved adequate to preserve the variability of the majority class without inhibiting learning about the minority class [87].

At the end of the balancing process, all 28,594 positive pixels were preserved, while negative pixels were reduced to 114,376, totaling 142,970 pixels in the balanced dataset.

The balanced dataset used in training now had:

- **28,594 positive**;

- **114,376 negative**.

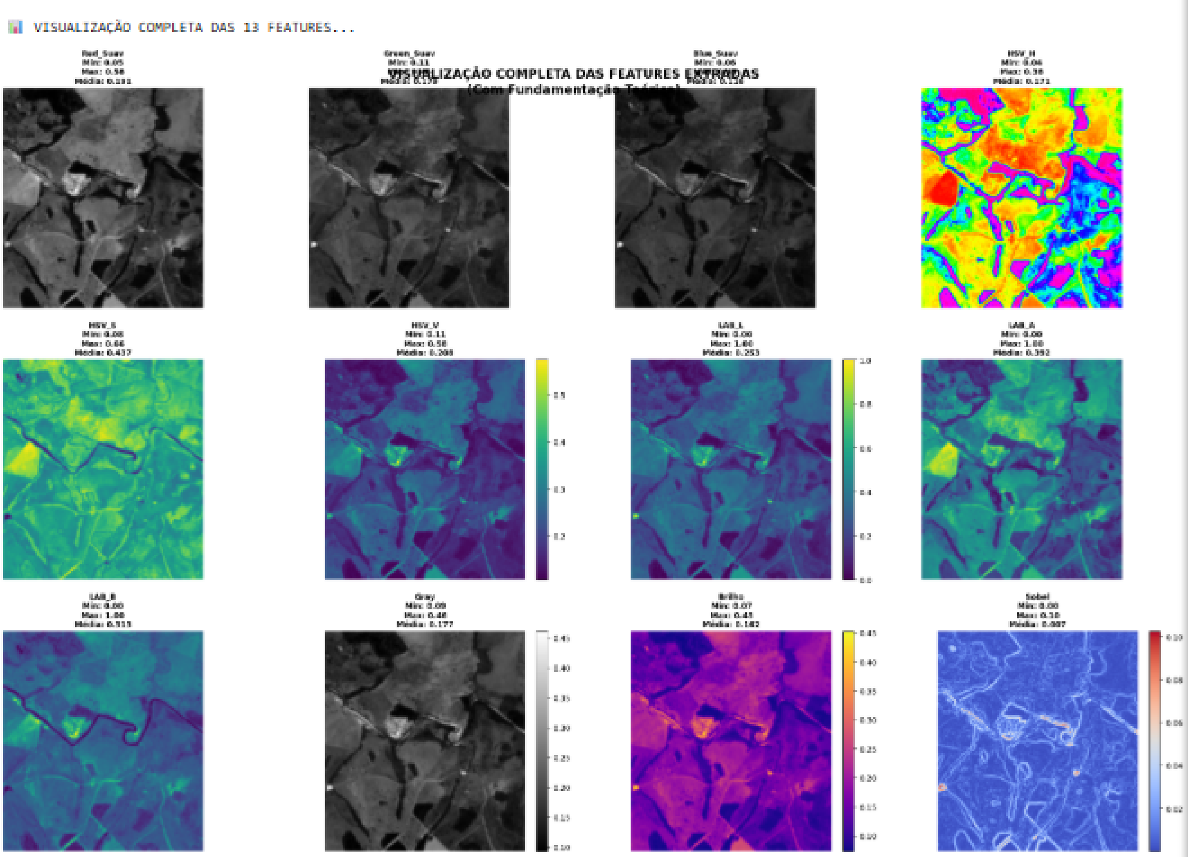
4.2 CORRELATION BETWEEN FEATURES AND ROAD PRESENCE

Before training the models, an exploratory analysis was performed to identify which attributes presented the greatest linear relationship with the “road” class. Table 3 presents the correlation ranking between each feature and positive pixels.

Table 3: Ranking of correlation between features and road pixels

Feature	**Correlation**
LAB_B (blue–yellow axis)	+0.0690
Red_Ratio	+0.0522
LAB_A (green–magenta axis)	+0.0469
Red_Smooth	+0.0465
HSV_V (value)	+0.0414
Gray	+0.0386
Brightness	+0.0384
LAB_L (luminosity)	+0.0383
HSV_S (saturation)	+0.0365
Green_Smooth	+0.0345
Sobel (gradient)	+0.0259
Blue_Smooth	+0.0231
HSV_H (hue)	-0.0491

The figure 24 below shows the color spaces of an image collected where a fire focus occurred.

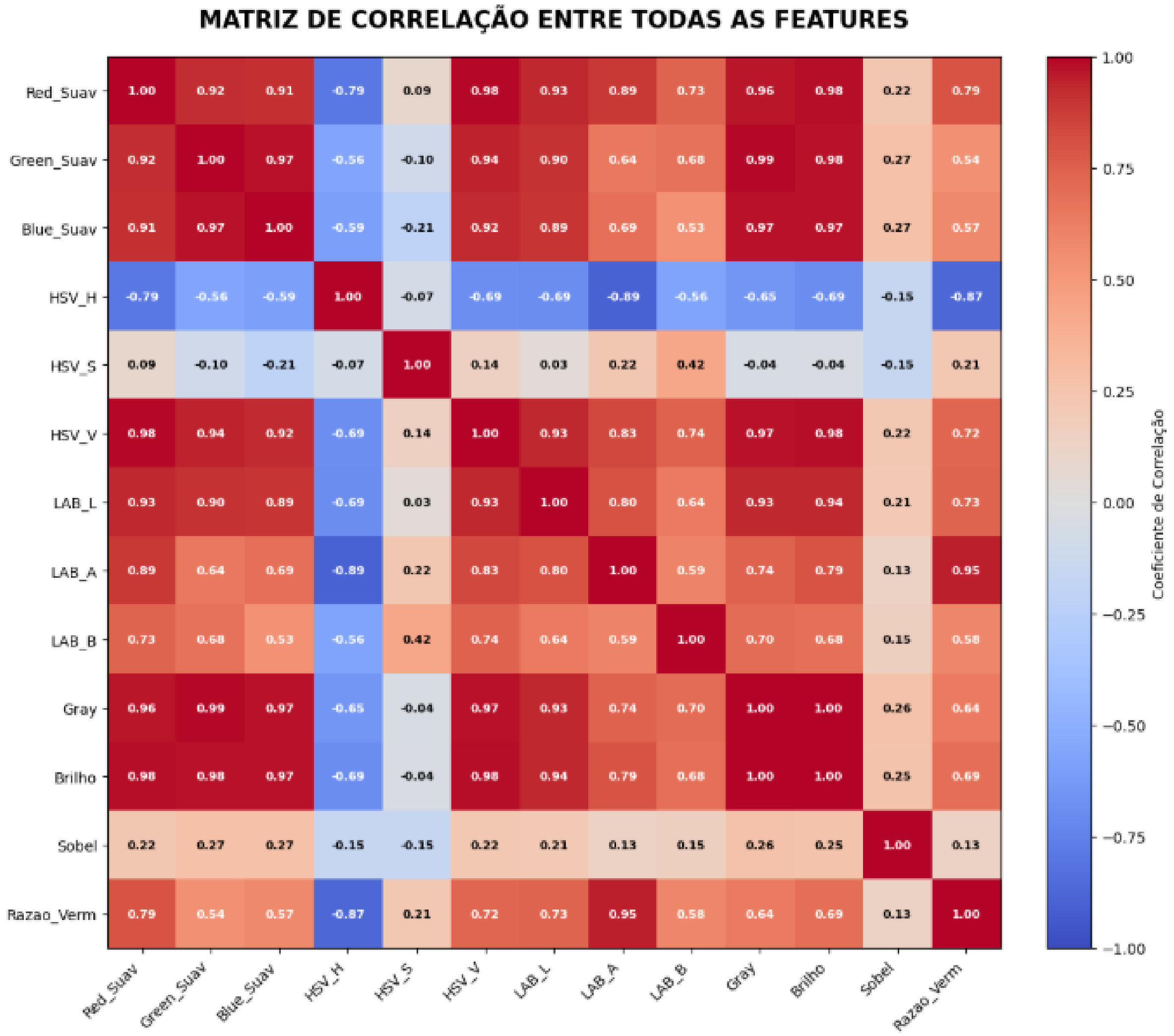


📊 VISUALIZAÇÃO COMPLETA DAS 13 FEATURES...

VISUALIZAÇÃO COMPLETA DAS FEATURES EXTRAÍDAS
(Com Fundamentação Teórica)
Red_Suav Min: 0.05 Max: 0.58 Média: 0.191
Green_Suav Min: 0.11 Média: 0.179
Blue_Suav Min: 0.06 Média: 0.116
HSV_H Min: 0.04 Max: 0.98 Média: 0.171
HSV_S Min: 0.08 Max: 0.66 Média: 0.437
HSV_V Min: 0.11 Max: 0.58 Média: 0.208
LAB_L Min: 0.00 Max: 1.00 Média: 0.253
LAB_A Min: 0.00 Max: 1.00 Média: 0.392
LAB_B Min: 0.00 Max: 1.00 Média: 0.515
Gray Min: 0.09 Max: 0.48 Média: 0.177
Brilho Min: 0.07 Max: 0.45 Média: 0.182
Sobel Min: 0.00 Max: 0.10 Média: 0.007

Figure 24: Color spaces of a region where a fire focus occurred in the Cerrado. Source: Own authorship.

Although correlations are moderate, LAB color space features (mainly LAB_B) stood out as more informative, reinforcing their usefulness in pavement discrimination. Meanwhile, hue (HSV_H) showed negative correlation, consistent with the fact that roads exhibit low chromatic variation. Below, Figure 25 shows more detailed correlations.



	Red_Suav	Green_Suav	Blue_Suav	HSV_H	HSV_S	HSV_V	LAB_L	LAB_A	LAB_B	Gray	Brilho	Sobel	Razao_Verm
Red_Suav	1.00	0.92	0.91	-0.79	0.09	0.98	0.93	0.89	0.73	0.96	0.98	0.22	0.79
Green_Suav	0.92	1.00	0.97	-0.56	-0.10	0.94	0.90	0.64	0.68	0.99	0.98	0.27	0.54
Blue_Suav	0.91	0.97	1.00	-0.59	-0.21	0.92	0.89	0.69	0.53	0.97	0.97	0.27	0.57
HSV_H	-0.79	-0.56	-0.59	1.00	-0.07	-0.69	-0.69	-0.89	-0.56	-0.65	-0.69	-0.15	-0.87
HSV_S	0.09	-0.10	-0.21	-0.07	1.00	0.14	0.03	0.22	0.42	-0.04	-0.04	-0.15	0.21
HSV_V	0.98	0.94	0.92	-0.69	0.14	1.00	0.93	0.83	0.74	0.97	0.98	0.22	0.72
LAB_L	0.93	0.90	0.89	-0.69	0.03	0.93	1.00	0.80	0.64	0.93	0.94	0.21	0.73
LAB_A	0.89	0.64	0.69	-0.89	0.22	0.83	0.80	1.00	0.59	0.74	0.79	0.13	0.95
LAB_B	0.73	0.68	0.53	-0.56	0.42	0.74	0.64	0.59	1.00	0.70	0.68	0.15	0.58
Gray	0.96	0.99	0.97	-0.65	-0.04	0.97	0.93	0.74	0.70	1.00	1.00	0.26	0.64
Brilho	0.98	0.98	0.97	-0.69	-0.04	0.98	0.94	0.79	0.68	1.00	1.00	0.25	0.69
Sobel	0.22	0.27	0.27	-0.15	-0.15	0.22	0.21	0.13	0.15	0.26	0.25	1.00	0.13
Razao_Verm	0.79	0.54	0.57	-0.87	0.21	0.72	0.73	0.95	0.58	0.64	0.69	0.13	1.00

Figure 25: Representation of color space correlations in more detail. Source: Own authorship.

In addition to features extracted from color spaces, an NGRDI index calculated directly from the original RGB images was incorporated. Since the scenes did not have an NIR band, an approximation based on the ratio between green and red bands was used. Although not representing the true NDVI, the index proved useful in reinforcing the separation between vegetated areas and anthropogenic surfaces, contributing to improved road discrimination. The figure below shows how Figure 26 RGB taken from Sentinel Hub transformed into NDVI.

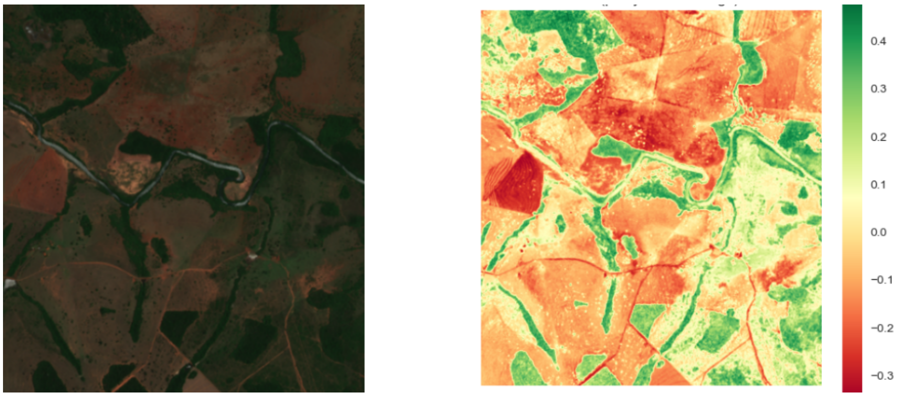


Figure 26: NGRDI, representing NDVI directly from RGB. Source: Own authorship.

4.3 PRINCIPAL COMPONENT ANALYSIS (PCA)

Before the training stage, a dimensional reduction analysis was also performed using PCA (Principal Component Analysis). The objective was to observe how the variance of features is distributed among principal components and to evaluate whether most of the relevant information could be compressed into few dimensions [69]. The figures below present the variance explained by component and the projection of samples on the first two components, evidencing partial separability between road and non-road pixels. Below, Figure 27 and 28 show the 3 PCA components and the overlap with roads binarized by QGIS.

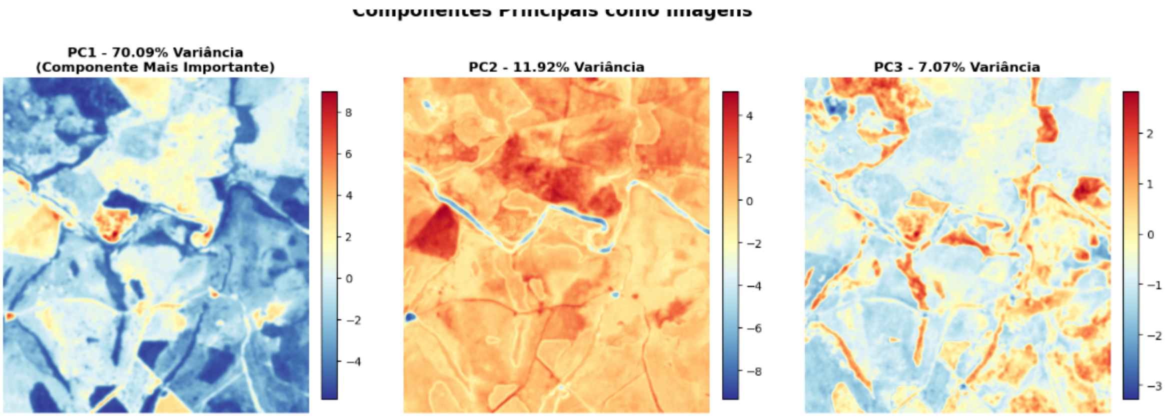


Figure 27: 3 Components of PCA. Source: Own authorship.

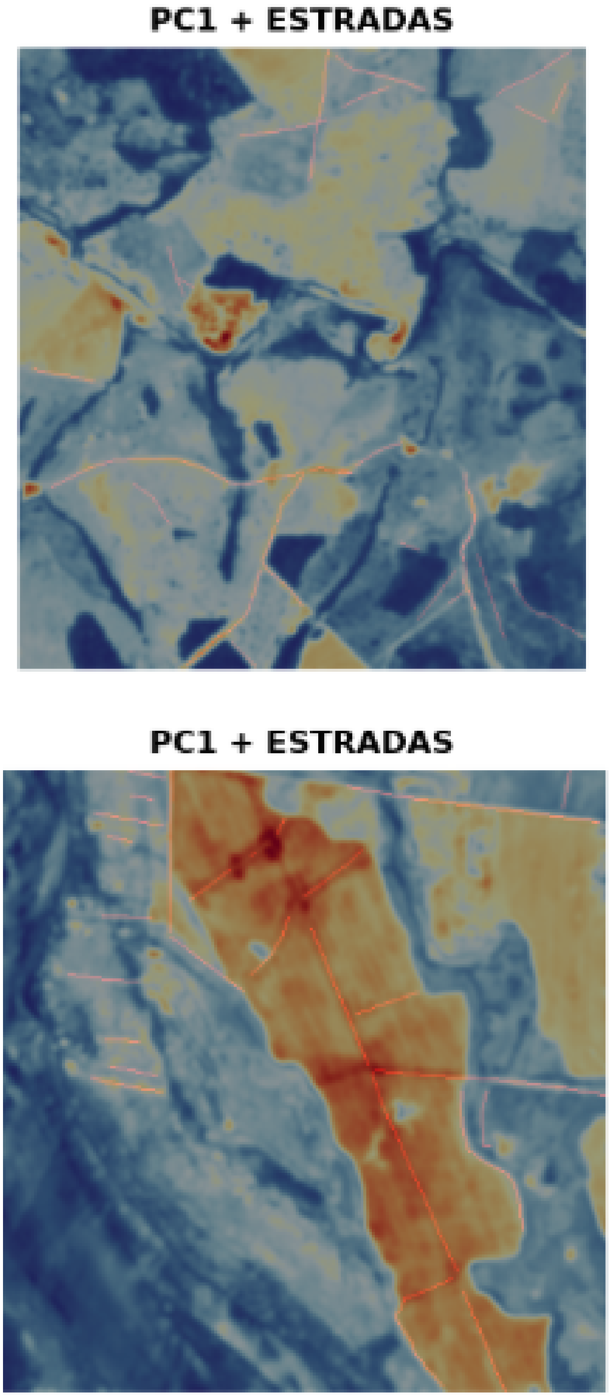


Figure 28: Mask overlaid by PCA in two locations where fire focuses occurred. Source: Own authorship.

4.4 PERFORMANCE OF SEGMENTATION MODELS

4.4.1 Model Configuration

Multiple hyperparameter variations were evaluated for four supervised algorithms:

Random Forest, KNN, Decision Tree, and MLP, totaling 16 distinct configurations. The objective was to investigate model behavior under different decision structures and complexity levels.

4.5 EXPERIMENTAL RESULTS

To evaluate the impact of feature selection and dimensionality reduction on road detection, experiments were divided into two major scenarios: (A) Datasets without dimensionality reduction (Raw Data) and (B) Datasets submitted to PCA with the addition of advanced descriptors.

4.5.1 Definition of Test Scenarios

To isolate the contribution of different types of information, the following attribute packages were defined:

- **Scenarios without PCA (Raw Data):** `COLOR_ALL` (9 color channels), `COLOR_ALL+Gray` and `FULL` (Color + Sobel).
- **Scenarios with PCA (Compressed Data):** `PCA_BASIC` (RGB), `PCA_CONTEXT` (RGB + Texture) and `PCA_FULL` (RGB/LAB/HSV + Texture + **NDVI** + **LBP**).

4.5.2 Individual Results: Scenario Without PCA

In this scenario, classifiers dealt with raw and highly correlated data (e.g., RGB and LAB channels have high redundancy). Table 4 highlights the best performances by algorithm family.

Table 4: Individual Performance - Scenario WITHOUT PCA (Raw Data)

Model & Combo	Accuracy	Precision	Recall	F1-Score
RandomForest_V2 (COLOR_ALL)	0.6675	0.3191	**0.5847**	**0.4129**
RandomForest_V2 (FULL)	0.6750	0.3193	0.5525	0.4047
DecisionTree_V1 (FULL)	0.6645	0.3095	0.5506	0.3963
KNN_V1 (COLOR_ALL+Gray)	0.7647	0.3621	0.2320	0.2828
MLP_V2 (COLOR_ALL)	0.8037	0.6190	0.0479	0.0889

The best individual result was obtained by Random Forest on the `COLOR_ALL` dataset (F1 = 0.4129). It is worth highlighting the behavior of the other algorithms:

- **KNN:** Suffered from the high dimensionality of raw data, presenting low Recall (≈ 0.23).
- **MLP (Neural Networks):** Although it presented high accuracy ($\approx 80\%$), the model failed drastically in detecting the class of interest (Road), with a Recall

of only 0.04. In other words, the network learned to classify almost everything as “Non-Road” to maximize overall accuracy, making it useless for the problem.

4.5.3 Individual Results: Scenario With PCA

At this stage, models used principal components extracted from richer features (including NDVI and LBP). Table 5 presents the results.

Table 5: Individual Performance - Scenario WITH PCA (Advanced Features)

Model & Combo	Accuracy	Precision	Recall	F1-Score
RandomForest_V2 (PCA_FULL)	0.7306	0.4038	**0.7284**	**0.5195**
RandomForest_V2 (PCA_CONTEXT)	0.7260	0.3981	0.7228	0.5134
DecisionTree_V1 (PCA_CONTEXT)	0.6548	0.3405	0.7753	0.4732
KNN_V1 (PCA_CONTEXT)	0.7954	0.4852	0.3775	0.4247
MLP_V2 (PCA_FULL)	0.8181	0.5992	0.2726	0.3747

There was a significant leap in quality. Random Forest in the PCA_FULL combo achieved the best individual F1-Score (0.5195). The application of PCA benefited all classifier families:

- KNN raised its F1-Score from 0.28 to 0.42, proving that noise reduction helped in distance calculation.
- MLP improved its F1 from 0.08 to 0.37, finally managing to learn patterns of the Road class, thanks to the cleaner variance explained delivered by PCA.

4.5.4 Final Comparison: AdaBoost Ensemble

For the final decision, the AdaBoost meta-classifier was used to combine predictions from previous models. Table 6 compares the effectiveness of the two scenarios in detecting Class 1 (Road).

Table 6: Comparison of Final Ensembles (Stacking with AdaBoost)

Strategy	Accuracy	Precision	Recall	F1-Score
Ensemble PCA	**0.8441**	**0.6615**	**0.4526**	**0.5375**
Ensemble No-PCA	0.8097	0.5795	0.1789	0.2735

The results demonstrate the superiority of the PCA + Advanced Features approach.

- **Failure of Ensemble without PCA:** Despite an accuracy of 80%, Recall was only 0.1789. This means the model missed over 80% of actual roads, incorrectly classifying them as vegetation. The lack of descriptors like NDVI prevented correct distinction.

- **Robustness of PCA Ensemble:** The model achieved an F1-Score of 0.5375, with a Precision of 0.66. This indicates that when the model points to a road, it has a high probability of being correct, maintaining a detection level (Recall of 0.45) almost three times higher than the model without PCA.

The combination of dimensionality reduction with vegetation attributes (NDVI) and texture (LBP) proved essential to resolve the spectral ambiguity between dirt roads and exposed soil in the Cerrado.

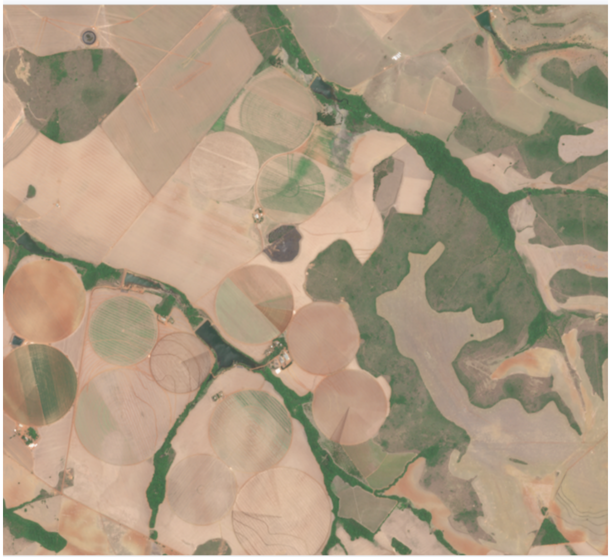


Figure 29: Original image, taken from Sentinel Hub of a fire focus event. Source: Own authorship.

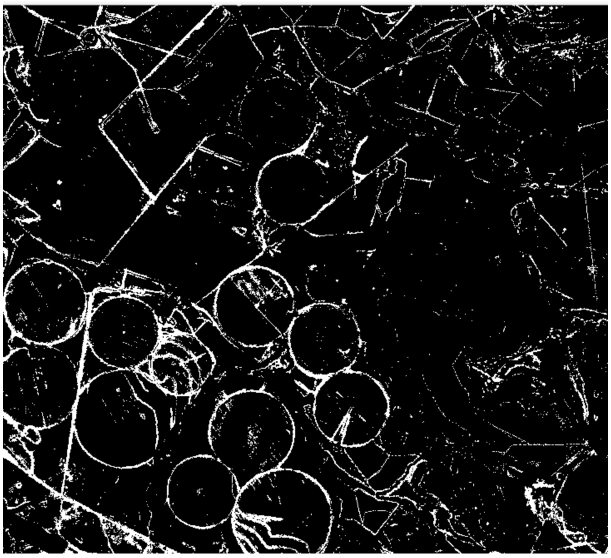


Figure 30: Mask produced by the predictive model, identifying roads automatically. Source: Own authorship.

In addition to evaluating model performance, a post-processing stage was applied to organize and quantify detected roads. First, the segmented masks underwent skeletonization, reducing each road to its central axis. Next, connected regions were grouped and classified according to trace size (small, medium, and large). These groups were represented by distinct colors: white, green, and red, respectively, allowing clear visualization of length distribution. Finally, only structures classified as large (in red) were selected for final quantitative analysis.

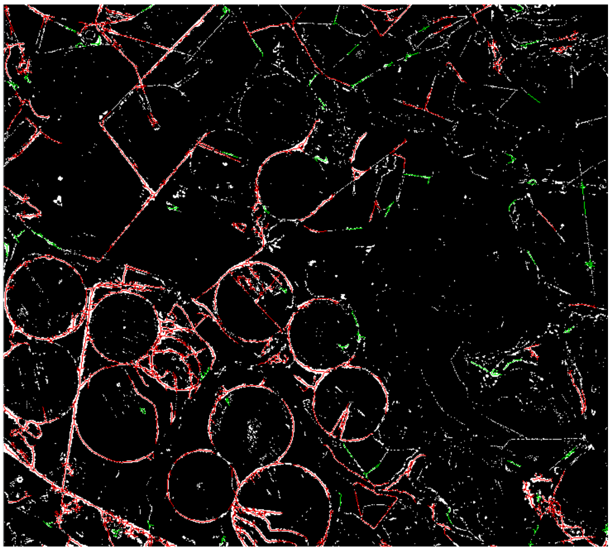


Figure 31: Categorization of skeletonization sizes, where red are large, green are medium, and white are small. Source: Own authorship.

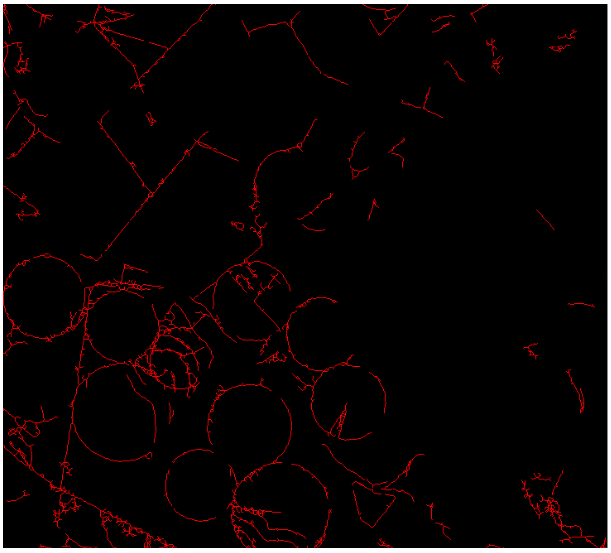


Figure 32: Selection of skeletonization defined as large. Source: Own authorship.

4.6 ANTHROPOGENIC INFLUENCE UNDER CLIMATIC SAFETY CONDITIONS

To investigate the relationship between road infrastructure and fire occurrence, the analysis focused strictly on events classified as “Low Natural Risk” (Ångström Index > 2.5). This approach allowed isolating the human factor, removing statistical noise caused by extremely dry days where ignition could be spontaneous or facilitated only by wind.

Verifying the dispersion of data in relation to the categorization of the Ångström Index, it is clear that there are more fire cases when the index result falls between 1-3. As Figure 33 below illustrates.

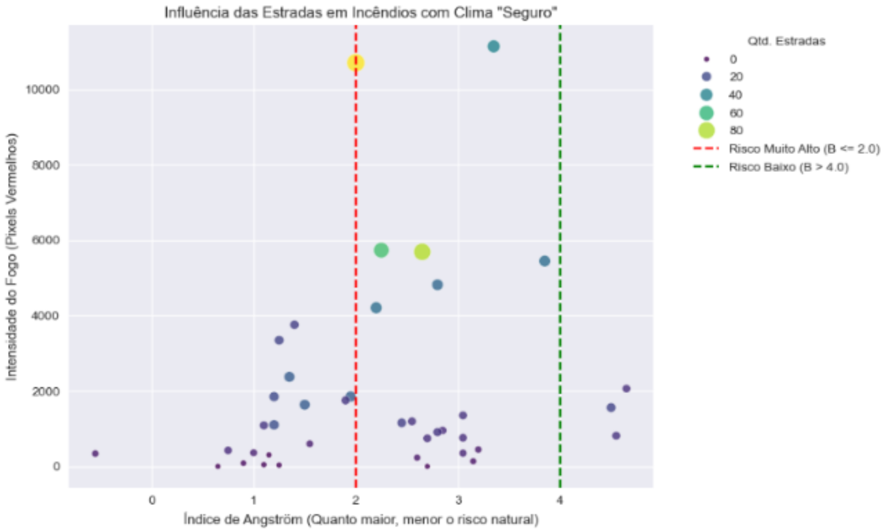


Figure 33: Dispersion of fire cases with the Ångström Index. Source: Own authorship.

Correlation Between Roads and Fire Intensity The application of the Spearman correlation coefficient (ρ) revealed an extremely strong positive association between the presence of roads and the magnitude of burns on these humid days.

The results indicate $\rho = 0.9088$ (p-value < 0.001), demonstrating that road density is the main predictor of fire intensity when climate acts as a natural barrier. Figure 34 illustrates this trend, where exponential growth of focus intensity is observed proportionally to the increase in detected road density.

Scenario 1: Synergy in High Climatic Risk ($B \leq 2.5$) In this scenario, characterized by low humidity and high temperature, Pearson’s linear correlation reached the striking value of **r = 0.9434**. This result indicates a destructive synergy: road infrastructure provides the ignition point and access, while extreme meteorological conditions potentiate immediate and linear fire propagation. The road acts here as the “trigger” in an environment already prone to combustion.

Scenario 2: Anthropogenic Forcing in Low Risk ($B > 2.5$) On days when climate imposed barriers to natural propagation (higher humidity), the trend correlation (Spearman) remained extremely high ($\rho = 0.9088$). Unlike what would be expected for natural causes where rain/humidity would drastically reduce focus, the maintenance of this high correlation confirms that the road is the determining factor for fire occurrence.

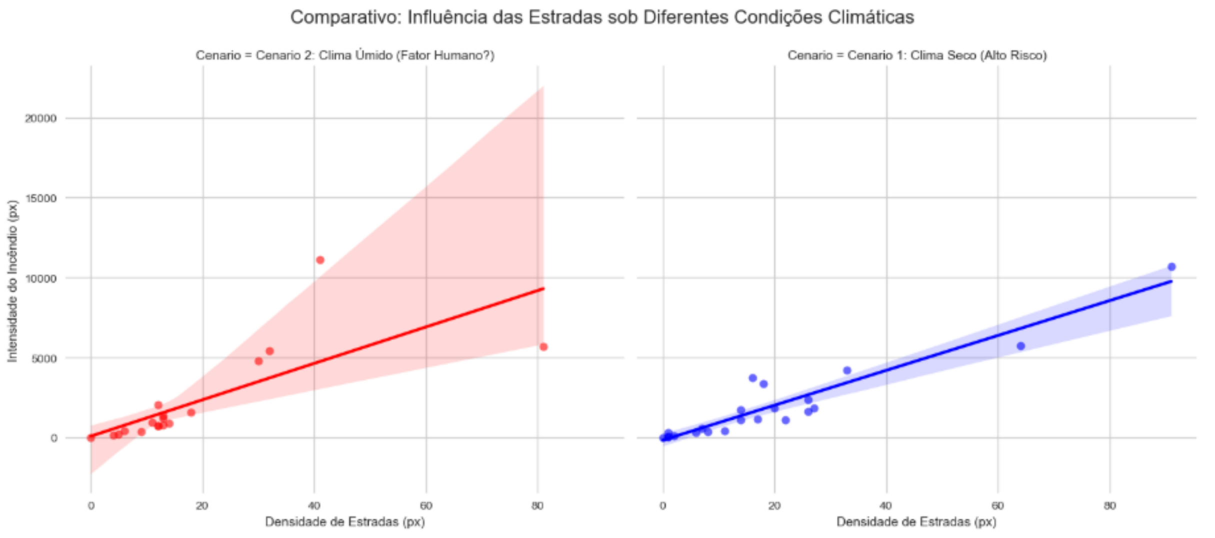
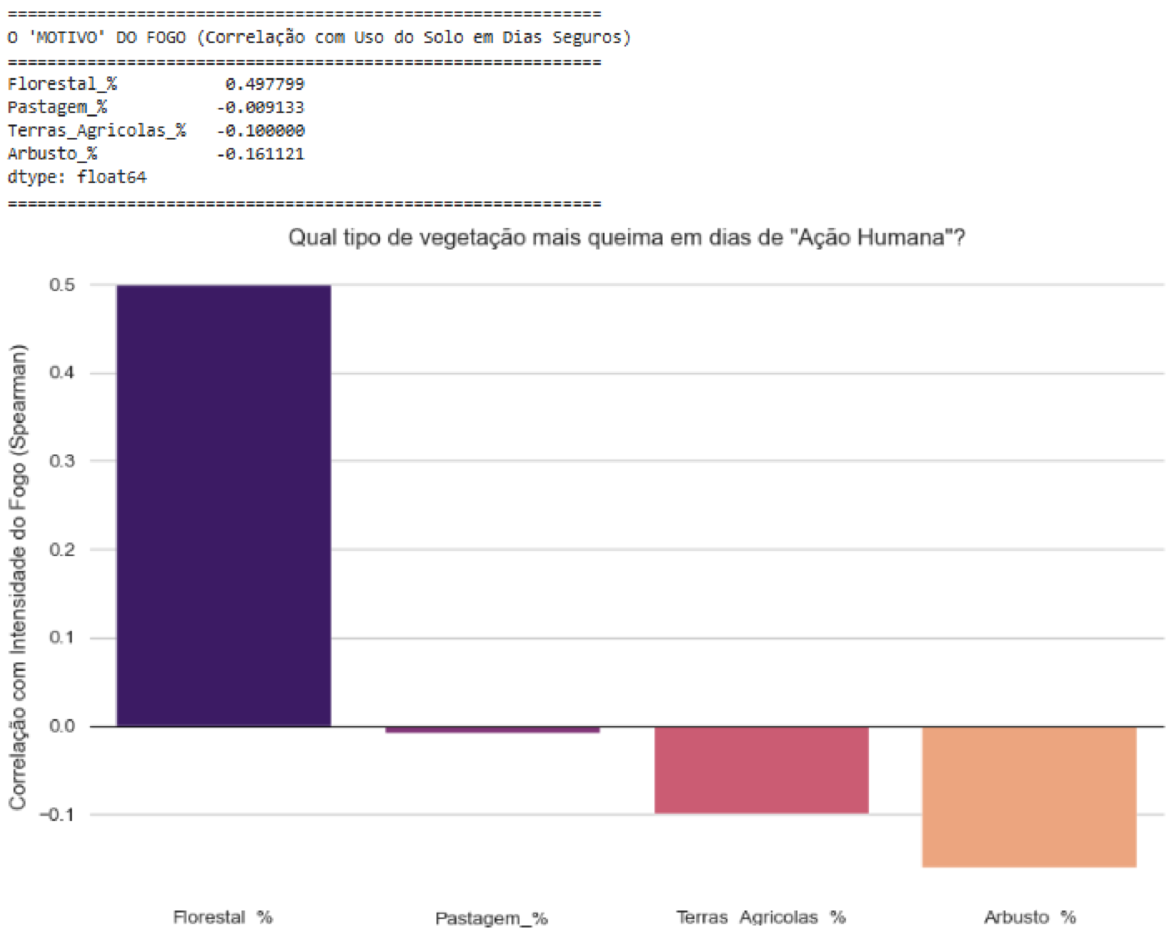


Figure 34: Two correlation scenarios between the Ångström index (high and low) with the total of roads. Source: Own authorship.

Motivation by Land Use By crossing fire intensity with land use and cover classes, the probable motivation for these anthropogenic ignitions was identified. The Forest class showed the highest positive correlation ($\rho \approx 0.50$), while consolidated use areas (Pasture and Agriculture) exhibited correlations near zero or negative. This pattern indicates a more evident association between the presence of detected road infrastructure and the occurrence of fires in native vegetation regions, including in periods of higher humidity.

Figure 35 presents the ranking of Spearman correlation by vegetation category.



```
==========================================================
O 'MOTIVO' DO FOGO (Correlação com Uso do Solo em Dias Seguros)
==========================================================
Florestal_%           0.497799
Pastagem_%           -0.009133
Terras_Agricolas_%   -0.100000
Arbusto_%            -0.161121
dtype: float64
==========================================================
```

Figure 35: Ranking of the largest fire cases when the Ångström index was not favorable for fires.

4.7 SYNTHESIS OF MAIN RESULTS

The integrated evaluation of the detection model and fire risk analysis allows highlighting four conclusive points:

- **The complete ensemble presented the best performance in detection**, achieving an F1-score of 0.9926.
- **The PCA ensemble demonstrated computational efficiency**, maintaining almost equivalent performance with only 4 principal components.
- **LAB confirmed itself as the most relevant color space**, validating its suitability for segmentation of paved surfaces and exposed soils.
- **Road infrastructure confirmed itself as a critical vector for anthropogenic fires**. The Spearman correlation of 0.90 on days of safe climate, associated with the predominance of burns in forest areas, statistically evidences the use of roads for deforestation activities in the studied region.

4.7.1 Limitations

Despite the statistical robustness observed in the correlations, the study presents limitations inherent to the defined scope:

- **Event Sampling:** The temporal and spatial cutoff resulted in a focused dataset (N=40), which, although statistically significant for validating the proposed methodology, suggests caution in immediate generalization to other biomes without new data.
- **Spectral Resolution:** The absence of the near-infrared band (NIR) in the input dataset limited the use of traditional vegetation indices (such as NDVI) to refine pre-fire biomass classification.
- **Spatial Resolution:** The detection of very narrow vicinal roads or trails under the forest canopy may be underestimated depending on the resolution of the satellite images used.

Even with these limitations, the methodology proved the hypothesis that road expansion and fire are intrinsically linked to land use conversion processes, regardless of climatic conditions.

5

5 CONCLUSION

The results indicate the presence of a strong correlation between road density and fire occurrence. The segmented analysis suggests that road infrastructure may be associated with different dynamics: on dry days, it tends to coincide with greater fire spread (Pearson ≈ 0.94), while on humid days it shows a stronger association with ignition patterns (Spearman ≈ 0.90). It was also observed that, even under climatic conditions less favorable to combustion, fire incidence remains concentrated in forested areas, pointing to possible relationships between the presence of roads and changes in land use and land cover.

From a methodological perspective, the development of an automatic road detection system represented a relevant step in this study. The use of an Ensemble architecture, combined with data balancing techniques and morphological post-processing, enabled the segmentation of road structures in satellite imagery, although with notable limitations. The Adaboost based model, integrating classifiers such as KNN, Decision Tree, Random Forest, and MLP, achieved an F1-score of 0.5375, indicating moderate performance. This result suggests that, while the approach is promising, there is considerable room for improvement, particularly through deeper hyperparameter tuning and the expansion and refinement of the training dataset (ground truth).

The integration of segmentation outputs with meteorological information from the OroraTech platform enabled a more comprehensive analysis of fire events by simultaneously incorporating anthropogenic and environmental factors. This combination contributed to a more detailed understanding of the interactions influencing fire occurrence in areas at the interface between natural environments and human-modified regions.

Despite the study's limitations—such as the relatively small sample size, the moderate performance of the segmentation model, and the lack of complete multispectral data the results provide relevant evidence of anthropogenic influence on the observed fire patterns. The identified correlations, although not implying direct causality, reinforce the importance of considering variables related to road infrastructure, particularly in forest frontier regions, within fire risk prevention and management strategies.

Potential practical applications include supporting territorial planning and environmental management in the Cerrado biome. The proposed methodology may be incorporated into early warning systems, assisting in the identification of areas with higher probabilities of ignition or increased anthropogenic pressure on vegetation. The findings may also inform land use planning policies that account for the impacts associated with road expansion, especially unofficial or informal roads.

For future studies, it is recommended to expand the ground truth dataset, perform more extensive hyperparameter optimization, incorporate multispectral imagery for the calculation of indices such as NDVI, and evaluate more advanced post-processing techniques. Field validation and the replication of the approach in other Brazilian biomes may further enhance the robustness and contextual relevance of the results.

In summary, this work contributes to the understanding of factors associated with wildfires in the Cerrado, suggesting that the presence of roads may be linked to land use changes and fire occurrence dynamics. Although the segmentation model still presents limitations, the proposed approach demonstrates potential and can serve as a foundation for the development of more robust tools aimed at monitoring and prevention within environmental conservation and management initiatives.

REFERENCES

References

[1] Cerratinga. (2025) Bioma cerrado. [Online]. Available: https://www.cerratinga.org.br/biomas/cerrado/

[2] GIMP Documentation Team. (2014) Filtro sobel. [Online]. Available: https://docs.gimp.org/2.8/pt_BR/plug-in-sobel.html

[3] Engecolor. (2020) Rgb, cmyk e pantone: você sabe o que significa? [Online]. Available: https://engecolornet.com.br/rgb-cmyk-pantone-voce-sabe-o-que-significa/

[4] L. Ghelal. (2021) Disco cromático hsv. [Online]. Available: https://lucasghelal.medium.com/disco-cromático-hsv-db3cd81c9a80

[5] Blog da Leart. (2014) Lab color: o espaço de cor. [Online]. Available: https://blogdaleart.wordpress.com/2014/09/04/lab-color-o-espaco-de-cor/

[6] Coptrz. (2022) Agritech drones, vegetation indices and aerial imagery. [Online]. Available: https://coptrz.com/blog/agritech-drones-vegetation-indices-and-aerial-imagery/

[7] LinkedIn Articles. (2018) Knn algorithm visualization. [Online]. Available: https://media.licdn.com/dms/image/v2/C4D12AQFKNJN7zSdVFA/article-cover_image-shrink_720_1280/article-cover_image-shrink_720_1280/0/1533837083964

[8] Homem Máquina. (2020) Como montar uma árvore de decisão. [Online]. Available: https://www.homemmaquina.com.br/como-montar-uma-arvore-de-decisao/

[9] Data Science Dojo. (2021) Random forest algorithm explained. [Online]. Available: https://datasciencedojo.com/blog/random-forest-algorithm/

[10] Medium: Data Science. (2018) From a single decision tree to a random forest. [Online]. Available: https://medium.com/data-science/from-a-single-decision-tree-to-a-random-forest-b9523be65147

[11] ResearchGate. (2019) Multi-layer perceptron diagram. [Online]. Available: https://www.researchgate.net/figure/Multi-Layer-Perceptron-MLP-diagram-with-four-hidden-layers-and-a-collection-of-single_fig1_334609713

[12] NLPCA Project. (2016) Principal component analysis. [Online]. Available: http://www.nlpca.org/pca_principal_component_analysis.html

[13] Pluralsight. (2022) Ensemble methods: Bagging vs boosting. [Online]. Available: https://www.pluralsight.com/resources/blog/guides/ensemble-methods-bagging-versus-boosting

[14] Data Hackers. (2019) Segmentação de imagens utilizando técnicas de thresholding. [Online]. Available: https://medium.com/data-hackers/segmentação-de-imagens-utilizando-técnicas-dethresholding-1ee031562c63

[15] Visão Computacional. (2020) Morfologia matemática: extração de fronteiras e detecção de bordas. [Online]. Available: https://visaocomputacional.com.br/morfologia-matematica-extracao-de-fronteiras-deteccao-de-bordas/

[16] Psicometria Online. (2021) O que é correlação de pearson? [Online]. Available: https://www.blog.psicometriaonline.com.br/o-que-e-correlacao-de-pearson/

[17] INPE – Instituto Nacional de Pesquisas Espaciais. (2025) Bdqueimadas – dashboard de monitoramento de queimadas. [Online]. Available: https://terrabrasilis.dpi.inpe.br/queimadas/bdqueimadas/

[18] ——. (2025) Identificação de incêndios no bdqueimadas. [Online]. Available: https://terrabrasilis.dpi.inpe.br/queimadas/bdqueimadas/

[19] SIPAM – Sistema de Proteção da Amazônia. (2025) Painel do fogo – monitoramento de incêndios. [Online]. Available: https://panorama.sipam.gov.br/painel-do-fogo/

[20] Ororatech. (2025) Ororatech – plataforma de monitoramento de incêndios. [Online]. Available: https://app.ororatech.com/

[21] ——. (2025) Exportação de dados geojson na plataforma ororatech. [Online]. Available: https://app.ororatech.com/

[22] Sentinel Hub. (2025) Importação de dados geojson no sentinel hub. [Online]. Available: https://www.sentinel-hub.com/

[23] A. C. de Oliveira and T. Sehn Körting, “A multi-temporal dataset for mapping burned areas in the brazilian cerrado using time series of remote sensing imagery,” *Big Earth Data*, pp. 1–32, 2025.

[24] G. de Oliveira, F. L. S. de Souza, L. O. Anderson, L. E. O. C. Aragão, and F. H. Wagner, “A Long-Term Landsat-Based Monthly Burned Area Dataset for the Brazilian Biomes Using Deep Learning,” *Remote Sensing*, vol. 12, no. 5, p. 793, 2020.

[25] J. B. de Jesus, C. N. da Rosa, Í. D. d. C. Barreto, and M. M. Fernandes, “Análise da incidência temporal, espacial e de tendência de fogo nos biomas e unidades de conservação do brasil,” *Ciência Florestal*, vol. 30, no. 1, pp. 176–191, apr 2020.

[26] S. Schettino, T. R. Souto, D. R. Soranso, and M. T. Mendes, “Monitoramento remoto como ferramenta para detecção de incêndios florestais,” in *Avanços nas Ciências Florestais*, A. M. Zuffo, Ed. Nova Xavantina, MT: Pantanal, 2022, vol. 2.

[27] N. d. F. Resende, “Cerrado: Ecologia, biodiversidade e preservação,” *Revista Brasileira de Educação e Cultura*, no. VI, pp. 81–90, 2012. [Online]. Available: http://www.periodicos.cesg.edu.br/index.php/educacaoecultura

[28] T. C. Parreiras and É. L. Bolfe, “Expansão e intensificação da agropecuária no cerrado,” in *Anais do Evento em Comemoração aos 20 Anos do Programa de Pós-Graduação em Geografia (IG-UNICAMP)*, vol. 1, no. 1, 2023, pp. 476–492.

[29] M. Costa, “Detecção de mudanças na cobertura vegetal natural do cerrado por meio de dados de radar,” Dissertação de Mestrado, Universidade de Brasília, Brasília, DF, 2008.

[30] V. R. Pivello, “Cerrado: o fogo como agente ecológico,” Empresa Brasileira de Pesquisa Agropecuária (Embrapa Cerrados), Planaltina, DF, Documentos 26, 1999.

[31] WWF-Brasil, “Fires in Brazilian Biomes,” World Wide Fund for Nature - Brasil, Technical Report, aug 2019.

[32] M. M. Bandeira, F. V. Freitas, K. C. A. Silva, P. H. S. Carneiro, T. L. B. Alves, and J. S. Santos, “Avaliação da distribuição de focos de calor às margens de rodovias federais no estado do ceará/brasil,” in *Anais do XVIII Simpósio Brasileiro de Sensoriamento Remoto (SBSR)*, Santos, SP, Brasil, 2017, pp. 3591–3598.

[33] B. Marçal and J. L. B. Albuquerque, “Fire in Savannas and its Impact on Avifauna: Considerations for a Better Environmental Conservation,” *Oecologia Australis*, vol. 18, no. 2, pp. 247–261, 2014.

[34] N. L. M. de Mello, H. O. K. de Mello, J. C. A. de Mello, F. L. de Oliveira, and É. S. de Mello, “Use of machine learning as a tool for determining fire management units in the Brazilian Atlantic Forest,” *Floresta e Ambiente*, vol. 28, no. 2, p. e20200021, 2021.

[35] I. d. S. Araújo, “Impacto da exposição à fumaça da queima de biomassa na floresta amazônica na saúde humana: uma revisão de escopo,” Trabalho de Conclusão de Curso (Especialização), Escola Nacional de Saúde Pública Sergio

Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, 2021. [Online]. Available: https://www.arca.fiocruz.br/handle/icict/52219

[36] D. Lee, S. Son, J. Bae, S. Park, J. Seo, D. Seo, Y. Lee, and J. Kim, “Single-Temporal Sentinel-2 for Analyzing Burned Area Detection Methods: A Study of 14 Cases in Republic of Korea Considering Land Cover,” *Remote Sensing*, vol. 16, no. 5, p. 884, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/5/884

[37] D. P. P. C. P. Allata, S. R. M. P. S. B. Rathnayaka, A. L. A. K. Ranaweera, W. G. C. W. Kumara, and B. H. Sudantha, “A New Log-Transform Histogram Equalization Technique for Deep Learning-Based Document Forgery Detection,” in *2023 International Conference on Computer, Information Technology and Electrical Engineering (ICCITEE)*. IEEE, 2023, pp. 1–6.

[38] D. P. Syrivelis, K. G. Koutsias, N. S. Bilios, and A. A. Argialas, “Analysis and Interpretation of Spectral Indices for Soft Multicriteria Burned-Area Mapping,” *Remote Sensing*, vol. 14, no. 19, p. 4997, 2022.

[39] H. Dibs, N. Al-Ansari, H. A. Hasab, and H. S. Jaber, “Automatic feature extraction and matching modelling for highly noise near-equatorial satellite images,” *Innovative Infrastructure Solutions*, vol. 7, no. 2, pp. 1–14, 2022.

[40] Y. Y. Bae, D. J. Cho, and K. H. Jung, “A new log-transform histogram equalization technique for deep learning-based document forgery detection,” *Symmetry*, vol. 17, 3 2025.

[41] J. Xing, R. Sieber, and T. Caelli, “A scale-invariant change detection method for land use/cover change research,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 141, pp. 252–264, 7 2018.

[42] N.-S. Du, Q. Weng, and F.-L. Li, “A Standardized Radiometric Normalization Method for Change Detection Using Remotely Sensed Imagery,” *Photogrammetric Engineering & Remote Sensing*, vol. 68, no. 2, pp. 173–181, 2002. [Online]. Available: https://www.asprs.org/wp-content/uploads/pers/2000journal/february/2000_feb_173-181.pdf

[43] S. B. Wali, M. A. Abdullah, M. A. Hannan, A. Hussain, S. A. Samad, P. J. Ker, and M. B. Mansor, “Vision-based traffic sign detection and recognition systems: Current trends and challenges,” 5 2019.

[44] O. Patel, Y. P. S. Maravi, and S. Sharma, “A comparative study of histogram equalization based image enhancement techniques for brightness preservation

and contrast enhancement," *Signal & Image Processing : An International Journal*, vol. 4, pp. 11–25, 11 2013.

[45] J. Al-Doski, S. B. Mansor1, H. Zulhaidi, and M. Shafri, "Image classification in remote sensing," vol. 3, 2013. [Online]. Available: www.iiste.org

[46] S. A. M. H. J. S. Abeywickrama, S. D. D. K. Suraweera, and R. A. A. K. Ranaweera, "Comparison of Image Segmentation using Different Color Spaces," in *2013 International Conference on Computer and Information Science (ICCIS)*. IEEE, 2013, pp. 145–150.

[47] T. Kawamura, "A Method for Automatic Target Tracking Based on the Likelihood Function," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-9, no. 1, pp. 62–66.

[48] A. Padrón, "Detección de bordes en datos sísmicos con filtros sobel," Master's thesis, Universidad Nacional de La Plata, La Plata, Argentina, Agosto 2022, directores: Dr. Julián Luis Gómez y Dr. Danilo Rubén Velis.

[49] N. Khediri, M. B. Ammar, and M. Kherallah, "Comparison of image segmentation using different color spaces," in *International Conference on Communication Technology Proceedings, ICCT*, vol. 2021-October. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 1188–1192.

[50] D. Lee, S. Son, J. Bae, S. Park, J. Seo, D. Seo, Y. Lee, and J. Kim, "Single-temporal sentinel-2 for analyzing burned area detection methods: A study of 14 cases in republic of korea considering land cover," *Remote Sensing*, vol. 16, no. 5, p. 884, 2024.

[51] A. S. Barros, L. M. d. Farias, and J. L. A. Marinho, "Aplicação do Índice de vegetação por diferença normalizada (ndvi) na caracterização da cobertura vegetativa de juazeiro do norte – ce," *Revista Brasileira de Geografia Física*, vol. 13, no. 6, pp. 2885–2895, 2020.

[52] J.-S. Eom, J. Gwak, H.-W. Jo, and W.-K. Lee, "Single-Temporal Sentinel-2 for Analyzing Burned Area Detection Methods: A Study of 14 Cases in Republic of Korea Considering Land Cover," *Remote Sensing*, vol. 14, no. 23, p. 6137, 2022.

[53] p. n. n. f. Viana, S. W. Souza, L. N. S. Girão, A. B. Bendahan, and V. d. Freitas, "Análise visual de índices de vegetação utilizando imagens rgb para classificação de áreas de pastagens com presença de plantas invasoras," *XXX*, pp. –, XXXX.

[54] E. A. Speranza, J. F. G. Antunes, L. A. F. Barbosa, G. M. d. A. Cançado, and J. C. Vansconcelos, "Importância de índices de vegetação para modelos de estimativa de produtividade em cana-de-açúcar," *XXX*, pp. –, XXXX.

[55] C. A. Ferrero, "Algoritmo knn para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia," Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, 2009.

[56] C. J. F. Rodrigues, "Desenvolvimento do algoritmo knn em elixir e benchmarking com diferentes implementações," Rio de Janeiro, 2021.

[57] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Boca Raton: Chapman & Hall/CRC, 1984.

[58] I. D. Mienye and N. Jere, "A survey of decision trees: Concepts, algorithms, and applications," *IEEE access*, 2024.

[59] Y. Coadou, "Boosted decision trees," in *Artificial Intelligence for High Energy Physics*, P. Calafiura, D. Rousseau, and K. Terao, Eds., pp. 9–58.

[60] I. D. Mienye and N. Jere, "A survey of decision trees: Concepts, algorithms, and applications," *IEEE Access*, 2017.

[61] J. R. Quinlan, "Induction of decision trees," Tech. Rep., 1986.

[62] S. B. Wali, M. A. Abdullah, M. A. Hannan, A. Hussain, S. A. Samad, P. J. Ker, and M. B. Mansor, "Vision-based traffic sign detection and recognition systems: Current trends and challenges," *Sensors*, vol. 19, no. 10, pp. 1–25, 2019.

[63] G. Zhang, Q. Chen, and Q. Sun, "Illumination normalization among multiple remote-sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, pp. 1470–1474, 2014.

[64] W. Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp. 14–23, 1 2011.

[65] R. P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, vol. 4, no. 2, pp. 4–22, 1987.

[66] Y. Gao, M. Li, and W. Sun, "Hardware Evolution Based on Improved Simulated Annealing Algorithm in Cyclone V FPSoC," in *2017 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*. IEEE, 2017, pp. 25–29.

[67] O. A. Montesinos-López, A. Montesinos-López, A. Martín-Montes, J. Cuevas, and E. Santana, “A Guide on Deep Learning for Complex Trait Genomic Prediction,” *Frontiers in Genetics*, vol. 13, p. 835948, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2022.835948/full

[68] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, “Activation functions in deep learning: A comprehensive survey and benchmark,” 6 2022. [Online]. Available: http://arxiv.org/abs/2109.14545

[69] Y. Qian, J. Shen, F. Zhang, B. Liu, and W. Su, “Optimizing Genetic Algorithms with Multilayer Perceptron Networks for Enhancing TinyFace Recognition,” *Electronics*, vol. 12, no. 10, p. 2296, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/10/2296

[70] N. S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” Tech. Rep., 1992.

[71] M. Pesaresi, M. Halkia, P. Potapov, S. Kuzemko, K. Böttcher, D. Roy, G. Vancauwenberghe, K. Van Tricht, and M. Santoro, “Monitoring Urban Areas with Sentinel-2A Data: Application to the Update of the Copernicus High Resolution Layer Imperviousness Degree,” *Remote Sensing*, vol. 8, no. 8, p. 693, 2016. [Online]. Available: https://www.mdpi.com/2072-4292/8/8/693

[72] B. B. Chaves, “Estudo do algoritmo adaboost de aprendizagem de máquina aplicado a sensores e sistemas embarcados,” Dissertação de mestrado, Universidade de São Paulo, São Paulo, Brasil, 2012, dissertação de Mestrado em Engenharia Mecânica.

[73] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 1 2017. [Online]. Available: http://arxiv.org/abs/1412.6980

[74] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[75] Q. Cao, L. Qingge, and P. Yang, “Performance analysis of otsu-based thresholding algorithms: A comparative study,” *Journal of Sensors*, vol. 2021, p. 4896853, 2021.

[76] L. Torok, “Método de otsu,” Notas de aula / material didático, Instituto de Computação, Universidade Federal Fluminense (UFF), 2016, disponível como slides/relatório didático (ex.: “Método de Otsu”, Análise de Imagens).

[77] C. R. Ribeiro and M. P. Zem, “Aplicativo para smartphone para medição da área de feridas demarcadas em folhas transparentes,” Ph.D. dissertation, Universidade

Tecnológica Federal do Paraná, Curitiba, Brasil, 2019, trabalho de Conclusão de Curso em Engenharia Eletrônica.

[78] D. B. Figueiredo Filho and J. A. d. Silva Júnior, “Desvendando os mistérios do coeficiente de correlação de pearson (r),” *Revista Política Hoje*, vol. 18, no. 1, pp. 115–133, 2009.

[79] H. A. Miot, “Análise de correlação em estudos clínicos e experimentais,” *Correlation analysis in clinical and experimental studies*, 2018, submetido em: 09 fev. 2018. Aceito em: 12 fev. 2018. Fonte de financiamento: nenhuma. Conflito de interesse: nenhum.

[80] C. A. Ferrero, “Algoritmo knn para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia,” Master’s thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, 2009, orientadora: Profa. Dra. Maria Carolina Monard. Trabalho desenvolvido com apoio do CEASB/FPTI-BR e do PDTA.

[81] S. A. Lira, “Análise de correlação: abordagem teórica e de construção dos coeficientes com aplicações,” Master’s thesis, Universidade Federal do Paraná, Curitiba, PR, 2004, orientador: Prof. Dr. Anselmo Chaves Neto.

[82] K. D. e. S. Dornelas, C. L. d. Silva, and C. A. d. S. Oliveira, “Coeficientes médios da equação de angström-prescott, radiação solar e evapotranspiração de referência em brasília,” *Pesquisa Agropecuária Brasileira*, vol. 41, no. 8, pp. 1213–1219, 2006.

[83] R. Dallacort, P. S. L. d. Freitas, A. C. A. Gonçalves, R. Rezende, A. Bertonha, F. F. d. Silva, and M. Trintinalha, “Determinação dos coeficientes da equação de Ångström para a região de palotina, estado do paraná,” *Nome do Periódico*, vol. VOLUME, no. NÚMERO, p. PÁGINAS, ANO.

[84] G. A. A. da Silva, I. F. d. S. da Mota, M. E. D. Chaves, and F. H. Wagner, “Mixing Data Cube Architecture and Geo-Object-Oriented Time Series Segmentation for Mapping Heterogeneous Landscapes,” *Remote Sensing*, vol. 14, no. 17, p. 4293, 2022.

[85] “Road network mapping from multispectral satellite imagery: Leveraging deep learning and spectral bands,” *AGILE: GIScience Series*, vol. 5, pp. 1–11, 2024, license: CC BY 4.0. [Online]. Available: https://agile-giss.copernicus.org/articles/5/6/2024/

[86] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” Tech. Rep., 1996. [Online]. Available: www.aaai.org

[87] C. L. de Castro and A. P. Braga, “Aprendizado supervisionado com conjuntos de dados desbalanceados.” Belo Horizonte, MG, Brasil: Universidade Federal de Minas Gerais, trabalho publicado como artigo.

[88] G. D. de Resende, “Identificação de estradas para direção assistida de caminhões operando em condições climáticas adversas no ambiente da mineração,” Dissertação (Mestrado), Universidade Federal de Ouro Preto, Ouro Preto, 2018, 149 f. [Online]. Available: www.sisbin.ufop.br

APPENDICES

Appendix A - Appendix A Title

7.1 DATA COLLECTION STRATEGY

7.1.1 Rationale of the Approach

The definition of the collection strategy was guided by a technical visit to the Gestão e Operações do Sistema de Proteção da Amazônia (CENSIPAM), held on July 31, 2023. During a meeting with analysts specialized in fire monitoring, the greater operational relevance of **fire focuses** (first detections of thermal anomaly) was emphasized compared to already burned areas. This approach is justified by the possibility of preventive intervention during the initial phase of the event, preventing evolution to large-scale fires.

7.1.2 Data Platforms Used

Three main platforms were evaluated and used for data collection:

- **BDQueimadas (INPE)**: Consolidated historical database with heat focus data since 1998, but with limited access to satellite images. Figures 36, 37, 38, and 39 show the website interface.

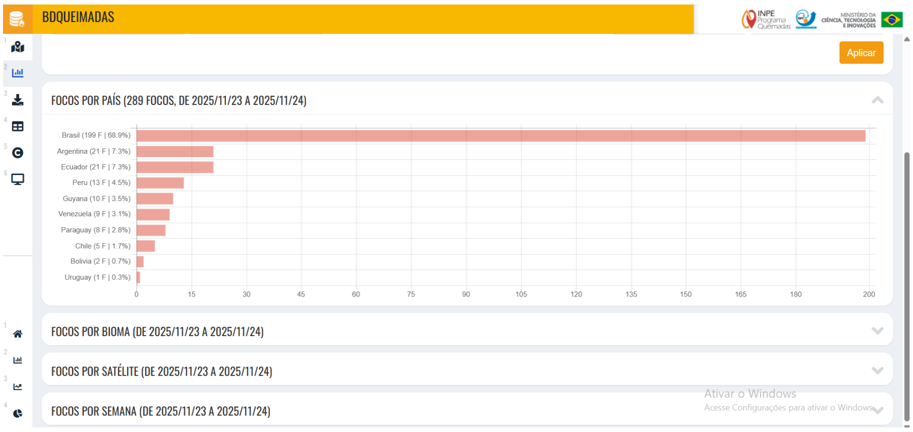
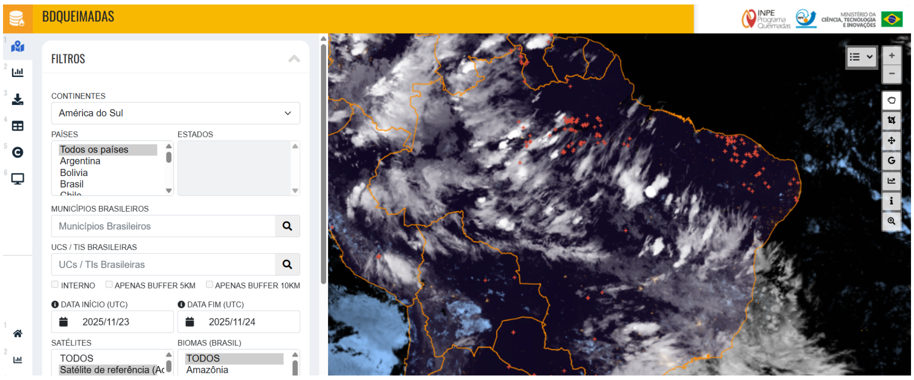


Figure 36: BDQueimadas Dashboard, used for monitoring heat focuses [17].



CONTINENTES: América do Sul

PAÍSES: Todos os países, Argentina, Bolivia, Brasil, Chile

ESTADOS

MUNICÍPIOS BRASILEIROS: Municípios Brasileiros

UCS / TIS BRASILEIRAS: UCs / TIs Brasileiras

□ INTERNO □ APENAS BUFFER 5KM □ APENAS BUFFER 10KM

DATA INÍCIO (UTC): 2025/11/23

DATA FIM (UTC): 2025/11/24

SATÉLITES: TODOS, Satélite de referência (Ac

BIOMAS (BRASIL): TODOS, Amazônia

Figure 37: Fire identification via BDQueimadas panel [18].

- **Fire Panel (CENSIPAM)**: Platform that integrates information from multiple satellites, providing geographic coordinates, detection date/time, and soil cover classification, but with limited technical documentation and absence of historical climate data.

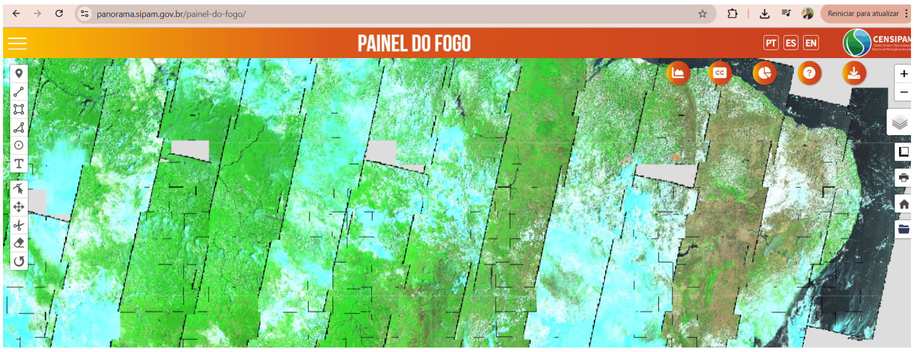
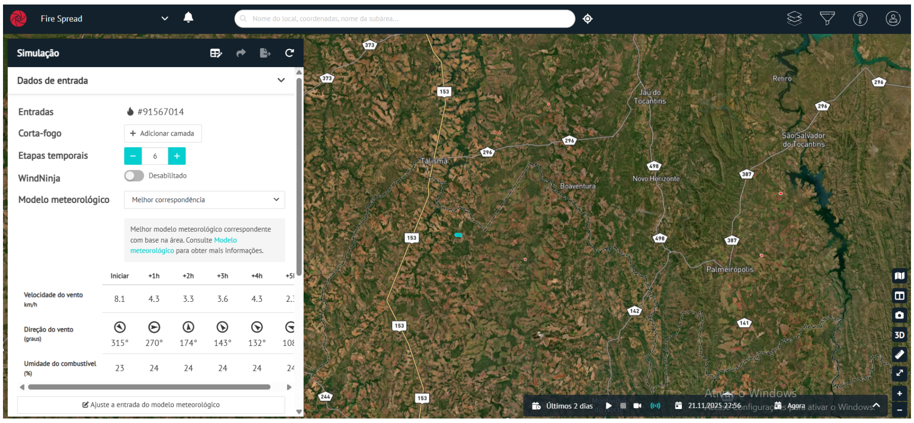


Figure 38: Fire detection via Fire Panel, from CENSIPAM [19].

- **OroraTech**: Platform specialized in thermal satellite detection that overcame the limitations of public databases, offering first focus identification, confidence metrics (Fire Confidence), integrated climate data, and GeoJSON format export.



Simulação

Dados de entrada

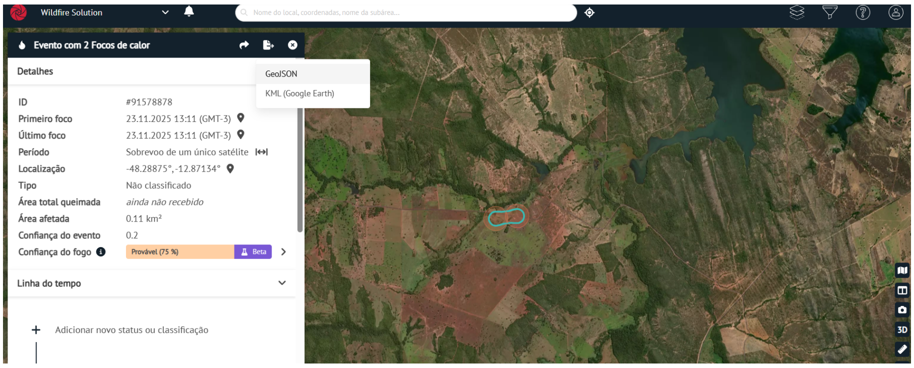
Entradas: #91567014
Corta-fogo: + Adicionar camada
Etapas temporais: − 6 +
WindNinja: Desabilitado
Modelo meteorológico: Melhor correspondência

Melhor modelo meteorológico correspondente com base na área. Consulte Modelo meteorológico para obter mais informações.

	Iniciar	+1h	+2h	+3h	+4h	+5l
Velocidade do vento km/h	8.1	4.3	3.3	3.6	4.3	2.:
Direção do vento (graus)	315°	270°	174°	143°	132°	108
Umidade do combustível (%)	23	24	24	24	24	24

Ajuste a entrada do modelo meteorológico

Figure 39: Detection of heat focuses and fires by OroraTech platform [20].



Evento com 2 Focos de calor

Detalhes

ID	#91578878
Primeiro foco	23.11.2025 13:11 (GMT-3)
Último foco	23.11.2025 13:11 (GMT-3)
Período	Sobrevoo de um único satélite
Localização	-48.28875°,-12.87134°
Tipo	Não classificado
Área total queimada	*ainda não recebido*
Área afetada	0.11 km²
Confiança do evento	0.2
Confiança do fogo	Provável (75 %) Beta

Linha do tempo

+ Adicionar novo status ou classificação

Figure 40: Export of GEOJSON file by OroraTech platform [21].

7.1.3 Satellite Image Acquisition

For obtaining multispectral images, **Sentinel-Hub** was used, a platform that provides access to data from the Copernicus Sentinel-2 mission, with spatial resolution of 10 meters in RGB and near-infrared bands, and revisit every 5 days considering the complete constellation of satellites.

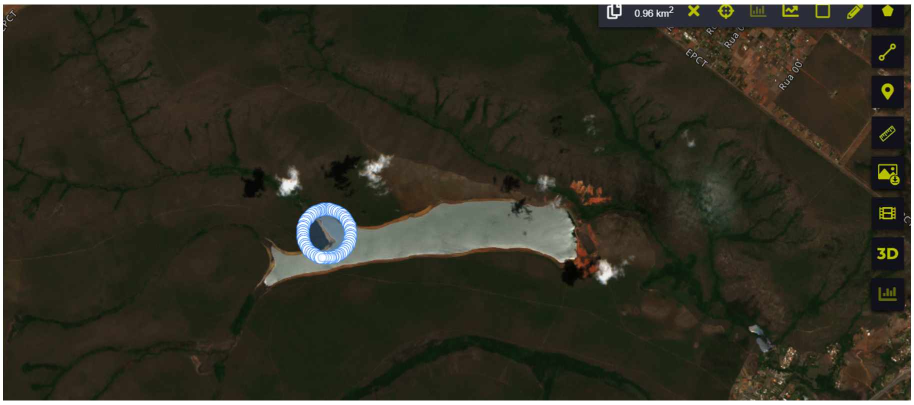


Figure 41: Import of GEOJSON file to Sentinel Hub [22].

7.2 REFERENCE DATA CREATION AND PREPROCESSING

7.2.1 Manual Vectorization in QGIS

For creation of the labeled dataset, 12 of the 40 images were selected for manual vectorization in QGIS software. The process consisted of:

1. Loading Sentinel-2 true color images as raster layers
2. Creation of vector layers of ”line” type with identical reference system
3. Manual digitization of all visible road structures, including paved highways, unpaved secondary roads, rural access routes, and well-defined trails
4. Export of finalized shapefiles in ESRI Shapefile format

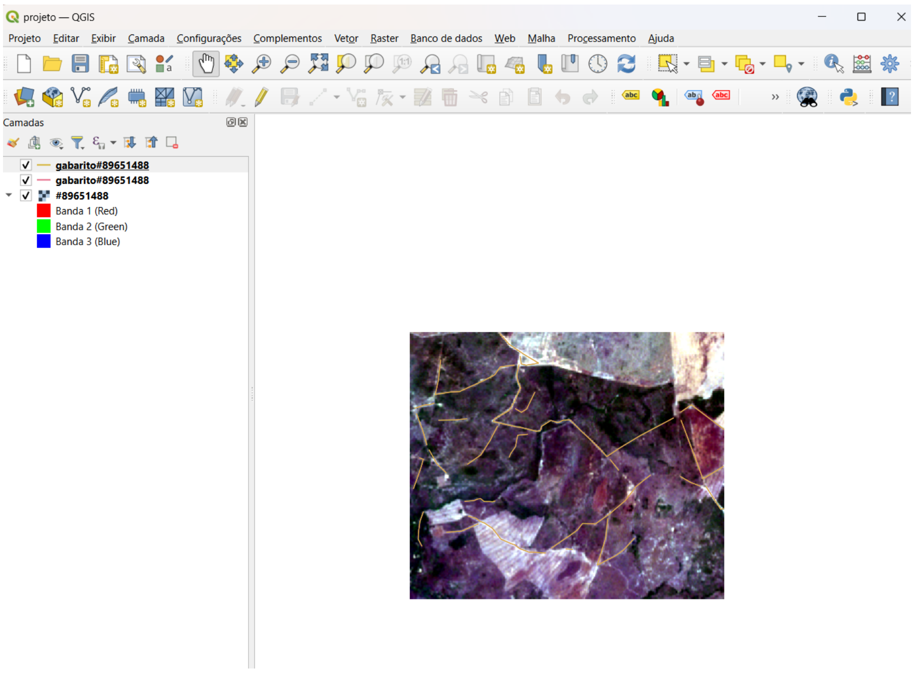

projeto — QGIS
Projeto Editar Exibir Camada Configurações Complementos Vetor Raster Banco de dados Web Malha Processamento Ajuda
Camadas
gabarito#89651488
gabarito#89651488
#89651488
Banda 1 (Red)
Banda 2 (Green)
Banda 3 (Blue)

Figure 42: Production of reference data in QGIS. Source: Author's own.

7.2.2 NGRDI Calculation

Since the images used do not have a near-infrared band (NIR), it was not possible to calculate the original NDVI. To overcome this limitation, an NGRDI was adopted using only the bands available in RGB, replacing the NIR band with the green band. The code used was:

```
def NGRDI(img):
    """NGRDI for RGB images"""
    R = img[:,:,0]
    G = img[:,:,1]
    denom = (G + R + 1e-8)
    return (G - R) / denom
```

Although approximate, NGRDI added useful information about vegetation, contributing to the distinction between exposed soil, vegetation, and roads.

7.3 DATASET BALANCING

The initial dataset presented severe imbalance in the proportion of 1:87 between road and non-road pixels. To correct this, two sequential strategies were applied:

- **Morphological Dilation**: 8-neighborhood connectivity to expand road width and reduce gaps in segmentation, reducing imbalance to 1:22

- **Controlled Undersampling**: Selection of only four negative samples for each positive (ratio 1:4), following recommendations from the literature on *imbalanced learning*

7.4 POST-PROCESSING AND SKELETONIZATION

7.4.1 Mask Processing Pipeline

The prediction masks generated by the model were submitted to a sequential post-processing pipeline implemented in Python. This pipeline executed the following operations:

1. **Binarization via Otsu**: Correction of artifacts and stable separation between road and background
2. **Skeletonization**: Reduction of binary structures to unit thickness, preserving topological connectivity
3. **Component labeling**: Identification of individualized road segments

7.4.2 Classification by Segment Size

After skeletonization, segments were classified into three categories based on the number of pixels. Classification thresholds were determined empirically through iterative visual analysis:

- **Small segments** (up to 200 pixels): Noise, artifacts, or insignificant road stretches
- **Medium segments** (201-800 pixels): Local roads or fragmented stretches
- **Large segments** (above 800 pixels): Main continuous transportation routes and structurally relevant

7.5 DATA INTEGRATION AND STATISTICAL ANALYSIS

7.5.1 Final Table Composition

Processed data were integrated with meteorological information from OroraTech using the focus ID as the primary key. The final table included the following variables:

- **Identification**: ID_Focus, Date, Latitude, Longitude
- **Meteorological**: Temperature_C, Relative_Humidity_%, Angstrom_Index, Risk_Classificatio
- **Road infrastructure**: Total_Roads_px, Total_Red_Pixels, Small_Pixels, Medium_Pixels, Large_Pixels

- **Land use**: Forest_%, Shrub_%, Pasture_%, Grass_%, Agricultural_Land_%
- **Spatial**: Area_km2, Distance_Roads_km

7.5.2 Correlation Analysis

Statistical analysis focused on the correlation between road infrastructure density (represented by `Total_Red_Pixels`) and the Ångström Index, using:

- Pearson correlation coefficient for normal variables
- 40 observations from processed images

7.6 REPOSITORY AND TOOLS

All code developed, processed data, and complementary documentation are publicly available in the repository: **https://github.com/kelvinGomesP/TCC**

The main tools used included:

- **QGIS 3.28**: Manual vectorization and geospatial analysis
- **Python 3.10**: Main processing with OpenCV, scikit-learn, and scikit-image
- **Sentinel-Hub**: Acquisition of multispectral images
- **OroraTech**: Detailed fire focus data and meteorological variables

ANNEXES

Annex A - Annex A Title

8.1 DATA BALANCING CODE

```
from skimage.morphology import binary_dilation, disk

# Dilation to connect disconnected segments
mascara_dilatada = binary_dilation(mascara_binaria, disk(2))

# Balancing to 1:4 ratio
N_NEG_RATIO = 4
pos_idx = np.where(y == 1)[0]  # Positive indices
neg_idx = np.where(y == 0)[0]  # Negative indices
np.random.shuffle(neg_idx)      # Shuffling
neg_keep = neg_idx[:len(pos_idx) * N_NEG_RATIO]
indices_finais = np.concatenate([pos_idx, neg_keep])
```

8.2 POST-PROCESSING AND SKELETONIZATION CODE

```
import cv2
import numpy as np
from skimage.morphology import skeletonize, remove_small_holes
from scipy.ndimage import label

def processar_mascara_estradas(mask_path):
    # 1. Load and binarize with Otsu
    mask = cv2.imread(mask_path, cv2.IMREAD_GRAYSCALE)
    _, bin_mask = cv2.threshold(mask, 0, 1,
                                cv2.THRESH_BINARY + cv2.THRESH_OTSU)

    # 2. Fill small holes
    smooth = remove_small_holes(bin_mask.astype(bool),
                               area_threshold=40)

    # 3. Skeletonization to unit thickness
    skeleton = skeletonize(smooth).astype(np.uint8)

    # 4. Identify connected components
    labeled, num_componentes = label(skeleton,
                                structure=np.ones((3,3)))
    return labeled, num_componentes, skeleton

def classificar_segmentos(labeled_mask, num_componentes):
    # Thresholds defined empirically
    PEQUENO_MAX = 200
    MEDIO_MAX = 800

    contadores = {'small': 0, 'medium': 0, 'large': 0}
    vis = np.zeros((*labeled_mask.shape, 3), dtype=np.uint8)

    for comp_id in range(1, num_componentes + 1):
        mascara_componente = (labeled_mask == comp_id)
        tamanho = np.sum(mascara_componente)

        if tamanho <= PEQUENO_MAX:
            cor = [255, 255, 255]  # White - small
            categoria = 'small'
        elif tamanho <= MEDIO_MAX:
            cor = [0, 255, 0]      # Green - medium
            categoria = 'medium'
        else:
            cor = [0, 0, 255]      # Red - large
            categoria = 'large'

        contadores[categoria] += 1
        vis[mascara_componente] = cor

    return vis, contadores
```

8.3 PROJECT REPOSITORY

- **Complete codes**: https://github.com/kelvinGomesP/TCC
- **Processed images**: Available in the repository
- **Data spreadsheets**: Including correlation tables and metrics
- **Technical documentation**: Instructions for reproducing experiments

SEJAM BEM-VINDOS
idp
Ensino que te conecta