



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Assessing Tree Crown Detection Methods for Urban and Rural Monitoring Using VHR Remote Sensing Imagery

TESI DI LAUREA MAGISTRALE IN
GEOINFORMATICS ENGINEERING - INGEGNERIA GEOINFORMATICA

Author: **Nataly Alejandra Sarmiento Ospina**

Student ID: 11002168

Advisor: Prof. Vasil Yordanov

Co-advisors: Federica Mezzani

Academic Year: 2025-26

*A mi padre, José Alvaro quien incluso con manos de
piedras las hizo sentir de algodón*

*A mi madre, Gloria, por su eterno optimismo y, por sus
brazos, que forman mi refugio más seguro*

*A Paula que me permitió siempre reír en voz alta e hizo
de la escucha nuestro portal.*

*A Nicolás que se ha mantenido en el proceso de encon-
trarnos y en la naturalidad de nuestra conexión*

*A Nicolás O por haber sido mi cómplice y más valiente
compañía, juntos construimos nuestro propio país*

A Raul por ser respaldo incondicional y guía firme y,

*A María Fernanda e Isabella por ser zona de confort en
mi zona fuera de confort*

Abstract

Individual tree crown detection from Very High Resolution (VHR) imagery is a core problem in remote sensing for urban and rural vegetation monitoring, yet published studies rarely compare representative methods under identical experimental conditions. The primary goal of this thesis is to systematically compare five representative detection paradigms. This thesis presents a controlled benchmark of five crown detection paradigms applied to a single multispectral scene (JILIN-1 KF01C approximately 0.5 m resolution, four spectral bands including near-infrared) acquired over Reggio Calabria, Italy, in June 2024, and evaluated against 800 manually delineated reference crown polygons distributed equally between urban and rural landscape strata.

The five methods span the full methodological spectrum: local maxima with watershed segmentation, object-based image analysis (OBIA) with Random Forest, a pixel-based Random Forest classifier, DeepForest (a pretrained RetinaNet detector), and U-Net (a semantic segmentation model with a ResNet-34 encoder). All methods share a common preprocessing pipeline, a fixed 512×512 px tile grid with spatial block partitioning, and a shared evaluation using Hungarian assignment IoU matching at thresholds of 0.25 and 0.50 on a fixed 302-crown test set. The threshold of 0.25 serves as a lenient criterion that captures coarse crown detections and tolerates annotation boundary uncertainty, while 0.50 provides a stricter standard reflecting more precise crown delineation.

As a benchmark comparison, no single method achieved satisfactory detection performance, and results are reported here as reference values rather than evidence of methodological superiority. Across the five methods, DeepForest produced the highest F1-score ($F_{1_{25}} = 0.313$; $F_{1_{50}} = 0.297$), the highest recall ($R_{25} = 0.503$), and the highest matched mean IoU ($\overline{\text{IoU}}_{25} = 0.666$). U-Net achieved the highest precision ($P_{25} = 0.307$) but the lowest recall among the supervised methods ($R_{25} = 0.235$). The three handcrafted methods (watershed, OBIA with Random Forest, and pixel-based Random Forest) clustered in a narrow F1 band ($F_{1_{25}}$ of 0.103–0.127). The rural stratum was consistently harder across all methods: DeepForest led in both strata but with a pronounced urban advantage ($F_{1_{25}} = 0.430$ urban versus 0.248 rural). The main failure modes were the spectral similarity between target crowns and surrounding Mediterranean scrub and olive vegetation,

and building cast shadow in the urban core. A recurring finding across supervised methods was a train, test distribution shift that required per-stratum threshold recalibration.

The benchmark demonstrates that detection accuracy and boundary fidelity are partly independent properties that must be reported jointly, and that a single model and threshold do not transfer optimally across heterogeneous urban and rural Mediterranean landscapes. The reproducible end-to-end pipeline provides a reference for method selection in precision forestry and urban vegetation monitoring under VHR satellite imagery.

Keywords: tree crown detection, very high resolution imagery, object-based image analysis, deep learning, remote sensing, benchmark evaluation

Abstract in lingua italiana

Il rilevamento delle chiome arboree individuali da immagini a Very High Resolution (VHR) rappresenta un problema centrale nel telerilevamento per il monitoraggio della vegetazione urbana e rurale; tuttavia, gli studi pubblicati raramente confrontano metodi rappresentativi in condizioni sperimentali identiche. L'obiettivo principale di questa tesi è confrontare sistematicamente cinque paradigmi rappresentativi di rilevamento. La tesi presenta un benchmark controllato di cinque paradigmi di rilevamento delle chiome applicati a una singola scena multispettrale JILIN-1 KF01C, con risoluzione di circa 0,5m e quattro bande spettrali, incluso il vicino infrarosso, acquisita su Reggio Calabria, Italia, nel giugno 2024, e valutati rispetto a 800 poligoni di riferimento di chiome delineati manualmente e distribuiti in parti uguali tra strati paesaggistici urbani e rurali.

I cinque metodi coprono l'intero spettro metodologico: massimi locali con segmentazione watershed, analisi di immagine basata su oggetti (OBIA) con Random Forest, classificatore Random Forest a livello di pixel, DeepForest, un rilevatore RetinaNet pre-addestrato, e U-Net, un modello di segmentazione semantica con encoder ResNet-34. Tutti i metodi condividono una pipeline di preprocessing comune, una griglia di tile fissa di 512×512 px con partizionamento spaziale a blocchi, e una valutazione condivisa basata su matching IoU mediante assegnazione ungherese alle soglie di 0,25 e 0,50 su un test set fisso di 302 chiome. La soglia di 0,25 funge da criterio permissivo, in grado di catturare rilevamenti approssimativi delle chiome e di tollerare l'incertezza nei confini delle annotazioni, mentre la soglia di 0,50 costituisce uno standard più rigoroso, associato a una delineazione più precisa delle chiome.

Come confronto di benchmark, nessun metodo ha raggiunto prestazioni di rilevamento soddisfacenti, e i risultati sono riportati come valori di riferimento piuttosto che come prova di superiorità metodologica. Tra i cinque metodi, DeepForest ha ottenuto il valore più alto di F1-score ($F_{1_{25}} = 0,313$; $F_{1_{50}} = 0,297$), il recall più elevato ($R_{25} = 0,503$) e la più alta IoU media sui match ($\overline{\text{IoU}}_{25} = 0,666$). U-Net ha raggiunto la precision più alta ($P_{25} = 0,307$), ma il recall più basso tra i metodi supervisionati ($R_{25} = 0,235$). I tre metodi artigianali, watershed, OBIA con Random Forest e Random Forest a livello di pixel, si sono concentrati in una fascia ristretta di $F_{1_{25}}$ compresa tra 0,103 e 0,127.

Lo strato rurale è risultato costantemente più difficile per tutti i metodi: DeepForest ha mantenuto le migliori prestazioni in entrambi gli strati, ma con un marcato vantaggio urbano ($F1_{25} = 0,430$ in ambito urbano contro 0,248 in ambito rurale). Le principali modalità di errore sono state la somiglianza spettrale tra le chiome target e la vegetazione circostante, composta da macchia mediterranea e uliveti, e le ombre proiettate dagli edifici nel nucleo urbano. Un risultato ricorrente nei metodi supervisionati è stato uno shift di distribuzione tra train e test, che ha richiesto una ricalibrazione delle soglie per ciascuno strato.

Il benchmark dimostra che l'accuratezza del rilevamento e la fedeltà dei confini sono proprietà parzialmente indipendenti che devono essere riportate congiuntamente, e che un singolo modello e una singola soglia non si trasferiscono in modo ottimale tra paesaggi mediterranei urbani e rurali eterogenei. La pipeline end-to-end riproducibile fornisce un riferimento per la selezione dei metodi nella selvicoltura di precisione e nel monitoraggio della vegetazione urbana tramite immagini satellitari VHR.

Parole chiave: rilevazione delle chiome arboree, immagini ad elevata risoluzione, analisi di immagini basata su oggetti, deep learning, telerilevamento, valutazione comparativa

Acknowledgements

I wish to thank the Politecnico di Milano and Italy for providing me the amazing opportunity to develop this master. I am grateful to Prof. Vasil Yordanov for his expert supervision and guidance throughout this work. I extend my appreciation to BIP Italia for offering me the chance to conduct this project in a professional setting, bridging academic research with practical applications. Special thanks go to Federica for her co-supervision and continuous support, and to Roberto and Daniele for their guidance. Finally, I would like to acknowledge all those who, though not explicitly mentioned, contributed to this achievement through their support, feedback, and encouragement.

Contents

Abstract	i
Abstract in lingua italiana	iii
Acknowledgements	v
Contents	vii
1 Introduction	1
1.1 Problem Statement	2
1.2 Research Objectives	3
1.3 Thesis Structure	4
2 State of the Art	5
2.1 Individual Tree Crown Detection	5
2.2 Methodological Evolution	5
2.3 Local Maxima and Watershed Segmentation	6
2.4 Object-Based Image Analysis	7
2.5 Machine Learning and Feature Engineering	8
2.6 Deep Learning Methods	9
2.7 Evaluation Frameworks in ITCD	12
2.8 Urban versus Rural Crown Detection	12
2.9 Research Gap and Thesis Positioning	13
3 Dataset Description	15
3.1 Study Area	15
3.2 Satellite Imagery	16
3.3 Reference Data	18
4 Methodology	19

4.1	Research Design	19
4.2	Tools and Software Environment	21
4.3	Reference Annotation Methodology	22
4.4	Preprocessing Pipeline	24
4.5	Evaluation Framework	27
4.6	Local Maxima and Watershed Segmentation	29
4.7	OBIA with Random Forest	32
4.8	Pixel-Based Random Forest	36
4.9	DeepForest	40
4.10	U-Net Semantic Segmentation	44
4.11	Analytical Hypotheses	48
5	Results	51
5.1	Reference Dataset Results	51
5.2	Preprocessing Results	53
5.3	Watershed and Local Maxima	64
5.4	OBIA with Random Forest	69
5.5	Pixel-Based Random Forest	72
5.6	DeepForest	74
5.7	U-Net	76
5.8	Cross-Method Comparison	78
6	Discussion	81
6.1	Interpretation of Main Findings	81
6.2	Methodological Trade-offs	85
6.3	Urban–Rural Robustness and Generalization	85
6.4	Practical Implications	86
6.5	Limitations	86
7	Conclusions and Future Work	89
7.1	Recommendations for Future Research	92
	Bibliography	95
	List of Figures	101
	List of Tables	105

1 | Introduction

Rural and urban tree populations are essential components of ecological resilience, climate regulation, and human well-being. At the same time, they are increasingly exposed to pressures arising from land-use change, heat stress, fragmentation, and biodiversity loss [16]. Reliable tree-level monitoring is therefore not only a scientific problem, but also a practical requirement for precision forestry and urban environmental management [6].

Tree crown detection from Very High Resolution (VHR) imagery sits at the intersection of remote sensing, computer vision, and ecological analysis. The task is deceptively simple in definition, but difficult in practice: an algorithm must identify where trees are, separate adjacent crowns, and maintain robustness across different canopy structures and background conditions. These difficulties become more pronounced in VHR data, where the increased spatial detail reveals shadow edges, internal canopy texture, mixed materials, and non-vegetated objects that can resemble crowns at first glance.

The methodological history of tree crown detection reflects the broader evolution of remote sensing analysis. Early studies relied on heuristic image processing and local geometric assumptions. Later work incorporated object-based image analysis and supervised classification. More recent approaches use deep learning to infer crown representations directly from annotated data. This progression has improved performance, but it has also made the field more fragmented, because studies often differ in imagery, labels, evaluation metrics, and application context. As a result, it remains difficult to compare methods in a way that is both fair and operationally meaningful.

This thesis addresses that gap by developing a unified benchmark for Individual Tree Crown Detection (ITCD) across representative methodological families, applied to a JILIN-1 KF01C multispectral scene acquired over Reggio Calabria, Italy, in June 2024. At approximately 0.5 m spatial resolution and four spectral bands including the near-infrared, the imagery provides sufficient detail for crown-level delineation while remaining representative of the commercial VHR data increasingly available for environmental monitoring. Reggio Calabria was selected because its landscape encompasses both a dense Mediterranean urban core and surrounding rural and peri-urban vegetation, offering a

single acquisition in which urban and rural canopy conditions can be evaluated under identical sensor and illumination characteristics. In this work, urban refers to areas dominated by built surfaces with fragmented, isolated tree cover, while rural refers to areas with continuous or semi-continuous canopy, including cultivated and semi-natural Mediterranean vegetation. Ground truth consists of manually delineated crown polygons produced through careful visual interpretation of the imagery, which serve as the reference for both training and evaluation across all methods. The objective is not only to measure which method performs best in a narrow sense, but to understand how each paradigm behaves in relation to data characteristics, interpretation needs, and computational requirements.

1.1. Problem Statement

Existing studies in individual tree crown detection have demonstrated that crowns can be identified from high-resolution imagery using a range of approaches, from rule-based geometric heuristics and object-based image analysis to classical machine learning classifiers and deep convolutional neural networks. However, the reported performance of these methods is difficult to compare across studies because experimental conditions differ substantially in sensor characteristics, spatial resolution, annotation type, crown definition, landscape context, and evaluation metric. A method reporting ninety percent overall accuracy on a well-separated coniferous stand in North America offers little guidance on how a detection pipeline will behave on Mediterranean olive groves at half-meter resolution. As a result, it remains unclear whether observed performance differences reflect the intrinsic capacity of the methods or the particular conditions under which they were evaluated.

This limitation is particularly consequential in mixed urban–rural Mediterranean environments, where isolated street trees, building shadows, roads, bare soil, cultivated orchards, and semi-continuous hillside canopy coexist within a single acquisition. Urban and rural landscapes generate structurally different error modes: in urban areas, the dominant failure is false detection on spectrally ambiguous built surfaces and shadow edges; in rural areas, the dominant failure is crown merging, boundary ambiguity, and the spectral similarity between target tree crowns and surrounding maquis or olive vegetation. A method optimized for one of these contexts may degrade substantially in the other, yet most published benchmarks focus on a single landscape type and report aggregate accuracy without disaggregating urban and rural performance. Furthermore, accuracy alone is insufficient to guide operational method selection. Interpretability and annotation effort constrain

which approaches can be deployed in practice, and these factors are rarely reported alongside detection metrics. The unresolved question is therefore not simply which method is best in a narrow sense, but which method is most appropriate under which landscape conditions and resource constraints.

This thesis addresses that gap by constructing a controlled experimental framework in which five methods spanning the full methodological spectrum, from local maxima watershed segmentation and object-based image analysis to Random Forest classification, DeepForest, and U-Net, are applied to the same imagery, evaluated against the same reference polygons, and compared using a consistent set of detection and segmentation metrics. Classical, object-based, machine learning, and deep learning approaches are rarely tested together under the same VHR satellite scene, and the present benchmark provides the first such comparison for a Mediterranean urban–rural environment using JILIN-1 KF01C imagery at approximately 0.5 m resolution.

1.2. Research Objectives

1.2.1. General Objective

The general objective of this thesis is to systematically evaluate and compare representative methods for individual tree crown detection using VHR remote sensing imagery, with particular emphasis on their suitability for urban and rural monitoring.

1.2.2. Specific Objectives

- Characterize the methodological evolution of tree crown detection from heuristic image processing to deep learning.
- Implement spectral, rule-based, object-based, classical machine learning, and deep learning approaches within a common experimental framework applied to a single VHR multispectral dataset.
- Assess how preprocessing, annotation design, and evaluation choices influence the comparability of crown detection methods.
- Quantify performance in terms of detection accuracy and segmentation quality across urban and rural landscape strata.
- Analyze the strengths and limitations of each method across rural and urban environments representative of the Mediterranean context.

1.3. Thesis Structure

The thesis is organised so that the reader can follow a single, continuous chain of steps from the definition of the problem to its empirical resolution. Chapter 2 reviews the state of the art in ITCD, using the literature to motivate each later design choice rather than as a detached survey. Chapter 3 describes the dataset: the study area, the JILIN-1 KF01C imagery, and the manually delineated reference crowns. Chapter 4 develops the methodology and implementation jointly: the preprocessing pipeline, the shared evaluation framework, and each of the five ITCD methods are described both at the conceptual level and with the concrete parameters, thresholds, and training settings used in the experiment. Chapter 5 then reports the outcomes in the same order, the preprocessing products and the per-method detection metrics, with a closing cross-method comparison. Chapter 6 interprets those results against the expectations established in the state of the art and discusses methodological trade-offs, robustness, practical implications, and limitations. Chapter 7 summarises the contributions and outlines directions for future work.

2 | State of the Art

2.1. Individual Tree Crown Detection

ITCD refers to the remote-sensing task of identifying individual tree instances, locating their crowns, and, when possible, delineating crown boundaries from image data [36]. The task is important because crown-level information supports tree counting, crown-size estimation, and the derivation of forest structural variables such as stem density, diameter-related proxies, and biomass indicators [1]. In Very High Resolution (VHR) imagery, ITCD becomes feasible because the spatial detail is sufficient to represent crown-scale patterns[27], but the same detail also introduces shadows, internal canopy texture, mixed background materials, and ambiguous crown borders [35].

The technical difficulty of ITCD arises from the fact that a tree crown is not always a spectrally or geometrically isolated object [10]. Crowns may overlap, branches may create several local peaks within the same tree, small trees may be suppressed by larger neighbouring crowns, and background surfaces may be confused with vegetation under complex illumination conditions [38]. These issues are especially relevant in heterogeneous landscapes, where isolated urban trees, built surfaces, shadows, continuous vegetation, and semi-natural canopy can occur within the same image scene [7, 12]. As a result, ITCD is not only a classification problem, but also an instance-separation problem in which the algorithm must distinguish neighbouring crowns that may share similar spectral responses [9, 22].

2.2. Methodological Evolution

The earliest tree crown detection workflows were shaped by practical constraints: limited computing power, coarse imagery, and a strong need for interpretable rules. Early methods typically began with manual or semi-automatic preprocessing and then used local intensity or shape assumptions to identify treetops and crown extents, often framing the problem as one of extracting salient peaks from a canopy surface model or from smoothed imagery [20]. A major limitation of these early approaches was their depen-

dence on strong assumptions about crown form and scene regularity: they could perform well in controlled rural conditions, but were sensitive to noise, parameter choice, and crown overlap. Nevertheless, they established a methodological principle that remains relevant: crown detection should be evaluated as an object extraction problem, not only as a pixel classification problem.

As imagery and computing resources improved, the field progressed through four broadly identifiable phases: geometric heuristics, object-based segmentation and classification, feature-engineered machine learning, and representation learning through deep neural networks [36]. This progression has improved performance, but it has also made the field more fragmented, because studies often differ in imagery, labels, evaluation metrics, and application context [36]. As a result, it remains difficult to compare methods in a way that is both fair and operationally meaningful, which is the central motivation for this thesis.

2.3. Local Maxima and Watershed Segmentation

The watershed transform became a foundational segmentation tool because it formalized the idea of splitting an image into catchment basins based on local minima and gradient structure [30]. In tree crown applications, watershed segmentation is frequently combined with local maxima detection, where the maxima serve as markers for likely treetops, translating crown detection into a geometric problem: identifying candidate treetops and expanding or separating crown regions from those markers [21, 27]. Their main advantage is that they do not require large labelled training datasets, which makes them useful as transparent baselines when annotation resources are limited [37].

The literature shows that the performance of this method family is highly dependent on imagery type, forest structure, and parameter configuration.[35]. In an urban RGB satellite study over the Beijing Forestry University campus, watershed-based delineation detected 886 trees but missed 290 and produced 137 false detections, reporting up to 85 % one-to-one treetop location accuracy with 15 % commission error and 23 % omission error, with omissions linked to image resolution, segmentation limitations, and overlapped crowns [27]. In a direct comparison study, watershed reached only 33.8 % accuracy and local maxima 36.6 % accuracy on a campus scene, with both methods incorrectly identifying branches as trees [38]. In litchi garden scenes, both methods produced extensive omission and commission errors because crown shapes were irregular, crowns were closely connected, and some background areas had high spectral values [38].

The sensitivity of this method family is most evident in its dependence on scale, marker

definition, and crown morphology. In mangrove environments, local maxima were considered less suitable because many mangrove crowns are relatively flat rather than conical, weakening the assumption that each crown has a clear radiometric peak [10]. Adding DSM information improved visual crown boundary quality but also increased false detections when multiple branches of the same tree were interpreted as separate crowns — a central trade-off of height-assisted watershed methods [10]. A particularly relevant case is the multi-scale segmentation method of Qiu et al. [21], which explicitly compared against traditional and marker-controlled watershed segmentation. In a rainforest plot, the proposed method reached 76.59 % accuracy compared with 56.10 % for marker-controlled watershed, reducing commission error from 35.61 % to 8.29 %. In a deciduous forest plot, the improvement was from 45.51 % to 63.20 %, attributed to multi-scale crown slices and the representation of treetops as maximum regions rather than single points [21].

The methodological value of watershed-based approaches is therefore not that they solve the problem completely, but that they provide an interpretable baseline whose failure modes are informative: over-segmentation indicates that the scene contains more fine-scale variation than the heuristic can handle, while under-segmentation indicates that adjacent crowns are not separable by local contrast alone. For this reason, local maxima and watershed segmentation are included in the benchmark as a reference point against which more advanced methods are compared.

2.4. Object-Based Image Analysis

As VHR imagery became more common, pixel-based classification began to look increasingly inadequate for object-level tasks. Object-Based Image Analysis (OBIA) emerged as a response by grouping pixels into segments prior to classification and by explicitly incorporating shape, texture, and contextual features [2, 5]. This shift was particularly important for tree crown detection because crowns are spatial objects whose boundaries matter. In practice, however, OBIA introduced new challenges: the segmentation step became highly sensitive to scale, shape, and compactness parameters [2, 19]. OBIA therefore solved one problem while creating another, reducing pixel noise but increasing dependence on manual tuning.

The features used in OBIA workflows are commonly defined from spectral, textural, geometric, and contextual information. In a date palm detection study, Al-Ruzouq et al. [2] optimized multiresolution segmentation through F1-score-based parameter selection and ant colony optimization, selecting 12 significant colour-invariant and textural features from an initial set of 31. GLCM homogeneity in the blue and green bands, together with

colour ratios, were identified as effective crown descriptors. In a Leipzig urban tree-cover study, OBIA combined digital orthophotos, DEM, DSM, LiDAR derivatives, and height measures to separate trees from buildings and other anthropogenic surfaces [3].

In a QuickBird-2 study at 0.6 m resolution, rule-based object classification used Normalized Difference Vegetation Index (NDVI) thresholds and brightness values to classify vegetation and trees [19]; the index and its role in this work are detailed in Section 4.4.2.

Reported performance is generally strong for tree canopy or target-species mapping under favourable conditions, as summarised in Table 2.1. However, even in successful cases, minor misclassifications occurred between crowns, other vegetation, built-up areas, and shadows, with causes including imperfect segmentation, limited spectral information, and shadow effects [2]. Random Forest appears in this literature either as a supervised object-based classifier or as a comparison target for other algorithms, and its performance degrades when class spectral signatures vary between acquisition conditions [33]. The main limitation of OBIA plus Random Forest is therefore not only the classifier, but the dependence of the full workflow on object generation, feature design, and scene-specific thresholds [2, 3].

Study	Sensor / GSD	OA (%)	Kappa	Target
Al-Ruzouq et al. [2]	VHSR aerial	91.9 / 87.0	0.90 / 0.85	Date palm
Kadu et al. [14]	WorldView-2	92.4	0.80	Urban trees
Kamal [15]	WorldView-2 / 0.5 m	68.0	—	Mangrove crowns

Table 2.1: Reported OBIA performance across selected ITCD studies. OA = overall accuracy. Dual values in Al-Ruzouq et al. [2] correspond to two study areas. No reference in the provided bibliography reports a standardized F1 or IoU range specifically for OBIA plus Random Forest in ITCD under object-level matching.

2.5. Machine Learning and Feature Engineering

Classical machine learning represented another step toward more flexible analysis. Instead of relying exclusively on fixed thresholds or segmentation heuristics, algorithms such as Random Forest learn decision boundaries from labelled data and engineered features [4]. For tree crown detection, the strength of classical machine learning lies in its balance between interpretability and adaptability: it can combine spectral indices, textural mea-

tures, and morphological descriptors in a single classifier, and feature importance can be examined more easily than in deep neural networks [2].

However, pixel-based classification alone does not automatically solve the instance-level separation of neighbouring crowns, so additional post-processing or object extraction is needed when the goal is individual crown detection rather than vegetation masking [31, 36]. In tropical forest, Wagner et al. [31] stated that it is difficult to select a single forest-wide spectral threshold for spectrum-based ITCD because of the complexity of tropical canopies. Al-Ruzouq et al. [2] also reported that misclassifications occurred between target crowns, other vegetation, built-up areas, and shadow, with causes including imperfect segmentation, limited RGB spectral information, and shadow effects. No reference in the provided bibliography covers a complete pixel-based Random Forest ITCD pipeline reporting precision, recall, F1, IoU, or object-level matching results, which supports the need for its explicit evaluation in the benchmark.

2.6. Deep Learning Methods

Deep learning changed the field by reducing the need for handcrafted features and enabling end-to-end learning from annotated imagery [39]. Convolutional neural networks can learn hierarchical representations that combine local texture, broader spatial context, and semantic structure. For tree crown detection, this matters because crowns are not uniform objects: they vary in shape, illumination, and density, and their boundaries are often ambiguous. The main limitation of deep learning methods is that they require annotated data, careful training configuration, and sufficient computational resources [7, 9, 26].

2.6.1. Detection-Oriented Models: DeepForest

Deep object detectors locate crowns as bounding boxes rather than exact masks, which is useful when the primary objective is to locate crowns and count trees. Unlike two-stage detectors such as Faster R-CNN, which separate region proposal from classification [23], DeepForest is built on a one-stage RetinaNet detector that predicts boxes directly, and it demonstrated that general-purpose detection models can be adapted to ecological imagery [32]. In an urban tree detection study, the original DeepForest model reached a recall of 0.69, precision of 0.61, and F1-score of 0.65. After single-tree fine-tuning for one epoch, test precision reached 0.794, recall 0.594, and F1 0.680 [7]. These results suggest that DeepForest can be reused in urban environments with moderate detection performance, but the balance between precision and recall remains sensitive to the training configuration

[7]. In a broader urban detection benchmark, anchor-free methods reached the best average precision of 0.686, while anchor-based two-stage methods remained competitive in more challenging images [34].

No reference in the provided bibliography covers DeepForest-specific transfer from RGB-trained models to multispectral VHR satellite imagery, nor a standardized IoU range for this method. This gap is directly relevant to the present thesis, which uses JILIN-1 KF01C RGB+NIR imagery at approximately 0.5 m resolution, making the unified benchmark necessary to test whether DeepForest can generalise to specific multispectral, urban, and rural conditions not covered by existing studies.

2.6.2. Semantic Segmentation: U-Net

Semantic segmentation models such as U-Net learn dense pixel-wise predictions and are widely used where object extent matters more than object identity [24]. U-Net’s encoder–decoder structure preserves spatial detail through skip connections, which helps recover fine crown boundaries. However, semantic segmentation assigns the same label to all tree pixels, meaning it cannot separate adjacent crowns without additional post-processing such as connected-component analysis or watershed-based instance separation [22].

Reported U-Net performance varies substantially across datasets, model configurations, and evaluation protocols, as summarised in Table 2.2. Schindler et al. [26] found that segmentation accuracy depended strongly on object size, with F1-scores of 0.665 for large objects, 0.451 for medium-sized objects, and 0.182 for small objects, a limitation directly relevant to crown detection where slight spatial mismatches reduce IoU below the matching threshold. Rahman et al. [22] showed that a self-attention U-Net dual-task framework improved instance-level segmentation accuracy over a U-Net baseline by 41.5 %, increasing Tree IoU from 0.2694 to 0.3810, suggesting that U-Net-style segmentation benefits from auxiliary outputs and attention mechanisms when individual crown separation is the objective. In urban tree canopy extraction, U-Net was strongly affected by canopy shadow, producing a high omission rate; Mask R-CNN increased overall accuracy by 3.39 % and F1-score by 0.128 over the U-Net baseline [35].

Study	Sensor / GSD	Prec.	Rec.	F1	IoU
Jin et al. [13]	— / —	0.886	0.783	0.827	0.713
Schindler et al. [26]	Aerial / 10 cm	0.517	0.546	0.531	—
Martins et al. [17]	Airborne RGB	—	—	0.914–0.899	0.739–0.700
Ulku et al. [29]	Multispectral	—	—	—	0.885 (Jaccard)

Table 2.2: Reported U-Net performance across selected ITCD and canopy segmentation studies. No reference in the provided bibliography covers U-Net performance on JILIN-1 KF01C RGB+NIR imagery or reports focal loss or weighted cross-entropy specifically for U-Net in ITCD.

2.6.3. Other Deep Learning Approaches

Other deep learning architectures for ITCD include YOLO-family detectors, transformer-based frameworks, Mask R-CNN, and domain-adaptation models [12, 34, 35, 37]. Performance can be high when training data, spatial resolution, and crown separability are favourable: YOLOv8m reached an F1-score of 0.932 and mAP@0.50 of 0.934 [28], while Mask R-CNN achieved an overall accuracy of 93.98 %, F1-score of 0.831, and IoU of 82.92 % for urban tree canopy extraction [35]. Jiang et al. [12] found that DINO-SWIN improved F1 by 9 % over DINO-Res50, and that the NIR1–R–G band composition reached F1 and AP values of 74 % and 76 %, respectively, demonstrating that band selection matters for multispectral tree detection.

The main reported limitations across this literature concern generalization, dense canopy, small crowns, and scene heterogeneity. Dimarco et al. [8] reported that trees with crown diameters under 2 m were frequently undetected, while dense deciduous canopies produced overlapping bounding boxes. Freudenberg et al. [9] found that only 40 % of trees smaller than 10 m^2 were detected, while crowns larger than 60 m^2 were over-segmented into an average of 2.4 polygons per tree. Jiang et al. [12] found that YOLO struggled with small roadside trees and clustered vegetation, while DETR produced the lowest F1-score of 17 % due to densely overlapping bounding boxes. These findings are relevant to the present benchmark because the Reggio Calabria scene contains both isolated urban trees and semi-continuous rural canopy within the same acquisition.

2.7. Evaluation Frameworks in ITCD

Evaluation is a key methodological issue because ITCD studies do not always measure the same output [36]. Some studies evaluate treetop localization, others evaluate polygon delineation, others evaluate canopy-mask segmentation, and others evaluate counting error. Detection-oriented studies commonly use precision, recall, and F1-score, while segmentation-oriented studies report Intersection over Union or related mask-overlap measures [9, 36]. Object-level evaluation usually requires a matching rule for predicted and reference crowns; one common strategy counts predicted polygons as true positives only when their IoU with a reference polygon exceeds a defined threshold [9]. Because different studies use different sensors, spatial resolutions, landscapes, crown definitions, and matching criteria, reported performance values are difficult to compare directly [28, 36]. Table 2.3 summarises the diversity of evaluation approaches across the reviewed literature.

Method family	Output type	Common metrics	Matching criterion
Local maxima	Point / polygon	Accuracy, omission, commission	Distance or overlap tolerance
Watershed	Polygon	Accuracy, F1	Polygon overlap
OBIA + RF	Polygon class	OA, kappa, F1	Object-level matching
RF Pixel	Mask	OA, F1	Not standardized
DeepForest	Bounding box	Precision, recall, F1	IoU threshold
U-Net	Pixel mask	F1, IoU, Jaccard	Mask overlap or polygon conversion

Table 2.3: Summary of output types and evaluation approaches across ITCD method families. The lack of a shared matching protocol across studies motivates the unified Hungarian-assignment IoU evaluation framework used in this thesis.

2.8. Urban versus Rural Crown Detection

Urban and rural environments present structurally different challenges for ITCD. Urban scenes contain built surfaces, roads, shadows, isolated trees, and fragmented vegetation, while rural or semi-natural scenes often contain more continuous canopy and stronger crown overlap [7, 12]. DeepForest was shown to be reusable in an urban environment, but challenges related to trees obstructed by buildings, shade, and overlapping crowns

required careful configuration and larger datasets [7]. Pantaleo et al. [18] reported that models trained on rural landscapes generated false positives in urban settings, suggesting that high accuracy in one environment does not guarantee robust generalization across the urban–rural gradient. In dense or smooth deciduous canopies, detection models produced overlapping bounding boxes or omission errors [8], while palm-category detection dropped to 40 % success in an urban context where trees were small and affected by spectral mixing [11]. These findings indicate that ITCD performance depends not only on algorithmic capacity, but also on crown separation, background complexity, and scene structure [8, 10].

2.9. Research Gap and Thesis Positioning

The literature demonstrates that ITCD methods can achieve high performance when imaging conditions, target structure, and evaluation settings are favourable. However, these results are tied to particular datasets, sensors, tree structures, and annotation settings, limiting their direct transferability [36]. Five interconnected gaps motivate the present benchmark.

First, limited cross-method comparability: the literature uses different assessment forms, visual inspection, treetop marking, omission and commission error, F1, mAP, and IoU, that evaluate different outputs and cannot always be compared directly [36]. Second, a gap between controlled and heterogeneous settings: methods validated on large, well-separated crowns degrade when crowns are small, overlapping, shadowed, or mixed with surrounding surfaces [11, 38]. Third, limited urban–rural transferability: high accuracy in one landscape type does not guarantee robustness across the urban–rural gradient [7, 18]. Fourth, no existing benchmark evaluates representative classical, object-based, machine-learning, detection-based, and segmentation-based methods under the same VHR satellite dataset, annotation protocol, and object-level matching criterion [28, 36].

This thesis addresses these gaps by applying different methods to the same JILIN-1 KF01C RGB+NIR scene over Reggio Calabria, evaluated against the same 800 manually delineated crown polygons using Hungarian-assignment IoU matching at thresholds of 0.25 and 0.50. By focusing on method behaviour, error modes, and operational trade-offs across urban and rural strata, the thesis responds to the broader limitation that existing ITCD studies often report strong results in isolated settings but provide limited evidence on cross-method comparability and real-world deployment suitability .

3 | Dataset Description

3.1. Study Area

The study area is located in Reggio Calabria, a city at the southernmost tip of mainland Italy in the Calabria region, positioned on the Strait of Messina facing Sicily. The area is representative of a Mediterranean environment characterized by warm, dry summers and mild winters, which shapes a distinctive vegetation structure combining drought-tolerant tree species, cultivated olive and citrus groves, and ornamental urban planting. Figure 3.1

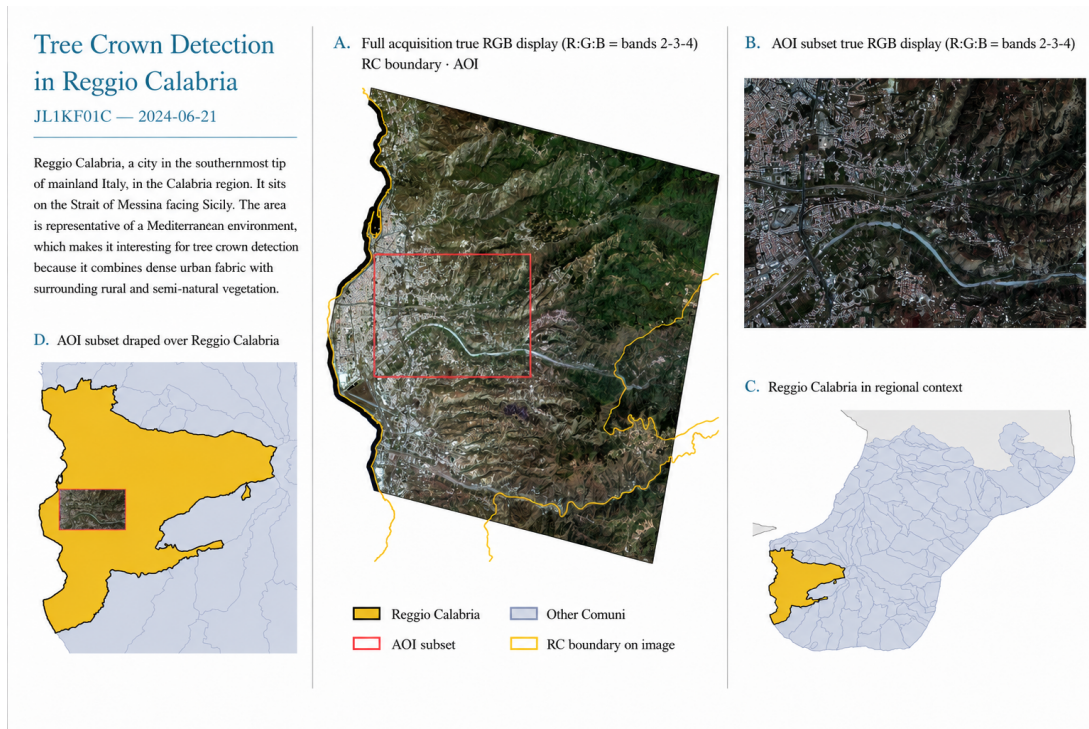


Figure 3.1: Overview of the study area. Panel A shows the full JILIN-1 KF01C acquisition footprint with the Reggio Calabria municipal boundary and the AOI subset extent (red rectangle). Panel B shows the AOI subset in true-colour composite (R:G:B = bands 2-3-4). Panel C places Reggio Calabria within its regional context in Calabria. Panel D shows the AOI subset draped over the municipal boundary.

3.2. Satellite Imagery

3.2.1. Very High Resolution Imagery

VHR imagery, typically defined as imagery with sub-meter spatial resolution, enables object-level analysis of urban and rural landscapes. Its strength is spatial detail, which makes it possible to observe canopy boundaries, crowns, roads, and small built structures in the same scene. Its main limitation is also tied to that detail: as resolution increases, intra-object variability, shadowing, and fine-scale clutter become more visible, and this can weaken simple pixel-based assumptions [5].

3.2.2. Sensor and Acquisition

The imagery used in this study was acquired by the JILIN-1 KF01C commercial sensor (scene identifier: JL1KF01C_PMSR2_20240621174656). The scene was captured on 21 June 2024 under summer illumination conditions, which correspond to peak foliar density for Mediterranean tree species and therefore represent a favorable phenological window for crown delineation. The main acquisition and sensor parameters are summarised in Table 3.1.

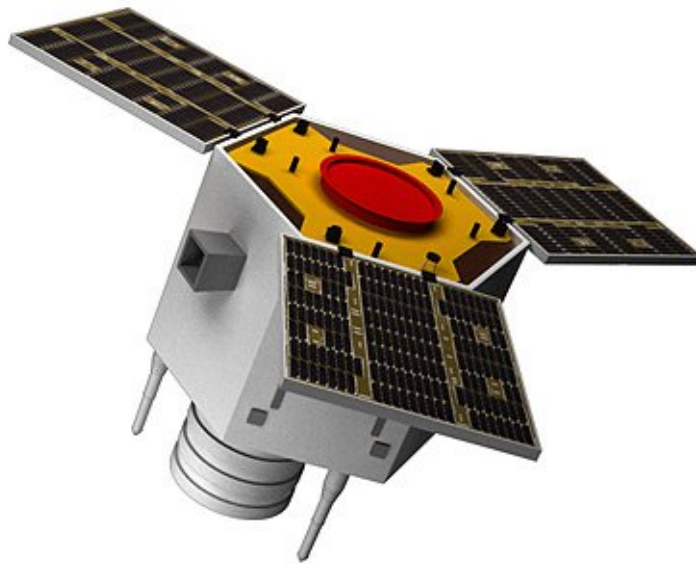


Figure 3.2: JILIN-1 Optical Satellite / KF01C sensor

JILIN-1

The JILIN-1 constellation, operated by Chang Guang Satellite Technology, comprises several platform families for optical, multispectral, video, and infrared imaging, most in

low-Earth sun-synchronous orbits. The first JILIN-1 optical satellite was launched on 7 October 2015 from the Jiuquan Satellite Launch Base [25]. Because revisit capability is a property of the constellation rather than of the KF01C platform in isolation, the present analysis treats the acquisition as a single-date observation and does not infer temporal change from it.

KF01C

KF01C is part of the Kuanfu (KF) wide-swath component of the JILIN-1 constellation. The supplied L4C product is a georeferenced four-band raster with approximately 0.5 m ground sampling distance. It includes red, green, blue, and near-infrared bands, allowing both true-colour interpretation and NDVI calculation.

The image covers approximately 21.6 km² and includes urban, peri-urban, and rural areas. This makes it suitable for evaluating tree-crown detection methods under varied background conditions while using the same sensor, acquisition, and image geometry for all workflows.

Parameter	Value
Sensor	JILIN-1 KF01C
Operator	Chang Guang Satellite Technology
Acquisition date	21 June 2024
Product level	L4C
Constellation / platform family	JILIN-1 / Kuanfu (KF) wide-swath component
Spatial resolution	≈ 0.5 m
Spectral bands	NIR (B1), Red (B2), Green (B3), Blue (B4)
Raster dimensions	11 021 cols $\times$ 7 840 rows
CRS	WGS 84 / UTM Zone 33N (EPSG:32633)
Spatial bounds (m)	E: 556 938.5–562 449.0; N: 4 214 960.0–4 218 880.0
Valid pixels	86 404 640 (100 %)
Delivery format	Integer DN (raw); float32 after scaling

Table 3.1: JILIN-1 KF01C acquisition and raster parameters for the Reggio Calabria scene.

3.3. Reference Data

3.3.1. Origin of the Reference Data

The reference dataset was generated specifically for this thesis through manual digitization of individual tree crown polygons. The primary annotations were produced by visual interpretation of the JILIN-1 KF01C scene, using the true-colour RGB composite as the main interpretation layer and NDVI as auxiliary support for vegetation confirmation.

A pre-existing set of approximately 9,000 urban crown polygons was also available from the ArcGIS REST service for Calabria.¹ This external inventory was not used as the full reference dataset because it covers only urban trees and would not support a balanced urban–rural comparison. Instead, it was used to check, confirm, and complete the primary urban annotations where it overlapped the study area. The final reference sample was balanced by design, with 400 urban crowns and 400 rural crowns.

3.3.2. Land-Use/Land-Cover Layer

The urban and rural strata were derived from the official land-use/land-cover (LU/LC) dataset published by the Calabria Region, rather than manually drawn for this thesis.² The LU/LC layer was aligned to the JILIN-1 image reference system and reclassified into two analytical categories: urban and rural. These categories were used both to stratify crown annotation and to assign each tile a dominant class.

Dataset	Source	Type / format	Scale / date	CRS
LU/LC layer	Official Calabria Region geoportal	GeoJSON vector layer	Native regional mapping scale; not acquisition-specific	Aligned to WGS 84 / UTM Zone 33N (EPSG:32633)
Manually digitized crown polygons	Created in QGIS from the JILIN-1 KF01C image; urban crowns checked and completed with the Calabria ArcGIS REST inventory	GeoJSON vector layer	Crown-level annotations; 800 polygons; digitized in 2026	WGS 84 / UTM Zone 33N (EPSG:32633)

Table 3.2: Summary of the reference datasets.

¹<https://services9.arcgis.com/Iko2iF79CuZQnhht/ArcGIS/rest/services/Calabria/FeatureServer>

²<https://geoportale.regione.calabria.it/dataset>

4 | Methodology

This chapter describes both the analytical rationale and the concrete implementation parameters behind every stage of the benchmark.

4.1. Research Design

This thesis adopts a comparative experimental design whose purpose is to benchmark five individual tree crown detection (ITCD) methods under a single, shared set of preprocessing, annotation, and evaluation conditions. Every method consumes the same imagery, is trained or tuned on the same reference crowns, and is evaluated against the same test polygons using the same object-level matching rule, so that any difference in measured performance can be attributed to the method itself rather than to inconsistencies in data preparation, label definition, or scoring [5].

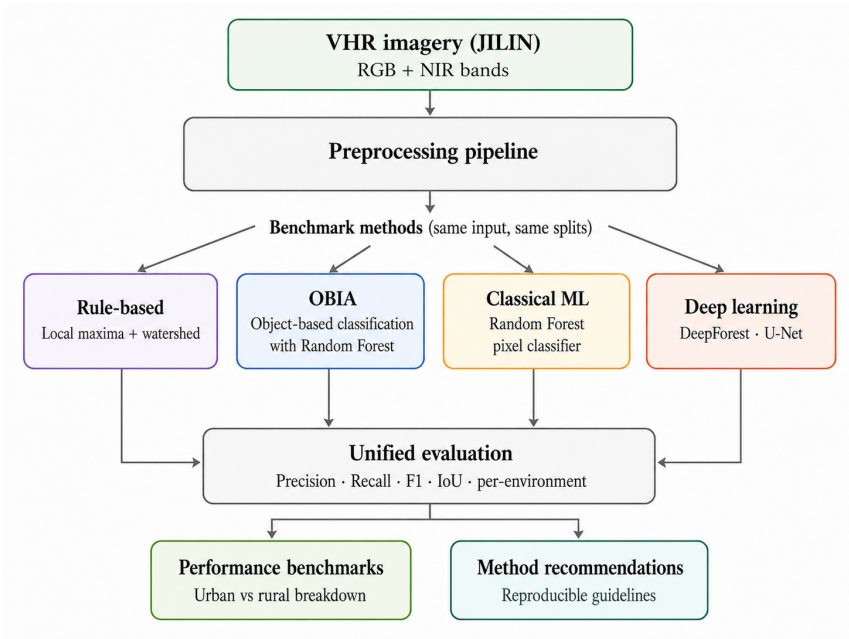


Figure 4.1: Research Design Pipeline

The five methods span the full methodological spectrum: a rule-based geometric baseline

(local maxima with watershed segmentation), an object-based classifier (OBIA with Random Forest), a pixel-based classifier (Random Forest on engineered pixel features), a pre-trained deep detection model (DeepForest), and a deep semantic segmentation model (U-Net). Together they represent four methodological families, geometric heuristics, object-based analysis, feature-engineered machine learning, and representation learning, chosen because each is prominent in the ITCD literature and offers a different set of strengths and limitations in relation to the scene complexity of the study area.

The complete implementation of all five methods, the preprocessing pipeline, the evaluation framework, and the supporting scripts are available in the public repository at https://github.com/nasarmientoo/tree_detection. The annotated reference crown polygons and all derived geospatial products are publicly available on Zenodo at <https://doi.org/10.5281/zenodo.21104150> to ensure full reproducibility and to allow independent reuse of the benchmark dataset.

4.1.1. Common Input and Output Structure

To make the comparison fair, the framework fixes a common input and output contract that every method must satisfy. On the input side, all methods draw from a shared pool of preprocessing products: a radiometrically scaled four-band image, an NDVI raster and vegetation mask, a fixed tiling grid, training crown polygons, and a set of size thresholds derived once from the training data. Methods differ only in which of these products they consume, not in how the products were generated. Table 4.1 records this mapping.

On the output side, every method must ultimately produce a set of predicted crown polygons in georeferenced vector form (GeoJSON), regardless of whether the method natively outputs markers, classified segments, pixel masks, or bounding boxes. This common output format is what allows methods with very different internal logic to be scored by a single evaluation procedure. Each method therefore ends with a method-specific conversion step that maps its native output to crown polygons, followed by a shared deduplication step and the shared evaluation.

Preprocessing output	Watershed	OBIA+RF	RF Pixel	DeepForest	U-Net
Scaled float32 image	✓	✓	✓	✓	✓
NDVI raster	✓	✓	✓	—	✓
NDVI mask	✓	✓	✓	—	—
Fixed tiles (512 px)	✓	✓	✓	✓	✓
Z-score normalized tiles	—	—	✓	—	✓
Crown polygons (train)	—	✓	✓	✓	✓
Area thresholds	✓	✓	✓	✓	✓
GLCM texture features	—	✓	✓	—	—
Object segments (Felzenszwalb)	—	✓	—	—	—
Spatial block split	✓	✓	✓	✓	✓
Deduplication rule	✓	✓	✓	✓	✓

Table 4.1: Mapping of preprocessing outputs to method-level input requirements. A checkmark indicates that the output is a direct input or dependency for that method.

4.2. Tools and Software Environment

The thesis was implemented entirely with open-source geospatial and scientific-computing software, chosen for reproducibility, transparency, and the availability of mature implementations of the methods under comparison. The choice of tools follows the two-stage nature of the work: a desktop GIS for the manual reference annotation and the inspection of geospatial layers, and a scripted Python environment for the automated preprocessing, modelling, and evaluation.

Reference crowns were digitized manually in **QGIS**, which was selected because it provides precise, interactive vector editing directly over the georeferenced imagery and raster layers, and because it reads and writes the same open formats (GeoPackage, GeoTIFF, GeoJSON) used by the rest of the pipeline. All remaining stages were implemented in **Python**, which offers a single ecosystem that spans raster handling, vector geometry, classical machine learning, and deep learning, avoiding format conversions and manual hand-offs between tools.

Because the deep learning models require GPU acceleration, DeepForest and U-Net were

trained on **Google Colab** using an NVIDIA Tesla T4 GPU, while the preprocessing and the handcrafted methods (watershed, OBIA, and pixel-based Random Forest) ran on a standard CPU. Table 4.2 summarises the tools and their role in the workflow.

Tool / library	Role in the workflow
QGIS	Manual crown digitization and visual inspection of geospatial layers
Python	Core language for preprocessing, modelling, and evaluation
rasterio	Reading/writing georeferenced rasters and tiles; coordinate transforms
GeoPandas / Shapely	Crown polygons, spatial joins, geometric IoU
NumPy / pandas	Array and tabular data processing
scikit-image	Segmentation, watershed, local maxima, region properties, GLCM texture
SciPy	Smoothing, morphology, Hungarian assignment
scikit-learn	Random Forest, cross-validation, metrics
imbalanced-learn	SMOTE oversampling for class imbalance
PyTorch	Deep learning training framework
segmentation-models-pytorch	U-Net architecture with pretrained encoder
DeepForest	Pretrained RetinaNet crown detector
Matplotlib	Figures and visual validation
Google Colab (Tesla T4 GPU)	GPU environment for deep learning training

Table 4.2: Software tools and libraries used in the thesis and their role in the workflow.

4.3. Reference Annotation Methodology

4.3.1. Annotation Protocol

Each reference polygon represents the visible outer boundary of one individual tree crown as observed from above. Polygons were drawn around the crown envelope that could be identified in the image, without attempting to delineate branches, internal canopy gaps, or other sub-crown structures as separate objects. NDVI was used only as an interpretative aid and did not automatically generate, modify, or select annotation polygons. A candidate object was annotated only when its crown structure could be visually distinguished

with sufficient confidence from the surrounding background and from neighbouring vegetation.

Manual digitisation and attribute compilation were performed in QGIS. Each polygon includes three attributes to preserve provenance and analytical role. The `image_id` field records the source image name, in this case JL1KF01C_PMSR2_20240621174656_200275828_101_0034_001_L4C_1067507. The `crown_status` field identifies whether the polygon belongs to the training or test subset, and the `date` field records the acquisition date of the source image (21 June 2024).

The sampling design intentionally balanced the two landscape strata. So that the reference data represent both continuous or semi-continuous vegetation and the fragmented, shadow-affected tree cover characteristic of the built-up environment. This balance does not imply that the two strata have the same natural tree density; rather, it ensures that neither landscape context dominates the training and evaluation sample.

4.3.2. Annotation Quality and Uncertainty

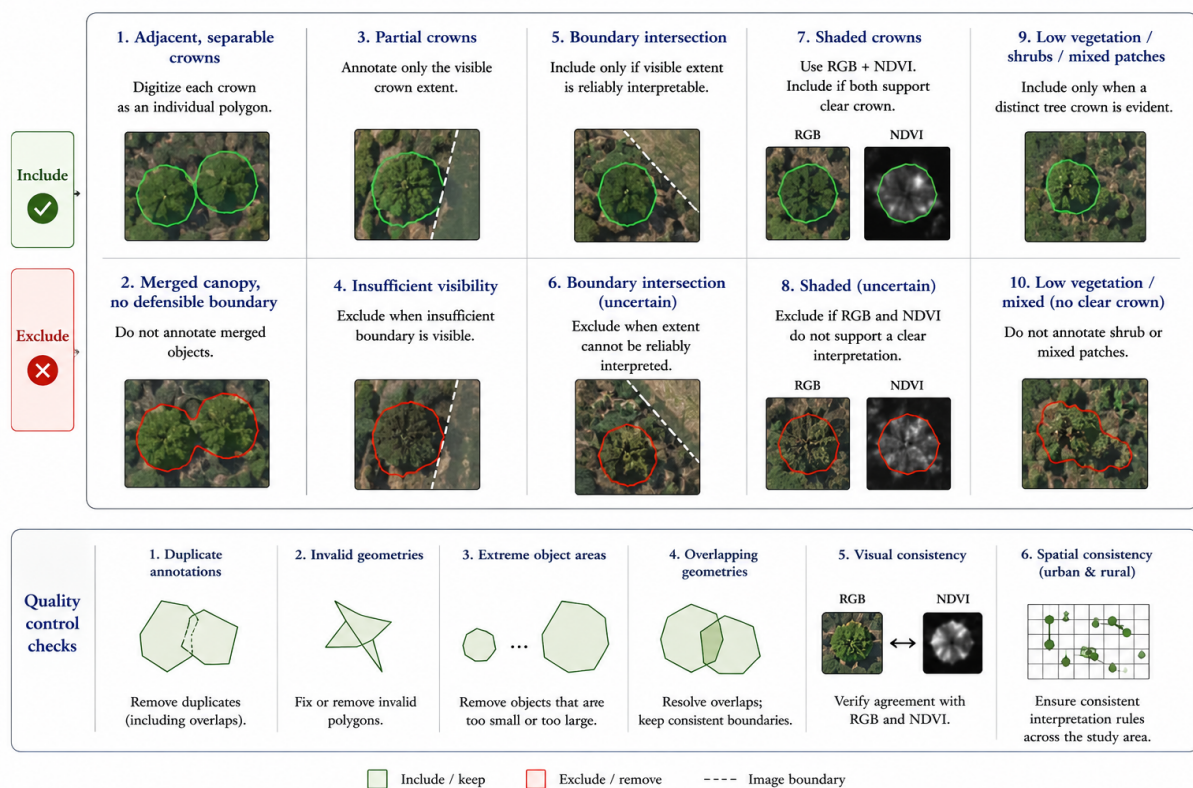


Figure 4.2: Annotation protocol/quality-control

Ambiguous cases were handled conservatively to maintain a reference set whose object definition is consistent across the study area. These rules reduce the number of uncertain examples, while recognising that the resulting annotations represent visible image crowns rather than a complete field inventory of all woody vegetation.

After digitisation, the polygons were passed for a quality-control review, used to identify topological errors and to ensure that crown boundaries followed the same interpretation rules in urban and rural samples.

4.4. Preprocessing Pipeline

Preprocessing establishes the common analytical foundation consumed by all five methods. It is designed to support fair comparison rather than to optimise any single model family, because different paradigms are sensitive to different input properties: spectral methods depend on radiometric consistency, object-based methods depend on segmentation stability, and deep learning methods depend on spatial context and annotation quality. All preprocessing parameters are derived from training data only and held fixed across every experiment.

4.4.1. Radiometric Foundation and Masking

The raw per-band value distributions must first be examined to understand radiometric properties and identify any systematic biases across landscape strata. This reconnaissance informs how each method will encounter the data: spectral classifiers and deep learning models are sensitive to the absolute scale of DN values, and differences in urban versus rural radiometry can create systematic variation that methods must accommodate. Therefore, the raw integer digital numbers are screened for outliers and invalid values — pixels whose DN values fall outside the physically admissible range $[0, 10,000]$ are flagged and masked, along with any nodata or background-fill values. The raw integers are then divided by 10,000 and cast to `float32`, producing a uniform reflectance-like scale shared by all downstream methods. This scaling is arbitrary in absolute terms but necessary to bring all methods into a common numerical regime.

Because the scene is constructed from satellite imagery that may be mosaicked from multiple passes or processed in segments, radiometric seams can occur at tile boundaries. These seams are not artifacts of radiometric error; they reflect genuine differences in acquisition or processing across the mosaic, and they should be documented but not necessarily corrected, since doing so might introduce artificial discontinuities. The scaled

image should therefore be inspected for such seams so that any discontinuities are known and can be monitored during evaluation.

4.4.2. Vegetation Separation and NDVI Thresholding

All methods that operate in the spectral domain or on segmented canopy require a vegetation mask to restrict processing to vegetated areas and exclude roads, rooftops, shadows, and bare soil. This masking is justified not only for efficiency but for interpretability: methods applied to non-vegetated areas will generate false detections or meaningless output. The mask is derived from the Normalized Difference Vegetation Index (NDVI), defined as

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}, \quad (4.1)$$

which measures the spectral contrast between red absorption (photosynthesis) and near-infrared reflectance (canopy structure) of vegetation. An NDVI threshold of 0.35 is selected as the canopy mask. This threshold value is empirical: it must be high enough to exclude non-vegetated surfaces (which have NDVI near or below zero) but not so high that it rejects valid crown areas or the heterogeneous understory vegetation that surrounds individual trees. The threshold is applied uniformly across all tiles and strata to ensure consistent masking. All downstream operations confine their attention to pixels inside this mask, and post-processing uses the mean NDVI of retained detections as a tiebreaker when duplicate detections must be removed.

4.4.3. Tiling and Spatial Partitioning

The scene must be divided into tiles for two reasons. First, many methods (watershed, OBIA, pixel RF) are sensitive to the spatial extent and context they operate over; tiling ensures consistent context across methods. Second, and more importantly, the scene is too large to process as a monolith for deep learning: GPU memory and dataset loading require dividing the data into manageable patches. A fixed tile size of 512×512 pixels is chosen as a compromise: large enough to capture spatial context (neighboring tree crowns, background variation) but small enough to be processed by standard deep learning hardware. The tiles overlap by a stride of 128 pixels (75% overlap), which ensures that crown polygons near tile boundaries are captured by at least one tile at full spatial context. This overlap is the source of duplicate detections across tile boundaries, which the shared deduplication step must later remove.

Tiles are not all retained indiscriminately. A tile filtering rule removes tiles that are predominantly nodata, empty of vegetation, or anomalously saturated, since these tiles

cannot contribute meaningful signal to any method. Tiles are retained if the nodata fraction is below 10%, the tree coverage (from the NDVI mask) is above 0.5%, and the NDVI range exceeds 0.10; these thresholds are conservative and permit nearly all tiles with any meaningful tree presence to pass.

Beyond the spatial tiling, the scene must be partitioned into training and test subsets. To prevent spatial leakage — where pixel neighborhoods in the training set inform model decisions on nearby test pixels — the partition is done at the block level, not the pixel level. The tiled scene is first divided into non-overlapping 1,024 pixel spatial blocks, and entire blocks are assigned to one subset or the other. This block-level split ensures that no pixel neighborhood crosses the train–test boundary, and it forces methods to generalize to truly held-out geographic areas.

4.4.4. Z-Score Normalization

Some methods (pixel-based Random Forest and U-Net) are sensitive to the absolute scale and distribution of input features. If one band has values in the range $[0, 0.15]$ and another band has values in the range $[0, 0.08]$, gradient-based optimizers and distance-based classifiers can be biased toward the larger-scale features. To level the playing field, z-score normalization (Eq. 4.2) transforms each band to have zero mean and unit variance:

$$\hat{x}_b = \frac{x_b - \mu_b}{\sigma_b}, \quad (4.2)$$

where μ_b and σ_b are the per-band mean and standard deviation. Crucially, these statistics are computed exclusively from the training tiles, and the same statistics are applied without refitting to the test tiles. This one-way application (train $\rightarrow$ test) ensures that no test-set information influences the normalization and prevents information leakage.

4.4.5. Feature Separability Analysis

Before designing the classifiers and selecting which features to use, the separability between tree and non-tree classes must be assessed across different feature spaces and strata. Separability is measured using the Jeffries–Matusita (JM) distance, a normalized measure on the scale $[0, 2]$ where values below 1.0 indicate poor class separation and values above 1.0 indicate useful discrimination. The JM distance is computed from the Bhattacharyya distance as

$$\text{JM} = 2 \left(1 - e^{-B} \right), \quad (4.3)$$

where B is the Bhattacharyya distance. This analysis is performed on the training data at the pixel and object levels across spectral indices, texture descriptors, and geometric measures. The results inform which features are worth including in the OBIA and Random Forest pipelines and which are likely to be uninformative.

4.4.6. Size Thresholds

Because objects of all sizes will be detected (ranging from single-pixel noise to large merged vegetation masses), size filtering is necessary to isolate crown-sized objects from clutter. Rather than selecting thresholds arbitrarily, they are derived empirically from the training data, using the per-stratum distribution of annotated crown areas. A minimum-area threshold is selected to retain 90% of training crowns (removing only the smallest 10%, which may be noise or ambiguous objects); a maximum-area threshold is selected to exclude extremely large blobs that cannot reasonably be single crowns. These per-stratum thresholds are applied identically to all five methods during post-processing so that size filtering is not a source of variation in the comparison.

4.4.7. Class-Balance Assessment

The final step in preprocessing design is to quantify the class imbalance that supervised methods will encounter. The tree-to-non-tree pixel ratio is computed at the level of the annotated training crowns (tree pixels) versus all other canopy pixels (non-tree pixels). This ratio informs the design of loss functions and sampling strategies for the classifier-based and deep learning methods: if the imbalance is severe (e.g., 1:50), then standard accuracy-based losses will produce poor recall because predicting “non-tree” for everything yields high accuracy. Therefore, methods employ class-weighting, oversampling, or specialized loss functions (such as focal loss) to compensate.

4.5. Evaluation Framework

A single evaluation framework is applied to every method so that results are directly comparable. Because the methods produce different native outputs, evaluation operates on the common crown-polygon representation rather than on any method-specific format. All methods were evaluated against the same fixed test set of 302 crown polygons, obtained from the 308 test-set annotations after removing those below the minimum per-stratum area threshold. The size filter applied uniformly is $[75, 786]$ px for urban tiles and $[86, 786]$ px for rural tiles.

4.5.1. Object-Level Matching

Predicted crown polygons are matched to reference crown polygons using a geometric Intersection over Union (IoU) criterion combined with optimal one-to-one assignment via the Hungarian algorithm. For a predicted polygon P and a reference polygon R , the geometric IoU is the ratio of their overlapping area to their combined area,

$$\text{IoU}(P, R) = \frac{\text{area}(P \cap R)}{\text{area}(P \cup R)}, \quad (4.4)$$

which ranges from 0 for disjoint polygons to 1 for a perfect overlap. A predicted polygon is counted as a true positive only when it is assigned to a reference crown whose IoU exceeds a defined threshold; unmatched predictions are false positives, and unmatched reference crowns are false negatives. Matching is performed at two complementary IoU thresholds. A threshold of 0.25 acts as a permissive criterion: it captures approximate crown overlap and tolerates annotation boundary uncertainty, counting coarse but spatially coherent detections as true positives and thereby providing an upper bound on the method's spatial reach. A threshold of 0.50 is a stricter standard that requires the predicted polygon to cover at least half of the combined predicted and reference area, reflecting more accurate crown delineation and a level of boundary fidelity comparable to established object-detection benchmarks. Reporting both thresholds jointly allows detection accuracy and boundary precision to be assessed as partly independent properties.

4.5.2. Deduplication

Because tiles overlap by 128 px out of 512 px, the same crown can appear in up to four adjacent tiles. A shared deduplication rule is therefore applied to every method before scoring: any pair of predictions whose IoU exceeds 0.5 is treated as a duplicate, and the prediction with the lower mean NDVI is removed. This rule is identical across methods so that deduplication does not advantage any particular pipeline.

4.5.3. Detection and Segmentation Metrics

From the matched sets, detection quality is quantified by precision, recall, and the F1-score. Let TP , FP , and FN denote the numbers of true positives (matched predictions), false positives (unmatched predictions), and false negatives (unmatched reference crowns), respectively. Precision measures how reliable the predicted crowns are,

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (4.5)$$

while recall measures how much of the reference crown set is successfully recovered,

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (4.6)$$

The F1-score, equivalent to the Dice coefficient in this binary setting, is the harmonic mean of precision and recall and summarises the balance between the two,

$$\text{F1} = \text{Dice} = \frac{2 TP}{2 TP + FP + FN}. \quad (4.7)$$

Segmentation quality is reported separately as the mean IoU of the matched true-positive pairs, where each pair contributes its geometric overlap from Equation (4.4). Equivalently, when the overlap is measured over foreground pixels rather than polygons, the Intersection over Union reduces to

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \quad (4.8)$$

Reporting both the detection metrics and the matched mean IoU prevents a method with good localisation but poor boundaries (or vice versa) from being judged on a single dimension. All metrics are reported separately for the combined test set and for the urban and rural strata.

4.6. Local Maxima and Watershed Segmentation

The watershed method treats crown detection as a geometric problem: candidate tree-tops are identified as local maxima on a smoothed vegetation surface and used as markers from which crown regions are delineated by the watershed transform. Because the method operates entirely on NDVI and requires no training labels, it provides a transparent baseline whose failure modes indicate where spectral contrast and canopy geometry alone are insufficient to separate individual crowns.

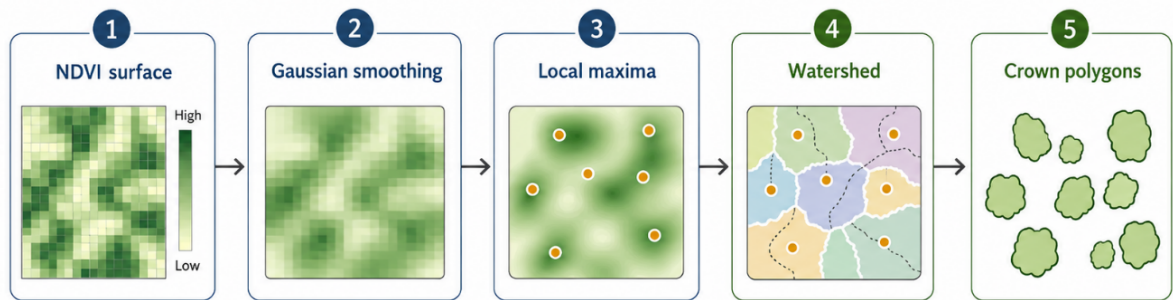


Figure 4.3: Overview of the local-maxima and watershed segmentation pipeline.

4.6.1. NDVI Surface Preparation

The input to the watershed pipeline is an NDVI surface computed from the scaled `float32` near-infrared and red bands. NDVI values are clipped to the valid range $[-1, 1]$ to remove computational artifacts, and a canopy mask is defined by the threshold $\text{NDVI} > 0.35$. Before smoothing, pixels outside this mask are temporarily filled with the threshold value 0.35 to prevent missing values from producing artificial gradients along canopy boundaries. After smoothing, the original mask is restored so that subsequent operations remain restricted to canopy pixels.

4.6.2. Smoothing and Scale Selection

Gaussian smoothing is applied to suppress fine-scale variation caused by shadows, gaps between leaves, branches, and changes in crown-interior density. Without smoothing, these features can create several local maxima within a single crown and consequently produce excessive watershed markers. The smoothing scale must therefore be large enough to remove intra-crown noise while remaining small enough to preserve distinct peaks associated with neighboring trees.

The smoothing scale is selected according to canopy structure. Urban tiles generally contain isolated trees with sharp boundaries and distinct NDVI peaks, so a scale of $\sigma = 2.0$ px is used. Rural tiles contain more continuous vegetation and stronger internal texture, requiring a larger scale of $\sigma = 3.0$ px. These values were selected from a sensitivity analysis over $\sigma \in \{1.5, 2.0, 3.0\}$ by examining both peak retention and the number of detected local maxima.

4.6.3. Local Maxima Detection

Local maxima are detected on the smoothed NDVI surface after the canopy mask has been restored. A canopy pixel is considered a candidate maximum when its NDVI value is greater than those of the surrounding pixels within the search neighbourhood. Candidate peaks are processed in descending order of smoothed NDVI, and a minimum spacing of 6 px is imposed so that weaker peaks lying too close to a stronger peak are suppressed. This non-maximum-suppression step reduces the likelihood that small variations within one crown generate multiple markers, while retaining spatially separated peaks that may represent different treetops.

The number of detected maxima is used during parameter selection to assess whether smoothing sufficiently suppresses canopy texture. However, the local-maxima analysis

showed that a global spacing rule alone is unreliable in large continuous vegetation regions: rural tiles produced substantially more maxima than urban tiles, even though their expected crown density was lower. Many of these maxima represented branches, crown edges, or local texture rather than separate trees. The local maxima are therefore combined with connected-component analysis before watershed markers are constructed.

4.6.4. Canopy Restriction and Blob Segmentation

The binary canopy mask is divided into connected components using 8-connectivity. Components with an area of at most 786 px are classified as small canopy blobs, whereas components larger than 786 px are flagged as large merged blobs. Local maxima and watershed segmentation are retained only within the small blobs, where an isolated canopy component can reasonably be interpreted as an individual crown candidate.

Large blobs are excluded because local spectral contrast cannot reliably recover crown boundaries within continuous vegetation. Although many local maxima may occur inside such a blob, they frequently correspond to within-crown texture rather than distinct treetops. Using all of them as watershed markers would cause severe over-segmentation, while retaining too few would merge adjacent crowns. Excluding these components makes this limitation explicit instead of producing unreliable crown polygons in regions where NDVI geometry is insufficient.

4.6.5. Marker Selection and Watershed Transform

For each small canopy blob, the local maxima detected within its boundary are examined and the maximum with the highest smoothed NDVI is retained as the blob's watershed marker. If the spacing-based detector returns no maximum because of boundary or plateau effects, the pixel with the highest smoothed NDVI in the blob is used as a fallback. This procedure preserves the local-maxima interpretation of the marker while enforcing at most one seed per small connected canopy component and preventing multiple texture peaks from fragmenting an isolated crown.

The watershed transform is applied to the inverted smoothed surface, $1 - \text{NDVI}$, so that high-NDVI treetop peaks become low-valued basins from which the algorithm can flood outward. Pixels outside the selected canopy blobs are masked and cannot be assigned to a watershed region. Using 8-connectivity, the transform expands each marker through its associated blob until it reaches the canopy boundary or a boundary shared with another marker, producing one labelled crown candidate for each retained seed.

4.6.6. Post-Processing and Output Format

The resulting segments are filtered using stratum-specific area limits of $[75, 786]$ px for urban tiles and $[86, 786]$ px for rural tiles. Accepted segments are vectorized into georeferenced polygons containing the shared attributes used by all detection methods, including tile identifier, land-cover class, detection method, area in pixels and square metres, centroid and bounding-box coordinates, and mean NDVI. One GeoJSON file is written for each tile. Because adjacent tiles overlap, detections with an intersection over union greater than 0.5 are deduplicated, retaining the polygon with the higher mean NDVI.

Parameter	Value
Canopy threshold	NDVI > 0.35
Fill value before smoothing	0.35
Smoothing σ (urban/rural)	2.0/3.0 px
Local-maximum minimum spacing	6 px
Local-maximum ranking	Descending smoothed NDVI
Blob-size cutoff	786 px
Marker-selection strategy	Strongest local maximum per small blob
Fallback marker	Maximum-NDVI pixel within the blob
Watershed surface	$1 - \text{NDVI}$
Watershed connectivity	8-connectivity
Size filter (urban/rural)	$[75, 786]/[86, 786]$ px
Area conversion	$\text{area}_{\text{px}} \times 0.25 \text{ m}^2$
Deduplication criterion	$\text{IoU} > 0.5$; retain higher mean NDVI
Output format	GeoJSON polygons with canonical attributes

Table 4.3: Implementation parameters for local-maxima detection and watershed segmentation.

4.7. OBIA with Random Forest

The object-based method was designed to address one of the main weaknesses of pixel-level classification: a single pixel has no spatial context. At the pixel scale, a classifier only sees spectral values, so it cannot determine whether a pixel belongs to a complete tree crown, a small patch of vegetation, or a spectrally similar background surface. Object-Based Image

Analysis (OBIA) approaches the problem differently by first grouping neighbouring pixels into spatially coherent image objects and then classifying those objects as tree or non-tree [5]. In this way, the detection decision is no longer based only on reflectance values, but also on the size, shape, internal consistency, and vegetation structure of each segment. This makes OBIA a useful intermediate method in the benchmark: it is more spatially informed than pixel-based classification, but still more interpretable than deep learning because every prediction can be traced back to explicit features and a known segment boundary.

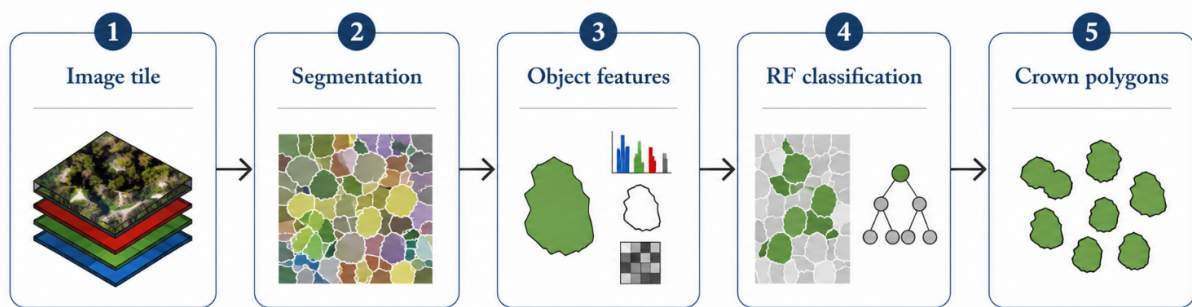


Figure 4.4: Object-Based Image Analysis with Random Forest workflow.

4.7.1. Tile-Level Input Preparation

The object-based pipeline was applied independently to each 512×512 px tile generated during preprocessing. Each tile was loaded as a scaled `float32` four-band raster containing the Blue, Green, Red, and NIR bands. The NDVI canopy mask, defined as $\text{NDVI} > 0.35$, was used to restrict the analysis to vegetated pixels. This restriction was applied to prevent the segmentation process from generating objects that combine canopy with roads, rooftops, shadows, bare soil, or other non-vegetated surfaces.

Before segmentation, all non-canopy pixels were set to zero. The segmentation input was then built from the NIR, Red, and Green bands. These bands were min-max normalized to the $[0, 1]$ range within the unmasked canopy area. This normalization was applied so that tile-level differences in brightness or radiometric range did not dominate the segmentation process.

4.7.2. Canopy Segmentation

Canopy objects were generated using the Felzenszwalb graph-based segmentation algorithm. The algorithm was applied to the normalized NIR, Red, and Green bands. These bands were selected because they capture the main spectral contrasts between vegetation,

shaded canopy, and surrounding surfaces. The Blue band was not used as a segmentation input because it contributed less to the spectral contrast required for crown boundary delineation.

The Felzenszwalb algorithm groups neighbouring pixels according to local spectral similarity. This makes it suitable for object-based crown detection because the resulting segments can follow image contrasts rather than fixed grid structures. The segmentation was parameterized using a scale value, which controls the amount of spectral variation allowed within each object. Three scale values were tested during parameter selection: 10, 15, and 50. The segmentation was implemented with `min_size`= 50px and $\sigma = 0.5$.

After segmentation, each object was filtered using the NDVI canopy mask. A segment was retained only if at least 50% of its pixels fell inside the canopy mask. This canopy-fraction criterion was used to remove mixed objects located along canopy boundaries before feature extraction.

4.7.3. Object Feature Extraction

Each retained segment was described using eleven object-level features. The spectral features consisted of the mean reflectance values of the NIR, Red, Green, and Blue bands, together with mean NDVI. These variables were included to describe the spectral behaviour of each object and its vegetation signal.

An additional spectral-spatial feature was computed using an NDVI top-hat response. This feature was calculated by subtracting a Gaussian-smoothed NDVI surface from the original NDVI image:

$$\text{NDVI}_{\text{th}} = \text{NDVI} - (G_{\sigma} * \text{NDVI}), \quad \sigma = 5, \quad (4.9)$$

where $G_{\sigma} * \text{NDVI}$ denotes convolution with a Gaussian kernel of scale $\sigma = 5$ px. This feature was included to represent local NDVI prominence within the canopy.

The geometric features were pixel area, perimeter, circularity, and elongation. Circularity and elongation were calculated as:

$$\text{circularity} = \frac{4\pi \cdot \text{area}}{\text{perimeter}^2}, \quad \text{elongation} = \frac{\ell_{\text{major}}}{\ell_{\text{minor}}}, \quad (4.10)$$

where ℓ_{major} and ℓ_{minor} are the major and minor axis lengths of the fitted ellipse. These descriptors were included to represent object size and shape. The final feature was canopy

fraction, defined as the proportion of pixels in each segment that belonged to the NDVI canopy mask.

4.7.4. Random Forest Classification

Training labels were assigned at the object level by comparing each segment with the manually digitized training crown polygons. A segment was labelled as tree (1) when its intersection-over-union with at least one training crown polygon was greater than or equal to 0.25. Segments that did not meet this condition were labelled as non-tree (0). Training crown polygons smaller than 75 px were excluded before label assignment.

Because tree objects were less frequent than non-tree vegetation objects, SMOTE over-sampling was applied to the minority tree class before model training. Oversampling targeted a 1:10 tree-to-non-tree ratio, using $k = \min(5, n_{\text{tree}} - 1)$ neighbours and random seed 42.

Separate Random Forest classifiers were trained for the urban and rural strata. This stratified training design was used because canopy structure, surrounding background, and object characteristics differ between urban and rural tiles.

Each Random Forest classifier was trained using 200 trees, a maximum depth of 15, a minimum of five samples per leaf, class weights of 0:1, 1:10, and random seed 42. Model checking was performed using 5-fold stratified cross-validation. During inference, each classified object received a tree probability score, which was converted into a binary tree/non-tree label using a probability threshold of 0.50 for both strata.

4.7.5. Post-Processing and Output Format

The resulting labelled basins are filtered by the per-stratum size thresholds ([75, 786] px urban, [86, 786] px rural). These thresholds remove noise (small fragments below the minimum) and large merged regions that slipped through the blob filter (though they should be rare). Each surviving segment is converted to a georeferenced polygon carrying a canonical attribute set shared by all methods: tile identifier, land cover class, detection method, area in pixels and square metres (at 0.5 m resolution, one pixel = 0.25 m²), centroid and bounding-box coordinates, and mean NDVI. One GeoJSON file is written per tile.

After processing all test tiles, predictions from overlapping tiles are merged. Because tiles overlap by 75% stride, the same crown may be detected in several adjacent tiles. The shared deduplication rule identifies overlapping predictions (IoU > 0.5) and retains

only the prediction with higher mean NDVI, under the assumption that a more vegetated detection is more likely to be well-centered on the actual crown. The surviving polygons then enter the shared evaluation framework.

Parameter	Value
Segmentation method	Felzenszwalb multiscale segmentation
Selected segmentation scale	15
Segmentation parameters	$\sigma = 0.5$, minimum segment size = 50 px
Segmentation input	NIR, Red, and Green bands; canopy pixels min-max normalised
Canopy-object retention	Canopy fraction ≥ 0.5
Min training-crown area	75 px
Object-label criterion	Crown annotation IoU ≥ 0.25
Feature count	11
Non-tree to tree imbalance	Urban 1:61.0 / rural 1:181.8
SMOTE target ratio	1:10 ($k = \min(5, n_c - 1)$, seed 42)
SMOTE achieved ratio	Urban 1:6.1 / rural 1:18.2
RF trees / depth / leaf	200 / 15 / 5
RF class weight / seed	{0:1, 1:10} / 42
Cross-validation	5-fold stratified
Probability threshold	0.50 for urban and rural
Size filter urban / rural	[75, 786] / [86, 786] px
Deduplication criterion	IoU > 0.5, retain higher mean NDVI
Output format	GeoJSON polygons with object-level attributes

Table 4.4: Object-based image analysis (OBIA) with Random Forest implementation parameters.

4.8. Pixel-Based Random Forest

The pixel-based classifier assigns a tree or non-tree label to every canopy pixel from a compact feature vector and then reconstructs crown objects from the classified raster. Unlike OBIA, it imposes no prior segmentation, the output boundaries are determined entirely by the pixel-level decision surface. This design choice isolates the contribution of spectral and texture information at the individual pixel scale, directly testing how far

hand-engineered pixel features can resolve individual crowns without any spatial grouping prior.

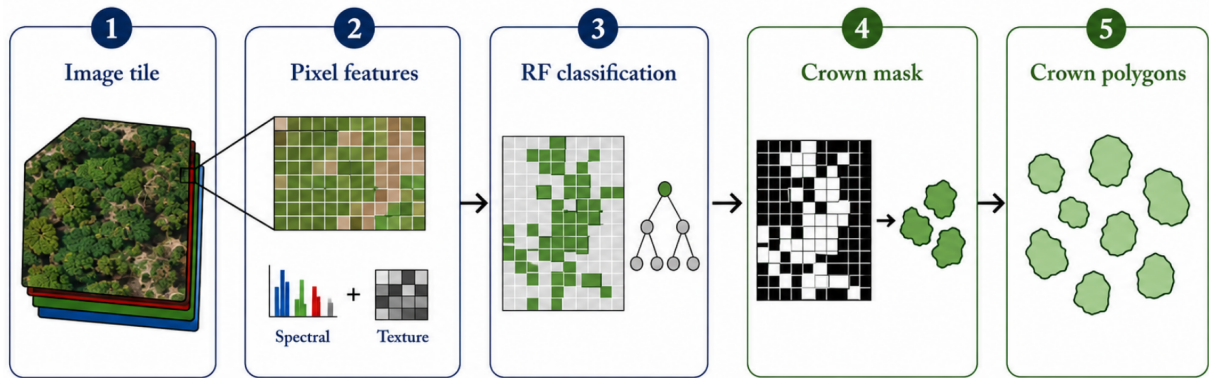


Figure 4.5: Random Forest workflow.

4.8.1. Label Rasterization and Training Target

Training supervision for the pixel-based classifier was derived from the manually digitized crown polygons. Each training crown polygon was rasterized onto the corresponding tile grid so that the labels matched the spatial resolution and alignment of the input raster. Canopy pixels whose centroids fell inside an annotated crown polygon were assigned the crown label (1), while canopy pixels outside the annotated polygons were assigned the non-crown label (0). Pixels located outside the NDVI canopy mask were assigned the ignore value (255) and were excluded from model training.

The use of the NDVI canopy mask restricted the training target to vegetated pixels only. This was done to focus the classifier on separating crown pixels from other canopy or vegetation pixels, rather than from clearly non-vegetated surfaces. Training polygons smaller than 75 px were excluded before rasterization in order to keep the training labels consistent with the minimum crown-size threshold used in the methodological framework.

4.8.2. Pixel Feature Construction

Each labelled canopy pixel was represented using a ten-dimensional feature vector. The first group of features consisted of the four spectral bands: NIR, Red, Green, and Blue. These features provided the direct radiometric information for each pixel. NDVI was included as an additional spectral feature because it combines the NIR and Red bands into a vegetation-sensitive index.

A local spectral-prominence feature was also computed using the NDVI top-hat response

with $\sigma = 5$ px, following Eq. 4.9. This feature was included to describe how much a pixel differs from the surrounding NDVI surface at a local scale.

The remaining four features were texture proxies computed from the NIR band using a 7 px sliding window. These consisted of a local-variance contrast proxy, a homogeneity proxy defined as $1/(1+\text{std})$, a range-based entropy proxy, and the local mean. These filter-based texture proxies were used instead of full GLCM descriptors to keep feature computation feasible at pixel scale. The 7 px window was selected to provide local neighbourhood information while keeping the computation manageable.

All features were computed from z-score normalized tiles. This normalization was applied so that features with different numeric ranges could be used together in the Random Forest model without allowing raw scale differences to dominate the split structure.

4.8.3. Sampling Strategy and Class Imbalance

The pixel-level training dataset was constructed separately for each land-cover stratum. All crown pixels were retained in the training sample. Non-crown canopy pixels were randomly sampled at a 1 : 10 crown-to-non-crown ratio using random seed 42. This sampling step reduced the number of majority-class pixels while preserving all available crown pixels.

Pixels with non-finite feature values were removed before model training. This filtering step was applied to ensure that the training matrix contained only valid numeric inputs.

After random sampling, SMOTE oversampling was applied to the minority crown class. The oversampling target was a 1:5 crown-to-non-crown ratio, using $k = \min(5, n_{\text{crown}} - 1)$ neighbours and random seed 42. SMOTE was applied after non-crown subsampling so that synthetic crown examples were generated within the sampled training set.

Separate feature matrices were produced for the urban and rural strata. This allowed the pixel-based classifier to be trained independently for each context using the same sampling and oversampling procedure.

4.8.4. Classifier Configuration

A Random Forest classifier was trained separately for the urban and rural strata. Each classifier used 200 trees, a maximum depth of 20, a minimum of ten samples per leaf, and class weights of 0:1, 1:5. Random seed 42 was fixed for reproducibility.

The classifier was trained to predict the binary pixel label, with crown pixels encoded as

1 and non-crown canopy pixels encoded as 0. Model checking was performed using 5-fold stratified cross-validation.

4.8.5. Probability Map

During inference, each trained Random Forest classifier produced a crown-probability value for every canopy pixel in each tile. These probabilities were assembled into a tile-level probability map restricted to the NDVI canopy mask. Pixels outside the canopy mask were excluded from prediction.

A fixed probability threshold of 0.85 was applied to convert the crown-probability map into a binary crown mask. The same threshold was used for both urban and rural strata. After thresholding, a morphological closing operation was applied to the binary mask. Urban tiles used a 3×3 closing kernel, while rural tiles used a 7×7 closing kernel. This post-processing step was used to connect nearby positive pixels and produce more spatially coherent crown-mask regions before object reconstruction.

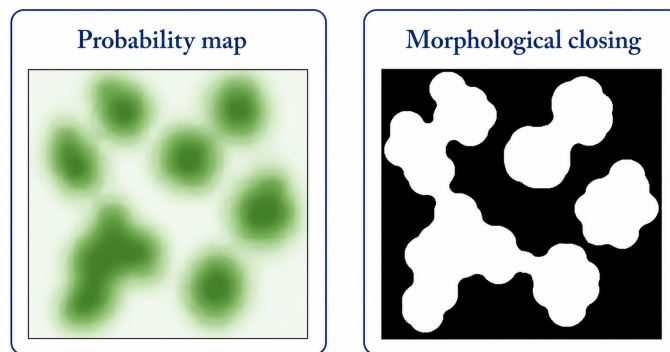


Figure 4.6: Probability Map

4.8.6. Output format

Connected components are extracted from the morphologically closed binary mask, filtered by the per-stratum size thresholds ($[75, 786]$ px for urban and $[86, 786]$ px for rural), and converted to georeferenced polygons carrying the canonical attribute set shared by all methods: tile identifier, land cover class, detection method, area in pixels and square meters, centroid and bounding-box coordinates, and mean NDVI. One GeoJSON file is written per tile. After all test tiles have been processed, predictions from overlapping tiles are merged and deduplicated using the shared rule ($\text{IoU} > 0.5$, retain higher mean NDVI), and the reconstructed polygons enter the shared evaluation framework.

Parameter	Value
Min training-crown area	75 px
Label encoding	1 crown / 0 non-crown / 255 ignore
Feature count	10
Features	4 band means, NDVI, NDVI top-hat ($\sigma = 5$), 4 NIR texture proxies
Texture proxy window	7 px
Non-crown sampling ratio	1:10 (seed 42)
SMOTE target ratio	1:5 ($k = \min(5, n_c - 1)$, seed 42)
RF trees / depth / leaf	200 / 20 / 10
RF class weight / seed	{0:1, 1:5} / 42
Cross-validation	5-fold stratified
Urban post-processing	no smoothing / threshold 0.85 / 3×3 closing
Rural post-processing	no smoothing / threshold 0.85 / 7×7 closing
Size filter urban / rural	[75, 786] / [86, 786] px
Deduplication criterion	IoU > 0.5, retain higher mean NDVI
Output format	GeoJSON polygons with canonical attributes

Table 4.5: Pixel-based Random Forest implementation parameters.

4.9. DeepForest

DeepForest is a pretrained object detection model based on a RetinaNet architecture with a ResNet-50 backbone, originally trained on large-scale airborne RGB tree-crown imagery [32]. Rather than engineering features by hand, the method inherits a learned hierarchical representation of crown appearance from its pretraining domain and adapts it to the study scene through fine-tuning. Including this method in the benchmark tests a specific hypothesis: whether a detection prior learned from a different sensor, altitude, and vegetation type can transfer to a Mediterranean VHR satellite scene, and whether bounding-box supervision is sufficient to localise individual crowns without requiring polygon-level annotations.

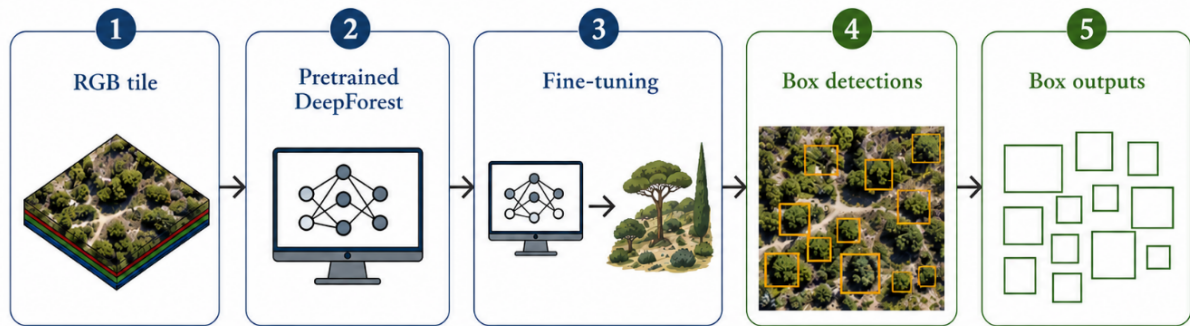


Figure 4.7: Deep Forest Pipeline Overview

4.9.1. Input Preparation and Annotation Conversion

The pretrained model expects three-band 8-bit RGB input, because it was trained on RGB aerial photography. The four-band tiles must therefore be adapted to this format before fine-tuning or inference. The NIR band is excluded and only the Red, Green, and Blue channels are used. Each band is contrast-stretched per tile using the 2nd–98th percentile range and scaled to $[0, 255]$, then saved as a 512×512 PNG image. Per-tile contrast stretching is applied rather than a global stretch because tile-level radiometric variation would otherwise cause some tiles to appear washed out or dark relative to the training distribution the model expects. The same conversion procedure is applied identically to training and test tiles to ensure that the radiometric domain seen at inference matches the one seen during fine-tuning; any systematic difference between training and test tile appearance would introduce a domain shift that would degrade detection confidence.

Crown polygon annotations must also be converted to the supervision format the model requires. Object detection models of this family are trained on bounding boxes rather than polygon outlines, so each training crown polygon is converted to its axis-aligned bounding box in pixel coordinates. Boxes are clipped to the tile extent to remove any portion extending beyond the tile boundary, and boxes whose shortest side is smaller than 5px are discarded because they are too small to carry meaningful spatial information for the detector. Because tiles overlap, a crown that appears in several adjacent tiles contributes a bounding box to each of those tiles, which provides implicit augmentation for the small training dataset without requiring additional annotations.

4.9.2. Fine-Tuning Strategy

The pretrained model is fine-tuned end-to-end without freezing the backbone. Freezing the backbone would preserve the pretrained feature representations exactly, which is appropriate when the target domain is very similar to the pretraining domain. Here, however, the pretraining domain (airborne RGB imagery of North American forests) differs from the target domain (VHR satellite imagery of Mediterranean olive and urban trees) in sensor altitude, spatial resolution, crown morphology, and background texture. Allowing all layers to update during fine-tuning lets the network shift its learned representations toward Mediterranean crown appearance, at the cost of requiring more epochs to converge. Training runs for 15 epochs with a batch size of 4 and a learning rate of 0.001. The annotated tiles are split 80/20 by image into training and validation subsets (seed 42) so that validation loss can monitor overfitting to the small fine-tuning set throughout training.

4.9.3. Post-Processing and Output Format

A practical constraint of the model version used is that the configured score threshold is not applied internally during the prediction call. Predictions are therefore generated at a permissive threshold of 0.01 to retain all candidate detections, and confidence filtering is applied in post-processing by discarding any detection below the deployed threshold. This two-stage approach ensures that the threshold can be adjusted without rerunning inference.

The deployed confidence threshold differs between strata: 0.20 for urban tiles and 0.10 for rural tiles. Both values are substantially lower than the defaults used in the original DeepForest model. This reflects a characteristic of fine-tuning on a small dataset: the score distribution is compressed relative to a model trained on thousands of examples, so genuine crown detections receive lower confidence scores than they would from a fully trained model. The rural threshold is lower than the urban threshold because rural crowns in continuous olive canopy are spectrally more ambiguous and the model assigns them lower confidence even when they are correctly located; a higher threshold would suppress valid detections in this stratum.

After confidence filtering, detections are further filtered by bounding-box area. The lower bound follows the per-stratum minimum crown area used by all methods (75px for urban and 86px for rural). The upper bound is set to four times the maximum crown area ($4 \times 786 = 3,144\text{px}$), which is intentionally looser than the polygon-based upper bound used in other methods. This relaxation is necessary because bounding boxes circumscribe

rather than trace crown contours: a correctly detected crown whose polygon area is near the 786px maximum will have a bounding box area that substantially exceeds that value, and applying a strict polygon-level maximum would discard genuine large-crown detections.

The surviving bounding boxes are georeferenced to their spatial coordinates and converted to rectangular polygons carrying the canonical attribute set shared by all methods: tile identifier, land cover class, detection method, area in pixels and square meters, centroid and bounding-box coordinates, and mean NDVI. One GeoJSON file is written per tile. After all test tiles have been processed, predictions from overlapping tiles are merged and deduplicated using the shared rule ($\text{IoU} > 0.5$, retain higher mean NDVI), and the box-derived polygons enter the shared evaluation framework.

Parameter	Value
Pretrained model	<code>weecology/deepforest-tree</code>
Architecture / backbone	RetinaNet / ResNet-50 (≈ 32 M)
RGB stretch	per-tile 2–98 percentile $\rightarrow$ uint8
Tile size	512×512 PNG
Min training-crown area	75 px
Min box side	5 px
Training boxes / tiles	3,144 / 634
Train / val split (seed)	508 / 126 tiles (42)
Train / val boxes	2,558 / 586
Backbone freezing	none
Epochs / batch / lr	15 / 4 / 0.001
Prediction score threshold	0.01 (filtered afterwards)
Deployed threshold (u / r)	0.20 / 0.10
Box-area filter (u / r)	[75, 3,144] / [86, 3,144] px
Deduplication IoU	> 0.5
Test tiles processed	348

Table 4.6: DeepForest implementation parameters. The committed model is a RetinaNet detector with a ResNet-50 backbone, fine-tuned without backbone freezing.

4.10. U-Net Semantic Segmentation

U-Net is a fully convolutional segmentation model with an encoder–decoder architecture connected by skip connections that propagate fine spatial detail from the encoding path back to the decoding path [24]. It predicts a per-pixel crown probability over the full tile and recovers individual crowns through connected-component analysis. The motivation for including a segmentation model alongside the detection-based DeepForest is that segmentation predictions are not constrained to bounding rectangles: the network can learn boundaries that follow crown contours. Relative to the classical methods, those boundaries are determined by learned spatial context rather than a fixed geometric criterion, which should in principle produce more accurate crown delineation in complex canopy.

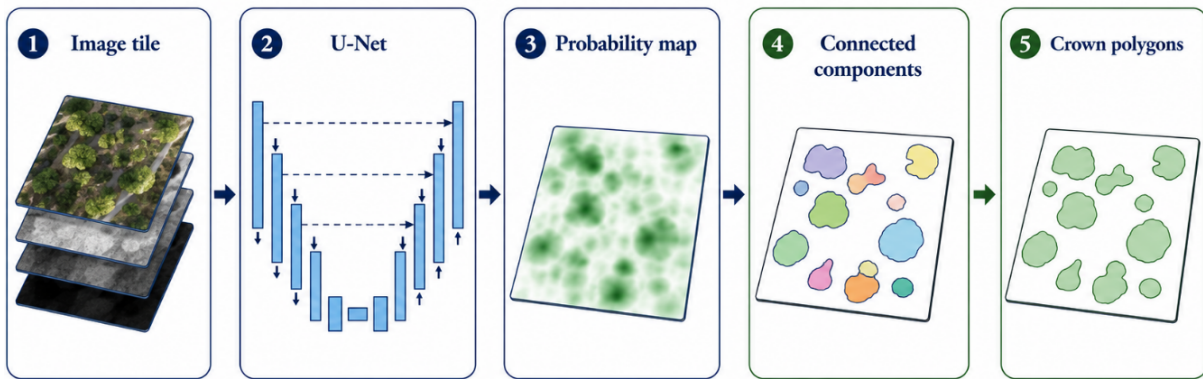


Figure 4.8: U-Net processing pipeline

4.10.1. Architecture and Input Configuration

The network uses a ResNet-34 encoder initialised from ImageNet weights paired with a standard U-Net decoder, built with the `segmentation_models_pytorch` library. Starting from ImageNet weights rather than random initialization is important even though ImageNet contains no remote sensing imagery: the lower encoder layers learn general edge and texture detectors that transfer across domains and accelerate convergence on the small training set available here. The input stem is modified to accept four spectral channels so that the near-infrared band contributes to the learned representation alongside the visible bands. This modification is motivated by the discriminative value of NIR for vegetation detection: excluding it would force the network to rely on visible-band contrast alone, discarding the strongest spectral signal available. The single-class output produces raw logits, with sigmoid activation applied at inference to convert logits to probabilities.

Each input band is z-score normalized using training-derived statistics and clipped to

$[-5, 5]$ to suppress extreme outliers that could destabilize the gradient computation. Non-finite values are replaced before any network operation. These normalization steps are essential for gradient stability and are applied identically to training and test tiles using statistics computed from training data only.

4.10.2. Training Target and Data Split

The supervision signal is derived from the crown polygon annotations by rasterizing each training polygon onto the tile grid to produce a binary crown mask: pixels inside an annotated crown receive label 1, all other canopy pixels receive label 0, and pixels outside the canopy mask receive an ignore value that is mapped to non-crown before training. This mapping is necessary because the network sees the full tile, not only the canopy, and assigning the ignore class to non-canopy pixels prevents the network from treating them as confident negative examples. Training polygons below 75px are excluded from the mask for the same reason as in the other methods: sub-threshold crowns introduce ambiguous supervision at crown boundaries.

The training tiles are divided into training and validation subsets at an 80/20 ratio by tile identifier (seed 42). Splitting by tile rather than by pixel ensures that geographically adjacent tiles are not simultaneously in both subsets, preventing spatial leakage. The validation subset is used exclusively to monitor overfitting and to select the best model checkpoint.

4.10.3. Data Augmentation

Augmentation is applied to the training tiles only to increase the effective diversity of the small annotated dataset without collecting additional data. Random horizontal and vertical flips (each with probability $p = 0.5$) and random 90° rotations expose the network to crowns in different spatial orientations, reducing overfitting to the specific orientation distribution of the training tiles. Brightness and contrast jitter (scale $\sim \mathcal{U}[0.8, 1.2]$, offset $\sim \mathcal{U}[-0.1, 0.1]$) is applied to the RGB channels only and deliberately excluded from the NIR channel. Applying radiometric jitter to the NIR band would corrupt the NDVI values derived from it and destroy the physical relationship between NIR and Red that the network needs to learn, so only the perceptual RGB channels receive this augmentation.

4.10.4. Loss Function and Optimization

The severe pixel-level class imbalance between crown and non-crown canopy pixels makes standard binary cross-entropy unsuitable: the loss would be dominated by the abundant non-crown pixels and the network would converge to a solution that under-detects crowns. Focal loss is used instead:

$$\mathcal{L}_{\text{focal}} = -\alpha, (1 - p_t)^\gamma, \log(p_t), \quad \alpha = 0.25, \gamma = 2.0, \quad (4.11)$$

where p_t is the predicted probability for the true class. The modulating factor $(1 - p_t)^\gamma$ down-weights easy examples: pixels that the network already classifies correctly with high confidence contribute little to the loss, and the gradient concentrates on the hard foreground crown pixels that the network is uncertain about. The parameter $\alpha = 0.25$ additionally up-weights the minority crown class relative to the background. Together these two terms direct the optimization toward difficult crown pixels rather than toward the easy non-crown majority.

The network is optimized with Adam at a learning rate of 10^{-4} . A reduce-on-plateau scheduler with patience 5 and reduction factor 0.5 automatically lowers the learning rate when the validation loss stops improving, allowing the optimizer to escape shallow local minima without manually tuning a learning rate schedule. Separate models are trained for the urban and rural strata for 30 epochs each with a batch size of 8. Training separate models per stratum is justified by the structural differences between the two environments: urban tiles have isolated trees against high-contrast built surfaces, while rural tiles have continuous or semi-continuous canopy with much weaker crown boundaries. A single model would need to simultaneously learn representations for both environments, which may lead to compromises that serve neither well. The checkpoint with the lowest validation loss is retained as the final model for each stratum.

4.10.5. Post-Processing and Output Format

At inference, the network produces a per-pixel crown probability via sigmoid activation. A threshold of 0.40 is applied to both strata to convert the probability map to a binary crown mask. The threshold is set below the conventional 0.50 default because focal loss with $\alpha = 0.25$ makes the model conservative: it assigns high probability only to pixels it is very confident about, and genuine crown pixels may receive probabilities in the range $[0.40, 0.60]$ that a 0.50 threshold would reject.

Morphological closing is applied to the binary mask before connected-component extrac-

tion to fill small intra-crown gaps. The kernel size differs between strata: 3×3 for urban tiles and 7×7 for rural tiles. The larger rural kernel serves the same purpose as in the pixel-based Random Forest: rural canopy produces spatially scattered positive pixels at this threshold, and the larger kernel is needed to merge these scattered predictions into crown-sized connected regions. Urban tiles have more spatially compact crown predictions, so a smaller kernel suffices without artificially enlarging the detected regions.

Connected components are extracted from the closed binary mask, filtered by the per-stratum size thresholds ($[75, 786]$ px for urban and $[86, 786]$ px for rural), and converted to georeferenced polygons carrying the canonical attribute set shared by all methods: tile identifier, land cover class, detection method, area in pixels and square meters, centroid and bounding-box coordinates, and mean NDVI recomputed from the raw scaled bands. One GeoJSON file is written per tile. After all test tiles have been processed, predictions from overlapping tiles are merged and deduplicated using the shared rule ($\text{IoU} > 0.5$, retain higher mean NDVI), and the mask-derived polygons enter the shared evaluation framework.

Parameter	Value
Architecture	U-Net / ResNet-34 encoder
Encoder initialization	ImageNet weights
Input channels	4 (NIR, Red, Green, Blue)
Input normalization	z-score (train-derived), clipped $[-5, 5]$
Training target	binary crown mask (ignore $\rightarrow$ non-crown)
Min training-crown area	75 px
Train / val split	80 / 20 by tile (seed 42)
Augmentation	H/V flip ($p = 0.5$), rot90, brightness/contrast jitter (RGB only)
Jitter scale / offset	$\mathcal{U}[0.8, 1.2]$ / $\mathcal{U}[-0.1, 0.1]$
Loss function	Focal loss ($\alpha = 0.25$, $\gamma = 2.0$)
Optimizer / learning rate	Adam / 10^{-4}
LR scheduler	Reduce-on-plateau (patience 5, factor 0.5)
Epochs / batch size	30 / 8 per stratum
Model selection	lowest validation loss check- point
Inference threshold	0.40 (both strata)
Urban post-processing	threshold 0.40 / 3×3 closing
Rural post-processing	threshold 0.40 / 7×7 closing
Size filter urban / rural	$[75, 786]$ / $[86, 786]$ px
Deduplication criterion	IoU > 0.5 , retain higher mean NDVI
Output format	GeoJSON polygons with canonical attributes

Table 4.7: U-Net implementation parameters. Separate models are trained per stratum. The inference threshold of 0.40 is set below the conventional default to account for the conservative probability outputs produced by focal loss under severe class imbalance.

4.11. Analytical Hypotheses

The benchmark is structured around three working hypotheses derived from the state of the art. First, simple spectral and rule-based methods will separate vegetation but will delineate individual crowns only weakly. Second, object-based and classical machine

learning methods will improve spatial reasoning and interpretability, yet remain sensitive to segmentation and feature design choices. Third, deep learning methods will provide the strongest detection and segmentation performance, particularly in heterogeneous urban scenes, at the cost of higher data and annotation demands. These hypotheses frame the comparison without predetermining its outcome.

5 | Results

This chapter presents the outputs of the experiments: the final reference dataset, the preprocessing products, the per-method detection metrics, and the supporting figures, tables, observed behaviours, and error categories. All detection metrics are computed on the fixed 302-crown test set using the shared Hungarian-assignment IoU matching at thresholds 0.25 and 0.50.

5.1. Reference Dataset Results

5.1.1. Annotation Protocol

The annotation protocol produced a final reference dataset of 800 manually delineated crown polygons, distributed equally between the two LU/LC-derived strata: 400 crowns in rural areas and 400 crowns in urban areas. Each polygon was digitized over the JILIN-1 RGB composite as the primary interpretation layer, with NDVI used only as an auxiliary aid, and carries the `image_id`, `crown_status`, and `date` attributes defined in the protocol. The complete digitization required approximately two weeks of interactive editing in QGIS and yielded a sample balanced across the urban–rural gradient.

Parameter	Rural	Urban	Total
Annotated crowns	400	400	800
Reference origin	Manual digitization for this thesis		
Urban inventory support	ArcGIS Calabria FeatureServer used to check and complete urban annotations		
Annotation type	Individual crown polygons		
Interpretation layers	JILIN-1 RGB composite (primary) + NDVI (auxiliary)		
Annotation effort	≈ 2 weeks		

Table 5.1: Summary of the reference crown dataset used in the benchmark.

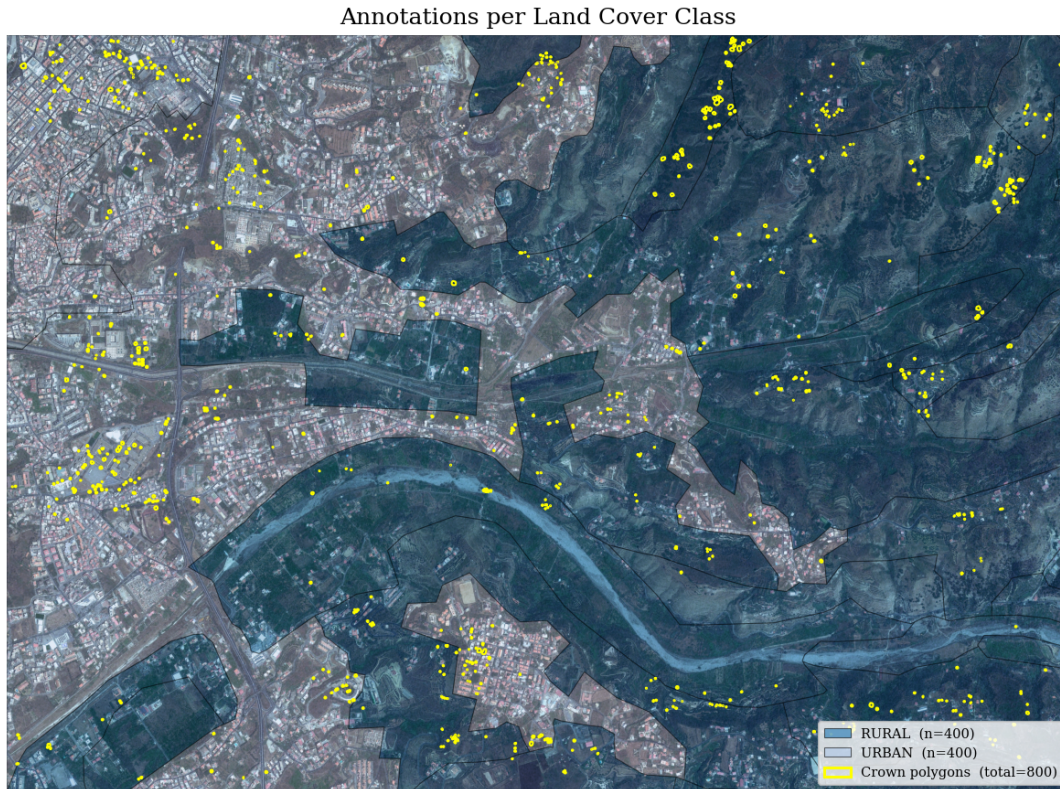


Figure 5.1: Spatial distribution of the 800 manually annotated crown polygons (yellow points) overlaid on the AOI true-colour composite. Rural sampling zones are shown in dark blue ($n = 400$) and urban zones in light blue ($n = 400$). Crown annotations are distributed across the full spatial extent of the AOI, covering both the dense urban fabric in the west and the more continuous vegetated areas to the north and east.

5.1.2. Annotation Quality and Uncertainty

The quality-control review confirmed that the annotated crowns span a wide size range while remaining centred on individual-crown scales. Across the 800 polygons the crown area distribution has a mean of 300.7 px and a median of 218.5 px, with a minimum of 32.9 px and a maximum of 2,211 px; the interquartile range is 143.8–360.7 px. When the minimum per-stratum crown-size threshold is applied, 777 crowns (97.1%) exceed 75 px and 23 (2.9%) fall below it. Of the 800 crowns, 799 fall inside at least one tile extent and one crown lies outside all tiles and is therefore excluded from every experiment. These figures indicate that the sub-threshold and out-of-extent cases together affect only a small fraction of the reference set, so the usable reference sample remains close to the full 800 annotations.

5.2. Preprocessing Results

5.2.1. Radiometric Foundation and Masking

Spectral distribution characterization

Figure 5.2 shows the spatial appearance of each of the four bands across the AOI. Band 1 (NIR) displays the highest mean DN value (1 462) and the broadest dynamic range. Bands 2–4 (Red, Green, Blue) show progressively lower mean values (713, 687, and 624 respectively). After screening, only 5 pixels out of 86,404,640 fall outside the admissible range $[0, 10,000]$, and the scaled `float32` product spans 0.0000–0.9795 with zero nodata pixels.

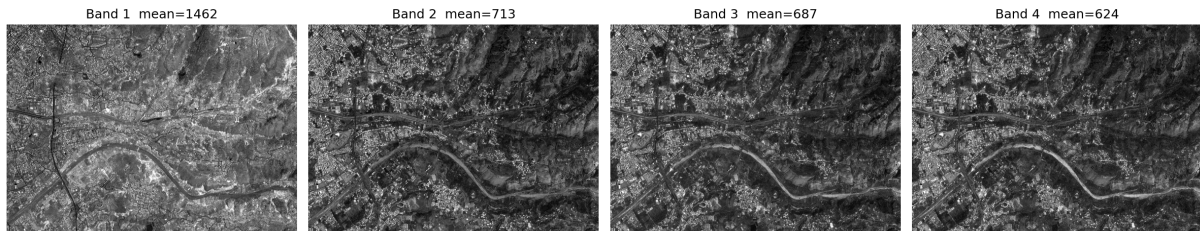


Figure 5.2: Spatial overview of the four spectral bands of the JILIN-1 KF01C image over the AOI. From left to right: Band 1 (NIR, mean DN = 1 462), Band 2 (Red, mean = 713), Band 3 (Green, mean = 687), Band 4 (Blue, mean = 624). The higher relative brightness of vegetated areas in Band 1 is visible in the hillside zones to the north and east.

Urban and Rural Spectral Differences

Figure 5.3 disaggregates the per-band DN distributions by land cover stratum. In all four bands, the rural distribution is narrower and more peaked than the urban distribution, which shows a heavier upper tail. This is particularly pronounced in Band 1 (NIR), where the urban stratum reaches a 99th-percentile value of 2 364 DN compared to 2 111 DN for the rural stratum, and in Bands 2 and 3, where urban pixels show substantially greater spread above the median. The two strata present different per-band DN distributions. The full per-band, per-class statistics are reported in Table 5.2.

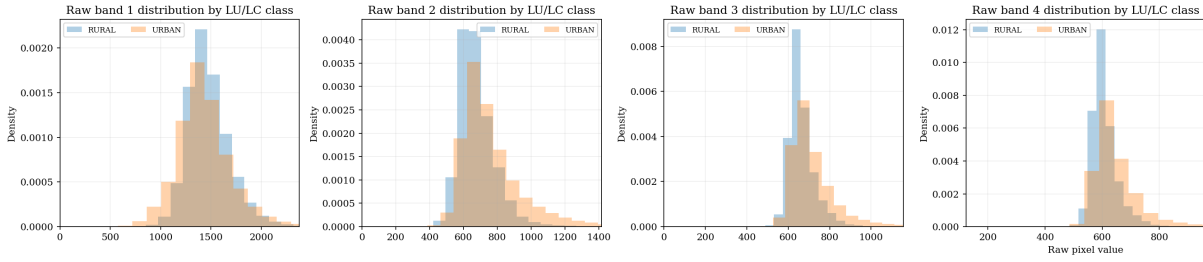


Figure 5.3: Raw DN distributions by land cover class (rural in blue, urban in orange) for all four spectral bands. Urban pixels show broader distributions and heavier upper tails relative to rural pixels, particularly in Bands 1–3.

Band	Class	Count	Min	p01	p50	p99	Max
1	Rural	52 203 225	0.0	1 057	1 447	2 111	9 762
1	Urban	34 201 415	0.0	852	1 408	2 364	11 490
2	Rural	52 203 225	0.0	495	658	1 045	5 628
2	Urban	34 201 415	0.0	510	716	1 434	6 202
3	Rural	52 203 225	231	555	652	885	3 665
3	Urban	34 201 415	0.0	565	688	1 168	4 685
4	Rural	52 203 225	296	531	598	768	2 819
4	Urban	34 201 415	123	539	627	970	4 264

Table 5.2: Per-band raw DN statistics stratified by land cover class. p01 and p99 denote the 1st and 99th percentiles respectively.

Data Quality: Nodata and Anomalous Pixels

Figure 5.4 maps the spatial distribution of nodata and out-of-range pixels (defined as DN values below 0 or above 10 000). Such pixels are sparse and spatially scattered across the scene, with no systematic concentration in any particular region. The dataset reports 100 % valid pixels with zero nodata entries in the delivered product; the scene is radiometrically complete within the AOI extent.

Anomalous pixels were further mapped across all 989 valid tiles (Figure 5.4). The mean anomaly density across all tiles is 0.0000 %, and no tile exceeds the 5 % anomaly coverage flag threshold. The maximum observed anomaly density in any tile is 0.0008 %, corresponding to at most 2 pixels per tile in the affected cases. No tile-level exclusion was applied on the basis of anomaly density.

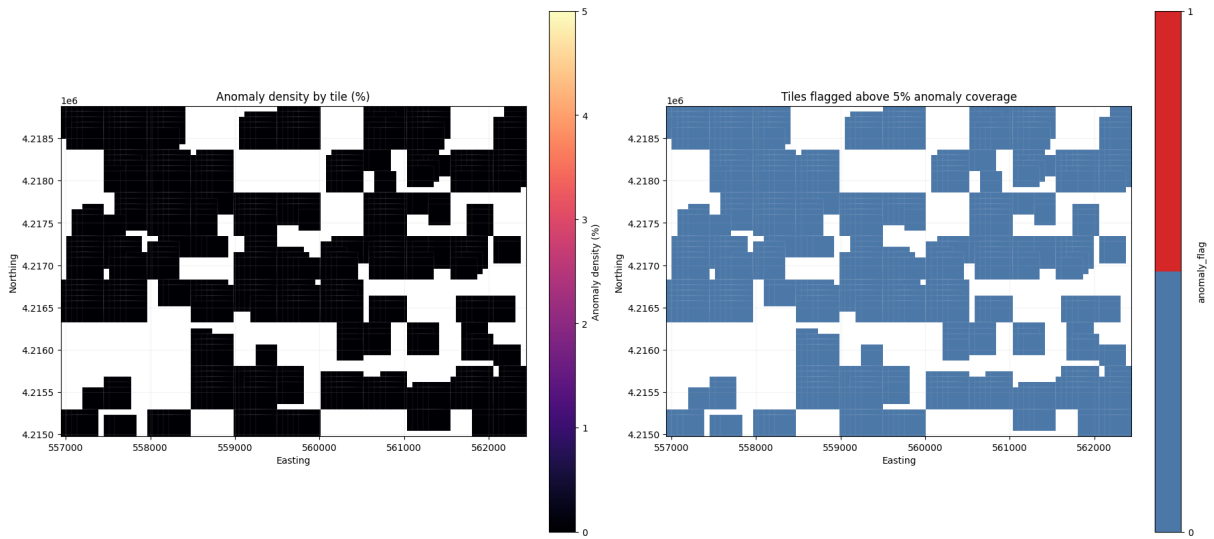


Figure 5.4: Spatial distribution of anomaly density by tile (left) and tiles flagged above the 5 % anomaly coverage threshold (right). No tiles are flagged (all shown in blue on the right panel). Anomaly density is effectively zero across the entire scene.

Shadow Coverage

Shadow pixels were identified using a brightness proxy mask defined as pixels where both NIR and Red scaled values fall below 0.1. Shadow coverage was computed per tile and stratified by land cover class (Figure 5.5). Urban tiles have a mean shadow coverage of 4.71 %, compared with a rural mean of 0.56 %. The maximum observed shadow coverage is 14.76 % in urban tiles, below the 20 % flag threshold (Figure 5.5). No tiles were excluded on the basis of shadow coverage. The substantially heavier shadow load in the urban stratum is consistent with the high-contrast built environment, where buildings and street furniture cast shadows adjacent to street trees.

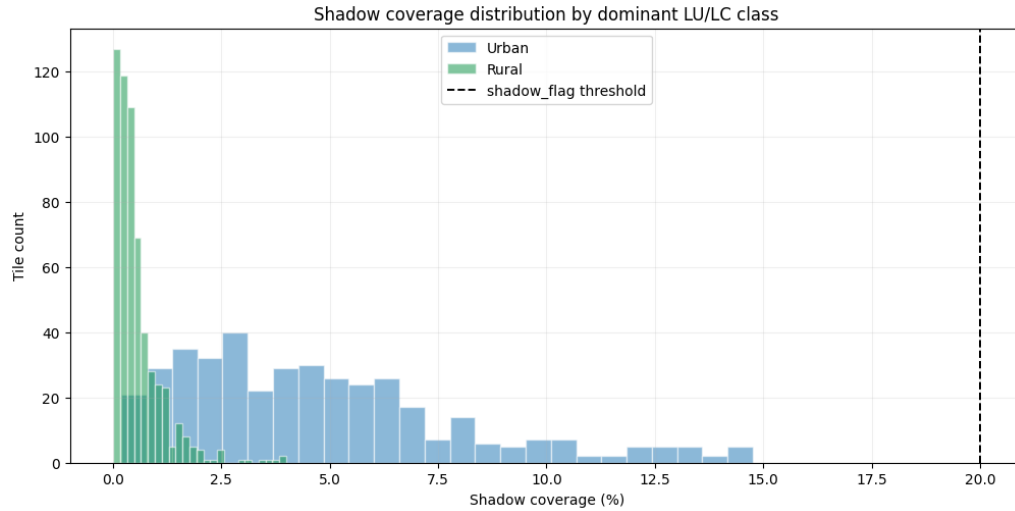


Figure 5.5: Shadow coverage distribution by dominant land cover class (urban in blue, rural in green). The shadow flag threshold of 20 % is shown as a black dashed vertical line. No tile exceeds this threshold. Urban tiles show a substantially broader and heavier-tailed distribution, with mean shadow coverage of 4.71 % compared to 0.56 % for rural tiles.

5.2.2. Vegetation Separation and NDVI Thresholding

NDVI Spatial Distribution

Figure 5.6 shows the spatial distribution of NDVI across the AOI, and Figure 5.7 shows its histogram. Across the scene, NDVI values range from -1 to 1 with a finite mean of 0.344 and a standard deviation of 0.077 . The distribution is unimodal and positively skewed, concentrated in the range $[0.2, 0.5]$, with no pixels outside the valid interval. In the spatial map, higher NDVI values occur in the vegetated hillside zones in the northern and eastern portions of the AOI, while the urban core in the west shows lower and more variable values.

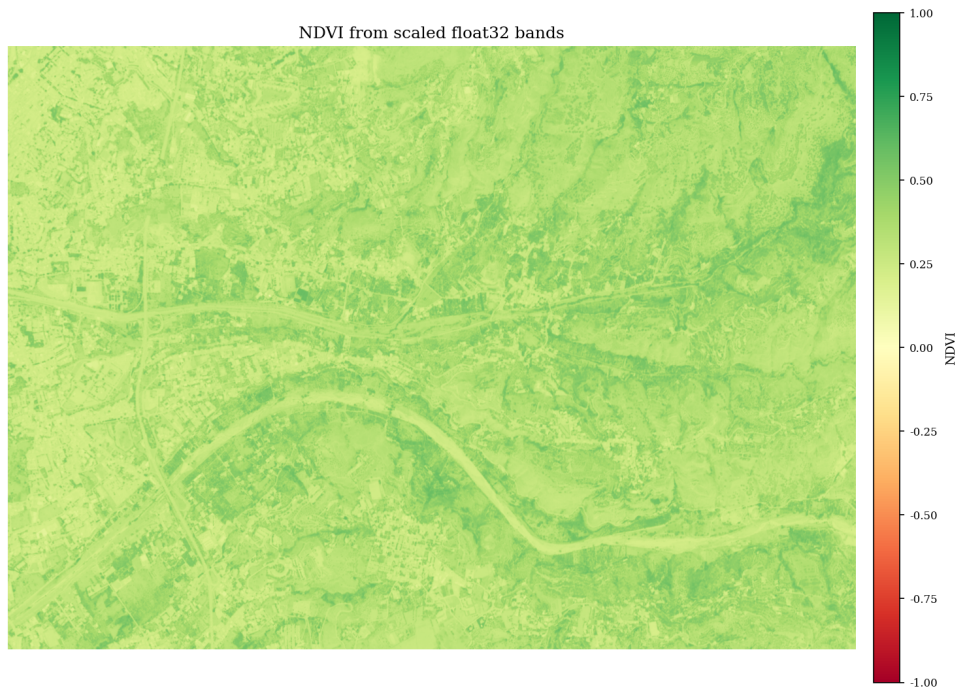


Figure 5.6: NDVI map computed from scaled `float32` bands over the AOI (mean = 0.344, std = 0.077). Higher values (darker green) correspond to dense vegetated cover on the hill-side zones; lower values (yellow-green) correspond to the urban core and the dry riverbed visible in the centre of the scene.

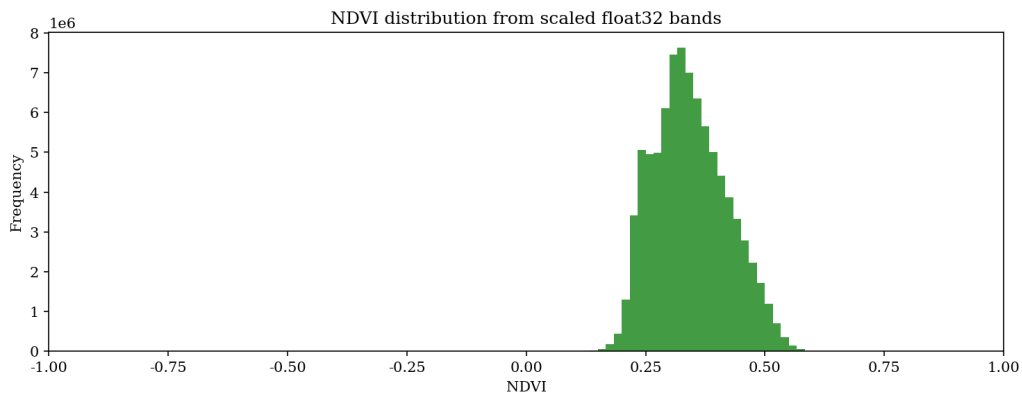


Figure 5.7: NDVI histogram computed from scaled `float32` bands. The distribution is unimodal and concentrated in the range $[0.2, 0.5]$, with mean = 0.344 and std = 0.077. There are no values below -0.25 .

NDVI Threshold Sensitivity

Threshold selection was performed empirically by examining NDVI histograms across random tiles from each split and land cover stratum. NDVI distribution across all four

subset–class combinations is unimodal and concentrated in the range $[0.2, 0.5]$, with the tile means clustering just below the 0.35 threshold. The threshold was evaluated at four candidate values (0.25, 0.30, 0.35, 0.40) on three representative tiles spanning the two landscape strata (Figures 5.8, 5.9, 5.10).

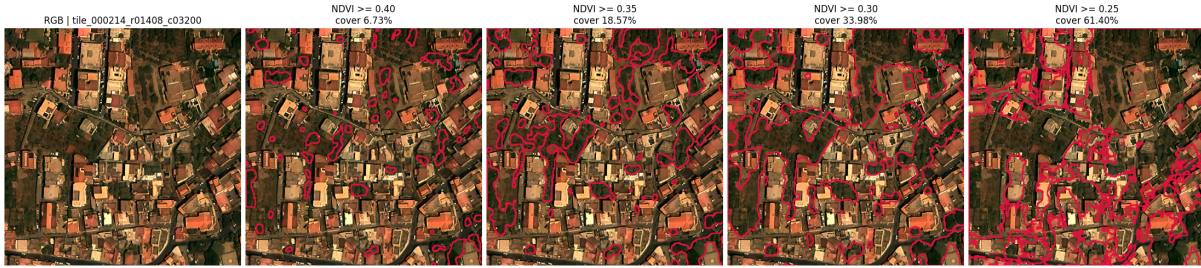


Figure 5.8: NDVI threshold sensitivity on urban tile `tile_000214_r01408_c03200`. Vegetation coverage increases from 6.73 % at threshold 0.40 to 61.40 % at 0.25. At 0.30 and 0.25, large non-vegetated areas such as roads and rooftops are included. The threshold of 0.35 retains 18.57 % coverage.

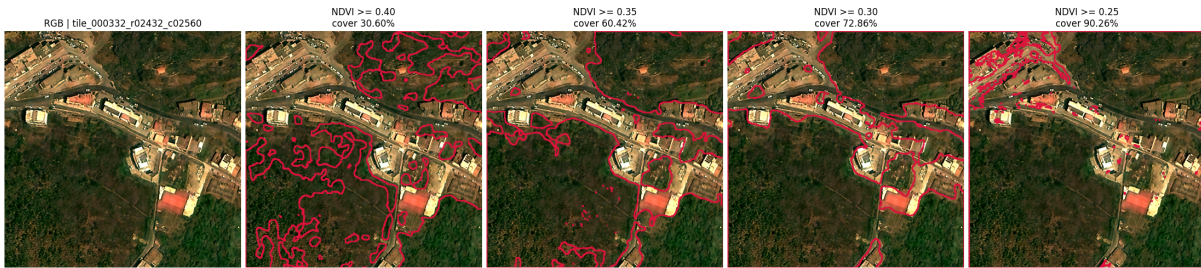


Figure 5.9: NDVI threshold sensitivity on peri-urban tile `tile_000332_r02432_c02560`. At threshold 0.35, 60.42 % of the tile is classified as vegetated.

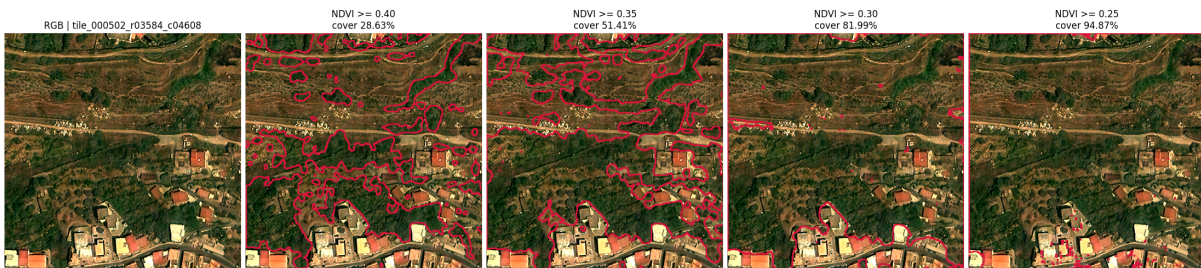


Figure 5.10: NDVI threshold sensitivity on rural tile `tile_000502_r03584_c04608`. Coverage at 0.35 reaches 51.41 %. At 0.25, 94.87 % of the tile is flagged.

Table 5.3 summarises the mean NDVI mask coverage across the three sample tiles at each candidate threshold. The value of 0.35 was selected as the operating threshold. At 0.35 the

mean vegetation coverage was 43.5 %, while lower thresholds produced higher coverage that included non-vegetated surfaces (Table 5.3). This threshold is applied uniformly across all tiles and both landscape strata.

Threshold	Mean coverage (%)	Min (%)	Max (%)
0.40	21.99	6.73	30.60
0.35	43.47	18.57	60.42
0.30	62.94	33.98	72.86
0.25	82.18	61.40	90.26

Table 5.3: Mean, minimum, and maximum NDVI mask coverage across three representative tiles at four candidate thresholds. The selected operating threshold is 0.35.

5.2.3. Tiling and Spatial Partitioning

Tiling and Spatial Split

The scene was subdivided into 512×512 pixel tiles and partitioned via a spatial block split to prevent leakage between training and test sets. The full scene yields 989 valid tiles, split into 641 training tiles (49 spatial blocks) and 348 test tiles (22 spatial blocks). The rural/urban proportions are stable across the partition (train 57.4%/42.6%; test 62.6%/37.4%). Of the 800 annotated crowns, 799 were successfully assigned to a tile; one crown falls outside all tile extents and is excluded from all experiments. By centroid assignment, the training subset receives 491 crowns (282 rural, 209 urban) and the test subset 308 crowns (150 rural, 158 urban). The test evaluation subset is further filtered to 302 crowns by removing those below the minimum area threshold derived from the object size analysis.

Subset	Tiles	Rural (%)	Urban (%)	Crowns
Train	641	57.4	42.6	491
Test	348	62.6	37.4	308
Total	989	—	—	799

Table 5.4: Tile and crown distribution across training and test subsets after spatial block partitioning. One crown falls outside all tile extents and is excluded. The rural/urban proportions are consistent across splits.

5.2.4. Z-Score Normalization

Per-band normalization statistics were computed exclusively from the 641 training tiles ($n = 168,034,286$ valid pixels per band). After applying the train-derived statistics, the resulting z-score normalized tiles exhibit per-band means of order 10^{-7} and standard deviations of exactly 1.0 on the training set (Table 5.5), confirming that the transform is correctly fitted and applied one-way from train to test.

Band		Mean (raw)	Std (raw)	Mean (norm.)	Std (norm.)
Band (NIR)	1	0.14531	0.02543	9.8×10^{-8}	1.0
Band (Red)	2	0.07163	0.01597	1.0×10^{-7}	1.0
Band (Green)	3	0.06882	0.00989	-3.6×10^{-7}	1.0
Band (Blue)	4	0.06256	0.00714	3.4×10^{-7}	1.0

Table 5.5: Per-band normalization statistics computed from 641 training tiles ($n = 168,034,286$ valid pixels per band). Raw values are in scaled `float32` units $[0, 1]$. Normalized means are effectively zero and standard deviations are exactly 1.0.

5.2.5. Feature Separability Analysis

GLCM texture metrics were computed on the NIR band across three window sizes (7, 11, and 15 pixels) for urban and rural training tiles. The urban stratum exhibits substantially higher contrast (≈ 0.48 vs ≈ 0.25), higher entropy (1.66–2.11 vs 1.32–1.63), and higher correlation (0.48–0.58 vs 0.46–0.53) than the rural stratum across all window sizes. Homogeneity is higher in rural tiles (≈ 0.898 vs ≈ 0.865). All four metrics are stable across window sizes, indicating that the urban–rural texture contrast is not an artefact of window-size choice.

JM Separability and Feature Ranking

Jeffries–Matusita distances were computed between tree and non-tree classes for all candidate features across both strata. Table 5.6 reports the full feature ranking. Object area and perimeter are the only features with JM distance above 1.0 in both strata (mean JM 1.495 and 1.365 respectively); every pixel-level texture and spectral-index feature falls below the 1.0 separability threshold, with the GLCM contrast features (mean JM ≤ 0.59)

the strongest of the sub-threshold group. This result indicates that, individually, only the two geometric size descriptors are reliably discriminative between tree and non-tree objects.

Rank	Feature	Urban JM	Rural JM	Mean JM	Useful
1	obj_area	1.396	1.594	1.495	Yes
2	obj_perimeter	1.267	1.463	1.365	Yes
3	nir_glcml_contrast=7	0.691	0.489	0.590	No
4	nir_glcml_contrast=11	0.633	0.348	0.490	No
5	obj_circularity	0.507	0.304	0.406	No
6	nir_glcml_contrast=15	0.486	0.300	0.393	No
7	ndvi_top_hat=10	0.409	0.282	0.345	No
8	obj_elongation	0.555	0.226	0.391	No
9	ndvi_top_hat=5	0.235	0.193	0.214	No
10	obj_mean_ndvi	0.031	0.168	0.099	No

Table 5.6: Top-10 feature ranking by mean Jeffries–Matusita distance between tree and non-tree classes across urban and rural training tiles. Features with JM distance ≥ 1.0 in at least one stratum are marked as useful. Object area and perimeter are the only features exceeding the 1.0 separability threshold.

5.2.6. Size Thresholds

Figure 5.11 shows the object area and perimeter distributions for tree and non-tree segments in urban and rural training tiles. Tree objects in both strata are concentrated in the range of 75–800 pixels in area, while non-tree objects span a much wider range including very small fragments and very large continuous regions. Applying the 90th-percentile tree-recall criterion, the minimum area thresholds are 75 pixels for urban (tree recall 90.42%) and 86 pixels for rural (tree recall 90.28%); a common maximum area of 786 pixels (tree recall 90.23%) is applied across both strata.

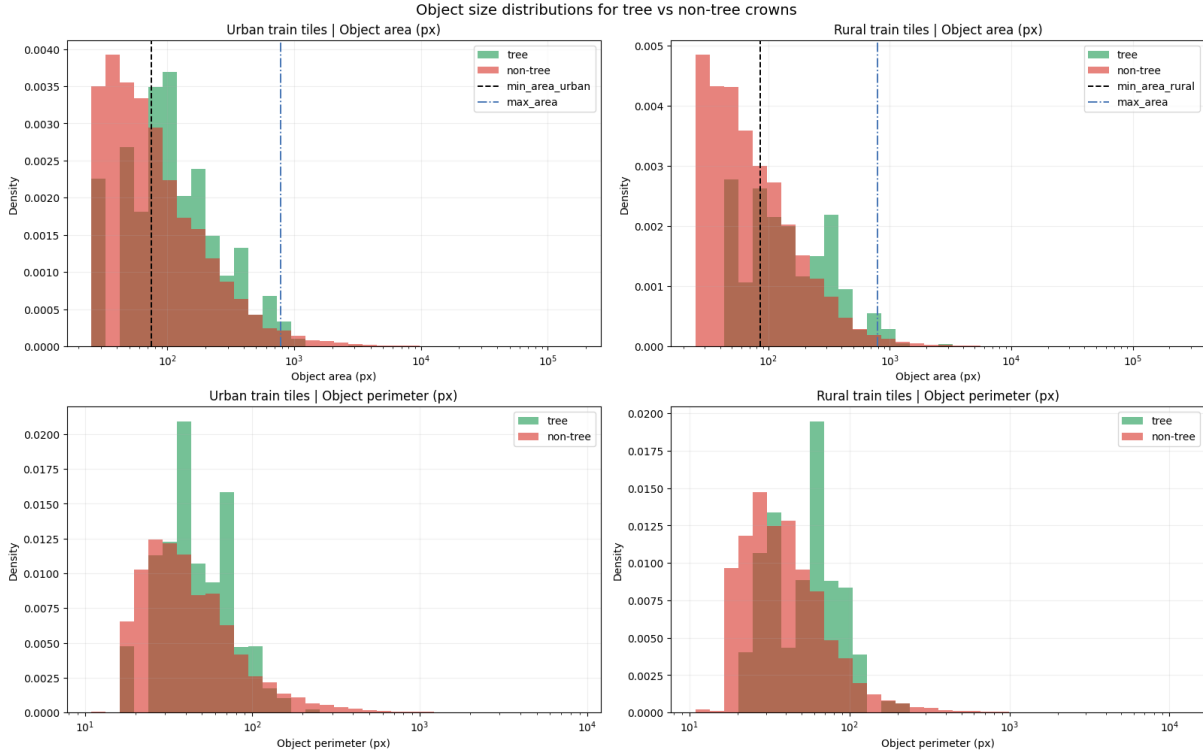


Figure 5.11: Object area (top) and perimeter (bottom) distributions for tree (green) and non-tree (red) segments in urban (left) and rural (right) training tiles, shown on a logarithmic scale. Vertical dashed lines indicate the minimum area threshold (black) and maximum area threshold (blue). Tree objects are concentrated between approximately 75 and 800 pixels in area, while non-tree objects span a substantially wider range in both directions.

5.2.7. Class-Balance Assessment

Class Imbalance

The tree-to-non-tree pixel imbalance was quantified at two levels: raw (all vegetation pixels above the NDVI threshold) and filtered (annotated crown polygons only). Table 5.7 reports the results by stratum. At the raw pixel level, the rural stratum is nearly balanced (1.16:1 tree to non-tree), while the urban stratum is imbalanced (0.38:1). Once filtered to annotated crown pixels only, both strata become severely imbalanced: 1:55.5 for rural and 1:42.5 for urban. Crown pixels therefore represent a small fraction of total canopy, which directly motivates the class-weighting, oversampling, and focal-loss strategies adopted by the supervised methods.

Class	Raw ratio	Raw (T:NT)	Filtered ratio (1:X)	Blobs (mean)
Rural	1.16:1	53.7%:46.3%	1:55.5	17.0
Urban	0.38:1	27.4%:72.6%	1:42.5	21.6

Table 5.7: Class imbalance summary by land cover stratum. Raw ratio refers to all pixels; filtered ratio refers to annotated crown pixels vs background. The filtered imbalance ranges from 1:42 to 1:55.

Retained Blob Inspection

Figures 5.12 and 5.13 show representative training tiles for the rural and urban strata respectively, with retained vegetation blobs (candidate crown regions after NDVI masking and area filtering) overlaid in red. The filtered ratios reported per tile range from 1:27 to 1:120 in rural tiles and from 1:28 to 1:143 in urban tiles. The number of retained blobs per tile ranges from 7 to 34 in rural and from 8 to 34 in urban. The most extreme per-tile imbalances correspond to dense urban fabric with minimal tree cover.

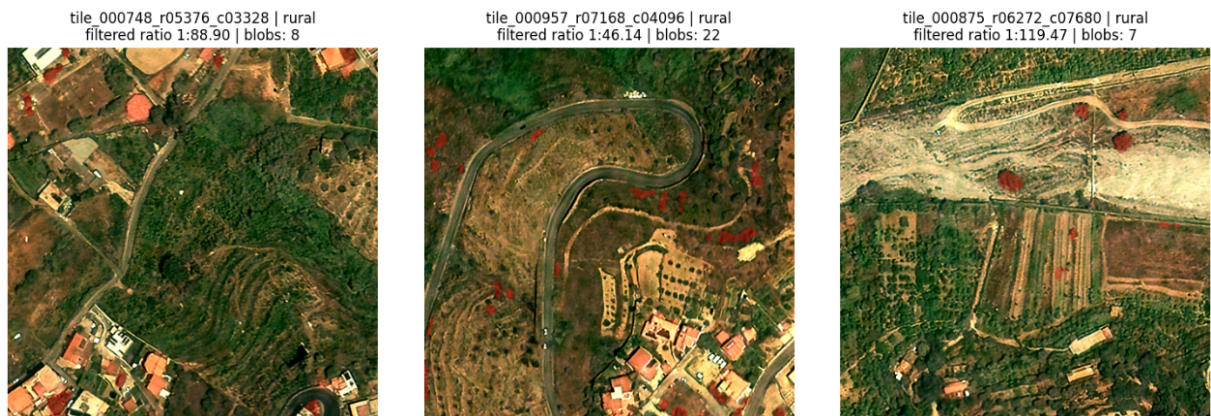


Figure 5.12: Sample of 3 rural training tiles with retained vegetation blobs (red markers) overlaid on the true-colour composite. Each tile caption reports the filtered non-tree to tree ratio and the blob count. Ratios range from 1:27 to 1:120.



Figure 5.13: Sample of 3 urban training tiles with retained vegetation blobs (red markers) overlaid on the true-colour composite. Filtered ratios range from 1:28 to 1:143. The highest-imbalance tiles correspond to dense urban fabric with minimal tree cover.

Table 5.8 consolidates the key parameters established during the preprocessing pipeline. These values are fixed for all five detection methods and are not re-optimised per method.

Parameter	Value
Radiometric scaling	DN / 10 000, float32
Tile size	512 × 512 px
Stride	128 px
Valid tiles	989
Training tiles	641 (49 spatial blocks)
Test tiles	348 (22 spatial blocks)
NDVI threshold	0.35
Min crown area	75 px urban; 86 px rural
Max crown area	786 px
Deduplication IoU	> 0.5 (keep higher mean NDVI)
Shadow flag threshold	20 % (no tiles excluded)
Anomaly flag threshold	5 % (no tiles excluded)
Normalization	z-score, train-derived parameters

Table 5.8: Summary of preprocessing parameters applied uniformly across all five detection methods. Parameters are derived from training data only and held fixed throughout all experiments.

5.3. Watershed and Local Maxima

5.3.1. NDVI Surface Preparation

On the six sample tiles used to develop the pipeline, per-tile NDVI computed from the scaled `float32` NIR and Red bands spanned the valid range after clipping, with tile-mean NDVI between 0.26 and 0.31 and canopy fraction ($\text{NDVI} > 0.35$) between 10.4% and 25.7%. After the fill-and-restore step, the minimum surviving values on the masked and smoothed canopy surface ranged from approximately 0.337 to 0.347 on most tiles. These values remained close to the 0.35 threshold, with small deviations caused by Gaussian smoothing near canopy boundaries, while out-of-canopy pixels were correctly restored as NaN.

5.3.2. Smoothing and Scale Selection

Gaussian smoothing preserved crown peaks while suppressing intra-canopy noise. For an isolated sharp peak, smoothing at $\sigma = 1.5$, 2.0, and 3.0 px retained 53.0%, 52.7%, and 52.0% of the raw NDVI maximum, respectively, whereas retention for typical canopy peaks remained between 97.7% and 99.2% across all three scales. Based on this sweep, the smoothing scale was fixed at $\sigma = 2.0$ px for urban tiles and $\sigma = 3.0$ px for rural tiles.

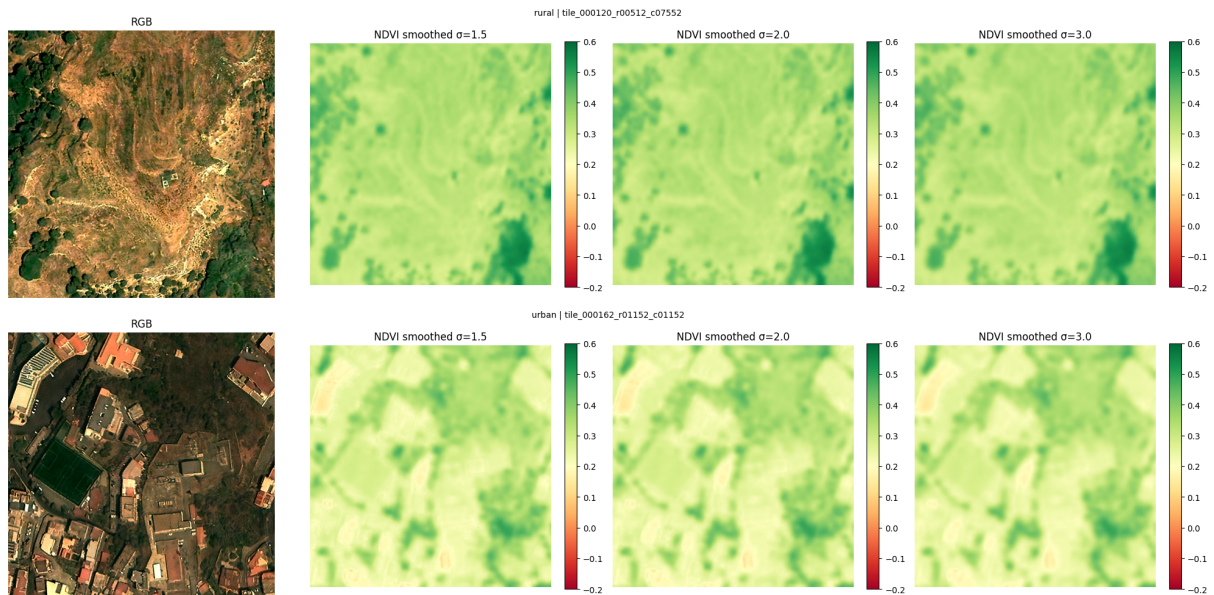


Figure 5.14: NDVI peak retention under Gaussian smoothing ($\sigma = 1.5$, 2.0, and 3.0 px).

5.3.3. Local Maxima Detection

Local maxima were detected on the smoothed NDVI surface after restoring the canopy mask. Candidate peaks were ranked by smoothed NDVI, and non-maximum suppression

with a minimum spacing of 6 px removed weaker peaks located too close to a stronger candidate. This step was intended to suppress multiple detections caused by fine-scale variation within the same crown while retaining spatially separated peaks that could represent distinct treetops.

The local-maxima counts revealed a substantial difference between the two strata. At the selected 6 px minimum spacing, rural tiles produced a mean of 174 candidate maxima per tile, compared with 92 in urban tiles. Relative to the annotated crown counts, the number of detected maxima was 133.8 times higher in rural tiles and 26.3 times higher in urban tiles. The much stronger over-detection in rural areas showed that a global local-maximum rule could not distinguish treetops from branches, crown edges, and internal texture within continuous vegetation. This result motivated the subsequent blob-aware restriction rather than the direct use of every detected maximum as a watershed marker.

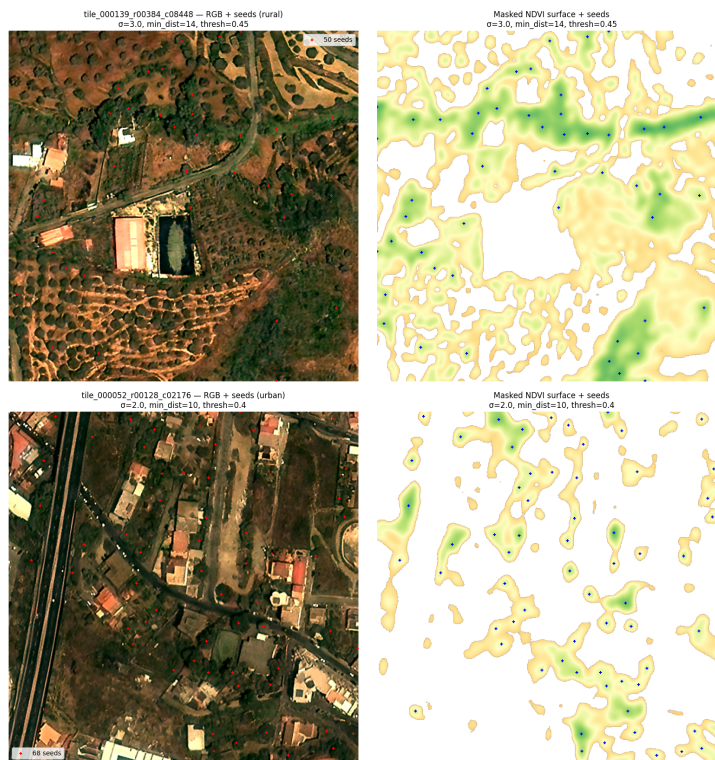


Figure 5.15: Mean local-maxima counts per tile by stratum and minimum-distance setting.

5.3.4. Canopy Restriction and Blob Segmentation

Connected-component analysis of the canopy mask confirmed that most vegetated area was concentrated in a few large, merged blobs. On the worst-case rural tile, the canopy was divided into 41 blobs with a median size of 146 px and a maximum size of 157,978 px. Only four blobs exceeded the 786 px crown-size cutoff, but these components contained

61.4% of all canopy pixels on the tile. Across the sample, small blobs (≤ 786 px) accounted for only 2.5%–17.8% of the canopy, while the flagged large blobs contained the remainder. Aggregated by stratum, large blobs contained an average of 93.4% of rural canopy and 82.2% of urban canopy, quantifying the proportion of vegetation for which local NDVI contrast could not reliably resolve individual crowns.

The candidate local maxima were intersected with the connected canopy components. Maxima located inside blobs larger than 786 px were rejected because the high candidate density in these components was dominated by within-canopy texture. For each eligible small blob, the candidate with the highest smoothed NDVI was retained as its marker. If no candidate was returned because of a plateau or boundary effect, the maximum-NDVI pixel within the blob was used as a fallback marker. This procedure converted the large set of spacing-based candidate maxima into at most one robust seed per small canopy component.

5.3.5. Per-Blob Seeding and Watershed Transform

After local-maxima detection and blob-aware restriction, the rural sample tiles retained a mean of 61.3 watershed seeds, while the urban sample tiles retained a mean of 39.3. Each retained seed was placed at the strongest local maximum of its corresponding small blob. The watershed transform was then applied to the inverted smoothed surface, $1 - \text{NDVI}$, causing the algorithm to expand outward from high-NDVI treetop locations while remaining confined to the selected canopy components.

After applying the stratum-specific size filters, rural tiles retained a mean of 21.3 segments and urban tiles retained a mean of 22.3. Depending on the tile, between 35% and 81% of the seeded regions were removed, with the removal rate generally higher in rural tiles because more markers produced segments outside the accepted size range. The similar final segment counts in the two strata, despite their different initial local-maxima and seed densities, demonstrate the strong effect of blob restriction and size filtering on the final watershed output.

5.3.6. Post-Processing and Output Format

All 348 test tiles were processed and vectorised, producing 354 per-tile GeoJSON files with no empty outputs. Before deduplication, the method generated 4,999 detections. The shared deduplication rule removed 3,755 detections (75%), primarily because of the 75% overlap between adjacent tiles, leaving 1,244 detections for evaluation. Of these, 80 matched a reference crown at IoU 0.25, while 1,164 were false positives and 228 reference

crowns remained undetected.

Performance by stratum and threshold is reported in Table 5.9. Recall was higher in the urban stratum (0.361) than in the rural stratum (0.153), whereas the mean overlap of matched detections was higher in rural tiles, with an mIoU of 0.611 compared with 0.536 in urban tiles. Watershed produced the lowest detection scores among the five evaluated methods. Its dominant error mode was unresolved continuous canopy: unrestricted local-maxima detection generated many texture-related candidates, while the blob-aware restriction excluded large merged vegetation components and therefore reduced recall. In the continuous vegetation that remained eligible for processing, weak NDVI gradients between neighboring crowns also caused under-segmentation.

Metrics are reported at two IoU thresholds: 0.25, which is a permissive criterion that counts approximately overlapping detections as correct and captures the method’s spatial reach, and 0.50, which is a stricter standard requiring more accurate crown delineation. Together they reflect how detection accuracy and boundary fidelity vary across the two matched sets.

Subset	P _{.25}	R _{.25}	F1 _{.25}	P _{.50}	R _{.50}	F1 _{.50}	mIoU
Urban	0.094	0.361	0.149	0.053	0.203	0.083	0.536
Rural	0.036	0.153	0.059	0.028	0.120	0.046	0.611
All	0.064	0.260	0.103	0.040	0.162	0.064	0.557

Table 5.9: Watershed detection performance by stratum at IoU 0.25 and 0.50. mIoU is the mean IoU of matched true positives at IoU 0.25.



Figure 5.16: Example watershed detections on representative urban and rural test tiles.

5.4. OBIA with Random Forest

5.4.1. Tile-Level Input Preparation

Each 512×512 px tile was loaded as a scaled four-band raster and restricted to the NDVI canopy mask before segmentation, with all non-canopy pixels set to zero and the NIR, Red, and Green bands min-max normalised within the canopy area. On the six development tiles this produced a clean canopy-only input on which segmentation operated without mixing vegetation and background surfaces.

5.4.2. Canopy Segmentation

The Felzenszwalb scale sweep showed a clear trade-off between object count and the fraction of objects falling in the target crown-size range. At scale 10, segmentation produced the most objects with 88%–91% of them in range; at scale 15 the in-range fraction was 86%–88%; and at scale 50 object counts dropped sharply with only 66%–74% in range. Scale 15 was selected as the operating value, balancing boundary adherence against fragmentation. The subsequent canopy-fraction restriction (≥ 0.5) retained the large majority of objects, discarding only boundary-mixed segments (for example 363 retained against 37 discarded on one urban tile) with zero zero-canopy violations across all sample tiles.

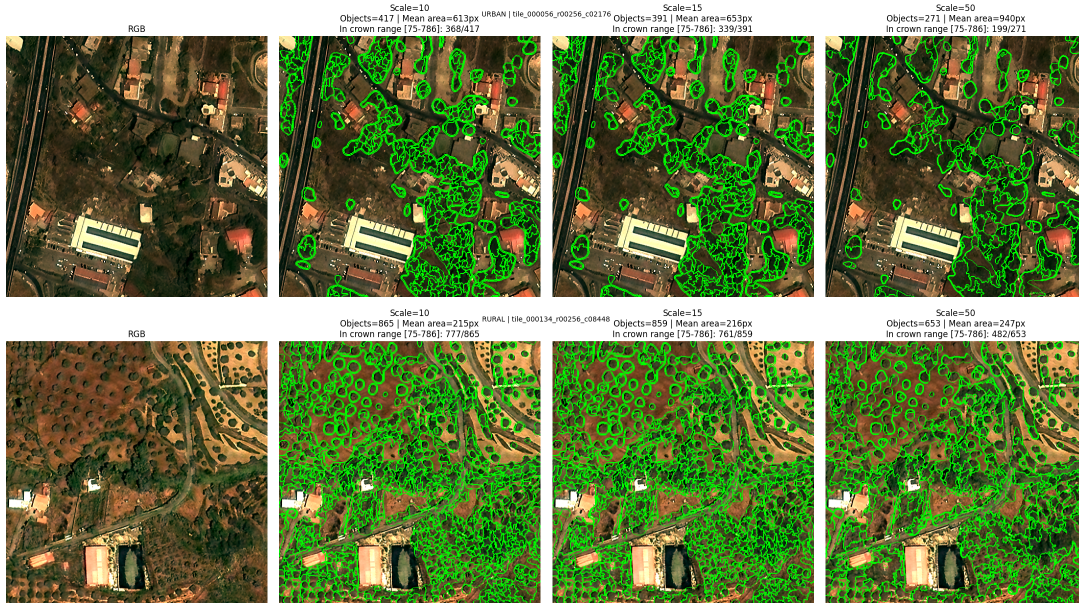


Figure 5.17: Felzenszwalb segmentation at scales 10, 15, and 50 on representative urban and rural tiles, with object count and percentage in the target size range.

5.4.3. Object Feature Extraction

The eleven-feature descriptor was computed for every retained object with no missing values across any feature. On the 3,407 objects of the six-tile sample, object area averaged 186.7px (median 150), mean NDVI 0.404, circularity 0.498, elongation 1.986, and the NDVI top-hat response 0.006. Extended to the full training set, the pipeline labelled 363,579 objects, of which 3,024 were tree and 360,555 non-tree, an overall imbalance of 1 : 119.2 (1 : 61.0 urban, 1 : 181.8 rural). The label assignment also revealed 32 training tiles that contained annotated crowns but produced no object meeting the $\text{IoU} \geq 0.25$ tree criterion, indicating cases where segmentation boundaries did not align with small or low-contrast crowns.

5.4.4. Random Forest Classification

SMOTE oversampling to a 1:10 target lifted the achieved training ratio to 1:6.1 (urban) and 1 : 18.2 (rural). On the SMOTE-balanced five-fold cross-validation the classifiers reached high apparent scores: urban precision 0.713, recall 0.975, F1 0.823; rural precision 0.715, recall 0.924, F1 0.806. In both strata the mean NDVI top-hat was the single most influential feature (importance ≈ 0.30 urban, 0.25 rural), followed by the geometric size descriptors (area, then elongation/circularity and perimeter); canopy fraction was effectively uninformative (≈ 0.001). Threshold tuning on held-out validation tiles placed

the best F1 at a probability threshold of 0.90 in both strata (0.759 urban, 0.790 rural).

5.4.5. Post-Processing and Output Format

Inference ran on all 348 test tiles (345 producing detections, 3 empty), generating 6,439 detections at a mean of 18.5 per tile. A train–test distribution shift appeared at inference: the maximum attainable classification probability on test tiles was only 0.84–0.89, below the validation-optimal threshold of 0.90, so the decision threshold had to be lowered to 0.50 to recover detections. After deduplication, 4,403 duplicates (68%) were removed, leaving 2,036 detections for evaluation: 134 true positives at IoU 0.25 against 1,902 false positives and 168 false negatives. OBIA achieved the highest recall among the handcrafted methods (0.444 overall), with comparable urban and rural F1 (Table 5.10) and cleaner object boundaries than the pixel classifier, while committing false detections on scrub patches.

Subset	P _{.25}	R _{.25}	F1 _{.25}	P _{.50}	R _{.50}	F1 _{.50}	mIoU
Urban	0.059	0.422	0.104	0.042	0.299	0.074	0.579
Rural	0.071	0.453	0.123	0.054	0.345	0.093	0.621
All	0.066	0.444	0.115	0.048	0.325	0.084	0.599

Table 5.10: OBIA detection performance by stratum at IoU 0.25 and 0.50.



Figure 5.18: Example OBIA detections on representative urban and rural test tiles.

5.5. Pixel-Based Random Forest

5.5.1. Label Rasterization and Training Target

Rasterizing the 474 size-filtered training crowns onto the tile grid produced the pixel-level supervision. The labelled set contained 801,583 crown pixels against 70,586,757 non-crown canopy pixels, an overall imbalance of 1:88.1. By stratum, the urban target held 370,327 crown pixels at a 1 : 51.9 ratio and the rural target 431,256 crown pixels at a 1 : 119.1 ratio, confirming that crown pixels are a small minority of canopy in both contexts and markedly scarcer in the rural stratum.

5.5.2. Pixel Feature Construction

Each canopy pixel was represented by the ten-feature vector `[nir, red, grn, blu, ndvi, ndvi_tophat, nir_contrast, nir_homogeneity, nir_entropy, nir_local_mean]`, all computed from the z-score normalized tiles. Feature construction completed with no errors over the 641 training tiles, with the only data-quality note being a small number of invalid casts on the NIR-derived proxies that were handled in the subsequent finiteness filter.

5.5.3. Sampling Strategy and Class Imbalance

Random subsampling of non-crown pixels reduced the training imbalance to 1:8.6 (urban) and 1:9.9 (rural) while retaining all crown pixels. Removing pixels with non-finite features dropped 67,076 urban pixels (1.89%) and none in the rural set. SMOTE oversampling of the minority crown class then brought the matrices close to balance, reaching 1:1.7 (urban, 1,822,000 synthetic-plus-real crown vs 3,120,658 non-crown) and 1:2.0 (rural, 2,156,280 vs 4,276,855).

5.5.4. Classifier Configuration

Five-fold cross-validation again returned high apparent scores with very high recall and lower precision: urban precision 0.664, recall 0.995, F1 0.796; rural precision 0.637, recall 0.992, F1 0.776. The NDVI top-hat was the most influential feature in both strata (importance 0.231 urban, 0.184 rural). In the urban model it was followed by the green band, local-mean NIR, and red band; in the rural model by NDVI, red, and green. The NIR GLCM contrast proxy was the least important feature in both cases.

5.5.5. Probability Map

Per-pixel crown probabilities reached up to ≈ 0.99 on test tiles. Threshold tuning on held-out validation tiles placed the best F1 at 0.85 for urban (F1 0.531) and 0.90 for rural (F1 0.403); a common deployment threshold of 0.85 was adopted, with morphological closing of 3×3 (urban) and 7×7 (rural). At this operating point a number of rural tiles produced no surviving component despite high local probabilities, because the closed positive regions did not reach the 86px minimum crown area.

5.5.6. Output format

Reconstructing connected components and applying the size filter yielded 4,816 detections across the 348 test tiles (344 with detections, 4 empty). Deduplication removed 3,827 duplicates (79%, the highest duplicate share of any method), leaving 989 detections: 82 true positives at IoU 0.25 against 907 false positives and 220 false negatives. The pixel classifier attained the highest F1 among the handcrafted methods at IoU 0.25 (0.127) but the lowest matched mean IoU (0.432), and its scores collapsed at the stricter IoU 0.50 (Table 5.11). Visually it produced the most jagged crown boundaries of all methods, consistent with the absence of any segmentation prior.

Subset	P. _{.25}	R. _{.25}	F1. _{.25}	P. _{.50}	R. _{.50}	F1. _{.50}	mIoU
Urban	0.097	0.214	0.134	0.035	0.078	0.049	0.440
Rural	0.076	0.331	0.123	0.022	0.095	0.035	0.427
All	0.083	0.272	0.127	0.026	0.086	0.040	0.432

Table 5.11: Pixel-based Random Forest detection performance by stratum at IoU 0.25 and 0.50.

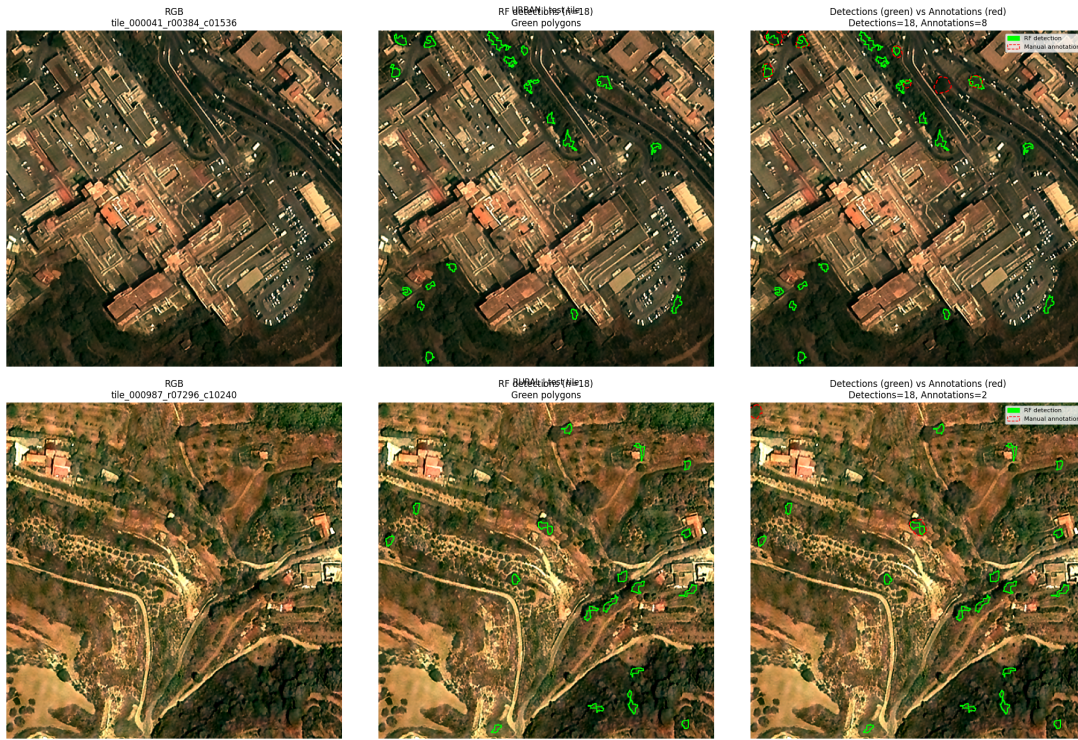


Figure 5.19: Example pixel-based Random Forest detections on representative urban and rural test tiles.

5.6. DeepForest

5.6.1. Input Preparation and Annotation Conversion

All 641 training and 348 test tiles were converted to 512×512 uint8 RGB images using the per-tile 2nd–98th percentile stretch, verified to span the full 0–255 range. Converting the crown polygons to bounding boxes produced 3,144 training boxes over 636 tiles, leaving 5 tiles without annotations. The boxes were small and approximately square, with mean width 18.7px and mean height 18.5px, a minimum side of 5px (the discard floor), and

a maximum box of 59×56 px, confirming that the supervision targets are crown-scale objects.

5.6.2. Fine-Tuning Strategy

The 634 annotated images were split 508/126 by image into training and validation subsets, carrying 2,558 and 586 boxes respectively. Fine-tuning the full RetinaNet/ResNet-50 detector (32.1 M parameters) for 15 epochs at batch size 4 and learning rate 0.001 on a Tesla T4 GPU converged to a final training loss of 0.1999 (classification 0.076, box-regression 0.124 at epoch level) and a validation loss of 0.3665. The gap between training and validation loss reflects the small fine-tuning set, but validation loss did not diverge, indicating the model adapted to the Mediterranean scene without catastrophic overfitting.

5.6.3. Post-Processing and Output Format

Because the configured score threshold was not applied internally, predictions were generated at 0.01 and filtered afterwards. On held-out tiles the rural detection threshold peaked in F1 around 0.50 (F1 0.725), but the deployed thresholds were set lower, at 0.20 (urban) and 0.10 (rural), to retain genuine low-confidence crowns produced by the compressed score distribution of a small-sample fine-tune. Inference over 348 test tiles left 29 tiles empty (20 rural, 9 urban) and produced 2,213 detections. Deduplication removed 1,544 duplicates (70%), leaving 669 detections: 152 true positives at IoU 0.25 against 517 false positives and 150 false negatives. DeepForest produced the highest detection scores of all five methods and the most threshold-stable F1 (from 0.313 at IoU 0.25 to 0.297 at IoU 0.50) together with the best matched mean IoU (0.666); its lowest precision occurred in the rural stratum, where boxes were committed on clustered non-target vegetation (Table 5.12).

Subset	P. _{.25}	R. _{.25}	F1. _{.25}	P. _{.50}	R. _{.50}	F1. _{.50}	mIoU
Urban	0.398	0.468	0.430	0.365	0.429	0.394	0.656
Rural	0.162	0.534	0.248	0.158	0.520	0.242	0.676
All	0.227	0.503	0.313	0.215	0.477	0.297	0.666

Table 5.12: DeepForest detection performance by stratum at IoU 0.25 and 0.50.



Figure 5.20: Example DeepForest detections on representative urban and rural test tiles.

5.7. U-Net

5.7.1. Architecture and Input Configuration

The U-Net with a ResNet-34 encoder was instantiated with a four-channel input stem and a single-class output, verified end-to-end (input $[2, 4, 512, 512]$ producing output $[2, 1, 512, 512]$) and totalling 24,439,505 trainable parameters. The per-band normalization statistics computed over canopy pixels of all 641 training tiles were NIR 0.1523 ± 0.0199 , Red 0.0632 ± 0.0070 , Green 0.0642 ± 0.0043 , and Blue 0.0591 ± 0.0032 , and were stored for one-way application to the test tiles.

5.7.2. Training Target and Data Split

The training tiles were divided 80/20 by tile identifier into 513 training and 128 validation tiles. Within each stratum the split gave 223 training and 50 validation tiles for urban and 290 training and 78 validation tiles for rural. A sample batch confirmed a crown pixel fraction of only $\approx 0.6\%$, restating at the model-input level the severe foreground scarcity that the loss function must counteract.

5.7.3. Data Augmentation

Augmentation was applied online to the training tiles only, comprising random horizontal and vertical flips, 90° rotations, and brightness/contrast jitter restricted to the RGB channels. No separate quantitative diagnostic is associated with this step; its effect is reflected in the training and validation loss trajectories reported below.

5.7.4. Loss Function and Optimization

Both stratum models were trained for 30 epochs with focal loss and Adam at learning rate 10^{-4} . The validation loss decreased monotonically and reached its minimum at the final epoch in both cases: the urban model converged to training/validation focal losses of 0.0013/0.0012 and the rural model to 0.0005/0.0008. The very low and closely tracking training and validation losses indicate stable optimization without overfitting, although such low focal-loss values are themselves a consequence of the dominant easy non-crown background.

5.7.5. Post-Processing and Output Format

Probability maps spanned up to 0.825 (urban) and 0.931 (rural). Threshold tuning on held-out tiles placed the best F1 at 0.40 in both strata (urban F1 0.562, rural F1 0.481), and 0.40 was deployed with 3×3 (urban) and 7×7 (rural) closing. Inference over 348 test tiles left 131 tiles empty (117 rural, 14 urban) and produced only 838 detections, the fewest of any method. Deduplication removed 607 duplicates (72%), leaving 231 detections: 71 true positives at IoU 0.25 against 160 false positives and 231 false negatives. U-Net achieved the highest overall precision (0.307) and the smoothest crown boundaries, but the lowest recall among the deep models, driven by a rural recall collapse to 0.122 (Table 5.13).

Subset	P. _{.25}	R. _{.25}	F1. _{.25}	P. _{.50}	R. _{.50}	F1. _{.50}	mIoU
Urban	0.302	0.338	0.319	0.163	0.182	0.172	0.530
Rural	0.305	0.122	0.174	0.203	0.081	0.116	0.548
All	0.307	0.235	0.266	0.173	0.132	0.150	0.532

Table 5.13: U-Net detection performance by stratum at IoU 0.25 and 0.50.

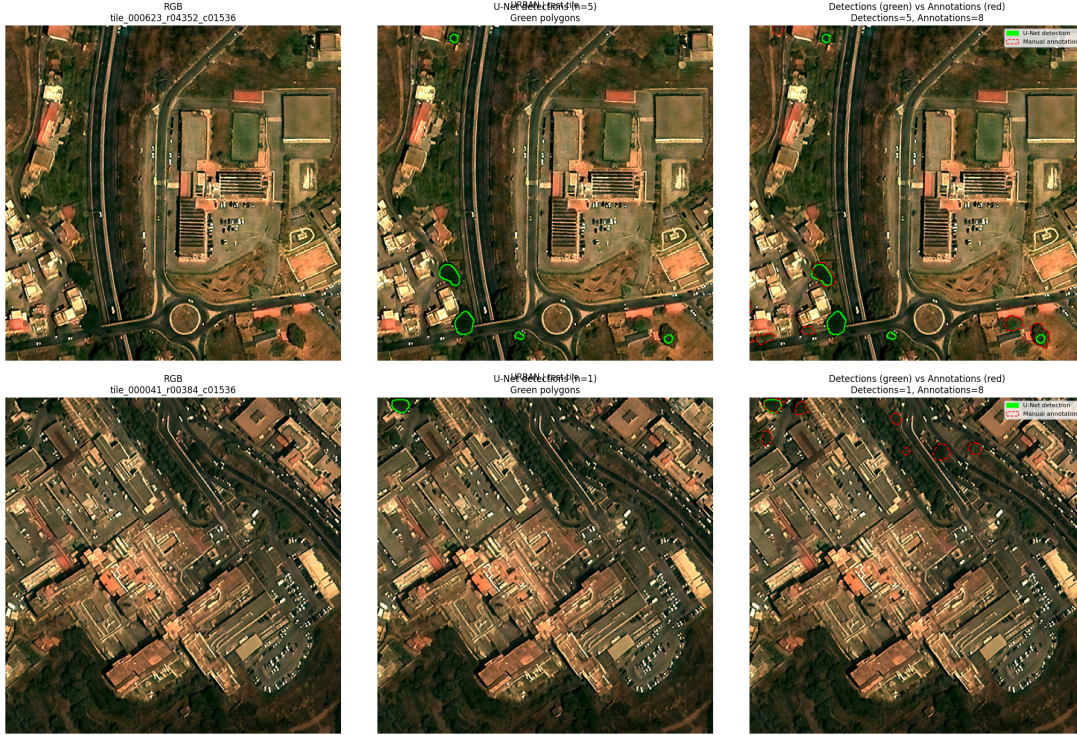


Figure 5.21: Example U-Net detections on representative urban and rural test tiles.

5.8. Cross-Method Comparison

5.8.1. Overall Performance

Table 5.14 consolidates precision, recall, F1-score, and matched mean IoU for all five methods at both IoU thresholds.

Method	P _{.25}	R _{.25}	F1 _{.25}	P _{.50}	R _{.50}	F1 _{.50}	mIoU
Watershed	0.064	0.260	0.103	0.040	0.162	0.064	0.557
OBIA + RF	0.066	0.444	0.115	0.048	0.325	0.084	0.599
RF Pixel	0.083	0.272	0.127	0.026	0.086	0.040	0.432
DeepForest	0.227	0.503	0.313	0.215	0.477	0.297	0.666
U-Net	0.307	0.235	0.266	0.173	0.132	0.150	0.532

Table 5.14: Overall performance of the five methods on the fixed 302-crown test set, at IoU thresholds 0.25 and 0.50. P = precision, R = recall, mIoU = mean IoU of matched true positives at IoU 0.25. Best value per column in bold.

DeepForest achieves the highest F1 at both thresholds (0.313 at IoU 0.25 and 0.297 at IoU

0.50) and the highest matched mean IoU (0.666). U-Net achieves the highest precision (0.307) but a lower recall (0.235). The three handcrafted methods cluster in a narrow F1 band of 0.103–0.127 at IoU 0.25.

5.8.2. Urban and Rural Performance

Table 5.15 disaggregates F1-score at IoU 0.25 by stratum.

Method	Urban F1 _{0.25}	Rural F1 _{0.25}
Watershed	0.149	0.059
OBIA + RF	0.104	0.123
RF Pixel	0.134	0.123
DeepForest	0.430	0.248
U-Net	0.319	0.174

Table 5.15: F1-score at IoU 0.25 disaggregated by landscape stratum. Best value per column in bold.

DeepForest leads in both strata, with a substantially higher urban F1 (0.430) than rural F1 (0.248). U-Net is the second-best method in both strata. The handcrafted methods remain below 0.150 in both strata.

5.8.3. Deduplication Behaviour

Because tiles overlap by 75%, each method produces many duplicate detections that the shared deduplication rule removes prior to scoring. Table 5.16 reports the raw and retained detection counts.

Method	Raw	Removed	Removed (%)
Watershed	4,999	3,755	75
OBIA + RF	6,439	4,403	68
RF Pixel	4,816	3,827	79
DeepForest	2,213	1,544	70
U-Net	838	607	72

Table 5.16: Raw detection counts and deduplication removals per method. Between 68% and 79% of raw detections are duplicates arising from the 75% tile overlap.

5.8.4. Detection Count Summary

Table 5.17 consolidates raw detection counts, deduplication removals, and object-level matching outcomes at IoU 0.25 for all five methods.

Method	Raw	Removed	Final	TP	FP	FN	Empty
Watershed	4,999	3,755	1,244	80	1,164	228	0
OBIA + RF	6,439	4,403	2,036	134	1,902	168	3
RF Pixel	4,816	3,827	989	82	907	220	4
DeepForest	2,213	1,544	669	152	517	150	29
U-Net	838	607	231	71	160	231	131

Table 5.17: Detection count summary for all five methods. Raw = detections before deduplication; Removed = duplicates eliminated by the shared IoU > 0.5 rule; Final = surviving detections entering evaluation; TP, FP, FN matched at IoU 0.25 against the 302-crown test set; Empty = test tiles producing no surviving detection.

5.8.5. Comparative Ranking

Ranked by overall F1 at IoU 0.25, the methods order as DeepForest (0.313), U-Net (0.266), RF Pixel (0.127), OBIA + RF (0.115), and Watershed (0.103). Ranked by matched mean IoU, the order is DeepForest (0.666), OBIA + RF (0.599), Watershed (0.557), U-Net (0.532), and RF Pixel (0.432); the two rankings do not coincide. These observations are interpreted in relation to the literature in Chapter 6.

6 | Discussion

This chapter interprets the results reported in Chapter 5 in relation to the state of the art reviewed in Chapter 2 and to the objectives stated in Chapter 1. It moves from the main per-method findings to the broader trade-offs they expose, the robustness of each paradigm across the urban–rural gradient, the practical implications for method selection, and the limitations that condition these conclusions. Throughout, each method is considered along six dimensions: detection accuracy, segmentation quality, interpretability, annotation needs, the computing setup it required, and urban/rural behaviour.

6.1. Interpretation of Main Findings

The central objective of the thesis was to determine which detection paradigms offer the best balance between accuracy, robustness, and interpretability for individual tree crown detection in VHR imagery. The benchmark addresses this question under a single, shared pipeline, so the observed differences can be attributed to the methods themselves. Two headline patterns emerge.

First, the deep learning methods clearly outperformed the handcrafted methods on detection accuracy, with DeepForest reaching the highest overall F1 (0.313 at IoU 0.25) and U-Net the highest precision (0.307), against a narrow handcrafted band of 0.103–0.127. Second, the F1 ranking and the segmentation-quality (matched mean IoU) ranking did not coincide, which indicates that detection success and boundary fidelity are partly independent properties and must be reported jointly, as argued in the evaluation literature [9, 36].

6.1.1. Preprocessing and Data Conditions

The preprocessing results frame every method’s behaviour. The Jeffries–Matusita separability analysis showed that only object geometry (area and perimeter) exceeded the usefulness threshold, while spectral and texture features did not, which anticipates the difficulty all spectrum-driven methods faced in distinguishing target crowns from surrounding

Mediterranean scrub and olive vegetation [2, 31]. The severe, stratum-dependent imbalance (up to 1:119 rural at the pixel level) and the higher urban shadow fraction (mean 4.71% urban versus 0.56% rural) align with the two dominant error modes observed downstream: false positives on non-target vegetation in rural tiles and shadow-related commission in urban tiles [7, 35]. These conditions justify the class-weighting and resampling strategies adopted by the supervised methods and explain why geometry-based size filtering was retained as a shared post-processing step.

6.1.2. Watershed and Local Maxima

In terms of *accuracy*, watershed was the weakest method (overall F1 0.103 at IoU 0.25), consistent with reports that local-maxima and marker-controlled watershed degrade where crowns are not radiometrically peaked and where canopy is continuous [27]. Its *segmentation quality* was nonetheless mid-range (mIoU 0.557), because the basins that were produced followed the NDVI surface reasonably well for isolated crowns.

Its main strength is *interpretability*: it is fully transparent and rule-based, with no learned parameters, so its failure modes are diagnostic of scene difficulty rather than of an opaque model. It has the lowest *annotation needs* of all methods, none, since it is unsupervised, and the simplest *computing setup*, running on CPU with no training phase, which makes it a useful reference where labels and compute are scarce [36]. Its *urban/rural behaviour* was strongly asymmetric (urban F1 0.149 versus rural 0.059): in the rural stratum, merged canopy blobs exceeding the maximum crown-size threshold were flagged and left unresolved, reproducing the under-segmentation failure documented for dense and flat-crowned canopies [10].

6.1.3. OBIA with Random Forest

OBIA improved *accuracy* over the baseline (F1 0.115) and achieved the highest recall among the handcrafted methods (0.444), supporting the literature argument that grouping pixels into objects and adding engineered features strengthens spatial reasoning [2]. Its *segmentation quality* was the best of the handcrafted methods (mIoU 0.599), reflecting that object boundaries are cleaner than pixel-level reconstructions.

Its *interpretability* is high: feature importances identified the NDVI top-hat and the geometric size descriptors as the dominant signals, in agreement with the separability analysis, although performance remained dependent on the segmentation scale. Its *annotation needs* are moderate, object labels obtained by intersecting segments with the reference crowns, and its *computing setup* runs on CPU (segmentation and Random Forest

training). The most informative result for OBIA was the large gap between its SMOTE-balanced cross-validation scores ($F1 \approx 0.81\text{--}0.82$) and its test performance, together with the lowered decision threshold (0.90 to 0.50): this exposes the central OBIA weakness of thresholds and segmentations that do not transfer across the train–test distribution shift [36]. Its *urban/rural behaviour* was the most balanced of all methods (urban $F1$ 0.104, rural 0.123).

6.1.4. Pixel-Based Random Forest

The pixel classifier obtained the best *accuracy* among the handcrafted methods at the lenient threshold ($F1$ 0.127) but its scores collapsed at IoU 0.50 ($F1$ 0.040). This is explained by its *segmentation quality*, the lowest of all methods (mIoU 0.432): without prior segmentation or spatial regularization, the reconstructed crowns have jagged boundaries that fall below the stricter overlap criterion, exactly the limitation expected of pixel-wise classification used for an instance-level task [4, 31].

Its *interpretability* is comparable to OBIA at the feature level, but it lacks object-level structure. Its *annotation needs* are effectively the same crowns rasterized to dense pixel masks, while its *computing setup* runs on CPU, with features extracted per pixel rather than per object. Its *urban/rural behaviour* was balanced (urban $F1$ 0.134, rural 0.123) but limited in both strata by boundary quality rather than by localization.

6.1.5. DeepForest

DeepForest delivered the best *accuracy* overall ($F1$ 0.313 at IoU 0.25 and 0.297 at IoU 0.50) and the highest recall (0.503), supporting the hypothesis that learned representations generalise better than handcrafted features and that a model pretrained on airborne RGB crowns can be transferred to a multispectral satellite scene [7, 32]. Its *segmentation quality* was the highest measured (mIoU 0.666), which is notable given that its output is a rectangular bounding box rather than a crown-shaped polygon: the boxes are coarse but consistently well centred, and their stability across the two IoU thresholds confirms good localization [34].

Its *interpretability* is low, as a deep detector, but this is partly offset by its reusability as a pretrained model. Its *annotation needs* are comparatively light, axis-aligned bounding boxes are cheaper to produce than masks, and transfer learning reduces the volume of labels required [7]. Its *computing setup* used GPU fine-tuning for 15 epochs on a single Tesla T4. Its *urban/rural behaviour* showed the largest stratum gap (urban $F1$ 0.430 versus rural 0.248); the model transferred most readily to the fragmented, well-separated street

trees of the urban core, while its lower rural precision reflects commission on clustered non-target vegetation, consistent with reports that detection models trained or validated in one landscape do not automatically generalise to the other [8, 18].

6.1.6. U-Net

U-Net achieved the highest *precision* (0.307) and an overall F1 of 0.266, but the lowest recall among the deep models (0.235), so its accuracy advantage is conditional on tolerating missed crowns. Its *segmentation quality* produced visually the smoothest boundaries, yet its matched mean IoU (0.532) was mid-range because many crowns were not detected at all rather than poorly delineated, echoing reports that semantic segmentation of crowns is sensitive to object size and imbalance [22, 26].

Its *interpretability* is low, as a dense encoder–decoder network. Its *annotation needs* are the highest of all methods, requiring per-pixel crown masks, and its *computing setup* comprises two separate models (urban and rural) trained for 30 epochs each on a Tesla T4. Its *urban/rural behaviour* was strongly asymmetric (urban F1 0.319 versus rural 0.174), with a rural recall collapse to 0.122; the combination of very low focal-loss training values and the extreme rural imbalance pushed the network toward the background class in continuous canopy, a known sensitivity of segmentation under imbalance and shadow [22, 35].

Table 6.1 summarises the dominant and secondary error modes observed across the five methods.

Method	Scrub FP	Built/shadow	Rural misses	Merged	Jagged bounds.	Low mIoU
Watershed	○	—	●	●	—	○
OBIA + RF	●	○	○	○	—	—
RF Pixel	●	○	○	—	●	●
DeepForest	●	○	○	—	—	○
U-Net	○	●	●	○	—	—

Table 6.1: Observed error modes per method. ● = primary error mode; ○ = secondary or partially observed; — = not observed. Scrub FP refers to false positives on scrub or non-target vegetation; built/shadow refers to errors associated with built surfaces or shadow; rural misses refers to undetected crowns in the rural stratum; jagged bounds. refers to fragmented or irregular crown boundaries.

6.2. Methodological Trade-offs

The benchmark makes the trade-offs among accuracy, segmentation quality, interpretability, annotation cost, and computing requirements explicit rather than assumed. The rule-based watershed baseline is maximally interpretable and label-free but least accurate; OBIA recovers accuracy and boundary quality while remaining interpretable, at the price of sensitivity to segmentation scale and non-transferable thresholds; the pixel-based Random Forest trades boundary quality for a marginally higher lenient-threshold F1; and the deep learning methods provide the strongest accuracy at the cost of GPU training, larger annotation requirements, and low interpretability.

The divergence between the F1 and matched-IoU rankings is itself a methodological finding: the method with the best handcrafted F1 (pixel Random Forest) had the worst boundaries, whereas a box-based detector (DeepForest) led on both detection and matched overlap. This confirms that no single scalar summarises method quality and that detection and delineation must be evaluated together, as the evaluation literature recommends [9, 36]. The annotation axis is equally consequential: moving from no labels (watershed) to boxes (DeepForest) to dense masks (U-Net) steadily increases both predictive ceiling and labelling effort, which is the practical cost rarely reported alongside accuracy in single-method studies.

6.3. Urban–Rural Robustness and Generalization

A central objective was to assess stability across the urban–rural gradient. Four of the five methods scored higher in the urban stratum than the rural stratum, and the two deep models showed the largest gaps (DeepForest 0.430 versus 0.248; U-Net 0.319 versus 0.174). The rural stratum was therefore the harder setting across paradigms, which is attributable to its continuous and semi-continuous canopy, the spectral similarity between target crowns and surrounding scrub and olive vegetation, and the merged blobs that exceed plausible single-crown sizes [7].

These are precisely the conditions that the preprocessing separability analysis flagged in advance. Only the two handcrafted classifiers (OBIA and pixel Random Forest) were approximately balanced across strata, but they were balanced at a low overall level rather than through genuine robustness[8]. The asymmetry observed for the deep models is consistent with reports that models perform best on the fragmented, well-separated trees typical of urban scenes and degrade on dense vegetation, and that performance does not transfer automatically between landscape types [18]. For operational monitoring across

heterogeneous Mediterranean landscapes, this implies that a single model and a single decision threshold are unlikely to be optimal for both strata, supporting the per-stratum calibration used in this work.

6.4. Practical Implications

The results translate into method selection guidance conditioned on the user's constraints rather than on accuracy alone. Where annotations and computing resources are scarce and transparency is required, the watershed baseline provides an immediately deployable, label-free reference, accepting low accuracy and poor performance on continuous canopy. Where interpretable, moderately accurate mapping is needed and a limited set of crown labels is available, OBIA with Random Forest offers the best handcrafted balance of recall and boundary quality on CPU hardware. Where the priority is detection accuracy and tree counting, and bounding-box labels plus a GPU are available, DeepForest is the strongest option and benefits from transfer learning, particularly in urban environments. Where precise per-crown delineation is required and dense mask annotations and substantial compute can be afforded, U-Net delivers high precision and smooth boundaries but must be paired with imbalance-aware training and, ideally, per-stratum models to avoid the rural recall collapse observed here.

The broader implication for precision forestry and urban vegetation monitoring is that method choice should be driven jointly by the target product (counts versus delineation), the available annotation type, the landscape context, and the available computing hardware.

6.5. Limitations

Several limitations condition the interpretation of these conclusions:

1. The reference data were produced through manual visual interpretation of a single scene. Annotation subjectivity and decisions concerning the boundary between separable and merged crowns therefore affect every metric. The evaluation represents crowns visible in the imagery rather than a complete field inventory.
2. The benchmark is based on a single JILIN-1 KF01C acquisition over Reggio Calabria on one date, limiting the generalisability of the results to other sensors, seasons, locations, and ecological settings.
3. The supervised methods required threshold recalibration on the test tiles because

of the documented train–test distribution shift. Their reported operating points are therefore specific to this scene rather than universally applicable.

4. The methods produce different native output types, including points, classified objects, pixel masks, and bounding boxes. Although these outputs were converted into a common crown-polygon representation and evaluated using a shared matching protocol, comparisons between the box-based detector and the segmentation methods remain intrinsically approximate [36].

These limitations should be considered alongside the reported results and motivate the research directions presented in the next chapter.

7 | Conclusions and Future Work

Published studies on individual tree crown detection rarely apply representative methods to the same imagery under identical experimental conditions. As a result, observed performance differences between approaches are difficult to interpret: they may reflect genuine algorithmic capability, or they may simply reflect differences in sensor resolution, annotation protocol, landscape context, or evaluation criterion [36]. This thesis was designed to address that limitation by constructing a controlled benchmark in which five methods, each representing a distinct paradigm in the ITCD literature, were applied to the same JILIN-1 KF01C VHR scene over Reggio Calabria and evaluated against a fixed test set of 302 manually delineated crowns using a unified Hungarian-assignment IoU protocol.

The five methods were selected to span the full methodological spectrum. Local maxima with watershed segmentation is a training-free geometric baseline that identifies candidate treetops as NDVI peaks and delineates crown regions using the watershed transform [30]; it was included as an interpretable reference point for rule-based heuristic approaches.

Object-based image analysis (OBIA) combined with Random Forest first segments the canopy into spatially coherent objects before classifying them using spectral, textural, and geometric features [5]; it represents the object-based paradigm that emerged in response to the limitations of pixel-level analysis. Pixel-based Random Forest classifies each pixel independently from engineered spectral and textural features [4], providing a direct comparison between pixel-level and object-level machine learning under the same classifier family. DeepForest is a pretrained RetinaNet one-stage detector fine-tuned on local crown annotations [32]; it was selected because it represents the application of domain-specific deep detection to ecological imagery. U-Net is an encoder–decoder segmentation network with skip connections trained to produce dense pixel-wise crown masks [24]; it was included to test whether learned boundary-aware segmentation recovers crown extent more accurately than detection-based approaches.

All five methods shared a common preprocessing pipeline, a fixed 512×512 px tiling grid, a spatial block train–test partition, a common crown-polygon output format, and a uniform evaluation protocol, ensuring that measured performance differences reflect the

methods rather than incidental experimental variation.

The main findings can be summarised as follows:

- **DeepForest achieved the best overall detection performance.** It obtained the highest F1-score at IoU 0.25 (0.313), the highest F1-score at IoU 0.50 (0.297), and the highest recall (0.503).
- **DeepForest also performed best in both landscape strata.** Its F1-score reached 0.430 in urban areas and 0.248 in rural areas, with U-Net ranking second in both strata.
- **The pixel-based Random Forest was the strongest handcrafted method in terms of detection.** It reached an F1-score of 0.127 at IoU 0.25, although its performance declined sharply at IoU 0.50 because of its irregular crown boundaries.
- **Rural crown detection was consistently more difficult.** Four of the five methods performed better in urban areas, and the two deep-learning methods showed the largest differences between the urban and rural strata.
- **No single decision threshold was equally suitable for both strata.** Differences in canopy density, vegetation structure, shadows, and spectral background required landscape-specific calibration.
- **Output quality depended on both detection and representation.** DeepForest obtained the highest matched mean IoU (0.666), although its detections were rectangular boxes. Among the methods producing crown-shaped segments, OBIA achieved the highest mIoU (0.599), while U-Net produced the visually smoothest boundaries.

The results confirm that learning-based methods outperform spectral and rule-based approaches on this dataset, but their advantage depends on sufficient training data, calibrated class balance, and landscape-adapted thresholds. The main obstacles across all methods were the spectral similarity between target crowns and surrounding Mediterranean vegetation, merged rural canopy, and building-cast shadows in urban areas. These findings indicate that high-resolution imagery alone does not guarantee reliable individual crown separation when neighbouring crowns have weak spectral or spatial boundaries.

Table 7.1 synthesises these findings, reporting the main strength, main limitation, and central conclusion for each of the five methods as evaluated under the shared framework.

Method	Main strength	Main limitation	Main conclusion
Local maxima and watershed	Transparent, training-free, and computationally light	Local maxima over-detect internal canopy texture; large merged rural blobs remain unresolved by local contrast alone	Provides an interpretable rule-based baseline whose failure modes are structurally informative about scene difficulty
OBIA with Random Forest	Interpretable object features and strongest crown boundary quality among the segmentation-based methods	Performance limited by the distribution shift between training and test objects, requiring threshold recalibration at inference	Object-level features improve boundary coherence over pixel classification, but spectral ambiguity between crowns and background vegetation constrains detection
Pixel-based Random Forest	Best detection F1 among the hand-crafted methods at the lenient IoU threshold	Fragmented masks and jagged boundaries cause a sharp performance drop at the stricter IoU threshold	Pixel-level spectral classification captures relevant vegetation patterns but is insufficient for accurate crown delineation without spatial regularization
DeepForest	Highest detection F1, recall, and matched mean IoU across both strata	Rectangular bounding boxes approximate crown extent but do not describe true crown shape; false detections increase in dense rural canopy	Pretrained domain-specific weights provide the strongest transfer to the Mediterranean scene, confirming the advantage of learned representations over hand-engineered features
U-Net	Smooth, crown-shaped segmentation masks and highest precision among all methods	Rural recall collapsed under severe class imbalance and limited rural training examples	Learned dense segmentation produces coherent boundaries but requires sufficient stratum-specific training data to generalise across structurally different vegetation types

Table 7.1: Summary of the main strength, main limitation, and main conclusion for each of the five evaluated tree crown detection methods.

The reproducible end-to-end workflow, shared preprocessing, consistent spatial block partitioning, and object-level matching through Hungarian assignment, provides a reference design for future benchmarks in precision forestry and urban vegetation monitoring under VHR satellite imagery.

Taken together, the benchmark demonstrates that detection accuracy and boundary fidelity are partly independent properties that must be reported jointly, and that a single model and threshold do not transfer optimally across heterogeneous urban and rural Mediterranean landscapes. DeepForest is preferable when crown localisation and detection rate are the main priorities, whereas OBIA or U-Net may be more suitable when crown-shaped boundaries are required. Rule-based and classical machine-learning approaches remain valuable when interpretability, limited training data, or low computational resources constrain method selection.

7.1. Recommendations for Future Research

Future research should first extend the benchmark to additional VHR sensors, acquisition dates, and geographical areas. A multi-sensor and multi-site evaluation would determine whether the method ranking observed for the JILIN-1 KF01C scene remains stable under different spatial resolutions, spectral configurations, seasons, and Mediterranean landscape conditions.

A larger and more balanced reference dataset would improve both model training and evaluation. Additional rural annotations are particularly important because the limited number of rural crowns and the dominance of background vegetation constrained the supervised methods. Repeated annotation by multiple interpreters could also quantify annotation uncertainty and improve the treatment of partially visible or merged crowns.

Multi-temporal imagery offers another important direction. Seasonal changes in vegetation condition may help distinguish target trees from spectrally similar scrub, olive vegetation, and other background classes. Repeated acquisitions would also extend the workflow from single-date crown detection to tree monitoring and change assessment.

Future models should also combine accurate localisation with crown-shaped instance boundaries. Instance-segmentation architectures such as Mask R-CNN, as well as attention- or transformer-based approaches, could preserve the detection strength of DeepForest while producing more realistic crown outlines. Landscape-aware models or separate urban and rural calibration strategies could further address the strong differences observed between the two strata.

Finally, field validation is needed to determine how well crowns identified in the imagery correspond to individual trees on the ground. Comparing image-derived detections with field measurements would clarify the difference between visible crown annotations and ecological ground truth, allowing the resulting products to support more reliable urban forestry, biomass assessment, and rural vegetation monitoring.

Bibliography

- [1] M. Abbasi and H. R. Riyahi Bakhtyari. Extraction of forest stands parameters from satellite data. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2012. metadata status: partial; evidence family pack(s): Other classical method; evidence pack source PDF: ab-basi_extraction_2012.pdf.
- [2] R. Al-Ruzouq, A. Shanableh, M. B. A. Gibril, and S. Al-Mansoori. Image segmentation parameter selection and ant colony optimization for date palm tree detection and mapping from very-high-spatial-resolution aerial imagery. *Remote Sensing*, 2018. doi: 10.3390/rs10091413. metadata status: verified; evidence family pack(s): OBIA plus Random Forest; Other classical method; Pixel-based Random Forest; evidence pack source PDF: al-ruzouq_image_2018.pdf.
- [3] E. Banzhaf and H. Kollai. Monitoring the urban tree cover for urban ecosystem services: The case of leipzig, germany. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2015. doi: 10.5194/isprsarchive-s-XL-7-W3-301-2015. metadata status: verified; evidence family pack(s): OBIA plus Random Forest; evidence pack source PDF: banzhaf_monitoring_2015.pdf.
- [4] M. Belgiu and L. Drăguț. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114: 24–31, 2016. doi: 10.1016/j.isprsjprs.2016.01.011.
- [5] T. Blaschke. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(1):2–16, 2010. doi: 10.1016/j.isprsjprs.2009.06.004.
- [6] J. B. Campbell and R. H. Wynne. *Introduction to Remote Sensing*. Guilford Press, New York, NY, 5 edition, 2011.
- [7] V. Dakov and D. Petrova-Antonova. Urban tree detection from remote sensing data based on deepforest model. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2024. doi: 10.5194/isprs-annals-X-4-W4-2024-35-2

024. metadata status: verified; evidence family pack(s): DeepForest; Other classical method; Other deep learning; evidence pack source PDF: dakov_urban_2024.pdf.
- [8] N. Dimarco, B. Bartiaux, L. Andreani, and R. Schlogel. Multi-model approach for tree detection and classification in wallonia region (belgium). In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2025. doi: 10.5194/isprs-archives-XLVIII-M-7-2025-259-2025. metadata status: verified; evidence family pack(s): DeepForest; Other deep learning; evidence pack source PDF: dimarco_multi-model_2025.pdf.
- [9] M. Freudenberg, P. Magdon, and N. Nolke. Individual tree crown delineation in high-resolution remote sensing images based on u-net. *Neural Computing and Applications*, 2022. doi: 10.1007/s00521-022-07640-4. metadata status: verified; evidence family pack(s): Local maxima plus watershed; Other classical method; Other deep learning; evidence pack source PDF: freudenberg_individual_2022.pdf.
- [10] M. K. Heenkenda, K. E. Joyce, and S. W. Maier. Mangrove tree crown delineation from high-resolution imagery. *Photogrammetric Engineering and Remote Sensing*, 2015. doi: 10.14358/PERS.81.6.471. metadata status: verified; evidence family pack(s): Local maxima plus watershed; OBIA plus Random Forest; Other classical method; evidence pack source PDF: heenkenda_mangrove_2015.pdf.
- [11] S. Idbraim, D. Mammass, L. Bouzalim, M. Oudra, and M. Arbelo. Palm trees detection from high spatial resolution satellite imagery using a new contextual classification method with constraints. *Image and Signal Processing*, 2016. metadata status: partial; evidence family pack(s): Unclear; evidence pack source PDF: idbraim_palm_2016.pdf.
- [12] Y. Jiang et al. Detection of individual trees in high-resolution remote sensing imagery using deep learning detectors, 2025. metadata status: provisional; evidence family pack(s): DeepForest; Other deep learning; U-Net; evidence pack source PDF: jiang_detection_2025.pdf.
- [13] D. Jin, J. Qi, N. B. Goncalves, J. Wei, H. Huang, and Y. Pan. Automated tree crown labeling with 3d radiative transfer modelling achieves human comparable performances for tree segmentation in semi-arid landscapes. *International Journal of Applied Earth Observation and Geoinformation*, 2024. metadata status: verified; evidence family pack(s): Local maxima plus watershed; Other deep learning; U-Net; evidence pack source PDF: jin_automated_2024.pdf.
- [14] S. R. Kadu, B. G. Hogade, and I. Rizvi. Detection of urban tree canopy from very

- high resolution imagery using an object based classification. *International Journal of Electrical and Computer Engineering*, 2022. metadata status: verified; evidence family pack(s): OBIA plus Random Forest; Other classical method; Other deep learning; evidence pack source PDF: kadu_detection_2022.pdf.
- [15] M. Kamal. Geographic object based image analysis (geobia) for mangrove tree crown delineation using worldview-2 image data, 2015. metadata status: partial; evidence family pack(s): Local maxima plus watershed; OBIA plus Random Forest; Pixel-based Random Forest; evidence pack source PDF: kamal_geographic_2015.pdf.
- [16] Y. Ke and L. J. Quackenbush. A review of methods for detecting individual tree crowns in remotely sensed imagery. *International Journal of Remote Sensing*, 32(21):6005–6020, 2011. doi: 10.1080/01431161.2010.515442.
- [17] J. A. C. Martins, K. Nogueira, L. P. Osco, F. D. G. Gomes, D. E. G. Furuya, W. N. Goncalves, D. A. Sant’Ana, A. P. M. Ramos, V. Liesenberg, J. A. dos Santos, P. T. S. de Oliveira, and J. Marcato Junior. Semantic segmentation of tree-canopy in urban environment with pixel-wise deep learning. *Remote Sensing*, 2021. doi: 10.3390/rs13163054. metadata status: verified; evidence family pack(s): DeepForest; Other deep learning; U-Net; evidence pack source PDF: martins_semantic_2021.pdf.
- [18] E. Pantaleo, V. Giannico, R. Cilli, G. A. Vivaldi, S. Camposeo, A. Monaco, R. Bellotti, N. Amoroso, et al. Automated olive grove classification and tree counting in very high resolution aerial imagery using deep learning. *Smart Agricultural Technology*, 2025. metadata status: compatibility_alias_for_pantaleo_automated_2025.
- [19] P. Patias and P. Stournara. Estimating wood volume for pinus brutia trees in forest stands from quickbird-2 imagery. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016. doi: 10.5194/isprsarchives-XLI-B7-329-2016. metadata status: verified; evidence family pack(s): Local maxima plus watershed; OBIA plus Random Forest; Pixel-based Random Forest; evidence pack source PDF: patias_estimating_2016.pdf.
- [20] S. C. Popescu and R. H. Wynne. Seeing the trees in the forest: Using lidar and multispectral data fusion with local filtering and variable window size for estimating tree height. *Photogrammetric Engineering and Remote Sensing*, 70(5):589–604, 2004.
- [21] L. Qiu, L. Jing, B. Hu, H. Li, and Y. Tang. A new individual tree crown delineation method for high resolution multispectral imagery. *Remote Sensing*, 2020. doi: 10.3390/rs12030585. metadata status: verified; evidence family pack(s): Local maxima plus

- watershed; OBIA plus Random Forest; Other deep learning; evidence pack source PDF: qiu_new_2020.pdf.
- [22] A. U. Rahman, E. Heinaro, M. Ahishali, and S. Junttila. Dual-task learning for dead tree detection and segmentation with hybrid self-attention u-nets in aerial imagery, 2025. metadata status: partial; evidence family pack(s): Local maxima plus watershed; Other deep learning; U-Net; evidence pack source PDF: rahman_dual-task_2025.pdf.
- [23] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99, 2015.
- [24] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.
- [25] Satellite Imaging Corporation. Jilin-1 optical satellite, 2025. URL <https://satimagingcorp.com/satellite-sensor/jilin-1/jilin-1-optical-satellite>. Accessed: 2026-06-22.
- [26] J. Schindler et al. Efficient tree mapping through deep distance transform learning, 2025. metadata status: provisional; evidence family pack(s): Local maxima plus watershed; Other deep learning; U-Net; evidence pack source PDF: schindler_efficient_2025.pdf.
- [27] Y. Shen, E. Yan, et al. Individual tree location detection by high-resolution rgb satellite remote sensing images. In *ACM International Conference Proceedings*, 2020. doi: 10.1145/3398329.3398351. metadata status: partial; evidence family pack(s): Local maxima plus watershed; Other classical method; evidence pack source PDF: shen_individual_2020.pdf.
- [28] S. N. Topgul, E. Sertel, S. Aksoy, C. Unsalan, and J. E. S. Fransson. Vhrtrees: A new benchmark dataset for tree detection in satellite imagery and performance evaluation with yolo-based models. *Frontiers in Forests and Global Change*, 2025. doi: 10.3389/ffgc.2024.1495544. metadata status: verified_key_year_mismatch; evidence family pack(s): DeepForest; OBIA plus Random Forest; Other deep learning; evidence pack source PDF: topgul_vhrtrees_2024.pdf.
- [29] I. Ulku, E. Akagunduz, et al. Deep semantic segmentation of trees using multi-spectral images, 2022. metadata status: provisional; evidence family pack(s): Local

- maxima plus watershed; Other deep learning; U-Net; evidence pack source PDF: ulku_deep_2022.pdf.
- [30] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991. doi: 10.1109/34.87344.
- [31] F. H. Wagner, M. P. Ferreira, A. Sanchez, M. C. M. Hirye, M. Zortea, E. Gloor, O. L. Phillips, C. R. de Souza Filho, Y. E. Shimabukuro, and L. E. O. C. Aragao. Individual tree crown delineation in a highly diverse tropical forest using very high resolution satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018. doi: 10.1016/j.isprsjprs.2018.09.013. metadata status: verified; evidence family pack(s): Other classical method; Pixel-based Random Forest; evidence pack source PDF: wagner_individual_2018.pdf.
- [32] B. G. Weinstein, S. Marconi, M. Aubry-Kientz, G. Vincent, H. Senyondo, and E. P. White. Deepforest: A python package for rgb deep learning tree crown delineation. *Methods in Ecology and Evolution*, 11(11):1732–1738, 2020. doi: 10.1111/2041-210X.13472.
- [33] S. Wickramarathna, J. Van Den Hoek, and B. M. Strimbu. Automated detection of individual juniper tree location and forest cover changes using google earth engine. In *Annals of Forest Research*, 2021. doi: 10.15287/afr.2020.2145. metadata status: partial; evidence family pack(s): Local maxima plus watershed; OBIA plus Random Forest; Other classical method; evidence pack source PDF: wickramarathna_automated_2021.pdf.
- [34] P. Zamboni, J. Marcato Junior, J. de Andrade Silva, G. T. Miyoshi, E. T. Matsubara, K. Nogueira, and W. N. Goncalves. Benchmarking anchor-based and anchor-free state-of-the-art deep learning methods for individual tree detection in rgb high-resolution images. *Remote Sensing*, 2021. doi: 10.3390/rs13132482. metadata status: verified; evidence family pack(s): DeepForest; Local maxima plus watershed; Other deep learning; evidence pack source PDF: zamboni_benchmarking_2021.pdf.
- [35] Y. Zhang et al. Individual tree detection and counting based on high-resolution remote sensing images. *Geo-spatial Information Science*, 2024. metadata status: provisional; evidence family pack(s): Local maxima plus watershed; Other classical method; Other deep learning; evidence pack source PDF: zhang_individual_2024.pdf.
- [36] H. Zhao, J. Morgenroth, G. Pearse, and J. Schindler. A systematic review of individ-

- ual tree crown detection and delineation with convolutional neural networks. *Current Forestry Reports*, 2023. doi: 10.1007/s40725-023-00184-3. metadata status: verified; evidence family pack(s): Local maxima plus watershed; OBIA plus Random Forest; Other deep learning; evidence pack source PDF: zhao_systematic_2023.pdf.
- [37] J. Zheng, H. Fu, W. Li, W. Wu, Y. Zhao, R. Dong, and L. Yu. Cross-regional oil palm tree counting and detection via a multi-level attention domain adaptation network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020. metadata status: verified; evidence family pack(s): Local maxima plus watershed; Other classical method; Other deep learning; evidence pack source PDF: zheng_cross-regional_2020.pdf.
- [38] Y. Zheng and G. Wu. YOLOv4-lite-based urban plantation tree detection and positioning with high-resolution remote sensing imagery. *Frontiers in Environmental Science*, 2022. doi: 10.3389/fenvs.2021.756227. metadata status: verified; evidence family pack(s): Local maxima plus watershed; Other classical method; Other deep learning; evidence pack source PDF: zheng_yolov4-litebased_2022.pdf.
- [39] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017.

List of Figures

3.1	Overview of the study area. Panel A shows the full JILIN-1 KF01C acquisition footprint with the Reggio Calabria municipal boundary and the AOI subset extent (red rectangle). Panel B shows the AOI subset in true-colour composite (R:G:B = bands 2-3-4). Panel C places Reggio Calabria within its regional context in Calabria. Panel D shows the AOI subset draped over the municipal boundary.	15
3.2	JILIN-1 Optical Satellite / KF01C sensor	16
4.1	Research Design Pipeline	19
4.2	Annotation protocol/quality-control	23
4.3	Overview of the local-maxima and watershed segmentation pipeline.	29
4.4	Object-Based Image Analysis with Random Forest workflow.	33
4.5	Random Forest workflow.	37
4.6	Probability Map	39
4.7	Deep Forest Pipeline Overview	41
4.8	U-Net processing pipeline	44
5.1	Spatial distribution of the 800 manually annotated crown polygons (yellow points) overlaid on the AOI true-colour composite. Rural sampling zones are shown in dark blue ($n = 400$) and urban zones in light blue ($n = 400$). Crown annotations are distributed across the full spatial extent of the AOI, covering both the dense urban fabric in the west and the more continuous vegetated areas to the north and east.	52
5.2	Spatial overview of the four spectral bands of the JILIN-1 KF01C image over the AOI. From left to right: Band 1 (NIR, mean DN = 1462), Band 2 (Red, mean = 713), Band 3 (Green, mean = 687), Band 4 (Blue, mean = 624). The higher relative brightness of vegetated areas in Band 1 is visible in the hillside zones to the north and east.	53

- 5.3 Raw DN distributions by land cover class (rural in blue, urban in orange) for all four spectral bands. Urban pixels show broader distributions and heavier upper tails relative to rural pixels, particularly in Bands 1–3. . . . 54
- 5.4 Spatial distribution of anomaly density by tile (left) and tiles flagged above the 5 % anomaly coverage threshold (right). No tiles are flagged (all shown in blue on the right panel). Anomaly density is effectively zero across the entire scene. 55
- 5.5 Shadow coverage distribution by dominant land cover class (urban in blue, rural in green). The shadow flag threshold of 20 % is shown as a black dashed vertical line. No tile exceeds this threshold. Urban tiles show a substantially broader and heavier-tailed distribution, with mean shadow coverage of 4.71 % compared to 0.56 % for rural tiles. 56
- 5.6 NDVI map computed from scaled `float32` bands over the AOI (mean = 0.344, std = 0.077). Higher values (darker green) correspond to dense vegetated cover on the hillside zones; lower values (yellow-green) correspond to the urban core and the dry riverbed visible in the centre of the scene. 57
- 5.7 NDVI histogram computed from scaled `float32` bands. The distribution is unimodal and concentrated in the range [0.2, 0.5], with mean = 0.344 and std = 0.077. There are no values below -0.25. 57
- 5.8 NDVI threshold sensitivity on urban tile `tile_000214_r01408_c03200`. Vegetation coverage increases from 6.73 % at threshold 0.40 to 61.40 % at 0.25. At 0.30 and 0.25, large non-vegetated areas such as roads and rooftops are included. The threshold of 0.35 retains 18.57 % coverage. 58
- 5.9 NDVI threshold sensitivity on peri-urban tile `tile_000332_r02432_c02560`. At threshold 0.35, 60.42 % of the tile is classified as vegetated. 58
- 5.10 NDVI threshold sensitivity on rural tile `tile_000502_r03584_c04608`. Coverage at 0.35 reaches 51.41 %. At 0.25, 94.87 % of the tile is flagged. . . . 58
- 5.11 Object area (top) and perimeter (bottom) distributions for tree (green) and non-tree (red) segments in urban (left) and rural (right) training tiles, shown on a logarithmic scale. Vertical dashed lines indicate the minimum area threshold (black) and maximum area threshold (blue). Tree objects are concentrated between approximately 75 and 800 pixels in area, while non-tree objects span a substantially wider range in both directions. . . . 62
- 5.12 Sample of 3 rural training tiles with retained vegetation blobs (red markers) overlaid on the true-colour composite. Each tile caption reports the filtered non-tree to tree ratio and the blob count. Ratios range from 1:27 to 1:120. 63

5.13	Sample of 3 urban training tiles with retained vegetation blobs (red markers) overlaid on the true-colour composite. Filtered ratios range from 1:28 to 1:143. The highest-imbalance tiles correspond to dense urban fabric with minimal tree cover.	64
5.14	NDVI peak retention under Gaussian smoothing ($\sigma = 1.5, 2.0$, and 3.0 px).	65
5.15	Mean local-maxima counts per tile by stratum and minimum-distance setting.	66
5.16	Example watershed detections on representative urban and rural test tiles.	69
5.17	Felzenszwalb segmentation at scales 10, 15, and 50 on representative urban and rural tiles, with object count and percentage in the target size range. .	70
5.18	Example OBIA detections on representative urban and rural test tiles. . .	72
5.19	Example pixel-based Random Forest detections on representative urban and rural test tiles.	74
5.20	Example DeepForest detections on representative urban and rural test tiles.	76
5.21	Example U-Net detections on representative urban and rural test tiles. . .	78

List of Tables

2.1	Reported OBIA performance across selected ITCD studies. OA = overall accuracy. Dual values in Al-Ruzouq et al. [2] correspond to two study areas. No reference in the provided bibliography reports a standardized F1 or IoU range specifically for OBIA plus Random Forest in ITCD under object-level matching.	8
2.2	Reported U-Net performance across selected ITCD and canopy segmentation studies. No reference in the provided bibliography covers U-Net performance on JILIN-1 KF01C RGB+NIR imagery or reports focal loss or weighted cross-entropy specifically for U-Net in ITCD.	11
2.3	Summary of output types and evaluation approaches across ITCD method families. The lack of a shared matching protocol across studies motivates the unified Hungarian-assignment IoU evaluation framework used in this thesis.	12
3.1	JILIN-1 KF01C acquisition and raster parameters for the Reggio Calabria scene.	17
3.2	Summary of the reference datasets.	18
4.1	Mapping of preprocessing outputs to method-level input requirements. A checkmark indicates that the output is a direct input or dependency for that method.	21
4.2	Software tools and libraries used in the thesis and their role in the workflow.	22
4.3	Implementation parameters for local-maxima detection and watershed segmentation.	32
4.4	Object-based image analysis (OBIA) with Random Forest implementation parameters.	36
4.5	Pixel-based Random Forest implementation parameters.	40
4.6	DeepForest implementation parameters. The committed model is a RetinaNet detector with a ResNet-50 backbone, fine-tuned without backbone freezing.	43

4.7	U-Net implementation parameters. Separate models are trained per stratum. The inference threshold of 0.40 is set below the conventional default to account for the conservative probability outputs produced by focal loss under severe class imbalance.	48
5.1	Summary of the reference crown dataset used in the benchmark.	51
5.2	Per-band raw DN statistics stratified by land cover class. p01 and p99 denote the 1st and 99th percentiles respectively.	54
5.3	Mean, minimum, and maximum NDVI mask coverage across three representative tiles at four candidate thresholds. The selected operating threshold is 0.35.	59
5.4	Tile and crown distribution across training and test subsets after spatial block partitioning. One crown falls outside all tile extents and is excluded. The rural/urban proportions are consistent across splits.	59
5.5	Per-band normalization statistics computed from 641 training tiles ($n = 168,034,286$ valid pixels per band). Raw values are in scaled <code>float32</code> units $[0, 1]$. Normalized means are effectively zero and standard deviations are exactly 1.0.	60
5.6	Top-10 feature ranking by mean Jeffries–Matusita distance between tree and non-tree classes across urban and rural training tiles. Features with JM distance ≥ 1.0 in at least one stratum are marked as useful. Object area and perimeter are the only features exceeding the 1.0 separability threshold.	61
5.7	Class imbalance summary by land cover stratum. Raw ratio refers to all pixels; filtered ratio refers to annotated crown pixels vs background. The filtered imbalance ranges from 1:42 to 1:55.	63
5.8	Summary of preprocessing parameters applied uniformly across all five detection methods. Parameters are derived from training data only and held fixed throughout all experiments.	64
5.9	Watershed detection performance by stratum at IoU 0.25 and 0.50. mIoU is the mean IoU of matched true positives at IoU 0.25.	68
5.10	OBIA detection performance by stratum at IoU 0.25 and 0.50.	71
5.11	Pixel-based Random Forest detection performance by stratum at IoU 0.25 and 0.50.	74
5.12	DeepForest detection performance by stratum at IoU 0.25 and 0.50.	75
5.13	U-Net detection performance by stratum at IoU 0.25 and 0.50.	77

5.14	Overall performance of the five methods on the fixed 302-crown test set, at IoU thresholds 0.25 and 0.50. P = precision, R = recall, mIoU = mean IoU of matched true positives at IoU 0.25. Best value per column in bold. .	78
5.15	F1-score at IoU 0.25 disaggregated by landscape stratum. Best value per column in bold.	79
5.16	Raw detection counts and deduplication removals per method. Between 68% and 79% of raw detections are duplicates arising from the 75% tile overlap.	79
5.17	Detection count summary for all five methods. Raw = detections before deduplication; Removed = duplicates eliminated by the shared IoU > 0.5 rule; Final = surviving detections entering evaluation; TP, FP, FN matched at IoU 0.25 against the 302-crown test set; Empty = test tiles producing no surviving detection.	80
6.1	Observed error modes per method. ● = primary error mode; ○ = secondary or partially observed; — = not observed. Scrub FP refers to false positives on scrub or non-target vegetation; built/shadow refers to errors associated with built surfaces or shadow; rural misses refers to undetected crowns in the rural stratum; jagged bounds. refers to fragmented or irregular crown boundaries.	84
7.1	Summary of the main strength, main limitation, and main conclusion for each of the five evaluated tree crown detection methods.	91

