Masters Program in Geospatial Technologies

Mapping urban tree species in a tropical environment using airborne multispectral and LiDAR data

Pablo Henrique Alves Cruz

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
UNIVERSITAT JAUME·I
Westfälische Wilhelms-Universität Münster

Mapping urban tree species in a tropical environment using airborne multispectral and LiDAR data

---

Dissertation supervised by

Jan R. K. Lehmann, PhD

Institute of Landscape Ecology,

University of Münster (WWU)

Münster, Germany

and co-supervised by

Joel Dinis Baptista Ferreira da Silva, PhD

Instituto Superior de Estatística e Gestão de Informação,

Universidade NOVA de Lisboa

Lisbon, Portugal

Filiberto Pla Bañón, PhD

Institute of New Imaging Technologies

Universitat Jaume I (UJI)

Castellón de la Plana, Spain

February 23, 2021.

Declaration of Originality

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisbon, February 23, 2021

Pablo Henrique Alves Cruz

[the signed original has been archived by the NOVA IMS services]

Acknowledgments

First of all, I need to express my gratitude and love to my family, who supported every decision of mine in life, who always believed in my potential and success. This is our accomplishment, our victory, it is for us!

I am extremely grateful for the Master in Geospatial Technologies program for providing me a unique and remarkable experience in life, with high quality education and incredible people to share this amazing period of my life.

Also, I would like to thank my supervisor Jan Lehmann for being supportive, encouraging and to believe in this research topic. I learned a lot from his insights and will always be thankful for the trusting in my potential. My gratitude also goes to my co-supervisors Joel Silva and Filiberto Pla for their suggestion and valuable feedbacks, and to Professor Marco Painho for his guidance through the thesis development and during the whole master program.

To all the colleagues and the friendships made in the program, for their outstanding companionship and the memorable moments shared. Special thanks to Eftychia, Yan, Filip and Mihail.

To my Brazilian friends in Lisbon with whom I felt like in home so many times, for bringing the warmth, love and understanding we only recognize in our own people.

To my friends and loved ones in Brazil, sending love and care from thousands of kilometers, special thanks to Nicole, Crislane, Ana Beatriz, Thais and Luma for always making my life better.

To Beto, for the daily support, his love and dedication, and the extraordinary relationship we built.

Mapping urban tree species in a tropical environment using airborne multispectral and LiDAR data

Abstract

Accurate and up-to-date urban tree inventory is an essential resource for the development of strategies towards sustainable urban planning, as well as for effective management and preservation of biodiversity. Trees contribute to thermal comfort within urban centers by lessening heat island effect and have a direct impact in the reduction of air pollution. However, mapping individual trees species normally involves time-consuming field work over large areas or image interpretation performed by specialists. The integration of airborne LiDAR data with high-spatial resolution and multispectral aerial image is an alternative and effective approach to differentiate tree species at the individual crown level. This thesis aims to investigate the potential of such remotely sensed data to discriminate 5 common urban tree species using traditional Machine Learning classifiers (Random Forest, Support Vector Machine, and k-Nearest Neighbors) in the tropical environment of Salvador, Brazil. Vegetation indices and texture information were extracted from multispectral imagery, and LiDAR-derived variables for tree crowns, were tested separately and combined to perform tree species classification applying three different classifiers. Random Forest outperformed the other two classifiers, reaching overall accuracy of 82.5% when using combined multispectral and LiDAR data. The results indicate that (1) given the similarity in spectral signature, multispectral data alone is not sufficient to distinguish tropical tree species (only k-NN classifier could detect all species); (2) height values and intensity of crown returns points were the most relevant LiDAR features, combination of both datasets improved accuracy up to 20%; (3) generation of canopy height model derived from LiDAR point cloud is an effective method to delineate individual tree crowns in a semi-automatic approach.

Keywords

Atlantic Rainforest

Airborne LiDAR

Airborne Multispectral

Individual Tree Crowns

k-Nearest Neighbor

Random Forest

Support Vector Machine

Urban Tree Species Classification

Acronyms

ARF - Atlantic Rainforest

Bagging – Bootstrap Aggregating

CHM – Canopy Height Model

FN – False Negative

FP – False Positive

gNDVI – Green Normalized Difference Vegetation Index

GS – Grid Search

ITC – Individual Tree Crown

k-NN – k-Nearest Neighbor

LAS – LASer format file

LiDAR – Light Detection and Ranging

NDVI – Normalized Difference Vegetation Index

OA – Overall Accuracy

OBIA – Object-based Image Analysis

OTB – Orfeo Toolbox

PA – Producer's Accuracy

PSI – Pixel Shape Index

RBF – Radial Basis Function

RF – Random Forest

RS – Random Search

SAGA-GIS - System for Automated Geoscientific Analyses

SEMAN – Secretaria Municipal de Manutenção de Salvador

SFS – Structure Feature Set

SIRGAS2000 – Sistema de Referência Geocêntrico para as Américas

SVM – Support Vector Machine

TN – True Negative

TP – True Positive

UA – User's Accuracy

UAV – Unmanned Aerial Vehicle

UTM – Universal Transverse Mercator

VHSR – Very High Spatial Resolution

WFS – Web Feature Service

Index of the text

Index of figures

Index of tables

1 Introduction

1.1 Background

The Atlantic Rainforest (ARF) is one of the most important biomes in America, hosting a huge diversity of tree species and animal species, yet it is still also the most endangered tropical biomes in the world. In Brazil, due to the impact of anthropogenic activities impact, urbanization and industrial activities, the ARF covers only 22% of its original area (1,3 million km²) in different stages of conservation, where only 7% of the reminiscent area is in good state of conservation and has over 100 hectares of area [1]. ARF fragments are also concentrated within urban centers, normally in conservation units protected by the Decree 11.428/2006, however the damages caused by human activities cannot be completed reverted, the preservation of urban forests and adequate management of urban trees (or trees outside of the forest, TOF) is extremely important for the biome's conservation and partial recovery. In this context, the development of trustable and robust mapping techniques aiming the creation of a detailed urban tree species inventory is an essential step to provide information in several application, such as biodiversity monitoring, proposal of public incentive policies of planting and preservation of native species trees and evaluation of urban sprawl effects on trees and green areas.

Traditionally, information collection regarding trees in urban areas and its respective species is related to tasks that include extensive field sampling, interpretation aerial or satellite imagery for manual classification of species by specialists, for example [2]. These methods, in general, are time-consuming, costly and, most of the times, not efficient to provide up-to-date information about the whole city tree coverage [3]. To overcome these limitation and issues, the adoption of remote sensing products such as aerial or satellite image provides highly detailed information to identify and to extract spectral information at the individual tree level. Additionally, from airborne LiDAR data it is possible to extract accurate measurements regarding tree height and other structural features useful

to differentiate tree species, above-ground biomass, and stand density for instance, based either on range or intensity of laser pulse returns [3]–[5]. Both sources of data are usually less expensive than a field sampling campaign.

The association of remotely-sensed data with machine learning classifiers stand as an efficient and affordable alternative, however most of the studies available employing this approach are focused in tree species classification of temperate and boreal forests [6], [7], and the majority of studies performed in tropical location are related to tree species classification in forested areas and/or to a limited number of species [8]–[12] instead of aiming to detect and classify tree species in urban centers. The lack of studies in tropical forest is normally associated to the challenges regarding spectral response similarity among species, difficulty to delineate tree crowns due to overlapping between canopies and the presence of predominant species leading to imbalanced training sample [9].

1.2 Problem Statement and Motivation

In urban centers, trees have vital importance in efforts to reduce the impacts of air pollution since they produce oxygen and, as consequence, improve air quality [13]. It also helps in the reduction of discomfort caused by urban heat islands, acting as temperature regulator, and reduces impacts of stormwater runoff [14]. Therefore, mapping individual trees and cataloguing their respective species is an urgent necessity to monitor the effects on urbanization in the city's natural landscape and to demand actions from authorities and population towards urban tree and biodiversity preservation.

Salvador, a city located in the northeast coast of Brazil, is considered as the ARF capital since the city is completely inserted in this biome and its associated ecosystems (restinga and mangrove). Nonetheless, in the past 20 years the city has been transformed with the intervention of infrastructure advances on transportation means, with the construction of two metro lines and more recently

the ongoing construction of bus rapid transit brought public attention to the notable impact of such interventions for the ARF reminiscent in Salvador. Real estate speculation is another pressure agent for deforestation in Salvador, causing reduction of 40% in Pituaçu Metropolitan Park' original area, due to the construction of both irregular housing and luxurious condos [15].

Góes and Oliveira (2011) analyzed 7 technical reports and studies regarding tree species inventory in Salvador, mainly in parkways in the city. From the studies on parkways, the authors found information of 2.469 trees from 82 different species, the analysis pointed to the predominance of exotic species (53,3%) and low representativeness of native regional species. Notwithstanding that these studies present a great overview regarding the city's biodiversity, none of them has exact and precise location of the inventoried trees which makes difficult to monitor their preservation and they are restrict to only few locations.

Given these circumstances and the fact that between 2016 and 2017 the City Hall of Salvador acquired airborne multispectral imagery (visible and near infrared) with high spatial resolution, and also high-density LiDAR data, the potential of these datasets associated with robust machine learning classifiers motivate the conduction of research to contribute to the mapping and conservation of Atlantic Rain Forest biome, as well as to understand the particularities of performing tree species classification in a tropical urban environment.

1.3 Aims and Research Questions

This study aims to assess the benefits of aerial imagery and airborne LiDAR data to perform tree species differentiation in urban environment by answering the outlined research questions:

1. What is the impact of different dataset combination, in each machine learning classifier performance, for tree species classification in a tropical urban environment?

2. Which are the most significant LiDAR-derived variables to distinguish the tropical tree species in this study area?
3. How effective it is to use LiDAR-derived Canopy Height Model to perform semi-automatic individual tree crown extraction?

1.4 Objectives

1) Evaluate and compare the performance of Random Forest, Support Vector Machine and k-Nearest Neighbor classifiers with features extracted from multispectral orthoimage and/or airborne LiDAR data.
2) Assess the model's ability to differentiate among tree species using different sets of variables (only multispectral features, only LiDAR features and combined features).
3) Process point cloud to generate LiDAR-derived Canopy Height Model, then extract individual tree crowns accurately.

2 Literature Review

2.1 Remote Sensing and Tree Species Classification

In their systematic literature review, Fassnacht *et al* (2016) in [6] selected and evaluated 101 peer-reviewed studies related to tree species classification published between January 1980 and December 2014. According to their analysis, hyperspectral sensors are the most used ones for tree species classification, followed by multispectral sensors that can range from moderate, high and super high spatial resolution (such as WorldView and IKNOS satellites or airborne sensors), the combination of data coming from both sensors type was also relevant in 28 of the evaluated studies. Besides, 99% of the studies have been conducted with the association of an active sensor (most of the studies using LiDAR) to the data coming from optical sensors.

Hamamura (2020) in [7] analyzed 33 papers and articles published between 2003 and 2018, stressing that the spatial distribution of such studies in the field of tree species classification is highly concentrated in places with temperate or boreal ecosystems, regions that present a homogeneous vegetation structure mainly aggregated in big urban parks. Unlike cities located in a tropical climate, where urban vegetation has a broader diversity of species and the spatial distribution is sparser.

More recently, sensors embedded in Unmanned Aerial Vehicles (UAV) have become one of the main sources of data in the forestry management field, mostly due to the low operational cost, the capacity to plan and obtain multitemporal information according to weather conditions, the possibility to use a diversity of sensors and finer spatial resolution information [9]. According to Guimarães *et al.* in [16] the use of UAV to distinguish tree species is performed using mostly multispectral sensors, with few studies adding color infrared (CIR sensors). To help the identification of tree species, the author usually calculates spectral indices (for instance, NDVI, NGDRI, VARI, etc) and UAV-derived point clouds to extract structural information. Guimarães et al. also highlight the predominant use

of two machine learning methods: Support Vector Machine and Random Forest. The first is more suitable for cases with a low number of samples and high variety of classes, while the latter is better for applications with high data dimensionality.

2.2 Object Based Image Analysis (OBIA) for Tree Species Classification

Due to the improvement in spatial resolution of remotely sensed data, the conventional classification method based on pixel-level became inadequate due to the reduced size of this unit when compared with the target under analysis, consequently, the spectral information contained in a single pixel could not represent properly the features of an individual target (e.g. building). In this scenario, object-based image analysis (OBIA) appears as an alternative to overcome the limitations found in the former method [17].

In OBIA, the basic unit of analysis is no longer the pixel but the 'image object', which corresponds to pixels grouped to form a shape that represents real-world objects. The method is divided into two main processes: (a)image segmentation, and (b) object extraction and classification [18]. Segmentation is a method to divide an image into distinguishable and homogenous regions that share similar properties such as color, shape, and texture. When compared to single pixels, the resulting objects of segmentation accumulate more spectral information such as mean values for each band, variance, minimum and maximum values, and, more importantly, it brings spatial information for each feature [19]. The success of object extraction and classification step is dependent on the performance and results of segmentation. The segmented objects will be used as input for training and testing the classification model, therefore the correct detection of features of interest plays an important role since the spectral information collected from those must be representative and consistent about the target.

For tree species classification in urban area, the method used by many authors is called Individual Tree Crown (ITC) delineation, which is an automatic procedure to identify the location, tree crown size and shape of each tree in a remote sensed

image. This approach is commonly applied in very high spatial resolution (VHSR) image associated with LiDAR data and allows to obtain estimation of variables such as height, biomass, diameter of tree crown, for instance. However, this approach fails to detect very small trees or those that are close to trees with larger canopy [20]. Aiming to enhance the accuracy of tree crowns segmentation, some authors apply masks to exclude non-tree object, such as thematic layer with buildings [21] or filters based on spectral index like NDVI (e.g. values above 0.2) and height threshold based on average height for tree species under study [3].

2.3 Light Detection And Ranging (LiDAR)

Light Detection and Ranging, also known as laser altimetry, is a cutting-edge remote sensing technology based on an active sensor that used light in the form of a laser to measure distances between sensor and target objects. LiDAR systems are composed by laser scanner device, onboard Global Positioning System (GPS) receiver and Inertial Navigation System (INS) which allows the system to acquire three-dimensional coordinates of targeted objects [22].

LiDAR carrying platforms can be divided into three main segments such as airborne, terrestrial, and space-borne, allowing this technology to fulfill the needs of different area coverage demands and levels of resolutions according to the subject of analysis. LiDAR observations are presented in two data types, point cloud and waveform. The first one is widely applied in forestry related studies and provide structural parameters such as tree height and canopy volume calculation. Whereas the waveform data brings distance information and also vertical distribution of targets and features about structure and physical properties [17].

Airborne LiDAR is the system most frequently used to collect data and to extract vegetation parameters, for instance, tree height, above-ground biomass, volume and Leaf Area Index at the stand level. Due to technology development, the LiDAR point cloud is becoming denser and enabling to recognize trees at the individual level, an essential asset for tree classification in urban environments

where the distribution is sparse and there is high spatial heterogeneity [3], [17], [21].

Individual tree crown (ITC) delineation using LiDAR datasets is usually conducted with the application of a Canopy Height Model (CHM), a raster product derived from the point cloud. A CHM is a digital elevation model which represents the canopy surface, it is obtained through the subtraction of a digital terrain model (DTM) from the digital surface model (DSM). Segmentation algorithms are applied in the CHM to extract the tree crowns, being marker-controlled watershed and region growing segmentation the most popular approaches, however the success of segmentation also relies in the pixel dimension of the CHM and its optimal size has to consider the average crown size and tree height (correlated variables about structural features of trees) [3], [23], [24].

2.4 Machine Learning

Machine Learning (ML) is under the domain of Artificial Intelligence which imitates the way a human brain process information and gain knowledge. ML aims to detect and take advantage of hidden patterns in the input training data, then applying these patterns to analyze unknown data. Due to its ability to treat data of high dimensionality, to model complex class signatures, and the fact it does not make assumptions about the data distribution (non-parametric algorithms), ML is very efficient and widely used in classification of remote sensing products [25]–[27].

2.4.1 General Machine Learning Workflow

The training of a machine learning algorithm follows a basic workflow (Figure 2.1) composed by the following main steps: data collection and preprocessing, dataset preparation, model building and model evaluation.

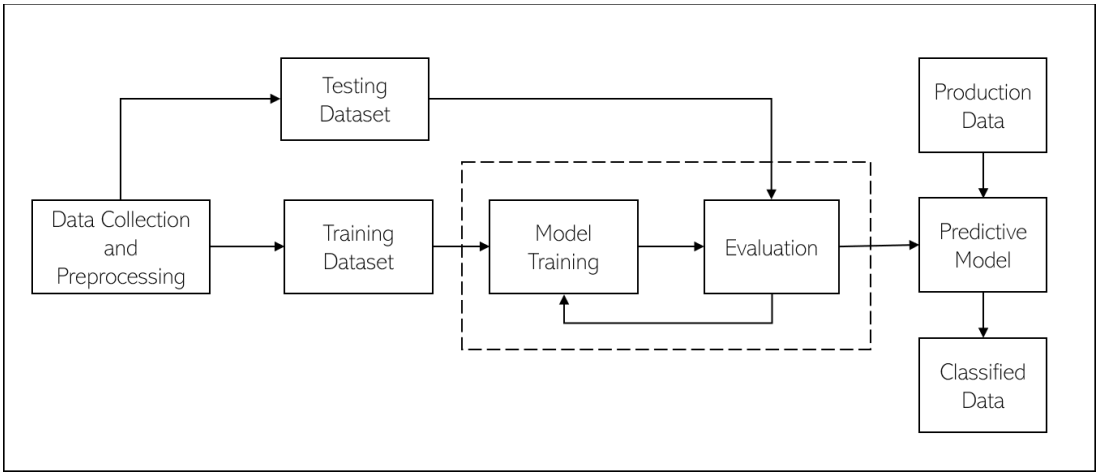


Figure 2.1 Machine Learning Workflow.

The quality of a good model based on machine learning algorithm is highly dependent on the quality of the dataset used during the model training. Since data are collected for plenty different reasons, it is necessary to identify and extract the information that will meet a project's need; therefore, data preprocessing is a key part in the machine learning process, often representing the most timing-consuming task during a project [28]. Data preprocessing is necessary since raw data usually comes from unprocessed, incomplete and noisy databases, containing problems such as redundant or obsolete fields, missing values, unsuitable data formats and inconsistent values [29].

Once the data is preprocessed and ready to use, it is divided into 60/20/20% for training dataset, testing data and validation dataset or 70/30% in case validation is not necessary. The training dataset is used as input to a learning system and should be able to provide consistent information and parameters from which the model will be created. Testing dataset contains the information used to assess the performance of chosen model. Validation dataset is only required when the machine learning model and its architecture are not pre-selected [27], [28].

During the "evaluation" step the model's performance will be assessed in terms of its predictive efficacy and to compute the cost function of training and validation datasets, which is important to detect problems due to high bias or high variance in the dataset. The first problem causes under-fitting in the model, i.e. the model is not able to generalize the relation between training features and outcomes.

While high variance in dataset causes the opposite, occurrence of over-fitting, when the model understands the detail and noise in training data too well, in such way that it impacts negatively the model's ability to generalize information and predict new data [25], [27], [28].

2.4.2 Random Forests (RF)

Random forests classifier is an ensemble classifier that combines a multitude of classification and regression trees (CARTs), performing a prediction through their combined results [30], [31]. Each tree is created using the combination of Bagging algorithm and Random Subspace Method. The first aims to generate subsets of training samples through replacement, which is applied to reduce variance. The latter reduces the bias between estimators by increasing the diversity of features used to grow each tree [28], [30], [32].

Bagging (acronym for Bootstrap AGGregatING) is the first step in the classification process, it allows the creation of multiple subsets derived from the training sample, this approach may select the same element more than once and include it in different subsets dataset while other elements may not be selected at all. The training process applies *in-bag samples* (about two thirds of the training sample) to create the trees, followed by an internal cross-validation step using *out-of-the bag* samples (the last one third from training sample) to estimate the random forest model performance [31].

The number of trees (*Ntree*) is defined by the user, each tree is developed from the bootstrapped subset and it is produced independently without any pruning and node splitting is based on user-defined number of features (*Mtry*) selected at random [30], [31], [33]. Finally, the prediction results from each tree are counted as a "vote" and the final classification is decided using the majority vote of the trees in the forest. Figure 2.2 illustrates the scheme of random forests classifier.

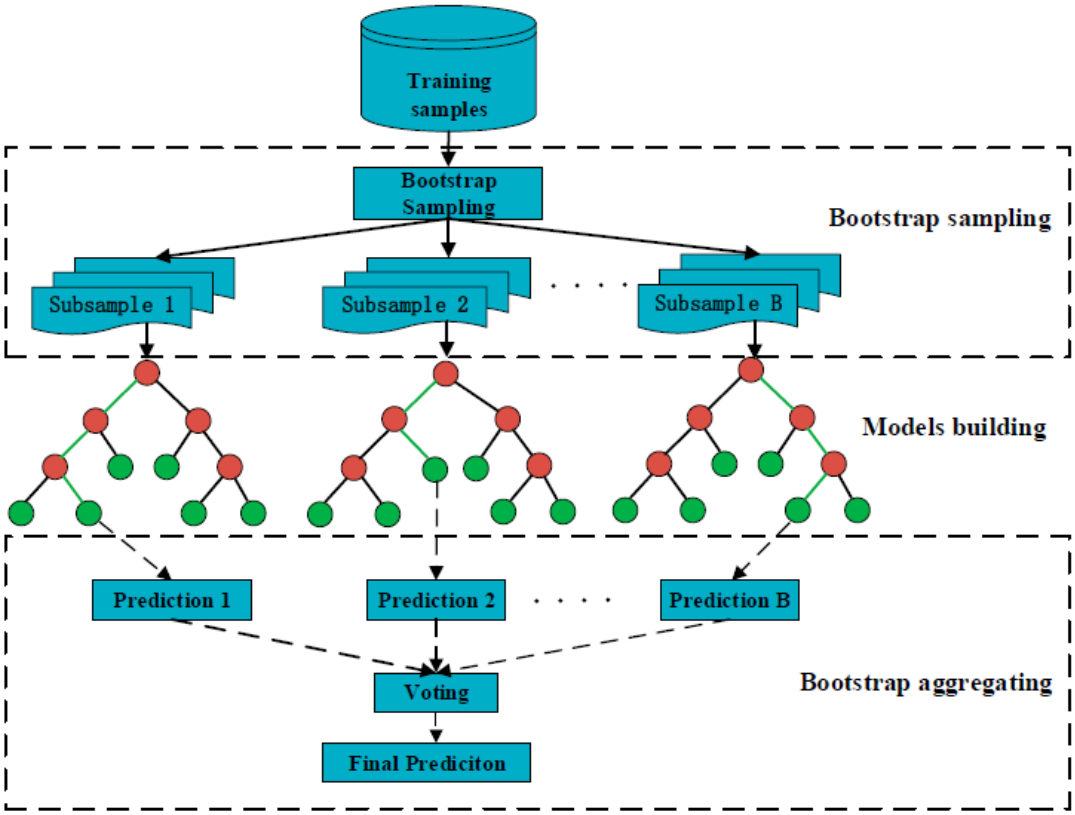


Figure 2.2 Random Forest Classifier Scheme. Source: Wang et al, 2019 [33].

2.4.2.1 RF classification applied for remote sensing

RF is among the top performing and most used classification algorithm for machine learning due to its flexibility on parameter optimization (only two user-defined variables), low computational complexity when compared to other algorithms, accurate results and the maturity proved by the numerous studies that have been applying it for different purposes.

RF has been utilized in different context of remote sensing-based analysis, such as wetland complex classification in the Avalon Peninsula (Canada) using synthetic aperture radar data [34], to evaluate annual deforestation dynamics in two Brazilian states between 1984 and 2014 using Landsat archive data [35] and land cover classification of a large area (30 x 30 km²) in Vietnam using Sentinel-2 imagery [36], these are few examples of the variety of studies and datasets where RF have been applied and showing accurate results.

In remote sensing studies, RF main advantages are related to (i) its ability to handle large data bases, (ii) provision of estimates of the most relevant variables in classification, (iii) it is relatively robust to outliers and noise, (iv) the algorithm generates internal unbiased estimation of the generalization error (out-of-bag error) [37]. However, the main drawbacks to be considered are the algorithm's sensitivity to imbalanced training sample, tending to favor the most representative classes, and to spatial autocorrelated training classes.

2.4.3 Support Vector Machines (SVMs)

SVMs are non-parametric statistical approaches applied to regression and supervised classification problems, therefore no assumption is made on the underlying data distribution. The method's principle is based on the classification of a set of data samples, the algorithm's goal is to determine a hyperplane that separates the dataset into a discrete predefined number of classes in a compatible way with the given training examples distribution, as shown in Figure 2.3 [38], [39].

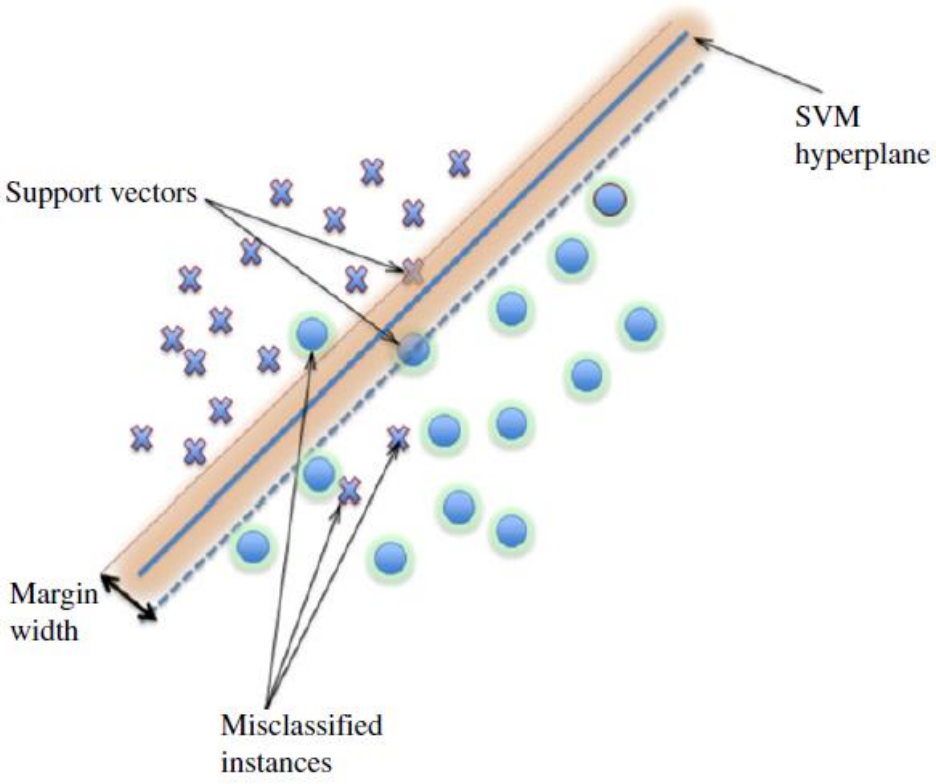


Figure 2.3 Example of linear SVM. Source: Mountrakis et al, 2011 [39].

The decision boundary, obtained during the training step, is known as optimal separating hyperplane and it is used to minimize misclassification. Its learning method consists in an iterative process aiming to find an optimal decision boundary able to detect training patterns and then apply it to test data under the same configuration. SVMs classifiers are binary, working to identify a single boundary between two classes, when more classes are involved the classifier is repeatedly applied to each possible combination of classes [25], [38], [39].

Since SVMs were initially developed to identify linear class boundary, and data under classification can present high dimensionality, Kernel functions are used to project the feature space to a higher dimension, assuming that a linear boundary potentially exist in this higher dimensional space [25], [40].

2.4.3.1 SVMs classification applied for remote sensing

SVMs have been successfully used in remote sensing images classification mainly due to its ability to produce good results even with small and/or imbalanced training samples, low sensitivity to the curse of dimensionality and, mainly by the fact that remote sensing data have unknown distribution and SVMs do not make assumptions on data distribution such as maximum likelihood classification does.

In remote sensing, SVMs have been addressed to classification tasks using data from different sensors, for instance Vohra (2020) used airborne hyperspectral and VIS sensors to compare the effectiveness of SVM and Artificial Neural Networks (ANN) classifiers in multilevel fusion for urban land classification, where SVM outperformed the ANN when using combination of spatial and spectral features [41]. Han (2016) applied UAV data for classification of land cover and irrigated area, SVM also showed better results (overall accuracy of 82,2%) than other classifiers such as decision tree and K-nearest nearest neighbor. Syifa et al (2019) applied SVM classification to detect flood distribution in Brazil using Landsat-8 and Sentinel-2 imagery [42].

Being a Kernel-based approach, the selection of correct kernel function and definition of parameter value (denoted by C) is a challenge in remotely sensed

imagery classification using SVMs. The parameter C controls the relationship between margin maximization and minimization of training error, this parameter tells the algorithm how much the user is concerned about misclassified points and has direct consequences on model overfitting.

2.4.4 k-Nearest Neighbors (k-NN)

k-NN is also a non-parametric classification and regression problems, it does not require a normal data distribution, it is a supervised machine learning with low training computational cost, and has no limitations in the number of independent variables and can be applied to generate estimates of both continuous and categorical variables [36], [43], [44]. The k-NN algorithm works under the assumption defined in Tobler's First Law of Geography, which states that near things are more similar than distant things, therefore the principle behind the algorithm is to search and find a predefined number of training samples closest in distance to the unlabeled data and assign a class from these [45], [46].

In a classification problem, an instance has its label assigned by a plurality of votes based on its neighbors, the object is classified as the most common class among its *k* nearest neighbors, for instance, if k=1 then the object is classified according to its very closest neighbor. There is no particular rule to determinate the value of *k*, it really depends on the type of data, and some authors claim that large values can be helpful to reduce the effects of noise in classification, however it reduce the distinction of boundaries between classes [47].

2.4.4.1 k-NN classification applied for remote sensing

The application of k-NN classifier for remote sensing problem is mostly motivated because it is easy-to-implement algorithm and it has been broadly in forest mapping studies, land cover and land use classification, as well as many other remote sensing related studies.

Tianwei et al (2020) [48] used k-NN to develop a method to identify seed maize production fields with the use time-series of Sentinel-2 Images, where this classifier achieved overall accuracy between 72,5% to 89,3% for different

seasons, outperforming classifiers as SVM and RF sometimes. Cao et al (2018) proposed an object-based approach using k-NN to classify mangrove species in China, with the support of hyperspectral UAV imagery and digital surfaces models in [49], where the classifier's overall accuracy reached 81,79%.

Even with good results showed in the above-mentioned studies, k-NN is usually outperformed by other non-parametric machine learning algorithms, the main reasons are the classifier's sensitivity to noise data and missing data, it does not work well with large dataset and performs poorly with high dimensionality.

2.5 Related Work

Some studies have been conducted to analyze the benefits of multi-seasonal data in tree species classification using WorldView-2 (WV-2) and/or WorldView-3 (WV-3) data, due to their very high spatial resolution and the availability of Shortwave Infrared bands from WV-3 which provides more detailed information of vegetation such as water content, cellulose and lignin. Li et al. [50] in their study used both satellites data to classify five tree species in two urban areas of Beijing (China), the results showed a higher overall accuracy (92,4% with SVM) when using bi-temporal, with differences up to 16,1% for SVM and 20% for RF on the overall accuracy for the same study area when compared to each image separately. In contrast, Ferreira et al. [51] study showed a depreciation in the accuracy when using WorldView-3 imagery from two different periods. However, this study analyzed the classification of 8 trees species in a tropical forest in Campinas (Brazil) and achieved higher overall accuracy of only 70 $\pm$ 8% also with SVM classifier, but in this case adding texture information from panchromatic band to imagery from wet season. Such differences in the results need to consider variables like the number of species to be classified (higher diversity, more complexity to differentiate them) and the surroundings of study areas is also an important factor. In urban areas trees are usually distributed individually or in small and scattered clusters, which affects the spectral characteristics of tree crowns with the influence of non-tree objects (e.g. asphalt, buildings); while in forests and parks, trees are densely grouped and the background is mostly

homogenous, which could make tree differentiation more difficult and less accurate.

Airborne hyperspectral sensors are preferred to perform analysis at the individual tree level due to its finer spatial resolution. In that case, airborne LiDAR data is usually associated to aggregate structural information about trees (e.g. height, crown shape, crown area, etc). Liu et al. [3] combined data from these sensors to classify 15 urban tree species in Surrey, British Columbia (Canada), even with high spatial resolution imagery of 1-m and dense LiDAR point cloud with 25 points/m² the higher accuracy achieved was 70 $\pm$ 3,1% using RF classifier. Authors attribute this result to the temporal distance between datasets and, as consequence, variability on tree conditions. Zhang et al. [21] combined these datasets to apply an object-based classification for 7 tree species in Seattle (USA) achieving 87% and 88,9% accuracies for RF and Multi Class Classifier, respectively. Authors reported that the coarse hyperspectral sensor' spatial resolution of 3-m introduced errors into the classification. Both studies report the importance of spatial resolution compatibility between LiDAR-derived CHM and hyperspectral data (usually $\geq$ 1m), due to its influence on the variable's extraction. Another common factor is the need to reduce dimensionality of hyperspectral data using techniques such as Principal Component Analysis (PCA) or Minimum Noise Fraction (MNF), ensuring the use of essential information from original dataset while discarding redundant and irrelevant information.

3 Methodology

This section describes and locates the area used to develop this research (3.1), lists and provides details about all employed datasets (3.2), and explains the methods implemented to achieve research's objectives (3.3).

3.1 Study Area

Salvador is the capital of Bahia, state in the northeastern coast of Brazil, located in the bounding box defined by the following coordinates: 13°00'58" S, 38°51'53" W (lower left corner) and 12°44'01" S, 38°18'15" W (upper right corner). Its administrative area, which includes continental territory and two islands, covers approximately 415,00 km². The area considered in this study has 6,28 km² (Figure 3.1).

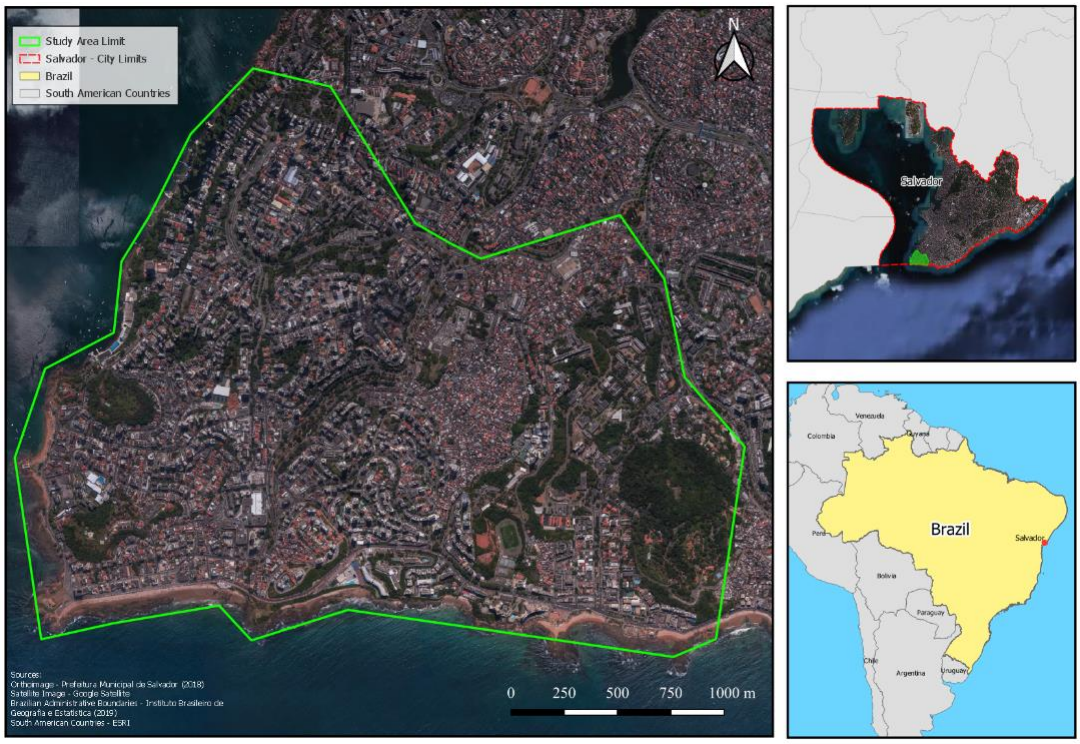


Figure 3.1 Study Area.

This specific area was chosen according to the concentration of trees registered in the municipality's tree inventory, as well as species diversity, its topography that has elevation ranging from -11,70m to 149,67m, and its diversity in urban occupation.

3.2 Data Description

3.2.1 Remote Sensing Data

This research is performed using remotely sensed data from the project "Mapeamento Cartográfico de Salvador", conducted by Salvador's city hall aiming to update the municipal cartographic database. Data collection was done between August 2016 and February 2017, including acquisition of multispectral (visible and near infrared) imagery and LiDAR data.

3.2.1.1 Multispectral Imagery

Aerial images were collected using Vexcel Ultracam-Lp multispectral sensor on-board of airplane, with 70% of overlapping area in both flight directions and average flight altitude of 1200m. From this product, an orthomosaic was generated with Red, Green, Blue and Near Infrared bands, 16 bits image and spatial resolution of 10cm. The data's reference system is "Sistema de Referência Geocêntrico para as Américas" (SIRGAS 2000) and it is projected in Universal Transverse Mercator (UTM) Zone 24S coordinate system.

3.2.1.2 LiDAR Data

The LiDAR dataset were acquired in the same period as multispectral imagery with maximum temporal distance of 48 hours between imagery and LiDAR data. The flight was done using airborne Laser Scanner RIEGL VQ-480 (Table 3.1) sensor on-board of a helicopter flying about 1000m above ground.

For the study area in analysis, the mean point density was 9 points/m² for all returns, nominal pulse spacing of 33cm and 70% area overlapping between flight lines. The final pre-processed point cloud was downloaded in LAS Format, version 1.2.

Acquisition Period	19/AUG/2016 - 13/FEB/2017
Laser scanner model	Riegl VQ480
Laser pulse repetition rate	300 kHz
Measurement rate	Up to 150 000 s-1
Laser wavelength	Near infrared
Beam divergence	0.3 mrad
Laser beam footprint	150mm at 500m
Field of view	60° (± 30°)
Scanning method	Rotating multi-facet mirror

Table 3.1 RIEGL VQ-480 Specifications

3.2.2 Tree Inventory

The data base with tree location and species was provided by the Secretaria Municipal de Manutenção (SEMAN), the municipality's bureau responsible for maintenance services in the city such as pruning trees and removing those who are affected by diseases or causing problems to sidewalks or to overhead electrical wiring, for instance. This data base is not specially designed for tree inventory, but to keep track of tree maintenance using geographical coordinates and listing the tree species when possible.

The data consists in a spreadsheet containing the geographical coordinates of trees in WGS84, scientific species name, popular species name, date of acquisition for each entry and other information regarding its localization (neighborhood, street name and point of reference). The extraction of information from database was done in August 25th, 2020.

3.3 Methods

The proposed methodology (**Figure 3.2**) consists in an evaluation of three different classifiers performance with three different dataset combinations to understand which of each will provide better results for tree species classification. Initially, the reference dataset containing the tree inventory data is examined to find possible inconsistencies and errors in data entry, followed by the determination of a study area according to sample distribution. After that, the individual tree crowns (ITC) are identified and manually delineated.

Still in the pre-processing stage, multispectral and LiDAR datasets are employed to extract spectral information, vegetation indices and texture information, and to generate a LiDAR-derived CHM as well as to extract metrics related to height and intensity. From the CHM, a semi-automatic extraction of tree crowns is done with watershed segmentation process. Both multispectral and LiDAR variables are assigned to manually delineated ITC and to the treetops found in the CHM.

The supervised learning step consists in preparing the entire dataset collected, dividing into three types: only multispectral variables, only LiDAR variables and both sources of variables combined. These data are used to train the models (RF, SVM and k-NN), which undergo through model tuning to find optimal parameters.

To evaluate and compare each model according to the dataset configuration, accuracy metrics such as overall accuracy, user's accuracy, producer's accuracy, and F1-score are applied. Confusion matrices are also used to enhance the results' comprehension through a visual assessment of how the classification corresponded or not the ground-truth data. The best model is then applied to production data (tree crowns segmented from the CHM) to create a final map.

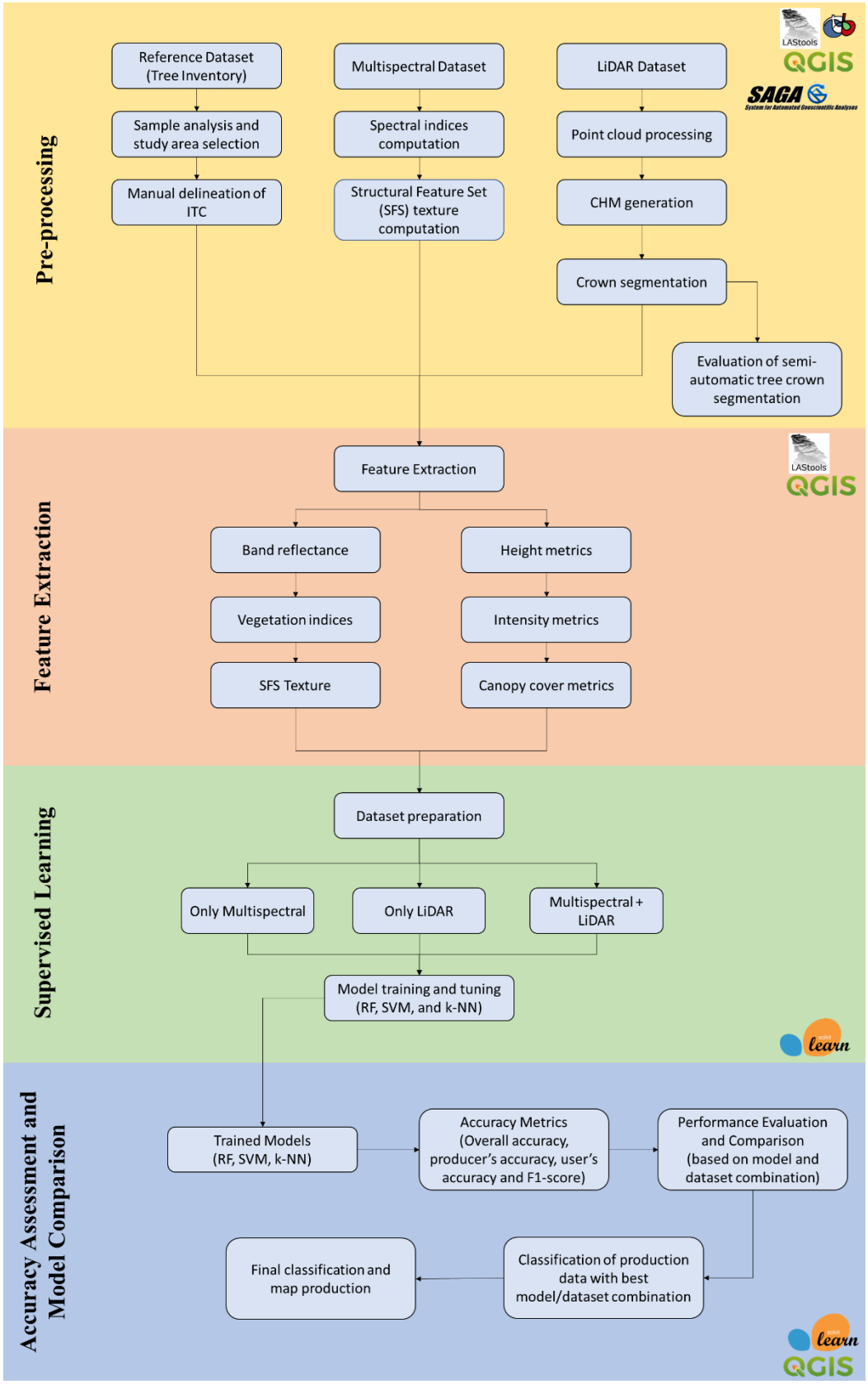


Figure 3.2 Methodology workflow.

3.3.1 Pre-processing

3.3.1.1 Reference dataset (Tree Inventory)

Initially, the dataset went through a thorough analysis where each relevant field was analyzed to detect potential inconsistencies. Firstly, the field "Científico", which holds the scientific Latin name of tree species, was reviewed since some of the names presented typos or had small differences in the writing, e.g multiple entries with *Pachira Aquatica Aubl* or *Pachira Aquatica AUBL.* concerning the same species. After that, the field "Cadastro", which shows the date of acquisition of a particular entry, presented incoherent values with entries registered between 2021 and 2024 while the data were extracted before these years. However, using or discarding such entries was decided through individual inspection checked using orthoimage as reference. The last analysis consisted in the frequency of each species to select the ones with a higher number of individuals, as shown in Figure 3.3.

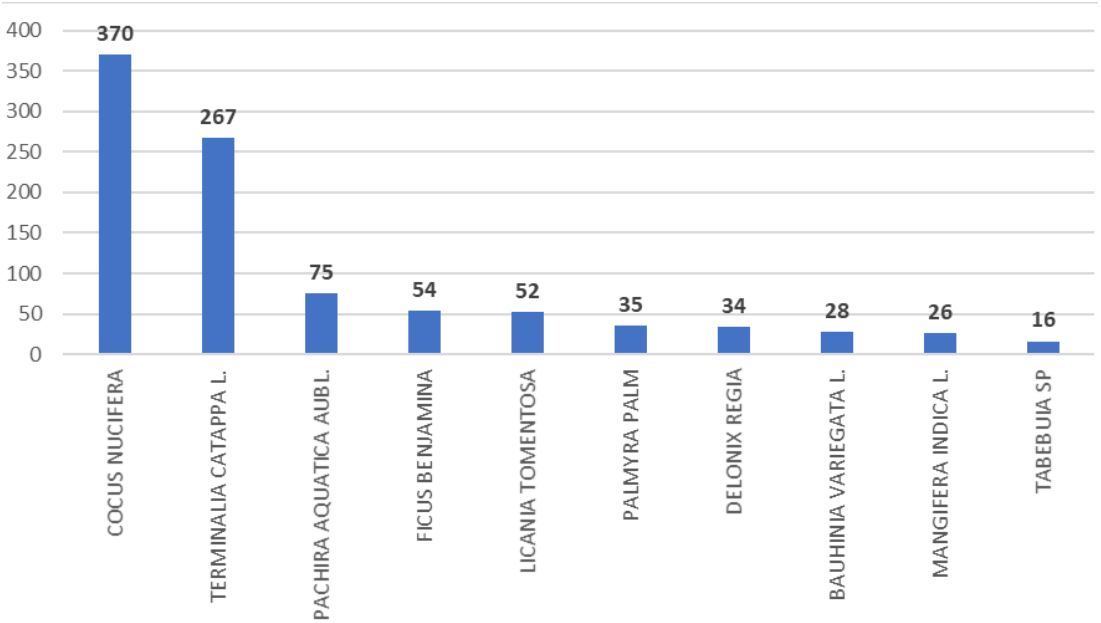


Figure 3.3 Occurrence of main tree species from tree inventory.

Once the tree inventory was evaluated and its consistencies were corrected, the corresponding tree crown of each entry was manually delineated using the 10cm-resolution orthoimage. During this step, some entries were discarded when found to be duplicated (according to date and visual inspection), or when tree crowns

of different species had overlapping area becoming impossible to distinguish which one belonged to a specific class. The manually delineated ITC (specified in Table 3.2) correspond to the ground-truth data from which variables will be extracted to feed train and test samplings for classification models.

Code	Species Name	Nº of ITC
50	*Delonix Regia*	15
70	*Ficus Benjamina*	21
90	*Licania Tomentosa*	33
130	*Pachira Aquatica AUBL.*	51
150	*Terminalia Catappa L.*	146

Table 3.2 Manually delineated ITC

3.3.1.2 Multispectral Imagery

This pre-processing step comprises the extraction of spectral information to support tree species differentiation during model development. Thus, reflectance of each band (red, green, blue and near infrared) was extracted, as well as the computation of two vegetation indices summarized in Table 3.3

Index	Band Combination	Reference
Normalized difference vegetation index (NDVI)	(NIR – Red) / (NIR + Red)	[52]
Green normalized difference vegetation index (gNDVI)	(NIR – Green) / (NIR + Green)	[53]

Table 3.3 Spectral indices calculated from multispectral imagery.

Given the availability of such high spatial resolution data, a total of 6 texture features were calculated using Structural Feature Set (SFS) application from Orfeo Toolbox (OTB): length, width, pixel shape index (PSI), weighted mean, ratio, and standard deviation. This statistical measures were proposed by Huang et al (2007) [54] and are based on direction lines, which can be understood as a series of predetermined number of equally spaced lines through the central pixel. The extension of extension line is based on the neighboring gray level similarity and the lines radiating from the central pixel in different direction. The spectral difference measured between a pixel and its central pixel defines whether this pixel lies in the homogeneous area [54], [55]. In this research, default parameters from OTB were adopted to define spectral and spatial thresholds, set to 50 and 100, respectively.

3.3.1.3 LiDAR data

The LiDAR point cloud was first processed using software LAS Tools, initially, the original tiles were tiled into smaller tiles, size of 200m and buffer of 50m, to reduce the amount of data and be able to use the free version of LAS Tools. Since the original file had very basic classification, including only three classes (unclassified, ground, and water), the first step was to apply LAS Ground to identify ground and non-ground points, followed by LAS Classify to perform classification in the points above ground level and determine if they are vegetation or buildings. Afterward, visual inspection throughout the study area helped identification of misclassified points that could be fixed manually using LAS View and LAS Layer modules. Subsequently, the elevation value of each point in the point cloud was normalized and transformed into height values using ground points as reference, this step was ran using LAS Height module.

To finalize point cloud classification, LAS Tiles was applied again in order to remove duplicated points created in the tiling and buffering procedure, followed by LAS Height process to remove points with heigh values above 30m and points belonging to class 6 (buildings) and class 9 (water), so that only relevant points would be used in the Canopy Height Model generation and LiDAR-derived variable extraction.

Once the point cloud was properly classified and reviewed, the next step consisted in the generation of a pit-free CHM, following the workflow proposed by Khosravipour et al (2014) [56]. This process was also carried out using LAS Tools packages and the final merging of the tiled image used QGIS' built-in GDAL functions. The final CHM has a spatial resolution of 0.33m, following the nominal pulse spacing distance to ensure accuracy compatible with the LiDAR sampling distance.

With the resulting processed point cloud, height, and intensity metrics, as well as canopy cover metrics, were assigned to each tree crown manually delineated. A total of 19 LiDAR-related variables were obtained and are defined and summarized in Table 3.4.

	Variable ID	*Definition*
Height Metrics	min	Minimum height value of crown return points
	max	Maximum height value of crown return points
	avg	Average height value of crown return points
	std	Standard Deviation of height values of crown return points
	p25	25th height percentile of crown return points
	p50	50th height percentile of crown return points
	p75	75th height percentile of crown return points
	b30	Percentage of points below 30% of tree height (calculated by height cutoff, known as breast height = 2m, and the maximum height)
	b50	Percentage of points below 50% of tree height (calculated by height cutoff, known as breast height = 2m, and the maximum height)
	b80	Percentage of points below 80% of tree height (calculated by height cutoff, known as breast height = 2m, and the maximum height)
Intensity Metrics	int_min	Minimum value of crown return intensity
	int_max	Maximum value of crown return intensity
	int_avg	Average value of crown return intensity
	int_std	Standard Deviation of crown return intensity
	int_p25	25th percentile of crown return intensity
	int_p50	50th percentile of crown return intensity
	int_p75	75th percentile of crown return intensity
Canopy Cover Metrics	cov	Number of first returns above the cover cutoff divided by the number of all first returns and output as a percentage
	dns	Number of all points above the cover cutoff divided by the number of all returns

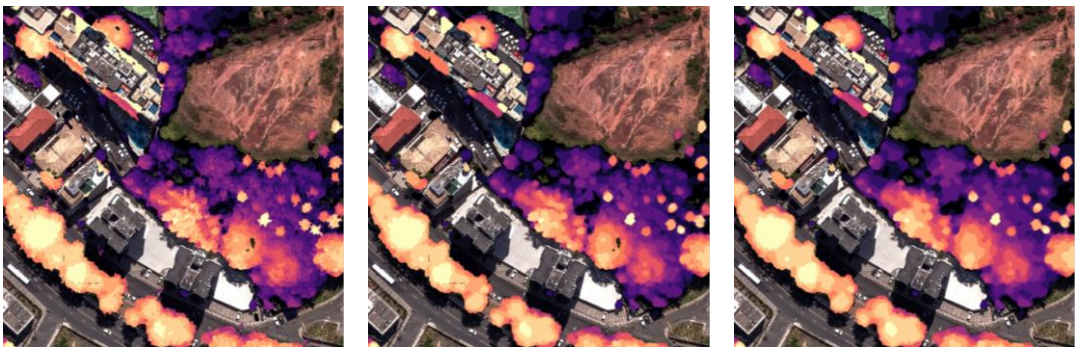
Table 3.4 Structural features derived from Airborne Laser Scanner data.

3.3.2 Crown Segmentation

Individual tree crown detection can be a challenging and costly operation in urban environments, mainly due to the existence of infrastructure elements and builds which can interfere in the performance of algorithms based on height ranges estimations to extract tree heights and, consequently, to delineate its crown. That is the reason why, prior to crown segmentation, points classified as buildings and water, and with elevation higher than 30 meters were removed from the point cloud used to generate the CHM.

Subsequently, the resulting CHM raster was submitted to morphological operations to remove noises the image and to smooth object outlines. Dilation

and erosion are basic image operations widely known in image processing, the first one is applied to remove cracks in objects and to eliminate "salt" noise inside an object, while erosion shrinks objects and removes "pepper" noise [57]. Secondary operations play a key role in image processing and they are created by combining erosion followed by dilation (known as Opening) or dilation followed by erosion (named as Closing). Opening operation is used to remove small object and to preserve the shape and size of larger objects in the image space, while Closing eliminates salt noise and fills small holes inside image objects [58]. Therefore, the CHM went through an opening operation followed by closing using a disk as structuring element with size 0.33m (corresponding to pixel size), successfully reducing the noise and filling gaps holes in image, as shown in **Figure 3.4**.



(a) (b) (c)

Figure 3.4 Morphological operations in CHM. (a) Original data, (b) Data after Opening operation, (c) Data after Closing operation.

3.3.2.1 Watershed Segmentation

A watershed segmentation algorithm was applied to extract tree crowns from the LiDAR-based CHM. This image operation is based on the idea of a grayscale image as a representation topographic relief, flooded with water, in which watersheds are represented by lines that divides the water from distinct basins [59]. The algorithm is implemented in System for Automated Geoscientific Analyses (SAGA-GIS) and it was ran using local maxima values method to identify seeds where the elevation is higher (treetops), the rule chosen to join segments is based on the difference between seed and saddle with a threshold value of 0.5m. This configuration was found after several trials with the different options to

join segments (e.g. do not join and seeds difference) and threshold values. The raster output was then converted into vector file for further operations.

Finally, the vector file went to a last filtering and cleaning, aiming to remove non-relevant objects, this step followed some constraints such as to remove objects with area values inferior to 3m², objects with negative NDVI values and tree crowns falling into polygons of highly concentrated vegetation (obtained from Salvador Mapping vector WFS service) which represents small forests and are not interesting for this research. This process represented a reduction of 55,73% of objects and assured that objects represent only tree crowns indeed.

3.3.2.2 Evaluation of semi-automatic tree crown segmentation

Crown segmentation results was assessed by comparing the tree crowns detected automatically with the manually delineated tree crowns using the multispectral imagery. Accuracy metrics for this task were errors of omission and commission, where the first occurs when no treetop is identified within the boundary of reference data, while commission error happens when more than one treetop is incorrectly detected, this procedure is similar to the one adopted by Khosravipour et al (2014) in [56].

3.3.3 Supervised Learning

This subsection covers the processes directly related to implementation and analysis of random forests, support vector machine and k-nearest neighbor classifiers. It includes dataset preparation to perform three tests with different set of features, model training and hyperparameter tuning strategy and definition of metrics adopted to evaluate each classifier's performance. Machine Learning classifiers and accuracy metrics were implemented using Scikit-learn [46].

3.3.3.1 Dataset preparation

A set of 32 features were obtained from the above-mentioned datasets after preprocessing, for each of the 266 samples representing five tree species. Multispectral-derived features were retrieved using QGIS 3.10.4 through Zonal Statistics Plugin, where the mean value of all pixels pertaining a tree crown was

assigned according to each variable. LiDAR-derived variables were assigned to manually delineated tree crown area using LAS Canopy's "plot metrics" functionality. After feature extraction, all information was stored in a tabular data frame to allow faster queries and operations using Python functions and libraries. The complete data frame is shown in Figure 3.5, however it was divided into three files: all variables, only multispectral variables, and only LiDAR variables, so each classifier would be run to these three distinct scenarios.

	id	area	min	max	avg	std	p25	p50	p75	b30	...	SFS_WMean	SFS_Ratio	SFS_Std	Red_mean	Green_mean
0	167	33.592	2.58	8.29	4.12	0.99	3.56	3.88	4.25	50.5	...	0.161	0.238	1.211	11242.546	11536.955
1	296	140.843	2.17	9.91	5.39	1.06	4.85	5.42	5.83	12.7	...	0.218	0.239	1.217	7810.218	8699.203
2	244	135.566	2.13	10.05	7.96	1.18	7.52	8.14	8.66	1.9	...	0.211	0.238	1.202	10772.376	11148.910
3	83	230.789	2.32	8.38	5.90	1.27	5.12	6.17	6.90	11.2	...	0.243	0.240	1.222	6635.670	7686.685
4	293	132.802	2.66	11.20	8.18	1.38	7.60	8.57	9.07	1.6	...	0.188	0.238	1.199	10828.621	11595.737
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
261	227	202.837	2.02	18.67	13.89	3.32	13.10	14.97	15.75	6.5	...	0.174	0.237	1.195	11988.424	13277.468
262	165	169.602	6.77	17.92	14.31	2.17	13.72	14.98	15.85	0.0	...	0.184	0.237	1.192	13987.484	13132.974
263	229	54.619	2.13	11.68	9.28	1.66	8.14	9.93	10.39	1.2	...	0.234	0.238	1.198	10880.568	11782.005
264	125	20.963	4.06	7.86	6.83	0.88	6.84	7.08	7.32	0.0	...	0.212	0.239	1.199	10680.206	11742.706
265	178	387.623	4.50	17.34	12.56	2.07	11.29	12.67	14.04	0.6	...	0.212	0.238	1.203	9689.291	11163.984

266 rows × 34 columns

Figure 3.5 Data frame with training and testing datasets containing 266 samples and 32 features.

3.3.3.2 Model training and tuning

One of this research's goals is to evaluate and compare the performance of three different ML classifiers, therefore RF, SVM and k-NN models were implemented and tuned accordingly.

Before creating each model, data is split into training and testing using *sklearn.model_selection.train_test_split* function, wherein the testing dataset is set to correspond to 30% of all features. This strategy is a traditional approach to evaluate the performance of classifiers, a crucial task when dealing with machine learning algorithm. Another relevant step to implement successful and reliable machine learning models is to perform model tuning, which consists in finding optimal values to a set hyperparameters that controls how a model will behave.

Hyperparameter optimization consists in the testing of several combinations of a model's-controlled parameters to evaluate which one is the best candidate to

provide the best result before the actual model training. This task can be done manually, but the most basic and straightforward technique is Grid Search (GS) in which a list of candidates for each hyperparameter is set and evaluated, creating a grid of possible combinations in the search space. Then, the combination that yields better results is selected and applied to the training model. However, depending on the number of hyperparameters and size of search space, GS can become very time consuming and demands a list of candidates set *a priori.*

Alternatively, Random Search (RS) is used to find optimal hyperparameters values to each model, this technique finds better models by effectively searching a larger, less promising configuration space, according to Bergstra and Bengio (2012) in [60]. Additionally, RS usually requires less computational cost. Tn this research, this technique is implemented using *sklearn.model_selection.RandomizedSearchCV* from Scikit-learn, setting 10-fold cross-validation, to find optimal values for Random Forest and SVM classifiers, k-NN does not need random search since it has only one hyperparameter to be set which can be done in a simpler way.

3.3.3.2.1 Random Forest

The RF classifier was implemented using sklearn.ensemble. RandomForestClassifier, model tuning for this algorithm aimed to find optimal values of hyperparameters such as number of trees, maximum depth and maximum number of features, the configuration for hyperparameter tuning are show in Table 3.5.

PARAMETER	VALUE(S)
CRITERION	Gini
N_TREES	100 to 500
MAX_DEPTH	None to 50

Table 3.5 User-defined RF hyperparameters for randomized search.

The Gini impurity is the default criterion used by scikit-learn RF classifier to measure the quality of a split, it is used to minimize the probability of

misclassification. The number of trees or number of estimators defines how many decision trees will be created from the dataset available, a large number of trees will generate more sub-samples and will help to reduce the bias in the data. Nevertheless, sometimes increasing the number of trees only will only spend more computational power for little or no performance gain [61]. Therefore, the number of trees varied from 100 to 500, which the literature states to be the recommended for RF [31]. Maximum depth represents the depth of each tree in the forest, a larger number implies deeper trees, in consequence, more splits to capture more information about the data. Here the maximum depth was set from None (it is possible to avoid pruning trees since RF does not overfit) to 50 splits.

3.3.3.2.2 Support Vector Classifier (SVC)

The sklearn.svm.SVC function supported the development of SVM classifier, which mainly relies in two hyperparameters: Regularization parameter (C) and gamma. The first one defines the amount of misclassification permitted for non-separable training data, allowing the adjustment of rigidity of training data. If C is large, SVM will try to minimize the number of misclassified examples and that results in a decision boundary with smaller margins[36], [39]. Kernel width parameter, also known as gamma, has direct relation with the smoothing of class-dividing hyperplane shape. Finally, the kernel type used in this research is Radial Basis Function (RBF) due to its usual good performance with remote sensing datasets. The input parameters for randomized search and respective ranges are summarized in Table 3.6.

PARAMETER	VALUES
C	0.1, 1, 10, 100, 1000
GAMMA	0.1, 1, 10, 100, 1000

Table 3.6 User-defined SVM hyperparameters for randomized search.

3.3.3.2.3 k-NN Classifier

For k-Nearest Neighbor, the only parameter to set is the k-value which defines the number of neighbors "voting" on the possible class for a specific data sample. For example, if k = 1 then the sample under evaluation will be assigned the same class the closest neighbor (or example from the validation dataset), when k = 3, then the three nearest neighbors are evaluated and the most common class among them is assigned to the sample being analyzed. For this hyperparameter, k-values ranging from 1 to 40 were tried for the dataset and a graph with k-value *versus* mean error was plotted to analyze the optimal value.

3.3.3.3 Feature Importance

Feature importance is a resource in Machine Learning used to measure and understand the impact of each feature in a model's performance, assigning scores based on how useful they are at predicting the target variable. It provides important information to perform dimensionality reduction and feature selection that can enhance the effectiveness of a classifier algorithm.

After the best hyperparameters were found, the RF classifier was trained accordingly and applied to the training dataset containing variables from both multispectral and LiDAR data. To understand which LiDAR variables are the most contributing in the model's performance, SHAP values were used to compute the feature importance in the RF model. This method is based on the Shapley values from game theory that represent the magnitude of the contribution of each feature to the model's prediction, as well as direction (sign) [62].

For SVM classifier, it would not be possible to compute feature importance given the fact that non-linear kernel functions (e.g. radial basis function, used in this research) projects the data to a space with higher dimensionality than the original feature space, being able to define boundaries in a non-linear decision surface.

3.3.3.4 Accuracy metrics

Assessing the performance of a classification model applied to remote sensing data includes the adoption of accuracy metrics, these are used to understand

how close to reality are the model's predictions. Therefore, accuracy assessment aims to compare the predicted labels assigned to an object using ML classifier and its actual label from the ground-truth data (test dataset). Table 3.7 illustrates an example of confusion matrix for a two-class problem, positive and negative. True positive (TP) values refer to samples correctly classified as Positive class, and False Positive (FP) are instances from Negative class but classified as Positive. Following the same concept, True Negatives (TN) are Negative samples correctly classified, and False Negatives (FN) represents Positive instances misclassified as negative [28].

		PREDICTED CLASS	
		Positive	Negative
ACTUAL CLASS	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 3.7 Binary confusion matrix

From the binary confusion matrix, it is possible to exemplify how to compute several metrics for each class. For this research, the adopted accuracy metrics are implemented in python using the functions available in Scikit-learn, such as: overall accuracy, user's accuracy, producer's accuracy, and F1-score.

- Overall Accuracy: represents the proportion of correctly classified reference sites (elements in diagonal) divided by the total number of reference sites. It is presented as percentage and calculated as follows:

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

- User's Accuracy (UA): shows the accuracy from the perspective of a map user, this metric tells us how often the class on the map will actually be present on the ground.

$$UA = \frac{TP}{(TP + FP)}$$

- Producer's Accuracy (PA): map's accuracy from the map maker point of view, it represents how often are actual features correctly represented on the predicted map.

$$PA = \frac{TP}{(TP + FN)}$$

- F1-score: weighted average of the precision and recall

$$F1 = 2\frac{(OA * PA)}{(OA + PA)}$$

4 Results

4.1 Semi-automatic crown segmentation evaluation

This evaluation consisted in the individual analysis of 276 tree crowns, manually delineated using the very high spatial resolution orthoimage, in comparison with the objects resulting of the watershed segmentation using the LiDAR-derived CHM. Figure 4.1 brings examples of the errors found.

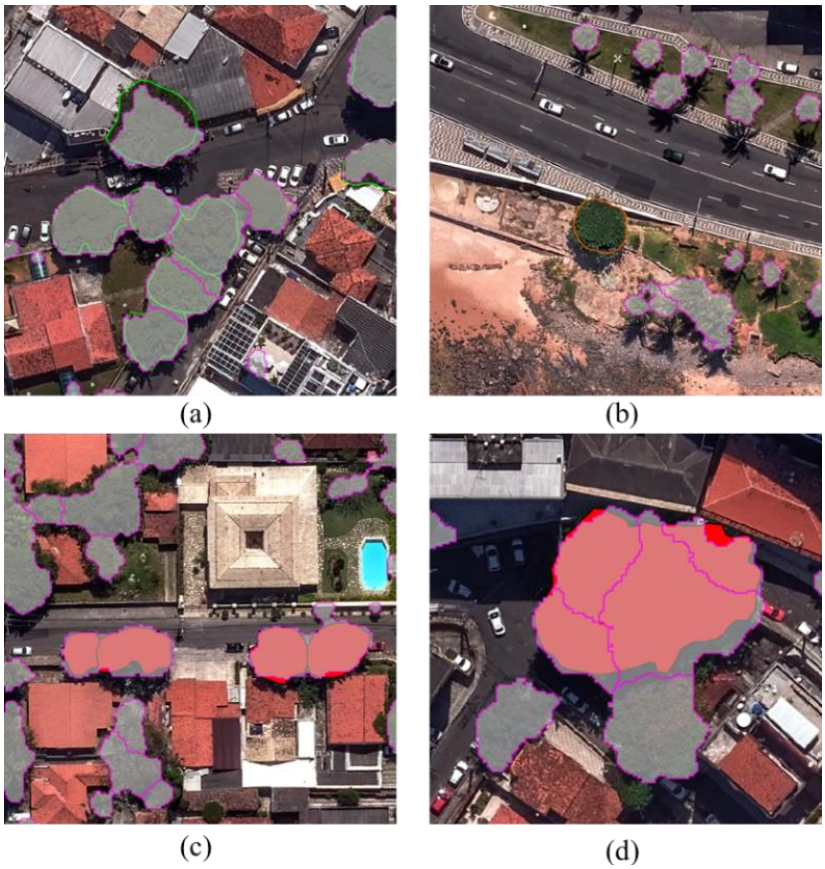


Figure 4.1 Evaluation of semi-automatic crown segmentation, gray objects with borders in magenta represent the segmentation results. (a) correctly segmented, (b) omitted objects, (c) under-segmented objects, (d) over-segmented objects.

The least common error was omission, only five treetops from the reference dataset were not detected in the segmentation process, perhaps these features could have been excluded in the filtering process if it were composed of multiple objects with area inferior to 3m² or even if the NDVI value for those were lower than the applied mask.

When analyzing commission error, it was proven the need of breaking down this error into two more specific classification: over-segmentation and under-segmentation. The first accounted for 4,71% of reference dataset, which exemplify the segmentation's inability to correctly identify single object and, as consequence, creating multiple polygons to represent an individual tree crown, as shown in Figure 4.1 d. In the opposite idea, under-segmented objects (35,15%) are representations of segmented feature that should have been separated into two or more objects to correctly objects to represent the tree crowns within its borders, as shown in (Figure 4.1 d).

Nevertheless, the 161 tree crowns correctly segmented instances (Figure 4.1 a) stand for the process's reliability to identify treetops, success supported also by the application of filters based on vegetation indices and height values from LiDAR point cloud. It is important to highlight that the CHM's resolution (33cm) is smaller than the orthoimage's (10cm), therefore slight differences in the shape and size of tree crowns is to be expected, which is also related to the fact that reference data was extracted manually. A visual assessment of the objects also points out to the outstanding capability to segregate overlapping tree crowns.

4.2 Hyperparameter Tuning and Feature Importance

For the RF classifier, initially all models were trained using nº of trees equals to 100, and maximum depth equals to none. When accounting only multispectral features, tuning RF's hyperparameters slightly increased the overall accuracy in 2,5%, however it remained unable to identify and label one the species (*Delonix Regia*). In the second test, the model could predict all classes using only LiDAR variables, the decrease in overall accuracy benefitted intra-class accuracy. The

best outcome of this model tuning appeared in the third dataset combination, in which the overall accuracy increased in 6,3%.

In the support vector classifier, the default parameters before model tuning corresponded to C equals to 1, gamma value was set to 'scale'[1]. SVM classifier had a similar issue as RF classifier, as it was not able to identify one class (*Ficus Benjamina* this time) even after model tuning when using only multispectral features. Slight increase in OA is noticed with the model ran with LiDAR features, whereas the third test increased OA in 4% when optimal parameters were set.

Finally, k-NN's performance upgraded 1,25% for the first test using only multispectral variables, being the only model to predict all classes in this case. The default k-value to train each model was 5.

Results of each classifier, including optimal hyperparameters and overall accuracy values pre- and post-model tuning, are summarized in Table 4.1.

	RF				SVM				K-NN		
Variables source	OA before tuning	nº of trees	Max depth	OA after tuning	OA before tuning	C	gamma	OA after tuning	OA before tuning	k value	OA after tuning
Multispectral	73,8%	144	25	72,5%	73,8%	100	0,01	68,8%	71,3%	6	72,5%
LiDAR	62,5%	100	35	62,5%	61,3%	1000	0,001	67,5%	62,5	12	63,75
Multispectral + Lidar	75,0%	144	10	82,5%	78,8%	10	0,01	81,3%	76,3%	12	77,5%

Table 4.1 Optimal hyperparameters for classifier models.

After finding optimal parameters, the most relevant features for RF model were ranked using SHAP values.

From the 20 most important features ranked in Figure 4.2, the 5 most important LiDAR variables are standard deviation of crown return intensity (int_std), minimum height value of crown returns (min), average height value of crown return points (avg), 75th height percentile of crown return points (p75) and percentage of points below 30% of tree height (b30). In relation to the multispectral-derived variables, near infrared band, vegetation indices (gNDVI

[1] **scale =** 1 / (n_features * X.var()) in Scikit-learn

and NDVI) and green band are in top of feature importance, along with the pixel shape index from the texture analysis.

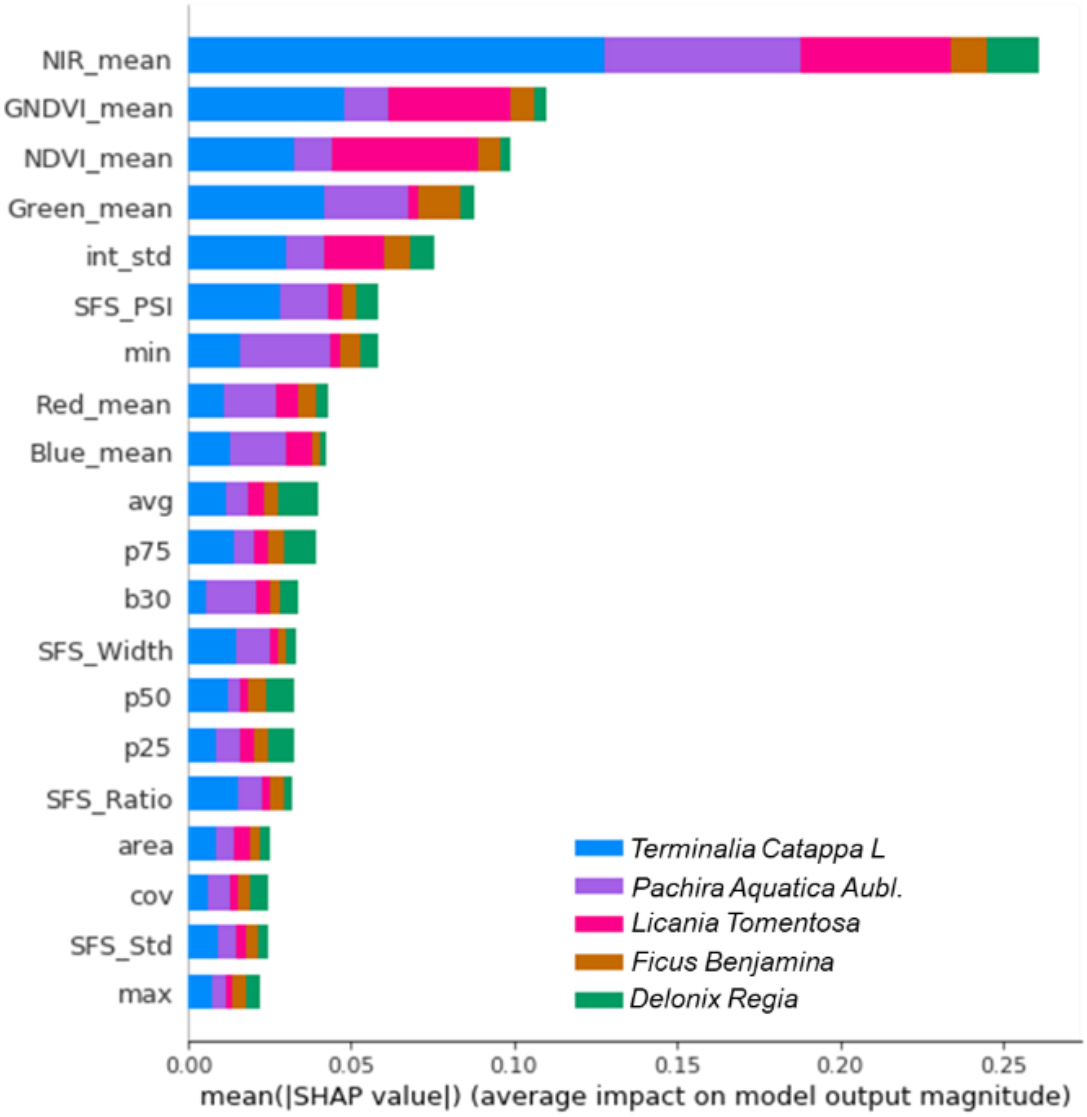


Figure 4.2 SHAP values to rank the most important features in the classification with RF applied to multispectral and LiDAR data combined.

4.3 Tree Species Classification

Confusion matrices for each classifier's performance is shown in Figure 4.3 grouped by dataset combination, while user's and producer's accuracies are displayed in three different tables briefly discussed in the following paragraphs.

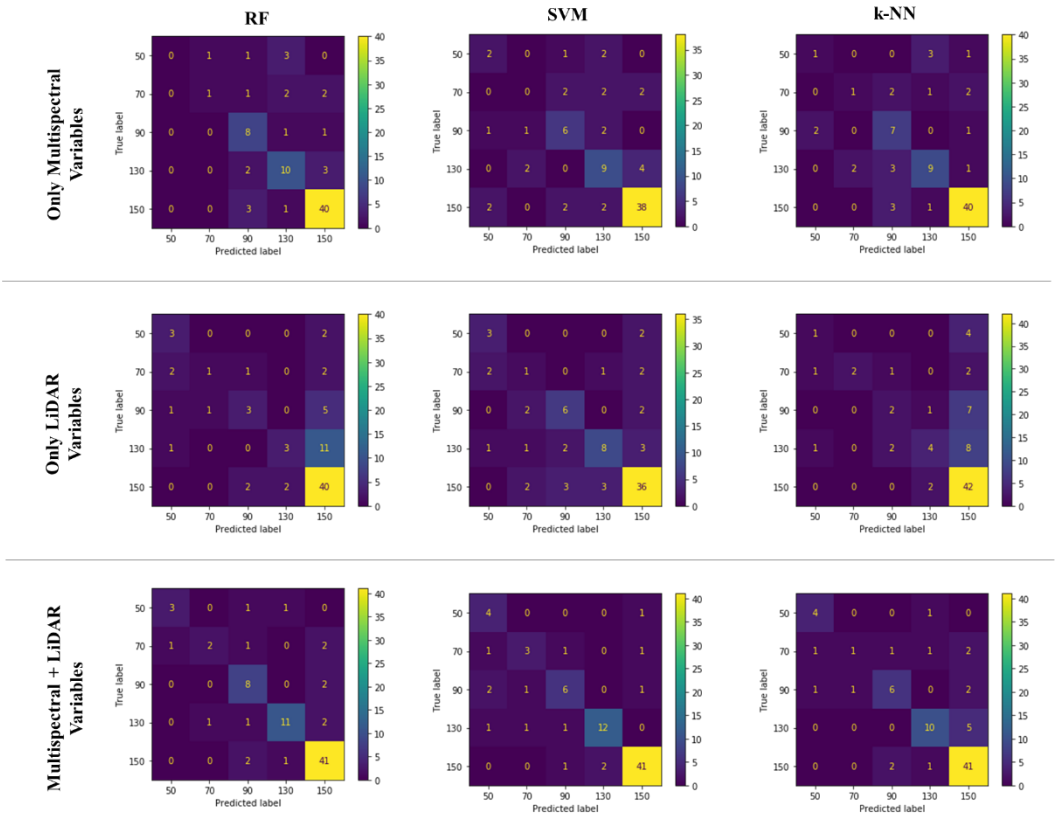


Figure 4.3 Confusion Matrices illustrating the performance of each model to classify 5 tree species using three datasets combinations. 50 – *Delonix Regia*, 70 – *Ficus Benjamina*, 90 – *Licania Tomentosa*, 130 – *Pachira Aquatica Aubl.*, and 150 – *Terminalia Catappa L.*

Table 4.2 compares the results for tree species classification using the three selected models trained using only multispectral variables, with highlights in gray to the species that an algorithm was not capable to obtain enough information to classify it correctly in testing dataset. k-NN showed a better performance in terms of overall accuracy, being also the only model able to detect all classes, even though species like *Delonix Regia* and *Ficus Benjamina* showed a poor performance when analyzing user's and producer's accuracy metrics values (lower than 35%), these are the classes with lower number of samples.

Figure 4.4 shows the average reflectance curves for each of the 5 selected species along the visible and near-infrared bands, as expected they show high similarity in the visible part of the spectrum with some considerable differences between reflectance values in the red band for *Pachira Aquatica Aubl* in comparison to all other species. From the graph, *Terminalia Catappa L.* shows higher values in the near-infrared portion and it is in accordance with the

contribution of that band in for that species' discrimination as shown in the variable importance analysis (Figure 4.2) for the RF classifier.

Species Name	RF OA=72,50%			SVM OA=68,75%			k-NN **OA=72,50%**		
	UA	PA	F1	UA	PA	F1	UA	PA	F1
Delonix Regia	0,0%	0,0%	0,0%	40,0%	40,0%	40,0%	33,3%	20,0%	25,0%
Ficus Benjamina	100%	16,7%	28,6%	0,00%	0,00%	0,0%	33,3%	16,7%	22,2%
Licania Tomentosa	50,0%	80,0%	61,5%	54,6%	60,0%	57,1%	46,7%	70,0%	56,0%
Pachira Acquatica Aubl.	62,5%	66,7%	64,5%	52,9%	60,0%	56,3%	64,3%	60,0%	62,1%
Terminalia Catappa L.	82,9%	88,6%	85,7%	86,4%	86,4%	86,4%	88,9%	90,9%	89,9%

Table 4.2 Classifiers' performance using only multispectral features (OA=overall accuracy, UA=user's accuracy, PA=producer's accuracy, and F1=F1-score).

When using only LiDAR-derived variables, the three models were able to perform classification to all 5 species (Table 4.3), even though the overall accuracy was lower in general. Large differences in accuracy could be noted between species, producer's accuracy ranged from 16,6% (*Ficus Benjamina* in SVM and k-NN) to 95,5% (*Terminalia Catappa L.*). Best performances in user's accuracy were 80,0% and 66,7%, both values registered for *Terminalia Catappa L.*, when using SVM and RF, respectively. In terms of overall accuracy, SVM had a greater value but a tad different from k-NN (-3,75%).

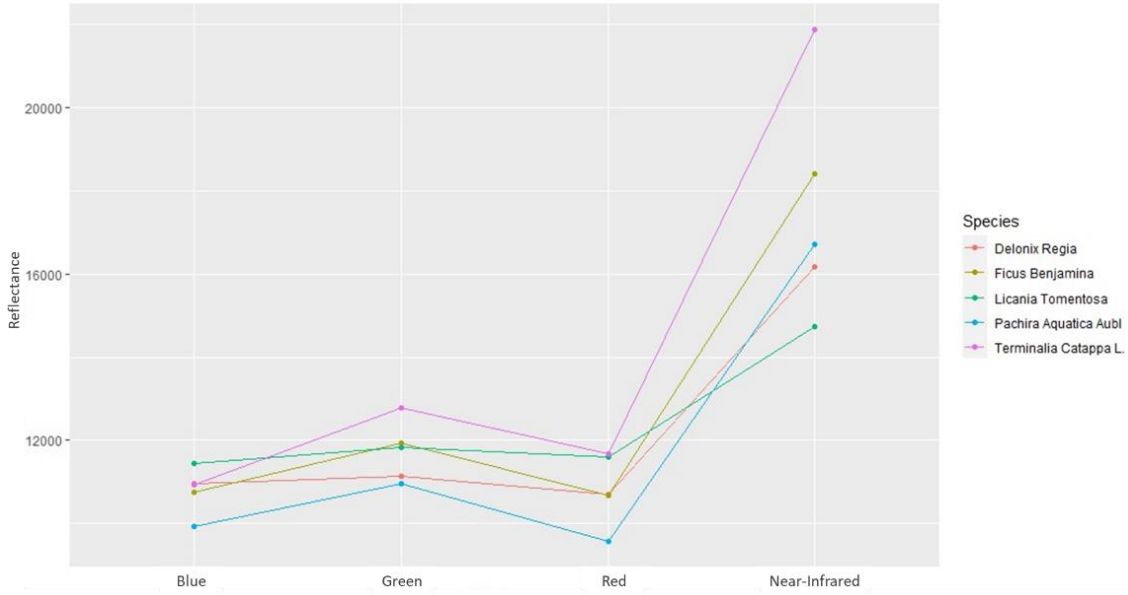


Figure 4.4 Average reflectance value (x10000) of all 5 species at visible and near-infrared bands.

Species Name	RF OA=62,50%			SVM **OA=67,50%**			k-NN OA=63,75%		
	UA	PA	F1	UA	PA	F1	UA	PA	F1
Delonix Regia	42,9%	60,0%	50%	50,0%	60,0%	54,5%	33,3%	20,0%	25,0%
Ficus Benjamina	50,0%	16,7%	25%	16,7%	16,7%	16,7%	100%	33,3%	50,0%
Licania Tomentosa	50,0%	30,0%	37,5%	54,5%	60,0%	57,1%	40,0%	20,0%	26,7%
Pachira Acquatica Aubl.	60,0%	20,0%	30%	66,7%	53,3%	59,3%	57,1%	26,7%	36,4%
Terminalia Catappa L.	66,7%	90,9%	76,9%	80,0%	81,8%	80,9%	66,7%	95,5%	78,5%

Table 4.3 Classifiers' performance using only airborne LiDAR features.

Lastly, the combination of both datasets brought a considerable improvement in performance, as shown in Table 4.4. RF classified improved 20% in its overall accuracy when compared to results using only LiDAR features, SVM classification accuracy increased from 68,75% with multispectral variables to 81,25% with combined datasets, while k-NN had a variation of 13,75% between the LiDAR and combination of both datasets. Not only the classification accuracies were benefited from this configuration of variables, but also user's and producer's accuracy for some of the species had a better performance such as *Delonix Regia* exceeded user's accuracy in all previous tests, reaching the mark of 75% for Random Forest. However, producer's accuracy did not improve for *Ficus Benjamina*, a class that persistently performed poorly throughout all possible sets of variables and models trained in this research.

Species Name	**RF OA=82,50%**			SVM OA=81,25%			k-NN OA=77,5%		
	UA	PA	F1	UA	PA	F1	UA	PA	F1
Delonix Regia	75,0%	60,0%	66,7%	50,0%	80,0%	61,5%	66,7%	80,0%	72,3%
Ficus Benjamina	66,7%	33,3%	44,4%	60,0%	50,0%	54,6%	50,0%	16,7%	25,0%
Licania Tomentosa	61,5%	80,0%	69,6%	66,7%	60,0%	63,2%	66,7%	60,0%	63,2%
Pachira Acquatica Aubl.	84,6%	73,3%	78,6%	85,7%	80,0%	82,8%	76,9%	66,7%	71,4%
Terminalia Catappa L.	87,2%	93,2%	90,1%	93,2%	93,2%	93,2%	82,0%	93,2%	87,2%

Table 4.4 Classifiers' performance using multispectral and airborne LiDAR features.

Given the best overall performance of RF classifier in the latter dataset combination, when compared to the two other classifiers and to previous variables configuration, this trained model was applied to the production data

(meaning the tree crowns segmented from the LiDAR-derived CHM) and an example of the result is illustrated in Figure 4.5.

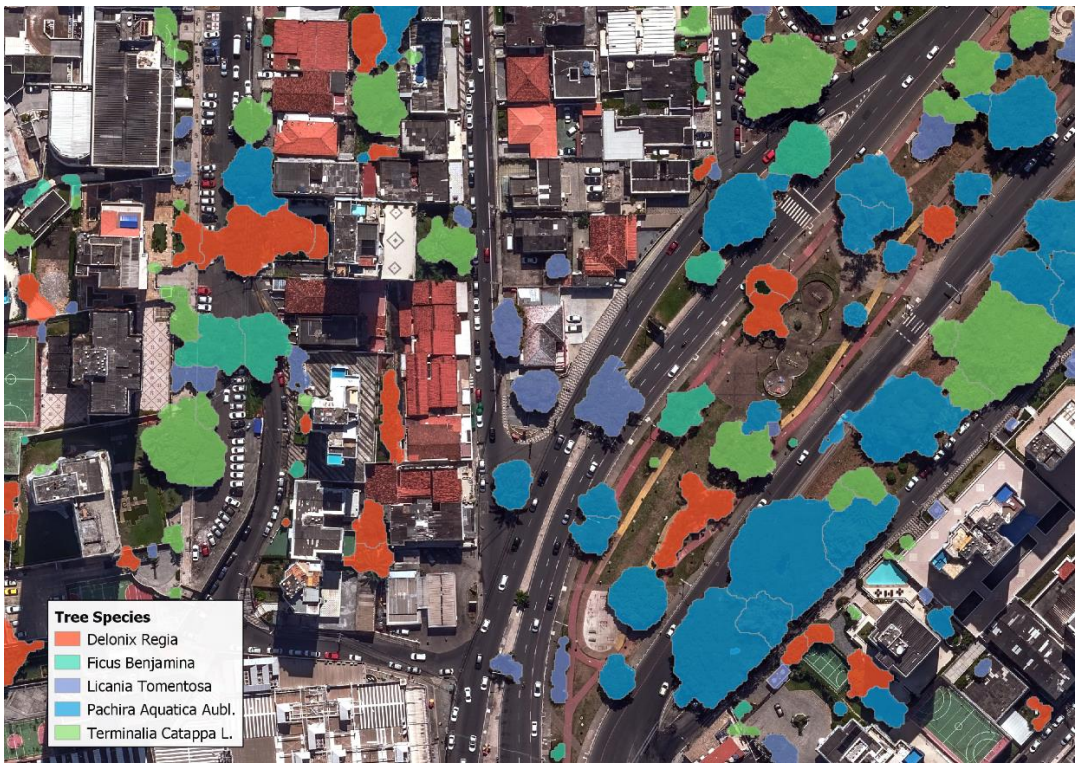


Figure 4.5 Random Forest classification applied to production data.

5 Discussion

The output from semi-automatic tree crown segmentation proved the efficiency of the process carried out using a LiDAR-derived CHM, even though its evaluation pointed to a tendency of aggregating multiple tree crowns into single objects, the watershed segmentation can be fine-tuned to meet the user's need. However, searching for optimal values is time-consuming and it is difficult to notice slight differences through visual inspection. In addition, the structure and shape of trees also had an impact in the CHM segmentation process to detect treetops, Zhen et al (2016, [63]) point out to the fact that most of the algorithms for individual tree crown detection assume a basic conical crown shape, which does not benefit the tree species in this research.

Manual delineation of individual tree crowns was a very sensitive task, even with very high-resolution images. The morphology of the trees did not favor this task, the constant occurrence of overlapping crowns made it harder to distinguish the limits, and the presence of shadows from buildings also had impact on the visual inspection to delineate individual tree crowns, which impacted in the final sampling size. Considering that, the OBIA approach with segmented objects from the LiDAR-derived CHM brings the advantage of using the point cloud distribution to detect tree crowns, which does not suffer from interference of shadows.

One of the main goals of this research was to understand the impact of dataset combination in the classification accuracy for three different machine learning models. The results revealed that the combination of both multispectral and LiDAR variables increased the performance of all classifiers, with improvements in overall accuracy up to 13% when comparing the findings with only one of the sources of information. The outstanding performance was shown by Random Forest classifier, yielding overall accuracy of 82,50% and user's and producer's accuracy higher than 60%, except for *Ficus Benjamina* which is the species with the worst performance regardless of the dataset combination or classifier.

This improvement in accuracy, brought by the combination of datasets, was expected since many studies related to tree species classification were benefited by the combination of datasets, however most of them are performed using hyperspectral data which provides more detailed information, therefore leads to better results found by Sothe et al (2019, [9]), Shen and Cao (2017,[23]),and Ferreira et al (2019, [51]), for example.

Using only LiDAR variables as input proved to be more efficient for all models to correctly identify the five tree species elected in this research. This can be related to the fact that LiDAR data provides information about the tree crown's structure, as it was stated by Liu et al (2017,[3]) about variables that relate to the characteristics of laser point distribution are valuable assets for tree species classification. Such features were ranked among the top 15 most important features in the RF classifier, including minimum and average height value of crown

return points, standard deviation and 75th percentile of crown return intensity. Canopy cover, canopy density and crown size variables did not contribute much to the model's performance, this fact can be attributed to the regular tree pruning that urban management and maintenance bureau performs to control tree growth and prevent problems with overhead electrical wiring, for instance.

The performance of classification for tree species such as *Delonix Regia* could have been compromised due the tree's structure, majorly composed by tree trunks and small leaves, which can lead to background reflectance effect where both spectral reflectance and laser pulse return can provide information about the bare ground and/or grass below the tree. Also, the data acquisition happened before its blooming season, when the red-orange flowers that characterize this tree appear and could have been a good way to easily distinguish this species from the other ones.

Ficus Benjamina was the species with worst performance, as highlighted earlier, and the reason for that could be associated to this species' similarity with every other species in terms of spectral reflectance values in both visible and near-infrared bands (**Figure 4.4**). The results shown by the confusion matrices (**Figure 4.3**) reveals that this class is persistently mislabeled and associated to other classes, showing no pattern in the misclassification. Lastly, from the feature importance rank using SHAP values, it is also possible to notice that none of the multispectral or LiDAR variables had a relevant contribution for *Ficus Benjamina*. Therefore, it is possible to say that multispectral data does not provide enough information to distinguish this species and it could perform better with the addition of physiological aspects, which cannot be computed with the available dataset.

Finally, the RF classifier (as well as the other two) yielded higher classification accuracies for the two most represented classes (*Terminalia Catappa L.* and *Pachira Aquatica Aubl*), this fact rises a red flag to the influence of the imbalanced dataset in the model's efficiency. Other studies also reported the limitation imposed by the difficulty in acquiring similar number of samples per tree species, given the high cost associated to ground-truth data collection [9], [64].

The results are consistent and promising, especially considering that most of similar studies have the advantage of using hyperspectral data to achieve similar results, however the reproducibility and application of the methodology proposed in this research to the whole city or similar environments face some challenges.

Firstly, the existing tree inventory is far from being representative of the species diversity held by Salvador, over 50 species were registered but the population is quite low and scattered all over the city, plus the data acquisition does not provide a precise location or has other inconsistencies that make the tree identification process harder. Therefore, it is important to dedicate some additional time to increase the sample size and represent better the diversity of species.

Moreover, the number of species selected for this research can also have an impact in the relatively high overall accuracy values found. Sothe et al (2019, [9]) performed classification using 12 tree species and achieved accuracy of 72,4% using UAV point cloud and hyperspectral data. Ferreira et al (2016, [65]) applied machine learning algorithms to classify 8 species in a Brazilian subtropical forest and accomplished 84% of accuracy when associating VNIR hyperspectral bands and shortwave infrared bands. In their pixel-based classification, Féret and Asner (2013, in [64]) mapped 17 tree species in a tropical forest located in Hawaii, with sampling ranging from 1 to 168 tree crowns, achieved overall accuracy of 73,2% using airborne hyperspectral data. Therefore, future work for this study area should include other species, even with small number of available samples, to analyze the performance of classifiers and consider the extrapolation of the models to other areas of the city.

Another challenge is the processing of LiDAR point cloud for an entire metropole like Salvador, it would require more computational power and perhaps another software to process such an amount of data considering also the specificities related to the city's elevation profile. Still concerning the point cloud, during the pre-processing stage, it was noticed that the nominal pulsing space is not the same for the whole city, the reason for that is the mapping's final elevation products were specified to have a spatial resolution of 50cm. Therefore,

generating the CHM with this spatial resolution will impact the information quality and segmentation outputs.

6 Conclusions

This thesis demonstrated the application of multispectral aerial image and airborne LiDAR data to identify and classify five urban tree species, in a tropical environment in the city of Salvador (Brazil), focusing on the detection of individual trees. The research was conducted using three different machine learning classifiers (random forest, support vector machine and k-nearest neighbor) assigned to three sets data inputs (multispectral variables, LiDAR variables and combination of both datasets) to evaluate the performance and which arrangement would yield better results. The highest overall accuracy found was 82,50% when applying random forest classifier to the combination of multispectral and LiDAR-derived features. The outlined research questions for this research have been discussed in more detail in the previous section and the findings are summarized in the following paragraphs.

Regarding the first research question, all the classifiers' performance were similar in terms of overall accuracy since the discrepancies did not exceed 5%, classification accuracy for most species were satisfactory except for *Ficus Benjamina* that consistently performed poorly in all scenarios and classifiers. Random forest and support vector machine classifiers outperformed k-nearest neighbor in most of the cases, except when using only multispectral variables, in which case k-NN had better overall accuracy as well as it was the only model able to detect every single tree species in the dataset (while RF was not able to identify *Delonix Regia* and SVM failed to distinguish *Ficus Benjamina*).

The answer to the second research question was supported by the feature importance analysis done with random forest model applied to both multispectral and LiDAR datasets, in which it was possible to notice that amongst the 19 metrics extracted from the LiDAR point cloud, the most contributing features were the standard deviation of crown return intensity, height-related metrics (minimum,

average and 75th percentile), and height bincentile that delivers the percentage of points between the breast height and a tree's maximum height.

The application of a LiDAR-derived CHM was proved to be an effective method towards the semi-automatic extraction of tree crowns in an urban environment, where the presence the of buildings and infrastructure elements make this task more complex. However, this approach relies heavily in the correctness of point cloud classification which is a time-consuming task that requires the setting of many parameters to be successful.

Lastly, the results found in this research are relevant for urban forestry inventory and management for many reasons. First, it can provide a consistent overview of predominant tree species in the city, with that information local authorities and specialists can define strategies to plant native species to replace invasive ones such as *Terminalia Catappa L.*, originally from Asian and Australian coastal environments, which is harmful to sidewalks and cause damages to both public roads and electrical wiring. Second, this approach can be used to support the implementation of a structured tree inventory using geographic database systems, with integrated use by many public sectors such as maintenance bureau, real state, environmental policy makers.

7 Bibliographic References

[1] B. Ministério do Meio Ambiente (Brasil), "Mapa de Vegetação Nativa na Área de Aplicação da Lei no. 11.428/2006 - Lei da Mata Atlântica (ano base 2009)," Brasília - DF, 2015.

[2] J. Aval, "Automatic mapping of urban tree species based on multi-source remotely sensed data," Universite de Toulouse, 2018.

[3] L. Liu, N. C. Coops, N. W. Aven, and Y. Pang, "Mapping urban tree species using integrated airborne hyperspectral and LiDAR remote sensing data," *Remote Sens. Environ.*, vol. 200, no. November 2016, pp. 170–182, 2017.

[4] C. Edson and M. G. Wing, "Airborne light detection and ranging (LiDAR) for individual tree stem location, height, and biomass measurements," *Remote Sens.*, 2011.

[5] S. Kim, T. Hinckley, and D. Briggs, "Classifying individual tree genera using stepwise cluster analysis based on height and intensity metrics derived from airborne laser scanner data," *Remote Sens. Environ.*, 2011.

[6] F. E. Fassnacht *et al.*, "Review of studies on tree species classification from remotely sensed data," *Remote Sens. Environ.*, vol. 186, pp. 64–87, 2016.

[7] C. Hamamura, "Sensoriamento remoto para identificação taxonômica e mapeamento de espécies arbóreas em ambiente urbano," Universidade de São Paulo, Piracicaba, 2020.

[8] C. A. Baldeck *et al.*, "Operational tree species mapping in a diverse tropical forest with airborne imaging spectroscopy," *PLoS One*, vol. 10, no. 7, 2015.

[9] C. Sothe *et al.*, "Tree species classification in a highly diverse subtropical forest integrating UAV-based photogrammetric point cloud and hyperspectral data," *Remote Sens.*, vol. 11, no. 11, 2019.

[10] W. C. Chew, A. M. S. Lau, and K. D. Kanniah, "Multi-level adaptive support vector machine classification for tropical tree species," *Int. J. Geoinformatics*, vol. 12, no. 2, pp. 17–25, 2016.

[11] M. Cross, T. Scambos, F. Pacifici, O. Vargas-Ramirez, R. Moreno-

Sanchez, and W. Marshall, “Classification of tropical forest tree species using meter-scale image data,” *Remote Sens.*, vol. 11, no. 12, pp. 1–18, 2019.

[12] D. Harrison, B. Rivard, and A. Sánchez-Azofeifa, “Classification of tree species based on longwave hyperspectral data from leaves, a case study for a tropical dry forest,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 66, no. December 2017, pp. 93–105, 2018.

[13] S. Janhäll, “Review on urban vegetation and particle air pollution - Deposition and dispersion,” *Atmos. Environ.*, vol. 105, pp. 130–137, 2015.

[14] R. Pu and S. Landry, “A comparative analysis of high spatial resolution IKONOS and WorldView-2 imagery for mapping urban tree species,” *Remote Sens. Environ.*, vol. 124, pp. 516–533, 2012.

[15] P. H. A. Cruz, M. Heimer, and J. C. Pedrassoli, “Ocupação indevida em unidades de conservação: estudo de caso no Parque Metropolitano de Pituaçu com uso de imagens orbitais disponíveis na nuvem,” in *Simpósio Brasileiro de Sensoriamento Remoto, 18. (SBSR)*, 2017, pp. 2287–2292.

[16] N. Guimarães, L. Pádua, P. Marques, N. Silva, E. Peres, and J. J. Sousa, “Forestry remote sensing from unmanned aerial vehicles: A review focusing on the data, processing and potentialities,” *Remote Sens.*, vol. 12, no. 6, 2020.

[17] K. Wang, T. Wang, and X. Liu, “A review: Individual tree species classification using integrated airborne LiDAR and optical imagery with a focus on the urban environment,” *Forests*, vol. 10, no. 1, pp. 1–18, 2018.

[18] M. D. Hossain and D. Chen, “Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective,” *ISPRS J. Photogramm. Remote Sens.*, vol. 150, no. February, pp. 115–134, 2019.

[19] T. Blaschke, “Object based image analysis for remote sensing,” *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, 2010.

[20] M. Dalponte, L. Frizzera, and D. Gianelle, “Individual tree crown delineation and tree species classification with hyperspectral and LiDAR data,” *PeerJ*,

vol. 2019, no. 1, 2019.

[21] Z. Zhang, A. Kazakova, L. M. Moskal, and D. M. Styers, "Object-based tree species classification in urban ecosystems using LiDAR and hyperspectral data," *Forests*, vol. 7, no. 6, pp. 1–16, 2016.

[22] K. Lim, P. Treitz, M. Wulder, B. St-Ongé, and M. Flood, "LiDAR remote sensing of forest structure," *Prog. Phys. Geogr.*, vol. 27, no. 1, pp. 88–106, 2003.

[23] X. Shen and L. Cao, "Tree-species classification in subtropical forests using airborne hyperspectral and LiDAR data," *Remote Sens.*, vol. 9, no. 11, 2017.

[24] W. S. W. M. Jaafar *et al.*, "Improving individual tree crown delineation and attributes estimation of tropical forests using airborne LiDAR data," *Forests*, vol. 9, no. 12, pp. 1–23, 2018.

[25] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review," *Int. J. Remote Sens.*, vol. 39, no. 9, pp. 2784–2817, 2018.

[26] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *Geosci. Front.*, vol. 7, no. 1, pp. 3–10, 2016.

[27] R. Boutaba *et al.*, "Comprehensive survey Machine Learning," *J. ofInternet Serv. andApplications*, vol. 9, no. 16, p. 99, 2018.

[28] 2011 Bruce, "Encyclopedia of Machine Learning (OUT)," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.

[29] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. 2005.

[30] L. Breiman, "Random forests," *Mach. Learn.*, 2001.

[31] M. Belgiu and L. Drăgu, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, 2016.

[32] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2005.

[33] H. Wang, M. Lei, Y. Chen, M. Li, and L. Zou, “Intelligent identification of maceral components of coal based on image segmentation and classification,” *Appl. Sci.*, vol. 9, no. 16, pp. 1–15, 2019.

[34] M. Mahdianpari, B. Salehi, F. Mohammadimanesh, and M. Motagh, “Random forest wetland classification using ALOS-2 L-band, RADARSAT-2 C-band, and TerraSAR-X imagery,” *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 13–31, 2017.

[35] P. Griffiths, B. Jakimow, and P. Hostert, “Reconstructing long term annual deforestation dynamics in Pará and Mato Grosso using the Landsat archive,” *Remote Sens. Environ.*, vol. 216, no. October 2017, pp. 497–513, 2018.

[36] P. Thanh Noi and M. Kappas, “Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery,” *Sensors (Basel).*, vol. 18, no. 1, 2017.

[37] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, “An assessment of the effectiveness of a random forest classifier for land-cover classification,” *ISPRS J. Photogramm. Remote Sens.*, vol. 67, no. 1, pp. 93–104, 2012.

[38] C. Kamusoko, “Remote sensing image classification in R,” no. march, p. 201, 2019.

[39] G. Mountrakis, J. Im, and C. Ogole, “Support vector machines in remote sensing: A review,” *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 3, pp. 247–259, 2011.

[40] V. Sharma, D. Baruah, D. Chutia, P. Raju, and D. K. Bhattacharya, “An assessment of support vector machine kernel parameters using remotely sensed satellite data,” *2016 IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. RTEICT 2016 - Proc.*, no. May, pp. 1567–1570, 2017.

[41] R. Vohra and K. C. Tiwari, “Comparative Analysis of SVM and ANN Classifiers using Multilevel Fusion of Multi-Sensor Data in Urban Land Classification,” *Sens. Imaging*, vol. 21, no. 1, 2020.

[42] M. Syifa, S. J. Park, A. R. Achmad, C. W. Lee, J. Eom, and J. Eom, "Flood mapping using remote sensing imagery and artificial intelligence techniques: A case study in Brumadinho, Brazil," *J. Coast. Res.*, 2019.

[43] H. Sun *et al.*, "Optimizing kNN for mapping vegetation cover of arid and semi-arid areas using landsat images," *Remote Sens.*, vol. 10, no. 8, 2018.

[44] R. Han, P. Liu, G. Wang, H. Zhang, and X. Wu, "Advantage of combining ObiA and classifier ensemble method for very high-resolution satellite imagery classification," *J. Sensors*, vol. 2020, 2020.

[45] Q. Meng, C. J. Cieszewski, M. Madden, and B. E. Borders, "K nearest neighbor method for forest inventory using remote sensing data," *GIScience Remote Sens.*, 2007.

[46] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, 2011.

[47] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Miscellaneous Clustering Methods*. 2011.

[48] T. Ren *et al.*, "Early identification of seed maize and common maize production fields using sentinel-2 images," *Remote Sens.*, vol. 12, no. 13, pp. 1–21, 2020.

[49] J. Cao, W. Leng, K. Liu, L. Liu, Z. He, and Y. Zhu, "Object-Based mangrove species classification using unmanned aerial vehicle hyperspectral images and digital surface models," *Remote Sens.*, vol. 10, no. 1, 2018.

[50] D. Li, Y. Ke, H. Gong, and X. Li, "Object-based urban tree species classification using bi-temporal worldview-2 and worldview-3 images," *Remote Sens.*, vol. 7, no. 12, pp. 16917–16937, 2015.

[51] M. P. Ferreira, F. H. Wagner, L. E. O. C. Aragão, Y. E. Shimabukuro, and C. R. de Souza Filho, "Tree species classification in tropical forests using visible to shortwave infrared WorldView-3 images and texture analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, no. January, pp. 119–131, 2019.

[52] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sens. Environ.*, 1979.

[53] A. A. Gitelson, Y. J. Kaufman, and M. N. Merzlyak, “Use of a green channel in remote sensing of global vegetation from EOS- MODIS,” *Remote Sens. Environ.*, 1996.

[54] X. Huang, L. Zhang, and P. Li, “Classification and extraction of spatial features in urban areas using high-resolution multispectral imagery,” *IEEE Geosci. Remote Sens. Lett.*, 2007.

[55] V. Thierion, S. Alleaume, C. Jacqueminet, C. Vigneau, K. Michel, and S. Luque, “The potential of Pléiades imagery for vegetation mapping: an example of grasslands and pastoral environments.,” *Rev. Fr. Photogramm. Teledetect.*, 2014.

[56] A. Khosravipour, A. K. Skidmore, M. Isenburg, T. Wang, and Y. A. Hussin, “Generating pit-free canopy height models from airborne lidar,” *Photogramm. Eng. Remote Sensing*, 2014.

[57] E. R. Davies, *Image filtering and morphology*. 2018.

[58] E. R. Dougherty and R. A. Lotufo, “Binary Opening and Closing,” *Hands-on Morphol. Image Process.*, vol. 25, pp. 25–44, 2009.

[59] A. S. Kornilov and I. V. Safonov, “An overview of watershed algorithm implementations in open source libraries,” *Journal of Imaging*. 2018.

[60] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, 2012.

[61] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[62] R. Rodríguez-Pérez and J. Bajorath, “Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions,” *J. Comput. Aided. Mol. Des.*, 2020.

[63] Z. Zhen, L. J. Quackenbush, and L. Zhang, “Trends in automatic individual tree crown detection and delineation-evolution of LiDAR data,” *Remote Sensing*. 2016.

[64] J. B. Feret and G. P. Asner, “Tree species discrimination in tropical forests

using airborne imaging spectroscopy," *IEEE Trans. Geosci. Remote Sens.*, 2013.

[65] M. P. Ferreira, M. Zortea, D. C. Zanotta, Y. E. Shimabukuro, and C. R. de Souza Filho, "Mapping tree species in tropical seasonal semi-deciduous forests with hyperspectral and multispectral data," *Remote Sens. Environ.*, 2016.

Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS