

UNIVERSIDADE PAULISTA

SAMIRA NASCIMENTO ANTUNES

HERBAI: A MULTI-STAGE DEEP LEARNING SYSTEM FOR AROMATIC HERB
RECOGNITION - FROM DATA-CENTRIC DETECTION TO MOBILE DEPLOYMENT

SÃO PAULO

2026

SAMIRA NASCIMENTO ANTUNES

HERBAI: A MULTI-STAGE DEEP LEARNING SYSTEM FOR AROMATIC HERB
RECOGNITION - FROM DATA-CENTRIC DETECTION TO MOBILE DEPLOYMENT

Thesis presented to the Graduate Program in
Production Engineering at Universidade
Paulista – UNIP, in partial fulfillment of the
requirements for the degree of Doctor of
Production Engineering.

Concentration Area: Operation Systems
Management.

Research Line: Business Networks and
Production Planning.

Research Project: Production Management and
Innovation in Agriculture, Industry, and
Services.

Advisor: Prof. Dr. Marcelo Tsuguio Okano.

SÃO PAULO

2026

Antunes, Samira Nascimento.

HerbAI: a multi-stage deep learning system for aromatic herb recognition - from data-centric detection to mobile deployment / Samira Nascimento Antunes. - 2026.

95 f.: il. color.

Tese de Doutorado Apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Paulista, São Paulo, 2026.

Área de concentração: Gestão de Sistemas de Operação.

Orientador: Prof. Dr. Marcelo Tsuguio Okano.

1. Computer vision. 2. YOLO. 3. EfficientNet. 4. Faster R-CNN. 5. SSD. 6. Aromatic herbs. 7. Open-set recognition. 8. Mobile deployment. 9. Artificial intelligence. 10. Plant identification. I. Okano, Marcelo Tsuguio (orientador). II. Título.

SAMIRA NASCIMENTO ANTUNES

HERBAI: A MULTI-STAGE DEEP LEARNING SYSTEM FOR AROMATIC HERB
RECOGNITION - FROM DATA-CENTRIC DETECTION TO MOBILE DEPLOYMENT

Thesis presented to the Graduate Program in
Production Engineering at Universidade
Paulista – UNIP, in partial fulfillment of the
requirements for the degree of Doctor of
Production Engineering.

Approved On: ____/____/____

DOCTORAL COMMITTEE

Prof. Dr. Marcelo Tsuguio Okano
Universidade Paulista – UNIP

Profa. Dra. Irenilza de Alencar Nääs
Universidade Paulista - UNIP

Prof. Dr. Wonsuk Daniel Lee
University of Florida

Prof. Dr. Steven Sargent
University of Florida

Prof. Dr. Rodrigo Franco Gonçalves
Universidade Paulista – UNIP

ACKNOWLEDGMENTS

Upon the completion of this doctoral thesis, I wish to express my deepest gratitude to everyone who, directly or indirectly, contributed to the results achieved.

I am deeply grateful to my advisor, Prof. Dr. Marcelo Tsuguio Okano, for the trust he placed in my work and for the patience, competence, dedication, and composure he demonstrated throughout this entire process. I am especially thankful for his guidance at every stage, from the conception of the central idea to the completion of this thesis, and for his continuous support throughout this research journey.

I would also like to express my sincere gratitude to Prof. Dr. Wonsuk Daniel Lee, to whom I owe a debt of gratitude beyond words. His guidance during my research at the University of Florida was fundamental to my academic and personal growth.

My sincere thanks also go to Prof. Dr. Irenilza de Alencar Nääs for the enriching discussions and exchanges of ideas that significantly contributed to the development of this thesis.

Finally, I thank the faculty and staff of Universidade Paulista (UNIP), as well as my colleagues in the program, for their collaboration, which greatly enhanced the quality of this work. I also gratefully acknowledge the support of the Support Program for Postgraduate Studies at Private Higher Education Institutions (PROSUP), partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), under Finance Code 001 and process number 88887.663535/2022-00.

ABSTRACT

This research presents the development and validation of HerbAI, a multi-stage deep learning system for the automated identification of aromatic herbs under real-world commercial conditions. The work is structured as a thesis by articles, integrating three studies into a single offline mobile recognition pipeline. The first stage employed YOLOv8 to establish an initial object detection baseline for three herb species, revealing limitations related to dataset size and over-detection. The second stage evaluated EfficientNet with transfer learning for the classification of visually similar herbs, complementing the detection approach. The third stage adopted a data-centric strategy, expanding the dataset to six visually separable classes through image curation, quality filtering, deduplication, and data augmentation, and comparatively evaluated YOLOv11s, Faster R-CNN, and SSD300/VGG based on mean average precision (mAP). YOLOv11s achieved the best overall performance, with mAP@0.5 of 0.802 and mAP@0.50:0.95 of 0.593 on the held-out test set. These findings were consolidated into HerbAIApp, a native Android application that integrates the YOLOv11s detector with an EfficientNet-B0 fallback classifier to handle herb groups outside the main detection scope, operating entirely offline. The application also incorporates multilingual support and optional text-to-speech output, extending its accessibility to individuals with anosmia and reinforcing the inclusion of non-specialist users in plant sensory recognition. Beyond its technical contributions to computer vision in agriculture, the study demonstrates the feasibility of moving from experimental model development to a deployable, field-ready solution for food distribution environments such as CEAGESP.

Key words: computer vision, YOLO, EfficientNet, Faster R-CNN, SSD, aromatic herbs, open-set recognition, mobile deployment, artificial intelligence, plant identification.

UTILITY - SOCIAL, ECONOMIC, AND ENVIRONMENTAL VALUE OF THE THESIS

This research aims to develop a solution that facilitates the identification of aromatic herbs through a convolutional neural network-based pipeline and a mobile application. By providing an accessible, efficient, and field-deployable tool, this thesis contributes to the broader adoption of automated herb identification in practical settings, supporting multiple sectors, including agriculture, food distribution, gastronomy, phytotherapy, and education.

Furthermore, the application of artificial intelligence in this context promotes technological innovation and offers a scalable, data-centric approach capable of assisting professionals, supply chain operators, and enthusiasts in recognizing visually similar herb species under uncontrolled conditions. By reducing classification errors and improving robustness against non-target and out-of-distribution inputs, HerbAI supports more reliable and sustainable practices in herb cultivation, distribution, and commercialization.

These contributions are aligned with global objectives outlined in the United Nations 2030 Agenda. By supporting accurate identification along agricultural and food distribution chains, this thesis contributes to Zero Hunger and Sustainable Agriculture (SDG 2) and to Good Health and Well-Being (SDG 3), given the relevance of correct herb identification for food safety and phytotherapeutic use. By improving operational efficiency and reducing classification-related losses for wholesalers, retailers, and producers, this work also relates to Decent Work and Economic Growth (SDG 8). The development of a deep learning-based, data-centric, and mobile-deployable identification system constitutes a direct contribution to Industry, Innovation, and Infrastructure (SDG 9), while the reduction of misclassification and product loss throughout the commercialization process supports Responsible Consumption and Production (SDG 12). Finally, by making herb recognition more accessible to non-specialist users, the proposed mobile application also has the potential to support Quality Education (SDG 4) through its applicability in vocational and botanical training contexts.

LIST OF FIGURES

Figure 1 HerbAIApp initial interface with multilingual menu.....	45
Figure 2 Examples of YOLOv11s detector outputs in the mobile application	46
Figure 3 Examples of EfficientNet-B0 fallback classifier outputs	47
Figure 4 HerbAIApp interface with optional text-to-speech enabled	48

SUMMARY

1. INTRODUCTION	11
1.1 Background.....	11
1.2 Problem Statement and Research Gap.....	12
1.3 Research Questions	12
1.4 Objectives	13
1.4.1 General Objective	13
1.4.2 Specific Objectives	13
1.5 Justification.....	14
1.6 Thesis Structure	15
 2. LITERATURE REVIEW	 16
2.1 Deep Learning for Aromatic Herb Recognition	16
2.2 Object Detection and Data-Centric Approaches in Agricultural Vision	18
2.3 Open-Set Recognition and Out-of-Domain Detection	21
2.4 Cascaded Classification and Detection Frameworks.....	25
2.5 Mobile and Offline Plant Recognition	27
 3. METHODOLOGY	 30
3.1 Methodology of Paper I - Model Development for Identifying Aromatic Herbs Using Object Detection Algorithm	30
3.2 Methodology of Paper II - Enhanced Classification of Aromatic Herbs Using EfficientNet and Transfer Learning	32
3.3 Methodology of Paper III - Data-Centric Object Detection for Aromatic Herbs: Image Quality Filtering, Class Selection, and Model Comparison	34
3.4 Methodology of the Mobile Deployment and Integrated Inference Pipeline	37
 4. RESULTS AND DISCUSSION.....	 39
4.1 Results of Paper I.....	39
4.2 Results of Paper II	40
4.3 Results of Paper III	42
4.4 Results of the Mobile Deployment and Integrated Inference Pipeline.....	44

5. FINAL CONSIDERATIONS 50

5.1 Conclusions 50

5.2 Recommendations for Future Work 51

REFERENCES 52

APPENDIX A – Paper I 57

APPENDIX B – Paper II 70

APPENDIX C – Paper III 84

1. INTRODUCTION

1.1 Background

Artificial intelligence has become increasingly central to plant and crop identification, yet fine-grained recognition of aromatic and medicinal herbs remains a particularly demanding case: species are often visually close, per-species labeled data is scarce, and the computational cost of deep architectures restricts deployment on the lightweight hardware typically available in the field [1]. Lightweight convolutional designs such as IPMCNet have addressed parts of this problem by cutting parameter counts while retaining accuracy on dozens of plant species [2], but the gap between curated, plain-background datasets and real field imagery remains substantial: accuracy drops of several percentage points have been documented once images move from public archives to uncontrolled commercial or field backgrounds [3].

A recognition system trained only on known classes, however, is not enough once it is exposed to spontaneous, real-world input. Models built under the closed-world assumption tend to force any unfamiliar image into one of the trained categories, and this has been shown to measurably degrade precision once a detector is exposed to images captured outside controlled conditions, whether irrelevant photographs submitted through a public chatbot [4] or visually related but untrained categories in agricultural classification [5]. This motivates treating the rejection of out-of-domain input as a necessary design element rather than an afterthought.

A second design trend relevant to this context is the move away from single end-to-end classifiers toward cascaded pipelines, in which an upstream detection or filtering stage narrows the search space before a dedicated classifier operates on it. This decomposition has been shown to reduce background noise and improve downstream precision across domains as different as medical image diagnosis, where a YOLOv8-based region detector preceding a Vision Transformer classifier produced systematic gains over feeding the full image directly to the classifier [6], and industrial waste sorting, where adding a classification stage after a one-stage detector raised mAP without retraining the detector itself [7].

Finally, translating any such pipeline into practical use requires moving inference onto the device itself. Recent plant-identification applications built on lightweight backbones such as EfficientNet Lite, converted to TensorFlow Lite or ONNX Runtime, have demonstrated fully offline operation in environments where network access cannot be guaranteed [8,9]. This

requirement is not a generic convenience in the present case: CEAGESP's facility directly borders the Centro de Detenção Provisória de Pinheiros, placing it within a radiocommunication signal-blocking area authorized under Resolução Anatel nº 760/2023 [10], which makes any cloud-dependent inference architecture unworkable regardless of network coverage elsewhere.

Taken together, these four threads - data-scarce fine-grained recognition, open-set rejection, cascaded detection-classification design, and offline mobile inference - define the space in which this thesis operates. Three prior studies began addressing pieces of this space individually for aromatic herbs specifically [11], [12], [13], and Section 1.2 frames the gap that remains once they are considered together.

1.2 Problem Statement and Research Gap

The three studies referenced above addressed this problem incrementally. Antunes et al. [11] established a baseline object detection model for three herb species using YOLOv8, surfacing limitations in recall and over-detection that pointed to dataset size and quality, rather than architecture, as the binding constraint. Antunes et al. [12] investigated image classification of the same species using EfficientNet and transfer learning. Antunes et al. [13] revisited the detection problem from a data-centric perspective, expanding the dataset to six classes and showing that quality filtering and visual-separability-based class selection produced larger gains than switching architectures.

None of these three studies, individually, incorporates a mechanism for rejecting images that do not correspond to a known herb class, combines detection and classification into a single cascaded pipeline, or has been validated as a deployable, fully offline mobile application, despite the operational necessity of offline inference at CEAGESP described in Section 1.1. The research gap addressed in this thesis is the absence of an integrated, multi-stage pipeline that brings together data-centric object detection, herb classification, and open-set awareness, validated for offline mobile deployment.

1.3 Research Questions

To what extent can a multi-stage deep learning system integrating data-centric object detection, herb classification, and open-set rejection reliably identify aromatic herbs

commercialized at CEAGESP using representative image datasets while operating entirely offline on a mobile device?

Secondary research questions:

- Among the detection architectures evaluated across this thesis [11], [13], which provides the most favorable trade-off between accuracy and inference latency for on-device deployment?
- Does combining a detection stage with a dedicated classification stage [12] improve recognition accuracy over either approach used in isolation?
- Can the resulting pipeline be deployed as a fully offline Android application without unacceptable loss of accuracy relative to its server-trained counterparts?

1.4 Objectives

1.4.1 General Objective

To develop and validate a multi-stage deep learning system for aromatic herb recognition, integrating data-centric object detection and image classification techniques into an offline mobile application designed for food distribution environments such as CEAGESP.

1.4.2 Specific Objectives

- To design and evaluate an object detection model for identifying aromatic herbs from images collected under commercial conditions [11];
- To evaluate EfficientNet-based image classification with transfer learning for distinguishing visually similar aromatic herbs [12];
- To investigate the impact of data curation, image quality filtering, and class selection on detection performance across multiple architectures [13];
- To consolidate the findings of these studies into HerbAI, a multi-stage recognition system, and assess its feasibility for offline mobile deployment.

1.5 Justification

The development of reliable automated systems for aromatic herb recognition is justified by both scientific and practical challenges. From a scientific perspective, aromatic herbs represent a fine-grained recognition problem in which visually similar species, high intra-class variability, limited labeled datasets, and uncontrolled image acquisition conditions make conventional plant identification approaches insufficient. In commercial environments such as CEAGESP, herbs are frequently presented in bunches, crates, or overlapping arrangements, under variable lighting and background conditions. These factors increase the complexity of detection and classification, requiring methods that go beyond controlled image datasets and account for the variability expected in food distribution contexts.

This research is also justified by the need to bridge the gap between model development and operational deployment. Although deep learning models have achieved strong results in agricultural image analysis, many approaches remain restricted to curated datasets, cloud-based inference, or closed-world classification settings. In practical use, however, a recognition system must not only identify known herb classes, but also handle uncertain, visually similar, or out-of-domain inputs without forcing them into incorrect categories. Therefore, the integration of object detection, classification, and open-set awareness is necessary to improve reliability under real use conditions.

The CEAGESP context further reinforces the relevance of an offline mobile solution. As a large food distribution environment, it brings together different actors in the supply chain, including wholesalers, retailers, workers, consumers, and professionals who may benefit from accessible plant identification support. At the same time, the operational environment may impose connectivity limitations, making cloud-dependent solutions less suitable. A mobile application capable of performing inference directly on the device can therefore increase accessibility, reduce dependence on network infrastructure, and support the practical use of artificial intelligence in agricultural and food distribution systems.

In the context of a thesis by articles, this study is justified by the need to consolidate and advance the contributions developed across the three papers. The first study established an initial object detection approach for aromatic herb identification; the second investigated image classification using transfer learning; and the third expanded the detection problem through a data-centric strategy involving dataset curation, class selection, and model comparison. However, these contributions remain incomplete if considered only as isolated experiments. This thesis brings them together into an integrated multi-stage deep learning

system, connecting data-centric detection, herb classification, out-of-domain handling, and offline mobile deployment.

Thus, the relevance of this research lies in its contribution to computer vision applied to agriculture, its practical orientation toward a real wholesale market environment, and its potential to support more accessible, robust, and field-deployable herb recognition. By addressing both methodological and operational challenges, the thesis contributes to the advancement of artificial intelligence solutions that are not only accurate in experimental settings but also designed to be feasible for use in production and distribution contexts.

1.6 Thesis Structure

This thesis is organized into five chapters. Chapter 1 introduces the research context, problem statement, research gaps, research questions, objectives, justification, and overall structure of the thesis. Chapter 2 presents the literature review, covering deep learning for aromatic herb recognition, object detection and data-centric approaches in agricultural vision, open-set recognition, cascaded classification and detection strategies, and mobile offline plant recognition. Chapter 3 describes the methodological procedures adopted in the three studies that compose this thesis, as well as the methodology used to integrate their findings into the proposed multi-stage system. Chapter 4 presents and discusses the results obtained in each study and analyzes the integrated HerbAI system in relation to the research questions. Finally, Chapter 5 provides the final considerations, summarizes the main conclusions, discusses the contributions of the thesis, and outlines recommendations for future work.

2. LITERATURE REVIEW

Given the growing role of artificial intelligence in agricultural and food-distribution systems, and the persistent challenge of distinguishing visually similar aromatic herbs under real, uncontrolled commercial conditions, deep learning-based recognition has emerged as a viable and strategic alternative to manual, expert-dependent identification. This chapter reviews the literature that supports this perspective, bringing together evidence on deep learning architectures for fine-grained herb recognition, object detection and data-centric approaches in agricultural vision, the challenges of open-set and out-of-domain recognition, the benefits of cascaded detection-classification frameworks, and the requirements for mobile and offline deployment.

2.1 Deep Learning for Aromatic Herb Recognition

The automatic recognition of aromatic herbs from digital images is a particularly demanding instance of fine-grained plant classification. Unlike common crops, aromatic and medicinal herbs are often confused with one another because their distinguishing traits - leaf venation, surface texture, glandular trichomes, overall silhouette - vary subtly between genera and are further obscured by growth stage, lighting, and cluttered natural backgrounds. Manual identification still relies heavily on trained botanists or experienced growers, a process that is slow, costly, and difficult to scale to the large number of species found in cultivation or in the wild [14].

Convolutional neural networks have progressively replaced handcrafted color, shape, and texture descriptors as the dominant strategy for this task, learning discriminative representations directly from raw images. Even so, classical pipelines combining handcrafted descriptors with conventional classifiers remain competitive specifically for medicinal and aromatic leaves, a pattern documented across more than twenty studies in a recent review reporting accuracies between 94% and 99% for SVM, random forest, and MLP classifiers built on color, texture, and shape features alone [14]. Three obstacles recur throughout this literature and are directly relevant to aromatic herb classification: (i) the limited number of labeled images available per species; (ii) the visual similarity among closely related species; and (iii) the computational cost of deep architectures, which restricts deployment on smartphones or embedded devices used for in-field identification.

Obstacle (i) is tackled both at the architecture level and at the level of data acquisition itself. Azadnia et al. replaced the customary flatten-and-fully-connected classifier head with a global average pooling layer feeding a single dense layer, lowering the risk of overfitting on a dataset of only 750 leaf images spanning five medicinal herbs and reaching up to 99.66% accuracy [15]. Where photographing more individual leaves is impractical, the training set can instead be assembled directly from the target growing environment: Tariku et al. built an orthomosaic from drone imagery over a one-hectare plot and fine-tuned EfficientNetV2 to 99.2% accuracy on seven species, while ResNet50 trained on the identical split collapsed to 60.2%, a gap that already anticipates obstacle (ii) [16].

Obstacle (ii), species-level visual similarity, is most often addressed through feature fusion or attention. Sharma et al. combined handcrafted texture and shape descriptors with deep features from VGG-16 and DenseNet-121, reducing the fused representation with UMAP before classification and reaching 99.18% accuracy on 30 Indian medicinal species [17]. Oppong et al. moved fusion inside the network itself, embedding a Log-Gabor filter bank into the transition layers of a DenseNet201 backbone; the resulting model generalized across four public leaf benchmarks (97%–100%) in addition to 98% on a self-collected set of 49 Ghanaian medicinal species [18]. Gao et al. addressed the same problem through a dual-branch channel-attention module embedded in ResNet, whose class activation maps confirmed that attention shifted the network's focus away from background clutter toward diagnostically relevant leaf regions [19].

The gap between curated training images and genuinely unseen field images is itself part of the problem being contextualized here. Islam et al. observed this directly: their fine-tuned DenseNet201 reached 96.30% validation accuracy on augmented data yet fell to 80.69% on held-out original test images, below the 90.09% reached by InceptionResNetV2, which had scored lower during validation [20]. Picek et al. examined the same generalization challenge at a much larger scale, benchmarking convolutional and transformer backbones on thousands of species under realistic field conditions, and showed that reframing classification as image retrieval in a learned embedding space outperformed equivalent softmax classifiers by up to 10 percentage points [21].

Obstacle (iii), computational cost, is addressed less by shrinking the network than by relocating where computation happens. Thanikkal et al. replaced raw images with a lightweight, rotation-invariant numeric shape descriptor, enabling a fully offline Android application that reached 96% accuracy on ten Ayurvedic species without network connectivity

[22]. Kavitha et al. instead moved computation off the device entirely, fine-tuning MobileNet on six medicinal species and serving inference from a cloud backend reached by a thin mobile client, reaching 98.3% accuracy [23].

Together, these studies frame aromatic herb recognition as a problem shaped less by any single architectural choice than by how scarce, similar, and field-variable the underlying data is: lightweight or fused architectures help with scarcity and similarity, but the validation-to-field gap shown by Islam et al. and Picek et al. indicates that accuracy on curated or augmented data should not be assumed to transfer directly to deployment, a methodological gap that any aromatic herb classifier built along these lines needs to account for.

2.2 Object Detection and Data-Centric Approaches in Agricultural Vision

Object detection performance in agricultural and plant-related imagery is often assumed to be primarily a function of network architecture. However, a growing body of evidence suggests that detection accuracy is equally, if not more, constrained by the intrinsic properties of the training data itself: how separable the target classes are from one another, how consistent samples within the same class remain across conditions, and how the dataset was curated, annotated, and represented in the first place.

A recurring observation in recent detection literature is that domain-specific imagery, including geospatial and agricultural scenes, tends to exhibit greater intra-class diversity and weaker inter-class separability than the generic benchmarks on which most detection backbones were originally validated [24]. This gap becomes particularly evident in low-data or few-shot regimes, which are common in agricultural settings where certain cultivars, growth stages, or pest occurrences are inherently rare or expensive to annotate; simply transferring a detector pretrained on abundant base classes to a small set of novel classes leads to degraded performance unless the training strategy explicitly accounts for class confusion. The same weak separability appears at a much smaller scale in pest monitoring on yellow sticky traps, where whitefly, *macrolophus*, and *nesidiocoris* individuals are small, low-resolution, and frequently share the yellow coloration of the trap itself. In that setting, the single architectural change that produced the largest accuracy gain was not an increase in raw network capacity but a module added specifically to recover long-range contextual information, so that pests could be told apart from a background of nearly identical color, which again points to the relationship between class and background, rather than the detector's

depth, as the limiting factor [25]. A related form of weak inter-class separability appears in orchard pest monitoring, where two invasive stink bug species, *Halyomorpha halys* and *Nezara viridula*, share similar size, shape, and coloration. Rather than asking a single architecture to resolve this ambiguity, one study trained five detectors of different families and combined their predictions through a decision fusion scheme that weights each model by its own per-class F1 score, so that the final decision leans on whichever detector best separates each species rather than forcing one network to carry the whole burden. This data-driven combination raised the ensemble F1 score to 0.902, exceeding any individual detector and most smaller subsets of the ensemble, again suggesting that the limiting factor was the relationship between the two confusable classes rather than the capacity of any single network [26].

This separability problem is not new to object detection alone; it has long been documented at the classification level. Even on a controlled benchmark such as CIFAR100, the visual similarity between certain categories makes them substantially harder to distinguish than unrelated pairs, regardless of which convolutional backbone is used, since the standard cross-entropy/softmax formulation does not explicitly penalize confusion between visually correlated classes [27]. By explicitly modeling the correlation structure between confusable categories and feeding it back into the loss function, that work shows consistent accuracy gains across multiple network families, reinforcing that the bottleneck is frequently the relationship between classes in the data, not the depth or capacity of the model.

Generalization across deployment conditions offers a further illustration of how much of the burden can be shifted away from the architecture and onto the data. In an anchor-free center-based detector built specifically for pig detection in crowded, heavily overlapping pens, the architecture itself already outperformed prior anchor-based and keypoint-based methods, but the largest remaining gain came from two purely data-side interventions: a negative training augmentation that exposes the network to out-of-distribution background imagery, and a test-time augmentation scheme that averages predictions over rotated copies of each frame. On a held-out set of pens never seen during training, these two interventions alone raised the F-measure from 80.5% to 87.6%, a gain comparable to, or larger than, the one obtained by introducing the new detector architecture in the first place, while the network weights and structure stayed exactly the same [28].

Dataset curation itself plays an equally decisive role, and this extends beyond simply collecting more images. In insect phenomics, citizen-science platforms already hold millions

of images but almost none of them carry bounding boxes; rather than treating this as an architecture problem, one study coupled an off-the-shelf, unmodified vision-language detector with a single generic text prompt to automatically localize insects across six million images spanning 2526 agriculturally relevant species, reaching agreement with human-drawn boxes comparable to detectors trained on millions of manually labeled samples, which shows that the curation method, not the detector, was what unlocked an already existing but unusable corpus [29]. The need for careful curation becomes even more pronounced when the target classes are visually close to begin with. In a smartphone-based scale insect detection system built for Taiwan's export agriculture, mealybugs, Coccidae, and Diaspididae are easily confused even by experienced observers, so the bounding boxes for the training set were drawn exclusively by annotators with more than two decades of field experience, on the understanding that unreliable labeling of such similar pests would undermine any detector regardless of its architecture. Even with this carefully curated dataset, accuracy still varied by class: Diaspididae, the category with the greatest internal morphological variation, was detected less reliably than the visually consistent mealybugs across Faster R-CNN, SSD, and YOLOv4 alike, again pointing to within-class diversity, rather than the detector, as the factor setting the ceiling on performance [30].

Curation problems are not limited to dataset size, balance, or annotator expertise; they can also stem from inconsistencies in the labeling process itself, which is easy to mistake for a modeling failure. In a strawberry receptacle detection study designed to predict fruit yield, a Faster R-CNN detector trained on manually annotated images reached a mean average precision of only 0.6587, an unusually low figure for that architecture. Closer inspection showed that part of this apparent gap came from receptacles that the human annotator had missed in the original labels but that the network still detected correctly, meaning that the measured error rate partly reflected the quality of the ground truth rather than a genuine limitation of the detector [31].

Perhaps the most direct demonstration that data quality can outweigh model choice comes from a coal-versus-gangue recognition study, in which the same downstream classifier and the same feature set were tested on two different image acquisition strategies: conventional natural images and laser speckle images of the same mineral samples [32]. Recognition accuracy rose from 78.7% to 96.7%, and the standard deviation across varying illuminance conditions dropped by more than 30%, purely because the way the data was captured and represented, with no modification to the classifier itself. The same pattern holds

for plant pathology detection: when the very same Faster R-CNN and single-shot detector architectures were trained on rice leaf disease images, simply replacing standard geometric and color-based augmentation with a curated set of GAN-generated synthetic images, filtered for blur before being added to the training pool, raised mean average precision from 0.84 to 0.93 for Faster R-CNN and from 0.81 to 0.91 for the single-shot detector, with the improvement confirmed as statistically significant across training and validation losses [33]. In both cases the network code never changed; only the data feeding it did.

Taken together, these findings, spanning image classification, livestock monitoring, mineral sorting, insect dataset curation, orchard pest monitoring, scale pest identification, fruit yield prediction, and plant disease detection, point to a common conclusion that is directly transferable to agricultural vision: detection and recognition accuracy is frequently bounded less by the sophistication of the underlying network than by how separable the classes are within the dataset and how carefully that dataset has been curated, balanced, annotated, and represented. This motivates the data-centric perspective adopted in this work, where improvements to model architecture are treated as complementary to, rather than a substitute for, addressing dataset quality and class separability in agricultural object detection.

2.3 Open-Set Recognition and Out-of-Domain Detection

Most deep learning-based classifiers are built under the closed-world assumption, according to which every sample presented to the model during inference necessarily belongs to one of the classes seen during training. This assumption proves fragile in real-world applications, where the system is frequently exposed to inputs that do not correspond to any known category, whether irrelevant objects, capture noise, or legitimate classes simply not included in the training set. This scenario, referred to as open-set recognition or out-of-domain (OOD) detection, has motivated a growing number of works across different application domains, reinforcing the relevance of explicitly considering the rejection of unknown samples in any classification pipeline intended to operate under uncontrolled conditions.

In image-based domains, Costa et al. used Monte Carlo Dropout combined with the Kolmogorov-Smirnov statistical test to quantify the uncertainty of ship-wake detection classifiers in SAR imagery, showing that samples rotated to angles not present during training

induce a distributional shift in the model's predictions, which allows the sample to be flagged as out-of-domain before an erroneous classification decision is even made [34].

In applications related to biology and agriculture, the problem takes a form that is even closer to the scenario faced in this work, in which images that do not contain the object of interest must be correctly rejected before any classification attempt.

Closer to the specific case of plant image classification, Dong et al. introduced an open-world setting for plant disease detection in which an unknown-aware region proposal network generates pseudo labels for disease instances that do not match any known category, while a class-agnostic detection head improves the recall of these unannotated regions, allowing the model to later incorporate newly labeled diseases without forgetting the ones learned previously [35]. This work operates directly at the detection level, locating the suspicious region rather than only classifying the whole image, and the authors further show that a detector trained on one species can correctly flag diseased regions of a different, unseen species as unknown instead of forcing them into one of the known classes, indicating that the rejection capability extends to categories that are visually related but absent from training.

Meng et al. addressed a closely related problem under the name Plant-relevant Open-Set Recognition (POSR), aiming explicitly at extending disease and species recognition from known to unknown classes across several plant datasets at once rather than on a single benchmark [36]. Their approach relies on a Vision Transformer pretrained on the large-scale PlantCLEF2022 dataset instead of on ImageNet, following the observation that a plant-related source dataset transfers better to disease and species tasks, and is further trained with additive margin Softmax in place of a generic Softmax loss so that features of the same known class are pulled closer together, leaving more room in the feature space for unknown samples to be separated. Tested on rice, tomato, rose disease, and cotton weed datasets, the method outperformed several CNN and ViT baselines, and the authors additionally show that openness, the relative proportion of unknown to known classes, hurts performance far more than the absolute number of classes, since accuracy on the unknown class drops sharply once the known classes become a minority of the test set.

In a similar vein, Jiang et al. compared a two-head out-of-distribution detector with semi-supervised OpenMatch for the recognition of unknown strawberry diseases, showing that the rejection performance of both approaches depends strongly on which surrogate outlier data is exposed during training, with healthy plant tissue and visually related but unannotated disorders proving far more effective negative examples than images unrelated to the domain

[37]. Their finding is directly applicable to the present problem, since it suggests that the quality of the out-of-domain images used to train a rejection mechanism, and not only the algorithm chosen, has a measurable effect on the final precision of the system.

Fuentes et al. approached an adjacent version of the same difficulty by framing it as open-set domain adaptation rather than plain classification. Working with tomato diseases collected at three greenhouse farms with different conditions, the authors built a system in which a bounding box detector first proposes candidate regions and a second module decides whether each region matches one of the source farm classes or should instead be labeled as unknown, with features transferred from a labeled source farm to an unlabeled target farm while a background class supplies contextual information about the greenhouse itself [38]. Across six combinations of source and target farms and four levels of openness, accuracy on both known and unknown classes declined gradually rather than collapsing as openness increased from a closed-set baseline to 73 percent, and the authors argue that reporting the harmonic mean of known and unknown accuracy gives a fairer picture of a model than either accuracy figure taken in isolation.

Fang et al. addressed a closely related botanical recognition task, proposing an open-set scheme for wood species identification from leaf images that combines deep feature extraction with a weighted Support Vector Data Description model for single-class recognition [39]. Their results show that conventional closed-set classifiers, including several modern convolutional architectures, are essentially unable to separate known from unknown species once the latter are introduced at test time, while the proposed two-stage scheme recovers most of the lost accuracy, reinforcing the conclusion already drawn from the agricultural literature that a dedicated rejection stage is indispensable rather than optional.

The same principle has also been explored outside image-based plant recognition. Gao et al. proposed an open-world fault diagnosis scheme in which the unknown probability of a sample is obtained directly from the Shannon entropy of the closed-set prediction, without requiring an auxiliary model, and is later reused to support incremental learning once the rejected samples are manually labeled [40]. Although applied to industrial process data rather than images, the work illustrates a lightweight rejection mechanism that can be derived from the classifier itself and subsequently linked to continual learning, a property that is desirable whenever a recognition pipeline is expected to evolve as new out-of-scope inputs are encountered in the field.

The nuclear power domain offers a particularly demanding test of this same idea, since a misdiagnosed fault can carry safety consequences well beyond a simple drop in accuracy. Kim et al. examined this challenge by applying OpenMax, a method that can be added to an existing classifier without changing its architecture, to a nuclear power plant accident diagnosis model that had been trained while deliberately leaving one of three accident scenarios out of the training set [41]. Without OpenMax, the excluded scenario was always forced into one of the trained classes, whereas fitting a Weibull distribution to the distances between activation vectors and class centroids allowed the model to flag most of the untrained scenario as unknown while losing only about one percent of accuracy on the scenarios it had been trained on. Their experiments further show that the activation function used in the hidden layers matters for this kind of post hoc rejection, since models built with ReLU were more prone to dead units that blurred the separation between known and unknown activation vectors, which weakened detection of the untrained scenario relative to models using ELU or tanh.

Finally, Allabadi et al. tackled novelty detection for wildlife and plant monitoring by combining a Local Outlier Factor module, applied to features extracted from a pretrained backbone, with the Segment Anything Model to localize and subsequently learn instances of species never seen during training [42]. Their experiments confirm that the difficulty of flagging an unknown category increases with its visual similarity to the known class, since species that are morphologically close to the training set are harder to separate than those that are clearly distinct, an observation consistent with the near-OOD and far-OOD distinction discussed above and that further motivates the inclusion of an explicit out-of-domain detection stage whenever the system may encounter visually related but unannotated categories.

Taken together, these works show that the problem of discriminating between the class of interest and any element that does not represent it, whether a background object, an unannotated class, or simply an image captured outside the expected context, is common across different application domains and cannot be treated as a negligible side effect of the classification stage. A recent survey by Kumar and Rama Krishna on open set recognition specifically for medicinal plant classification reaches a similar conclusion from a broader vantage point, observing that the large majority of methods reviewed still pair a conventional closed-set classifier, most often convolutional or transformer based, with a thresholding or distance based rejection stage added on top rather than a network designed from the outset to

anticipate novelty, and that public benchmarks suited to this open-set protocol remain scarce, since most leaf datasets in circulation were curated for closed-set classification and only adapted afterward [43]. The absence of mechanisms for rejecting unknown samples tends to manifest as a drop in precision and an increase in false positives precisely under the conditions in which the system is most in demand, that is, outside the controlled test set. This body of evidence motivates the need to incorporate, in the problem addressed in this work, explicit mechanisms for detecting samples that do not correspond to the class of interest, so that irrelevant images or images belonging to unforeseen categories are correctly rejected prior to any classification decision.

2.4 Cascaded Classification and Detection Frameworks

Recent literature increasingly favors multi-stage or cascaded architectures over single end-to-end models, particularly when an initial detection or coarse decision step can constrain the search space for a subsequent, more specialized classifier. Khodabakhshi et al. proposed a three-stage hybrid framework for depression severity detection in social media text, in which an initial binary classifier separates depressed from non-depressed posts, a clustering stage (Deep Embedded Clustering) groups depressed posts by linguistic similarity, and a final classifier determines the severity level within each cluster. This cascaded design consistently outperformed both a one-stage (direct four-class) and a two-stage (binary followed by three-class) baseline across three independent datasets, achieving the highest maximum and third-quartile scores in most statistical comparisons performed [44].

A similar rationale appears in fault diagnosis. Saeed et al. designed a hierarchical pipeline for rolling-bearing fault diagnosis composed of three sequentially dependent models: a fault detection model that flags abnormal vibration signals and localizes the onset sample, a fault classification model that determines the fault category (ball, inner race, or outer race), and a fault-severity model specific to each detected category. This cascade not only improved overall classification accuracy (99.62%) but also enabled the accumulation of timestamped detections, allowing precise time-to-detection measurements that a flat, single-stage classifier could not natively provide [45].

A foundational example of this cascaded paradigm appears in early neuromorphic vision systems. Khosla et al. proposed NEOVUS, a video object recognition architecture in which a form- and motion-based detection module first localizes candidate object regions

using bio-inspired saliency and background-subtraction techniques, after which a separate convolutional neural network classifier assigns class labels to each detected region. In a systematic DARPA-led evaluation against two state-of-the-art computer-vision baselines, this detection-then-classification cascade achieved the highest recognition accuracy while consuming energy three orders of magnitude lower than the baselines, indicating that decoupling localization from classification can yield not only higher accuracy but also substantial efficiency gains suitable for real-time, low-power deployment [46].

A more recent, explicitly efficiency-driven cascade inverts the typical ordering of these stages. Chen et al. designed a two-stage framework for railway foreign-object detection in which a lightweight binary classifier first labels each incoming frame as "normal" or "intruded," and only intruded frames are forwarded to a heavier YOLOv3 detector for localization and classification of the specific object. Because the large majority of frames captured during normal railway operation contain no intrusion, this coarse-to-fine cascade allows the costly detection stage to be invoked only when necessary, raising the overall system frame rate from 32.6 FPS (YOLOv3 alone) to 51.77 FPS while preserving a mean average precision of 85.89% on the detection task. This shows that the auxiliary stage in a cascaded architecture need not always follow the detector but can instead serve as an upstream filter that gates whether the more expensive detection model should run at all [47].

Other works adopt a related, though less strictly sequential, strategy of chaining specialized sub-networks that each refine the representation before a final decision layer. Gupta and Jadon's PlantDetectNet pipeline applies feature extraction (GRU and Depthcat CNN) followed by three complementary refinement modules (Global VGG16, Gated ConvNeXt, and Residual DenseNet) whose outputs are concatenated prior to the softmax layer. An ablation study confirmed that removing any individual stage caused substantial accuracy degradation (from 98.8% down to as low as 79.8%), indicating that each stage performs a distinct and necessary filtering function rather than redundant computation [48]. Likewise, in the medical imaging domain, Karpakam and Kumaresan structured their brain-tumor diagnosis system as a sequential pipeline of denoising (Adaptive Bilateral Filtering), segmentation (Otsu–Gannet Segmentation), and feature extraction (GLCM and Enhanced Grasshopper Optimization Algorithm) before the final DIR-GAN classification stage, arguing that this staged refinement allows the downstream generative classifier to operate on clinically relevant, denoised, and well-segmented regions rather than on raw, noisy imaging data [49].

A similar staged-refinement strategy underlies the colorectal cancer detection framework of Raju et al., which sequentially combines integrated CNN feature extraction (AlexNet+DenseNet-201), a Transformer attention module, and an SVM classifier. Their ablation study confirmed that testing accuracy rose monotonically from 95.40% (CNN alone) to 98.10% (CNN+Transformer) and finally to 99.00% (CNN+Transformer+SVM), indicating that each added stage contributes a necessary, non-redundant improvement rather than redundant computation. The framework further appends an unsupervised K-means clustering and bounding-box stage to visually delineate malignant regions only after the classification decision has been made, illustrating how a cascaded design can yield auxiliary localization outputs that a single end-to-end classifier would not natively provide [50].

Collectively, these studies support the design choice of separating detection (or coarse filtering) from fine-grained classification: a dedicated upstream stage—whether a binary detector, an object localizer, or a denoising/segmentation module—consistently reduces the burden placed on the downstream classifier, improves its precision on the relevant data subset, and, in several cases, enables auxiliary outputs (e.g., localization coordinates or detection timestamps) that a single monolithic classifier cannot produce. This motivates the cascaded architecture adopted in the present work, which combines a primary detector with an auxiliary classifier/filter stage.

2.5 Mobile and Offline Plant Recognition

The deployment of deep learning-based plant recognition systems on mobile devices, rather than relying on cloud inference, has become a recurring design choice in recent agricultural and botanical computer vision literature, particularly when target environments cannot guarantee stable network access.

In the context of smallholder agriculture, Appiah et al. introduced *PlanteSaine*, a mobile application for maize, tomato, and onion pest and disease diagnosis in Burkina Faso, combining EfficientNetB3 feature extraction with an Artificial Neural Network classifier converted to TensorFlow Lite (.tflite) format for Android deployment [51]. The authors emphasize that "the application's offline functionality ensures accessibility even in remote areas with limited Internet connectivity," while an optional messaging feature allows synchronization with agricultural authorities only when connectivity becomes available [51].

This hybrid offline-first, online-when-possible design is directly analogous to the connectivity pattern expected in our deployment scenario.

A related hybrid design is reported by Chung et al., who developed a dual-path convolutional network (an InceptionV3 backbone combined with a central-attention path) for the recognition of 14 tree species, coupled with a YOLOv3 detector for real-time, in-camera object detection, and packaged the resulting models into a TensorFlow Lite-based Android application leveraging the Android Neural Networks API (NNAPI) for hardware-accelerated inference [52]. The authors describe the resulting system as an "online/offline available" identification tool and confirm that no Internet connection is required to use the app, while keeping an optional online database mode [52]. They further report that TensorFlow Lite's post-training quantization reduced an InceptionV3-based model from 95.7 MB to 23.9 MB and cut latency from 1130 ms to 845 ms at a cost of roughly 0.8% accuracy [52], illustrating the same accuracy-versus-footprint trade-off that motivates the lightweight architectural choices (Ghost modules, C2f-UIB) discussed throughout this section. Although focused on whole-plant images rather than disease symptoms, this offline-capable, hybrid-mode design reinforces the broader pattern observed across the works above.

Beyond pure on-device inference, Chamara et al. proposed AICropCAM, which performs edge image processing with a Raspberry Pi 4B and transmits only compact inference results (rather than raw images) via a low-power, long-range LoRaWAN link, explicitly motivated by "the lack of computing, electric power, and connectivity infrastructure in agricultural fields" [53]. Although AICropCAM does not target smartphones, it reinforces the broader principle adopted in the present work: conventional, continuously connected cloud-inference architectures are unsuitable for many real-world agricultural and logistics environments, and that processing should be moved as close as possible to the point of image capture.

More recently, Wang et al. proposed APlight-YOLOv8n, a lightweight enhancement of YOLOv8n for the real-time recognition of twelve aquatic plant species across four life forms, which the authors deployed and field-tested directly on a commodity Android smartphone [54]. By replacing the detection head with a Faster Detect module tuned to small- and medium-sized targets and substituting part of the neck's C2f blocks with Universal Inverted Bottleneck (C2f-UIB) modules, the model's parameter count was reduced to 2.74 M (a 13.3% reduction relative to the YOLOv8n baseline) and its computational cost to 5.5 GFLOPs (a 32.9% reduction), while mAP50 improved to 74.4% and a stable 32.70 FPS was

maintained during on-device field inference [54]. The model was exported to the ONNX format and executed entirely on the handset.

In the specific context of disease and pest detection, Deng et al. proposed two lightweight YOLO-based detectors - an Improved YOLOv5s incorporating the Ghost module and an Improved YOLOv7-tiny combining CBAM attention with an SIoU loss - for the recognition of six rice diseases and insect pests and transplanted both models onto an Android smartphone (Android 9.0, 4 GB RAM) using the NCNN forward-inference framework [55]. After NCNN conversion, the best-performing model occupied only 14.3 MB of storage and required 49% CPU usage and 262.9 MB of RAM at runtime, while sustaining an on-device inference speed of 374 ms/frame under FP16 precision [55]. The explicit goal stated by the authors was to achieve offline field detection on mobile devices rather than to rely on a connected server [55]. This work adds NCNN to the set of mobile-inference frameworks already discussed alongside TensorFlow Lite and ONNX Runtime, and its reported CPU/RAM figures offer a concrete quantitative reference for the resource budget needed to run lightweight detectors entirely offline on commodity Android hardware - directly relevant to the feasibility argument for the CEAGESP deployment.

Finally, although not focused on mobile deployment, Ouf compared Faster R-CNN and YOLOv4 for leguminous seed detection and reported inference times of 53.1 ms and 47.2 ms per image, respectively, on GPU hardware [56]. The substantially lower latency of the one-stage YOLOv4 detector, together with its reported mAP@0.5 of 98.52% against 84.56% for Faster R-CNN [56], supports the broader trend observed across the works above of favoring lightweight, single-stage architectures when migrating from server-side training to resource-constrained, offline execution.

A further factor specific to the present application reinforces the choice of a fully offline Android-based solution: Android is by far the dominant mobile operating system in Brazil, holding a 77.93% market share compared to 22.07% for iOS among the two dominant platforms as of May 2026 [57], which justifies targeting Android exclusively rather than developing a cross-platform or iOS-first solution.

3. METHODOLOGY

Considering the structure of this thesis by articles, the methodological approach adopted in this research combines experimental, quantitative, and data-centric procedures, organized according to the specific objectives and research questions addressed in each study. Unlike a purely theoretical thesis, this work is based on the development, training, evaluation, and integration of deep learning models for aromatic herb recognition using representative image data and offline mobile deployment experiments.

The three studies that compose this thesis were designed as complementary stages of the same research trajectory. Paper I established an initial object detection approach for aromatic herb identification using images collected in commercial and uncontrolled environments. Paper II investigated the use of transfer learning and image classification techniques to distinguish visually similar herb species. Paper III expanded the detection problem through a data-centric strategy, including image curation, quality filtering, class selection, annotation procedures, and the comparative evaluation of different object detection architectures.

Together, these methodological stages supported the development of the proposed HerbAI system, a multi-stage deep learning solution that integrates object detection, herb classification, and offline mobile deployment. Therefore, the methodology of this thesis is not limited to the isolated execution of the three articles but also includes the articulation of their results into an integrated framework suitable for practical use in food distribution environments such as CEAGESP.

The following sections describe the methodological procedures adopted in each article and explain how these procedures are connected to the general objective of the thesis and to the development of the integrated recognition system.

3.1 Methodology of Paper I - Model Development for Identifying Aromatic Herbs Using Object Detection Algorithm

Connection with the specific objective: this paper contributed directly to the first specific objective of the thesis, which was to design and evaluate an object detection model for identifying aromatic herbs from images collected under commercial and uncontrolled conditions.

Method: experimental study using an object detection model based on YOLOv8.

General objective: to develop and evaluate a deep learning model capable of detecting three aromatic herbs commercialized at CEAGESP: rosemary, mint, and bay leaf.

Research question: can a YOLOv8-based object detection model correctly identify rosemary, mint, and bay leaf in images obtained from different sources and acquisition conditions?

The study followed the CRISP-DM methodology, organizing the research into stages of business understanding, data understanding, data preparation, modeling, evaluation, and deployment analysis. In the business understanding stage, the problem was defined from the need to identify aromatic herbs in the context of CEAGESP, a real food distribution environment. In the data understanding and preparation stages, images were gathered from different sources, including photographs captured with cell phones, cameras, and tablets, images retrieved from web sources, and images collected during visits to CEAGESP.

After image consolidation, a cleaning process was carried out to remove images with significant blur or inadequate visual quality. The final dataset was divided into training and validation subsets, using approximately 90% of the images for training and 10% for validation. The dataset included 234 training images and 27 validation images for rosemary, 302 training images and 35 validation images for mint, and 262 training images and 31 validation images for bay leaf. The classes were defined as bay leaf, rosemary, and mint, and the images were annotated with bounding boxes using a computer vision annotation tool.

The modeling stage used the standard YOLOv8 architecture, without architectural modifications. This decision was important because the purpose of the study was not to propose a new neural network architecture, but to test whether an existing object detection model could learn to identify aromatic herbs from the available dataset. The training was performed in a cloud environment using Google Colab resources. First, individual models were trained for each herb class using 300 epochs. After this stage, a consolidated model containing the three classes was trained, also totaling 300 epochs.

The evaluation was based on object detection metrics commonly used in YOLO models: precision, recall, mAP50, and mAP50-95. Precision was used to assess how many detections made by the model were correct, while recall measured the model's ability to find the expected objects in the images. The mAP50 metric was adopted as the main performance

indicator, considering an IoU threshold of 0.50, and the study used mAP50 greater than or equal to 0.7 as the target reference for model performance.

The results showed that the model was able to detect the three aromatic herbs, with the consolidated model reaching mAP50 and precision values above 0.7. However, the recall values indicated over-detection and recognition confusion, especially when the model was exposed to visually similar herbs that were not part of the training classes. Rosemary presented the best performance, while mint and bay leaf showed greater need for dataset improvement and bounding-box refinement.

This paper was important for the thesis because it established the first experimental baseline for aromatic herb detection using deep learning. It also showed that model performance was strongly affected by dataset size, image variability, annotation criteria, and the visual similarity between herb species. These limitations guided the next stages of the thesis, especially the use of classification models, dataset expansion, data-centric curation, and the later development of an integrated recognition system.

3.2 Methodology of Paper II - Enhanced Classification of Aromatic Herbs Using EfficientNet and Transfer Learning

Connection with the specific objective: this paper contributed to the second specific objective of the thesis, which was to evaluate image classification with transfer learning for distinguishing visually similar aromatic herbs.

Method: experimental study using the EfficientNet convolutional neural network for image classification.

General objective: to evaluate the performance of EfficientNet in the classification of three aromatic herbs: rosemary, mint, and bay leaf.

Research question: can EfficientNet, with and without transfer learning, classify visually similar aromatic herbs from image data with acceptable accuracy?

This paper addressed a different task from Paper I. While the first study used object detection to locate herbs in images, the second study focused on image classification, assigning each image to one of the predefined herb classes. This distinction was important for the thesis because classification could support a later recognition pipeline, especially in cases where the object had already been detected or cropped.

The methodological procedure was organized into two main stages: dataset preparation and experimental model training. The dataset contained images of rosemary, mint, and bay leaf. These classes were selected because they present visual similarities in color, texture, and shape, which make classification more difficult. The images included both individual leaves and herbs arranged in bunches or mixed with other objects. This choice preserved part of the variability found in real situations, but also introduced classification noise, since the model sometimes had to learn from images with different visual compositions.

The experiments were conducted in Google Colab using Python and libraries commonly adopted in image classification tasks, including TensorFlow, NumPy, Pandas, Matplotlib, and OpenCV. The EfficientNet architecture was selected because it allows different combinations of network depth, width, and image resolution, making it possible to compare classification performance under different computational costs.

Two training strategies were evaluated. In the first, EfficientNet was trained using transfer learning, with ImageNet pre-trained weights. Variants from EfficientNetB0 to EfficientNetB7 were tested, with image resolutions ranging from 224 to 600 pixels. Each model was trained for 50 epochs. In the second strategy, EfficientNet was trained without pre-trained weights, using only the aromatic herb dataset. This custom training was carried out to verify whether the model could learn more stable features directly from specific herb images. Due to computational limitations, this second strategy evaluated variants from EfficientNetB0 to EfficientNetB4, also using 50 epochs.

The main evaluation metric was accuracy, defined as the proportion of correctly classified images. Loss was also monitored during training to assess convergence and possible overfitting. The same general training conditions were maintained across the experiments to allow comparison between model variants and between the transfer learning and custom training approaches.

The transfer learning experiments produced different results across EfficientNet variants. EfficientNetB4 obtained the highest accuracy among the pre-trained models, reaching 0.80, followed by B2 and B3, both with 0.77. However, the results also showed instability across resolutions, with some higher-resolution models requiring substantially longer training time without producing better accuracy. EfficientNetB7, for example, required 9 h 44 min and reached 0.65 accuracy.

In the custom training experiments, EfficientNetB0 achieved the best result, with 0.88 accuracy, followed by B1 and B3, both with 0.80. Although training without transfer learning required more time, the results indicated more stable learning behavior in some configurations. These findings suggested that the pre-trained ImageNet features were useful for reducing training time but were not always sufficient to handle the specific visual characteristics of aromatic herbs.

This paper was important for the thesis because it showed that classification could complement object detection in the recognition of aromatic herbs. It also reinforced a limitation already observed in Paper I: model performance was strongly affected by the composition and quality of the dataset. The presence of mixed images, bunches, isolated leaves, and visually similar species affected accuracy and produced classification confusion. These results supported the later decision to refine the dataset, improve class selection, and combine detection and classification in a multi-stage recognition system.

3.3 Methodology of Paper III - Data-Centric Object Detection for Aromatic Herbs: Image Quality Filtering, Class Selection, and Model Comparison

Connection with the specific objective: this paper contributed to the third specific objective of the thesis, which was to investigate the impact of dataset curation, image quality filtering, class selection, annotation consistency, and model comparison on the detection of aromatic herbs.

Method: experimental and data-centric study using object detection models trained on the same curated dataset.

General objective: to evaluate whether a data-centric strategy could improve aromatic herb detection under real-world visual conditions, using a curated and balanced dataset across six herb classes.

Research question: does systematic dataset construction, including image quality filtering, class selection, annotation review, and class balancing, improve the detection performance of different object detection architectures for aromatic herbs?

This paper expanded the problem investigated in Paper I. Instead of training a detector on three herb classes with a limited dataset, the third study adopted a data-centric approach, in which the focus was the quality, composition, and visual coherence of the dataset. The

purpose was to evaluate whether better data construction could improve detection performance without depending only on architectural changes.

The dataset was built from three sources: original photographs captured at CEAGESP, images collected in controlled indoor settings, and web-scraped images retrieved through structured keyword searches in Portuguese and English. The web search terms included common herb names, commercialization contexts, cultivation settings, bundles, loose leaves, and market presentation formats. After collection, all images were standardized, consolidated by class, and converted to a consistent image format.

A deduplication step was performed to remove duplicate and near-duplicate images. This was necessary to reduce redundancy and avoid inflated performance caused by similar images appearing in different dataset splits. After deduplication, the images underwent manual visual curation based on the CEAGESP context. Images were excluded when they had excessive blur, poor lighting, high noise, inadequate botanical representation, or an appearance that did not match the product as commercialized. Scientific illustrations, heavily edited images, and images representing non-commercial growth stages were also removed.

The class selection process was based on visual separability. Since the final system depends only on RGB images, species that require smell, touch, or non-visual cues for reliable distinction were not retained. The selected classes were basil, mint, sage, bay leaf, leek, and rosemary. Other herbs were excluded when their visual distinction was not reliable under uncontrolled image conditions. This was the case for species such as chives, dill, fennel, tarragon, marjoram, oregano, thyme, coriander, and parsley, which presented either strong visual overlap with retained classes or inconsistent visual representation in market images.

Data augmentation was applied only to the training set, using the Albumentations library. Five transformations were used: Gaussian blur, brightness adjustment, Gaussian noise, random cropping, and horizontal flipping. These transformations were selected to simulate common variations in real image acquisition, such as focus problems, uneven illumination, sensor noise, framing differences, and changes in object orientation. For each class, 100 images were randomly selected from the curated pool and processed through the augmentation pipeline. The augmented images were manually reviewed, and only visually acceptable samples were retained.

After curation and augmentation, the final dataset contained 5,679 images and 13,490 annotated herb instances, since several images contained more than one visible herb object or bundle. The dataset was approximately balanced across the six classes and divided into training, validation, and test subsets using a stratified split of approximately 70%, 15%, and 15%. The test set was held out during training and used only for final evaluation.

The annotation process combined automatic pre-annotation and manual review. First, a lightweight auxiliary YOLOv11s model was trained on a small subset of 480 training images and 120 validation images. This auxiliary model was not part of the final comparison; it was used only to generate initial bounding-box proposals. The pre-annotations were imported into CVAT and manually reviewed image by image. During review, incorrect labels, imprecise boxes, false positives, and missed objects were corrected. The final annotations were exported in YOLO format and validated using Python scripts to check image-annotation correspondence, class identifiers, coordinate ranges, and empty annotation files.

Three object detection architectures were trained and evaluated on the same dataset: YOLOv11s, Faster R-CNN, and SSD300/VGG. YOLOv11s was trained using the Ultralytics framework with pretrained weights, image size of 640 x 640 pixels, and 100 epochs. Faster R-CNN was implemented in Detectron2 using the `faster_rcnn_R_50_FPN_3x` configuration, with a ResNet-50 backbone and a Feature Pyramid Network, trained for 20,000 iterations after converting the annotations from YOLO to COCO format. SSD300/VGG was trained using a VGG backbone for 20,000 iterations, after converting the annotations to Pascal VOC format.

The evaluation protocol considered the differences between the frameworks. $AP@0.5$ was adopted as the common metric for cross-model comparison, since it was available for all three models. YOLOv11s and Faster R-CNN were also evaluated using $mAP@0.50:0.95$, which provides a stricter assessment of localization quality across multiple IoU thresholds. SSD300/VGG was evaluated using the VOC 2007 protocol, so $mAP@0.50:0.95$ was not available for this model.

The results showed that YOLOv11s achieved the best overall performance on the held-out test set, with $mAP@0.5$ of 0.802 and $mAP@0.50:0.95$ of 0.593. Faster R-CNN obtained a similar $AP@0.5$ of 0.798, but with lower $mAP@0.50:0.95$ and less consistent per-class behavior. SSD300/VGG reached $mAP@0.5$ of 0.726, with lower overall performance but a more uniform distribution across classes. These results indicated that the curated dataset

allowed different model families to reach competitive performance, while YOLOv11s presented the best balance between accuracy, localization quality, and deployment feasibility.

This paper was important for the thesis because it showed that the main improvement over the first detection study did not come only from changing the model. The most relevant methodological change was the construction of a better dataset: larger, cleaner, more balanced, visually coherent, and manually reviewed. The findings reinforced the central argument of the thesis that aromatic herb recognition depends not only on model architecture, but also on data quality, annotation consistency, and class definition.

3.4 Methodology of the Mobile Deployment and Integrated Inference Pipeline

Connection with the general objective: this stage contributed to the integration of the thesis results into a functional mobile prototype capable of performing offline aromatic herb recognition.

Method: applied development of a native Android application integrating object detection, image classification, multilingual interaction, and offline inference.

General objective: to implement and validate the feasibility of HerbAI as a mobile application for offline aromatic herb recognition.

Research question: can the models developed throughout the thesis be integrated into a mobile application capable of performing offline inference using both object detection and fallback classification?

This stage was not developed as a separate article. It was carried out as an additional applied contribution of the thesis, connecting the experimental results obtained in the previous studies to a functional mobile prototype. The purpose was to verify whether the trained models could be converted, embedded, and executed locally on an Android device, without requiring cloud inference or an internet connection.

The mobile application was developed natively for Android using Kotlin and Jetpack Compose. The app allows the user to select an image from the device gallery or capture a new image using the camera. After image acquisition, the input image is resized to 640 x 640 pixels and submitted to the inference pipeline. The application was designed to operate offline, which is a relevant requirement for use in environments where network access may be unavailable, unstable, or restricted.

The integrated inference pipeline combines two complementary models. The first model is the YOLOv11s object detector, converted to TensorFlow Lite and embedded in the Android application. This detector is responsible for identifying the six herb classes defined in Paper III: basil, mint, sage, bay leaf, leek, and rosemary. When valid detections are returned, the app displays the predicted class, confidence score, and bounding box over the input image. For mobile readability, the app displays only the highest-confidence detection per class, avoiding excessive overlapping boxes on small screens.

The second model is an EfficientNet-B0 classifier exported to ONNX and executed using ONNX Runtime Android. This classifier is used as a fallback stage when the detector does not return a valid detection. Its role is not to replace the detector, but to identify selected herb groups outside the six-class detection scope. In the current prototype, the classifier supports oregano and the parsley/coriander group. A minimum confidence threshold of 0.70 was adopted for fallback classification. When the classifier reaches this threshold, the app reports that the herb was identified outside the detection group.

The app also includes interface and accessibility features required for practical use. The interface supports Portuguese, English, and Spanish. Text outputs are localized according to the selected language. An optional text-to-speech function was implemented to read the recognition result aloud, allowing the app to support users who may benefit from auditory feedback. The interface also includes buttons for image selection and camera capture, visual display of the selected image, prediction overlays, and a message indicating that the application operates offline.

The Android development environment included Android Studio, Gradle, Kotlin, Jetpack Compose, TensorFlow Lite, TensorFlow Lite Support, and ONNX Runtime Android. The app was tested using both an Android emulator and a physical Samsung Galaxy S21 device. The repository also includes screenshots and a demonstration video documenting the main app behaviors: image selection, camera input, multilingual interface, optional audio output, detector results, and fallback classifier results.

This stage was important for the thesis because it moved the work from model evaluation to practical deployment. Papers I, II, and III evaluated detection, classification, and dataset construction strategies. The mobile application connected these results into a single operational pipeline. In this sense, HerbAI demonstrates that the proposed approach can be implemented as an offline mobile system, supporting the use of deep learning for aromatic herb recognition beyond the experimental environment.

4. RESULTS AND DISCUSSION

This chapter presents the main results obtained throughout the thesis, organized according to the three articles and the additional mobile deployment stage developed as part of this research. The results are discussed in relation to the general objective of the thesis: to develop and validate a multi-stage deep learning system for aromatic herb recognition using representative image data and offline mobile deployment experiments.

The first three sections summarize the findings of the articles that compose the thesis. Paper I presents the initial object detection baseline using YOLOv8 for three aromatic herbs. Paper II evaluates image classification using EfficientNet and transfer learning. Paper III expands the detection problem through a data-centric approach, including dataset curation, class selection, annotation review, and comparison among YOLOv11s, Faster R-CNN, and SSD300/VGG. The final section presents the results of the HerbAI mobile application, which integrates the detection and classification models into an offline Android prototype.

Together, these results show the progression of the research from initial model development to dataset refinement, comparative evaluation, and mobile deployment. This organization allows the thesis to discuss not only model performance, but also the methodological decisions that affected the feasibility of applying deep learning to aromatic herb recognition in a practical environment such as CEAGESP.

4.1 Results of Paper I

The first paper evaluated the feasibility of using YOLOv8 for the automatic detection of three aromatic herbs: rosemary, mint, and bay leaf. The study represented the first experimental stage of this thesis and established an initial baseline for herb detection using images obtained from different acquisition sources and under heterogeneous visual conditions.

The individual training results showed that YOLOv8 was able to learn relevant visual patterns for the three selected classes. For rosemary, the model reached mAP50 of 0.739, precision of 0.724, and recall of 0.786. For mint, the model obtained mAP50 of 0.738, precision of 0.739, and recall of 0.714. For bay leaf, the model reached mAP50 of 0.700, precision of 0.611, and recall of 0.727. These results indicated that the detector could identify the target herbs but also showed differences in performance among classes.

The consolidated model, trained on all three classes together, achieved mAP50 values above the minimum target adopted in the study. From the third training stage onward, the variation in the main metrics became smaller, indicating stabilization of the learning process. Rosemary presented the strongest behavior among the three classes, with mAP50, precision, and recall values above 0.8 in the consolidated training. Mint showed intermediate results, with mAP50 and precision close to 0.65 and recall close to 0.6. Bay leaf presented the weakest performance, with greater recognition confusion during testing.

The results also revealed an important limitation: the model tended to produce over-detections and confusion when exposed to herbs outside the trained classes. This was observed especially in tests with visually similar or untrained aromatic herbs. The recall values above 0.5, combined with the qualitative testing, suggested that the detector was able to find herb-like regions, but not always with enough discrimination between target and non-target classes.

Another relevant finding was the influence of dataset quality and annotation strategy on detection performance. The study showed that the number of images, the diversity of acquisition conditions, and the definition of bounding boxes directly affected the model results. In some cases, the bounding-box decision was not trivial, especially for herbs arranged in bunches or with small leaves, such as mint. This indicated that model performance depended not only on the YOLOv8 architecture, but also on the consistency and representativeness of the training dataset.

Overall, Paper I demonstrated that object detection could be applied to aromatic herb recognition but also showed that the initial dataset was still limited for robust real-world deployment. The main contribution of this paper to the thesis was therefore twofold: it confirmed the technical feasibility of training a YOLO-based detector for aromatic herbs, and it identified the main weaknesses that guided the following studies, especially the need for dataset expansion, improved annotation criteria, class refinement, and strategies to reduce false detections.

4.2 Results of Paper II

The second paper evaluated the use of EfficientNet for the classification of three aromatic herbs: rosemary, mint, and bay leaf. Unlike Paper I, which focused on object detection and bounding-box localization, this study analyzed image classification, in which

each input image is assigned to one herb class. This distinction was important for the thesis because classification could later be used as a complementary stage in a broader recognition pipeline.

The results showed that EfficientNet could classify aromatic herbs with different levels of performance depending on the model variant, image resolution, and training strategy. In the transfer learning experiments, using ImageNet pre-trained weights, EfficientNet-B4 achieved the best result, with accuracy of 0.80 and loss of 0.73 after 50 epochs. EfficientNet-B2 and EfficientNet-B3 both reached accuracy of 0.77, while EfficientNet-B0 reached 0.71. Higher-resolution variants did not necessarily improve performance. EfficientNet-B7, for example, required 9 h 44 min of training and reached accuracy of 0.65, showing that increasing model scale and resolution did not automatically produce better classification results.

The custom training experiments, performed without pre-trained weights, produced the highest accuracy in the study. EfficientNet-B0 reached accuracy of 0.88, followed by EfficientNet-B1 and EfficientNet-B3, both with accuracy of 0.80. Although these models required longer training time than the transfer learning models, they showed more stable learning behavior in some configurations. This result indicated that the model could learn discriminative features directly from the herb dataset, even with the visual similarity among the classes.

The comparison between transfer learning and custom training showed a trade-off between training efficiency and stability. Transfer learning reduced training time and allowed faster experimentation but produced stronger oscillations in accuracy across EfficientNet variants. Custom training demanded more computational time, but produced competitive and, in some cases, superior accuracy. These findings suggested that ImageNet pre-training was useful, but not sufficient to fully solve the classification problem when the target classes presented subtle visual differences.

The results also confirmed that dataset composition affected classification performance. The dataset included both isolated leaves and herbs arranged in bunches or mixed with other objects. This diversity helped approximate real-world conditions but also introduced visual noise. The study identified that classification confusion was influenced by the mixture of image types and by the similarity between rosemary, mint, and bay leaf in some acquisition conditions. Therefore, the paper reinforced the need for better image curation and more consistent input representation.

Overall, Paper II showed that EfficientNet is a viable architecture for aromatic herb classification, especially when model resolution and training strategy are selected carefully. Its main contribution to the thesis was to demonstrate that classification can complement object detection, particularly when the recognition task does not require localization. The findings also supported later decisions in the thesis, including the use of EfficientNet-B0 in the mobile application as a lightweight classifier for herb groups outside the detector scope.

4.3 Results of Paper III

The third paper evaluated the impact of a data-centric strategy on aromatic herb detection. Unlike Paper I, which used an initial three-class dataset, this study expanded the detection problem to six visually selected herb classes: basil, mint, sage, bay leaf, leek, and rosemary. The main result of this paper was not only the comparison among detection architectures, but the demonstration that dataset quality, class definition, and annotation consistency directly affected model performance.

The final dataset contained 5,679 images and 13,490 annotated herb instances. This difference occurred because several images contained more than one visible herb object or bundle. The dataset was built through field photography, indoor photography under controlled conditions, web scraping, deduplication, visual curation, data augmentation, and manual annotation review. After curation, the dataset was divided into training, validation, and held-out test subsets, with the same splits used for all models.

The validation results showed that YOLOv11s achieved the best performance among the models evaluated on the validation set. It reached $\text{mAP}@0.5$ of 0.819, while SSD300/VGG reached 0.752. Per-class results also showed that bay leaf obtained the highest AP among the classes, while rosemary remained one of the most challenging classes due to its thin, needle-like morphology and greater susceptibility to background confusion. Faster R-CNN was not included in the per-class validation table because these values were not extracted during the original Detectron2 evaluation; however, it was included in the held-out test set comparison.

On the independent test set, YOLOv11s again achieved the best overall result, with $\text{mAP}@0.5$ of 0.802 and $\text{mAP}@0.50:0.95$ of 0.593. Faster R-CNN obtained a very close $\text{AP}@0.5$ of 0.798, but a lower $\text{mAP}@0.50:0.95$ of 0.541. SSD300/VGG reached $\text{mAP}@0.5$

of 0.726. These results showed that YOLOv11s and Faster R-CNN were similar when evaluated at IoU 0.5, but YOLOv11s performed better under the stricter localization metric.

The per-class test results showed different behavior among architectures. YOLOv11s presented more consistent results across the six classes, with AP@0.5 values of 0.792 for basil, 0.816 for mint, 0.799 for sage, 0.857 for bay leaf, 0.839 for leek, and 0.713 for rosemary. Faster R-CNN reached AP@0.5 of 0.740 for basil, 0.790 for mint, 0.780 for sage, 0.840 for bay leaf, 0.878 for leek, and 0.759 for rosemary. SSD300/VGG obtained AP@0.5 of 0.714 for basil, 0.759 for mint, 0.715 for sage, 0.776 for bay leaf, 0.751 for leek, and 0.638 for rosemary.

The comparison showed that YOLOv11s had the best balance between accuracy and deployment feasibility. Faster R-CNN performed competitively, especially at AP@0.5, but its two-stage architecture is computationally more demanding. SSD300/VGG had the lowest overall mAP but still produced usable results across all classes. Since SSD300/VGG was evaluated using the VOC 2007 protocol, mAP@0.50:0.95 was not available for this model, which limited direct comparison with YOLOv11s and Faster R-CNN under stricter IoU thresholds.

The speed-related results were treated with caution. YOLOv11s was the only model explicitly profiled for inference latency, reaching approximately 1.2 ms per image for model inference and 21.1 ms per image for the complete preprocessing-inference-postprocessing pipeline on the RTX 4090. SSD300/VGG processed the validation set of 849 images in approximately 13 seconds, which provides only an approximate throughput estimate under its evaluation pipeline. Faster R-CNN inference time was not explicitly logged. Therefore, conclusions about speed were restricted to measured or directly observed values.

Overall, Paper III showed that the main improvement over the initial detection study came from the dataset construction process. By expanding the number of classes, filtering low-quality images, removing duplicates, defining classes based on visual separability, balancing the dataset, and manually reviewing annotations, the detection performance became more stable across different architectures. The paper confirmed that aromatic herb recognition is not only a model selection problem, but also a data quality and class definition problem.

This paper was central to the thesis because it provided the strongest detection model used in the mobile application. The YOLOv11s model selected from this study became the

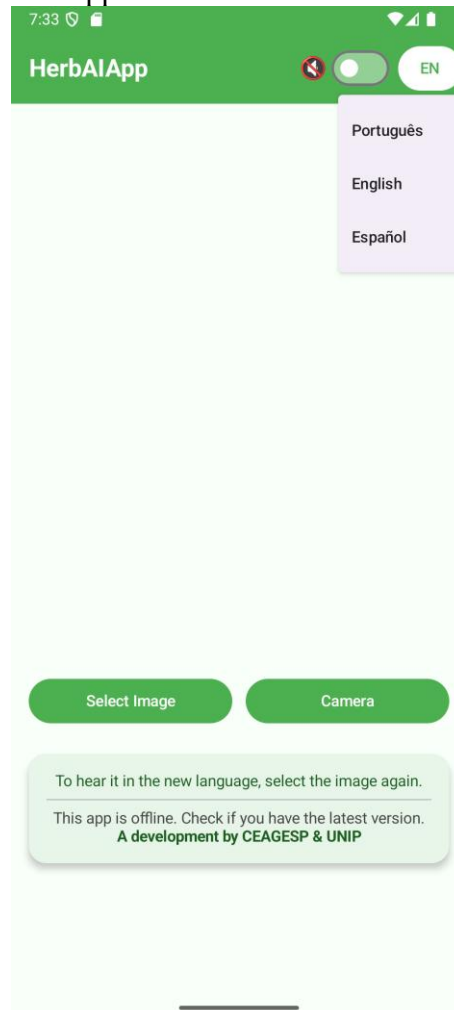
detector integrated into HerbAI, supporting the six-class recognition stage of the offline mobile inference pipeline.

4.4 Results of the Mobile Deployment and Integrated Inference Pipeline

The mobile deployment stage resulted in a functional Android prototype named HerbAIApp. This stage was not developed as an independent article, but as an additional applied contribution of the thesis. Its purpose was to verify whether the models evaluated in the previous studies could be integrated into a single offline mobile inference pipeline.

The application was implemented natively in Android using Kotlin and Jetpack Compose. The final prototype allows the user to either select an image from the device gallery or capture a new image using the camera. After image input, the app runs the recognition pipeline locally on the device, without requiring internet access or cloud processing. Figure 1 shows the main interface of the application, including the image input options and the multilingual menu.

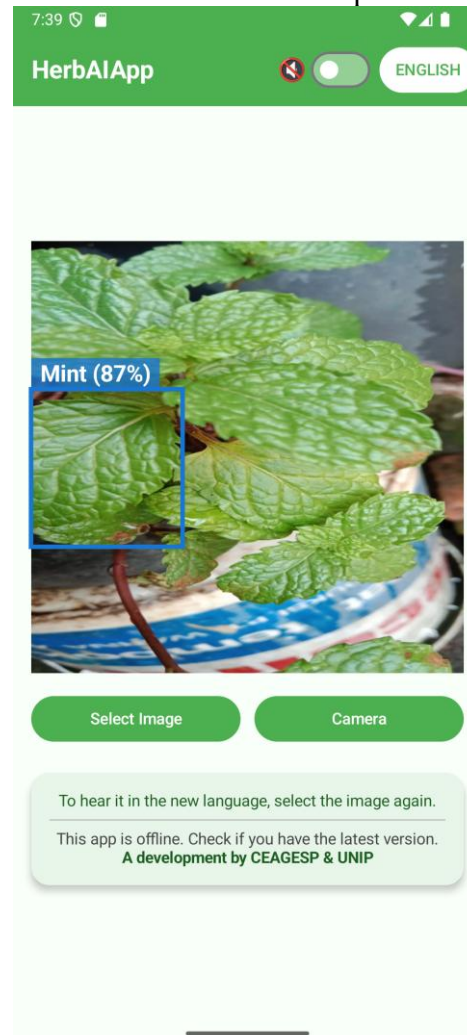
Figure 1. HerbAIApp initial interface with multilingual menu.



Source: Prepared by the author (2026).

The first stage of the mobile pipeline uses the YOLOv11s detector converted to TensorFlow Lite. This model detects the six herb classes selected in Paper III: basil, mint, sage, bay leaf, leek, and rosemary. When a valid detection is found, the application displays the predicted class, confidence score, and bounding box over the image. In the visual tests performed with the prototype, the detector returned valid predictions for all six target classes. Figure 2 presents examples of detector-based outputs generated by the mobile application.

Figure 2. Examples of YOLOv11s detector outputs in the mobile application.



Source: Prepared by the author (2026).

The app interface was designed to show only the highest-confidence detection per class. This decision improved readability on mobile screens, especially for images with overlapping herbs or repeated leaves. Instead of displaying multiple overlapping boxes, the app presents a cleaner output focused on the most relevant detection result.

The second stage of the pipeline uses an EfficientNet-B0 classifier exported to ONNX. This classifier is activated when the detector does not return a valid detection. Its purpose is to identify selected herb groups outside the six-class detector scope. In the current prototype, the fallback classifier recognizes oregano and the parsley/coriander group. A confidence threshold of 0.70 was used for this stage. When the classifier reaches this threshold, the app informs the user that the herb was identified outside the detection group. Figure 3 shows examples of fallback classification results.

Figure 3. Examples of EfficientNet-B0 fallback classifier outputs.



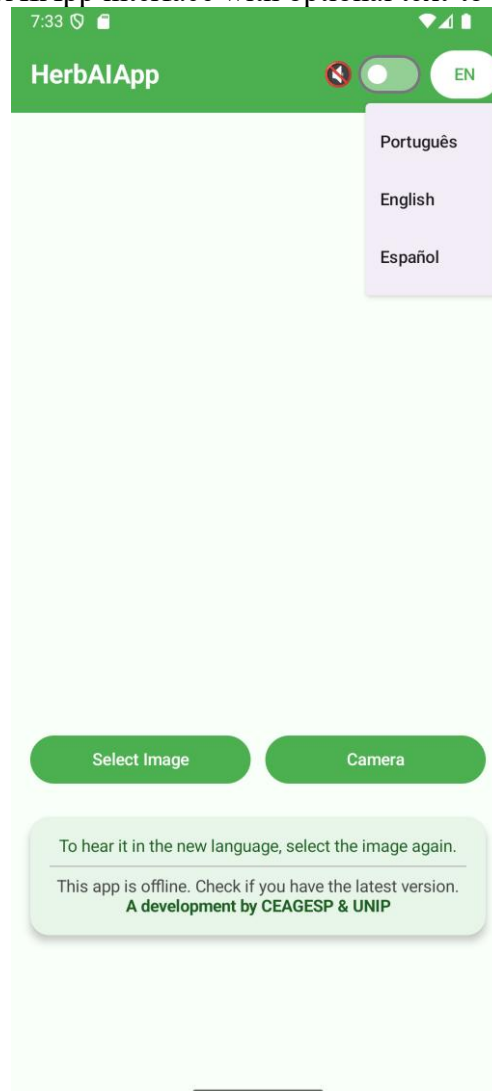
Source: Prepared by the author (2026).

The fallback classifier achieved high quantitative performance on its held-out test set. The model reached test accuracy of 0.9810, test loss of 0.104545, and best validation accuracy of 0.9905. For oregano, the classifier obtained precision of 0.9574, recall of 1.0000, and F1-score of 0.9783. For the parsley/coriander group, it obtained precision of 1.0000, recall of 0.9667, and F1-score of 0.9831. These results supported its use as a complementary classification stage in the mobile prototype.

The app also implemented multilingual and accessibility features. The interface supports Portuguese, English, and Spanish, allowing the same prediction output to be presented in different languages. In addition, an optional text-to-speech function was added so that the recognition result can be read aloud. This feature is relevant for accessibility and

reinforces the practical orientation of the application. Figure 4 illustrates the interface with the text-to-speech option enabled.

Figure 4. HerbAIApp interface with optional text-to-speech enabled.



Source: Prepared by the author (2026).

The prototype was tested on an Android emulator and on a Samsung Galaxy S21 device. The tests confirmed the execution of the main app functions: image selection, camera input, offline inference, bounding-box visualization, fallback classification, multilingual output, and optional audio feedback. The screenshots produced during testing documented both detector-based predictions and classifier-based fallback results.

Overall, this stage showed that the proposed approach could move beyond model training and be implemented as a working mobile system. The results demonstrated the

feasibility of combining object detection and classification in a single offline Android application. In the context of the thesis, HerbAIApp represents the practical integration of the three studies: YOLO-based detection from Papers I and III, EfficientNet-based classification from Paper II, and mobile deployment as the final applied outcome.

5. FINAL CONSIDERATIONS

This thesis developed and evaluated HerbAI, a multi-stage deep learning system for aromatic herb recognition designed for practical food distribution contexts. The research was organized into three articles and an additional mobile deployment stage. Together, these stages addressed object detection, image classification, data-centric dataset construction, model comparison, and offline mobile inference.

The results showed that aromatic herb recognition is not only a model selection problem. Dataset quality, visual class definition, annotation consistency, and deployment constraints directly affected the feasibility of the proposed system. The progression from the initial YOLOv8 detector to the curated six-class YOLOv11s model and the final Android prototype demonstrated that technical performance improves when model development is combined with careful data preparation and practical deployment requirements.

Overall, the thesis confirmed the feasibility of using deep learning to identify aromatic herbs in uncontrolled visual conditions, especially when detection and classification are treated as complementary stages. The HerbAIApp prototype showed that the proposed approach can be implemented as an offline mobile application, supporting the use of artificial intelligence beyond the experimental environment.

5.1 Conclusions

The general objective of this thesis was achieved. A multi-stage deep learning system for aromatic herb recognition was developed, evaluated, and deployed as an offline Android prototype.

Paper I showed that YOLO-based object detection could identify rosemary, mint, and bay leaf, but also revealed limitations related to over-detection, dataset size, and bounding-box consistency. Paper II showed that EfficientNet could classify visually similar herbs and supported the use of classification as a complementary recognition stage. Paper III demonstrated that dataset curation, image quality filtering, class selection, and annotation review improved detection stability across different architectures.

Among the evaluated detection models, YOLOv11s presented the best balance between accuracy and deployment feasibility. It achieved the highest test performance in Paper III and was selected for integration into the mobile application. The EfficientNet-B0

classifier was used as a fallback stage for herb groups outside the detector scope, expanding the practical behavior of the app without forcing all inputs into the six detection classes.

The final mobile prototype confirmed that the proposed pipeline could operate offline, using image selection or camera input, multilingual output, bounding-box visualization, fallback classification, and optional text-to-speech. Therefore, the thesis contributes both experimentally and practically: it advances the study of aromatic herb recognition using deep learning and demonstrates a working mobile implementation designed for practical use in food distribution contexts.

5.2 Recommendations for Future Work

Future work should expand the dataset with new images collected directly at CEAGESP and in other commercialization environments. Increasing the number of real images, especially for the most difficult classes, may improve model robustness under variations in lighting, occlusion, background, and product presentation.

Another recommendation is to evaluate the mobile application with end users, including CEAGESP workers, retailers, students, and non-specialist users. This would allow the assessment of usability, response time, language support, and the practical value of offline herb recognition in real operating conditions.

Future versions of HerbAI may also include additional herb classes, if they meet visual separability criteria. Classes that are difficult to distinguish using RGB images alone should be treated carefully, either through grouped labels, open-set rejection, or additional data sources.

Finally, future research should investigate model compression, quantization, and broader on-device benchmarking. These steps may reduce model size, improve inference speed, and support deployment on lower-cost Android devices.

REFERENCES

- [1] W. Ding, M. Abdel-Basset, I. Alrashdi, and H. Hawash, "Next generation of computer vision for plant disease monitoring in precision agriculture: A contemporary survey, taxonomy, experiments, and future direction," *Information Sciences*, vol. 665, p. 120338, 2024, doi: 10.1016/j.ins.2024.120338.
- [2] Y. Chen et al., "IPMCNet: A lightweight algorithm for invasive plant multiclassification," *Agronomy*, vol. 14, no. 2, p. 333, 2024, doi: 10.3390/agronomy14020333.
- [3] B. Dey, J. Ferdous, R. Ahmed, and J. Hossain, "Assessing deep convolutional neural network models and their comparative performance for automated medicinal plant identification from leaf images," *Heliyon*, vol. 10, no. 1, p. e23655, 2024, doi: 10.1016/j.heliyon.2023.e23655.
- [4] N. Penpong, Y. Wanna, C. Kamjanlard, A. Techasen, and T. Intharah, "Attacking the out-of-domain problem of a parasite egg detection in-the-wild," *Heliyon*, vol. 10, no. 3, p. e26153, 2024, doi: 10.1016/j.heliyon.2024.e26153.
- [5] M. Saadati et al., "Out-of-distribution detection algorithms for robust insect classification," *Plant Phenomics*, vol. 6, Art. no. 0170, 2024, doi: 10.34133/plantphenomics.0170.
- [6] H. AbuAlkebash, R. A. A. Saleh, and H. M. Ertunç, "Automated explainable deep learning framework for multiclass skin cancer detection and classification using hybrid YOLOv8 and vision transformer (ViT)," *Biomedical Signal Processing and Control*, vol. 108, p. 107934, 2025, doi: 10.1016/j.bspc.2025.107934.
- [7] D. Yudin, N. Zakharenko, A. Smetanin, R. Filonov, M. Kichik, V. Kuznetsov, D. Larichev, E. Gudov, S. Budenny, and A. Panov, "Hierarchical waste detection with weakly supervised segmentation in images from recycling plants," *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107542, 2024, doi: 10.1016/j.engappai.2023.107542.
- [8] M. Flórez, O. Becerra, E. Carrillo, M. Villa, Y. Álvarez, J. Suárez, and F. Mendes, "Deep learning application for biodiversity conservation and educational tourism in natural reserves," *ISPRS International Journal of Geo-Information*, vol. 13, no. 10, p. 358, 2024, doi: 10.3390/ijgi13100358.
- [9] Y. Ren, G. Li, Z. Li, W. An, J. Yang, and H. Ni, "A mobile identification system for invasive plants based on adaptive hybrid feature correction-guard network model, supporting fully offline decision-making," *Journal of King Saud University - Computer and Information Sciences*, vol. 38, no. 1, Art. no. 13, 2025, doi: 10.1007/s44443-025-00396-6.
- [10] Agência Nacional de Telecomunicações (Anatel), Resolução nº 760, de 6 de fevereiro de 2023, Regulamento sobre Bloqueador de Sinais de Radiocomunicações (BSR). Brasília, DF: Diário Oficial da União, 8 fev. 2023. [Online]. Available: <https://informacoes.anatel.gov.br/legislacao/resolucoes/2023/1842-resolucao-760>
- [11] S. N. Antunes et al., "Model development for identifying aromatic herbs using object detection algorithm," *AgriEngineering*, vol. 6, no. 3, pp. 1924-1936, 2024, doi: 10.3390/agriengineering6030112.

- [12] S. N. Antunes, M. D. O. B. Divino, L. D. S. Cordeiro, F. P. L. Aguiar, and M. T. Okano, "Enhanced classification of aromatic herbs using EfficientNet and transfer learning," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 14, no. 5, pp. 4123-4136, Oct. 2025, doi: 10.11591/ijai.v14.i5.pp4123-4136.
- [13] S. N. Antunes, W. S. Lee, I. A. Nääs, and M. T. Okano, "Data-centric object detection for aromatic herbs: Image quality filtering, class selection, and model comparison," submitted for publication.
- [14] S. Reena and R. Rathinasabapathy, "A review on identification of plant leaf image classification using machine learning algorithms," *International Journal of Pattern Recognition and Artificial Intelligence*, to be published, doi: 10.1142/S0218001426310014.
- [15] R. Azadnia, M. M. Al-Amidi, H. Mohammadi, M. A. Cifci, A. Daryab, and E. Cavallo, "An AI based approach for medicinal plant identification using deep CNN based on global average pooling," *Agronomy*, vol. 12, no. 11, p. 2723, 2022, doi: 10.3390/agronomy12112723.
- [16] G. Tariku, I. Ghiglieri, G. Gilioli, F. Gentilin, S. Armiraglio, and I. Serina, "Automated identification and classification of plant species in heterogeneous plant areas using unmanned aerial vehicle-collected RGB images and transfer learning," *Drones*, vol. 7, no. 10, p. 599, 2023, doi: 10.3390/drones7100599.
- [17] M. Sharma, N. Kumar, S. Sharma, S. Kumar, S. Singh, and S. Mehandia, "Medicinal plants recognition using heterogeneous leaf features: an intelligent approach," *Multimedia Tools and Applications*, vol. 83, pp. 51513-51540, 2024, doi: 10.1007/s11042-023-17639-1.
- [18] S. O. Oppong, F. Twum, J. B. Hayfron-Acquah, and Y. M. Missah, "A novel computer vision model for medicinal plant identification using log-Gabor filters and deep learning algorithms," *Computational Intelligence and Neuroscience*, vol. 2022, Art. no. 1189509, 2022, doi: 10.1155/2022/1189509.
- [19] R. Gao, R. Wang, L. Feng, Q. Li, and H. Wu, "Dual-branch, efficient, channel attention-based crop disease identification," *Computers and Electronics in Agriculture*, vol. 190, p. 106410, 2021, doi: 10.1016/j.compag.2021.106410.
- [20] S. Islam et al., "BDMediLeaves: A leaf images dataset for Bangladeshi medicinal plants identification," *Data in Brief*, vol. 50, p. 109488, 2023, doi: 10.1016/j.dib.2023.109488.
- [21] L. Picek, M. Šulc, Y. Patel, and J. Matas, "Plant recognition by AI: Deep neural nets, transformers, and kNN in deep embeddings," *Frontiers in Plant Science*, vol. 13, Art. no. 787527, 2022, doi: 10.3389/fpls.2022.787527.
- [22] J. G. Thanikkal, A. K. Dubey, and M. T. Thomas, "An efficient mobile application for identification of immunity boosting medicinal plants using shape descriptor algorithm," *Wireless Personal Communications*, vol. 131, pp. 1189-1205, 2023, doi: 10.1007/s11277-023-10476-3.
- [23] S. Kavitha, T. Satish Kumar, E. Naresh, V. H. Kalmani, K. D. Bamane, and P. K. Pareek,

"Medicinal plant identification in real-time using deep learning model," *SN Computer Science*, vol. 5, no. 1, Art. no. 73, 2024, doi: 10.1007/s42979-023-02398-5.

[24] T. Zhang, X. Zhang, P. Zhu, X. Jia, X. Tang, and L. Jiao, "Generalized few-shot object detection in remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 195, pp. 353-364, 2023, doi: 10.1016/j.isprsjprs.2022.12.004.

[25] S. Wang, D. Chen, J. Xiang, and C. Zhang, "A deep-learning-based detection method for small target tomato pests in insect traps," *Agronomy*, vol. 14, no. 12, p. 2887, 2024, doi: 10.3390/agronomy14122887.

[26] M.-A. Dinca, D. Popescu, L. Ichim, N. Angelescu, and C. M. Pinotti, "Decision fusion-based system to detect two invasive stink bugs in orchards," *Smart Agricultural Technology*, vol. 9, p. 100548, 2024, doi: 10.1016/j.atech.2024.100548.

[27] X. Yu, Y. Chen, G. Zhou, Y. Liu, F. Li, and Z. Wang, "Error refactor loss based on error analysis in image classification," *IET Computer Vision*, vol. 16, no. 2, pp. 192-203, 2022, doi: 10.1049/cvi2.12079.

[28] M. Mattina, A. Benziou, K. Nasreddine, and F. Richard, "An efficient anchor-free method for pig detection," *IET Image Processing*, vol. 17, no. 3, pp. 613-626, 2023, doi: 10.1049/ipr2.12659.

[29] B. Feuer et al., "Zero-shot insect detection via weak language supervision," *The Plant Phenome Journal*, vol. 7, no. 1, Art. no. e20107, 2024, doi: 10.1002/ppj2.20107.

[30] J.-W. Chen, W.-J. Lin, H.-J. Cheng, C.-L. Hung, C.-Y. Lin, and S.-P. Chen, "A smartphone-based application for scale pest detection using multiple-object detection methods," *Electronics*, vol. 10, no. 4, p. 372, 2021, doi: 10.3390/electronics10040372.

[31] S. Yoon, J. S. Jo, S. B. Kim, H. S. Sim, S. K. Kim, and D. S. Kim, "Prediction of strawberry yield based on receptacle detection and Bayesian inference," *Heliyon*, vol. 9, no. 4, p. e14546, 2023, doi: 10.1016/j.heliyon.2023.e14546.

[32] H. Li, Q. Wang, L. Ling, Z. Lv, Y. Liu, and M. Jiao, "Research on recognition of coal and gangue based on laser speckle images," *Sensors*, vol. 23, no. 22, p. 9113, 2023, doi: 10.3390/s23229113.

[33] Y. Haruna, S. Qin, and M. J. Mbyamm Kiki, "An improved approach to detection of rice leaf disease with GAN-based data augmentation pipeline," *Applied Sciences*, vol. 13, no. 3, p. 1346, 2023, doi: 10.3390/app13031346.

[34] M. Costa, D. Sobien, R. Garg, W. Cheung, J. Krometis, and J. A. Kauffman, "Uncertainty quantification in data fusion classifier for ship-wake detection," *Remote Sensing*, vol. 16, no. 24, p. 4669, 2024, doi: 10.3390/rs16244669.

[35] J. Dong, A. Fuentes, S. Yoon, H. Kim, Y. Jeong, and D. S. Park, "A new deep learning-based dynamic paradigm towards open-world plant disease detection," *Frontiers in Plant Science*, vol. 14, Art. no. 1243822, 2023, doi: 10.3389/fpls.2023.1243822.

- [36] Y. Meng, M. Xu, H. Kim, S. Yoon, Y. Jeong, and D. S. Park, "Known and unknown class recognition on plant species and diseases," *Computers and Electronics in Agriculture*, vol. 215, p. 108408, 2023, doi: 10.1016/j.compag.2023.108408.
- [37] K. Jiang, J. You, U.-O. Dorj, H. Kim, and J. Lee, "Detection of unknown strawberry diseases based on OpenMatch and two-head network for continual learning," *Frontiers in Plant Science*, vol. 13, Art. no. 989086, 2022, doi: 10.3389/fpls.2022.989086.
- [38] A. Fuentes, S. Yoon, T. Kim, and D. S. Park, "Open set self and across domain adaptation for tomato disease recognition with deep learning techniques," *Frontiers in Plant Science*, vol. 12, Art. no. 758027, 2021, doi: 10.3389/fpls.2021.758027.
- [39] T. Fang, Z. Li, J. Zhang, D. Qi, and L. Zhang, "Open-set recognition of wood species based on deep learning feature extraction using leaves," *Journal of Imaging*, vol. 9, no. 8, p. 154, 2023, doi: 10.3390/jimaging9080154.
- [40] F. Gao, W. Zhong, Q. Jiang, X. Peng, and Z. Li, "OWFD-UCPM: An open-world fault diagnosis scheme based on uncertainty calibration and prototype management," *Knowledge-Based Systems*, vol. 286, p. 111403, 2024, doi: 10.1016/j.knosys.2024.111403.
- [41] S. G. Kim, Y. H. Chae, and S. R. Koo, "Application of an open-set recognition method for detecting untrained accident scenarios in a nuclear power plant accident diagnosis model," *Nuclear Engineering and Design*, vol. 427, p. 113421, 2024, doi: 10.1016/j.nucengdes.2024.113421.
- [42] G. Allabadi, A. Lucic, Y.-X. Wang, and V. Adve, "Learning to detect novel species with SAM in the wild," *International Journal of Computer Vision*, vol. 133, pp. 2247-2258, 2025, doi: 10.1007/s11263-024-02234-0.
- [43] B. A. Kumar and R. Krishna D, "Research survey on open set recognition for medicinal plant classification: Discussion on methodologies, datasets, and challenges," *Operations Research Forum*, vol. 7, no. 1, Art. no. 11, 2025, doi: 10.1007/s43069-025-00591-1.
- [44] P. Khodabakhshi, M. Mahootchi, and H. Mosadegh, "A novel hybrid clustering and classification framework for multi-level depression detection in social media texts," *Engineering Applications of Artificial Intelligence*, vol. 160, p. 111952, 2025, doi: 10.1016/j.engappai.2025.111952.
- [45] A. Saeed, M. U. Akram, M. Khattak, and M. B. Khan, "An interpretable hybrid framework combining convolution latent vectors with transformer based attention mechanism for rolling element fault detection and classification," *Heliyon*, vol. 10, no. 21, p. e38993, 2024, doi: 10.1016/j.heliyon.2024.e38993.
- [46] D. Khosla, Y. Chen, and K. Kim, "A neuromorphic system for video object recognition," *Frontiers in Computational Neuroscience*, vol. 8, Art. no. 147, 2014, doi: 10.3389/fncom.2014.00147.
- [47] W. Chen, S. Meng, and Y. Jiang, "Foreign object detection in railway images based on an efficient two-stage convolutional neural network," *Computational Intelligence and Neuroscience*, vol. 2022, Art. no. 3749635, 2022, doi: 10.1155/2022/3749635.

- [48] P. Gupta and R. S. Jadon, "PlantDetectNet: a hybrid IoT and deep learning framework for secure plant disease detection and classification," *Evolving Systems*, vol. 16, Art. no. 55, 2025, doi: 10.1007/s12530-025-09685-x.
- [49] S. Karpakam and N. Kumareshan, "A multi-modality framework for precise brain tumor detection and multi-class classification using hybrid GAN approach," *Biomedical Signal Processing and Control*, vol. 104, p. 107559, 2025, doi: 10.1016/j.bspc.2025.107559.
- [50] A. S. N. Raju et al., "Colorectal cancer detection with enhanced precision using a hybrid supervised and unsupervised learning approach," *Scientific Reports*, vol. 15, Art. no. 3180, 2025, doi: 10.1038/s41598-025-86590-y.
- [51] O. Appiah, K. O. Hackman, B. A. A. Diallo, K. O. Ogunjobi, S. Diakalia, O. Valentin, D. Abdoul-Karim, and G. Dabire, "PlanteSaine: An artificial intelligent empowered mobile application for pests and disease management for maize, tomato, and onion farmers in Burkina Faso," *Agriculture*, vol. 14, no. 8, p. 1252, 2024, doi: 10.3390/agriculture14081252.
- [52] Y. Chung, C.-A. Chou, and C.-Y. Li, "Central attention and a dual path convolutional neural network in real-world tree species recognition," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, p. 961, 2021, doi: 10.3390/ijerph18030961.
- [53] N. Chamara, G. Bai, and Y. Ge, "AICropCAM: Deploying classification, segmentation, detection, and counting deep-learning models for crop monitoring on the edge," *Computers and Electronics in Agriculture*, vol. 215, p. 108420, 2023, doi: 10.1016/j.compag.2023.108420.
- [54] D. Wang et al., "A real-time mobile aquatic plant recognition algorithm based on deep learning for intelligent ecological monitoring," *Scientific Reports*, vol. 16, Art. no. 5075, 2026, doi: 10.1038/s41598-026-35310-1.
- [55] J. Deng et al., "Deep-learning-based rice disease and insect pest detection on a mobile phone," *Agronomy*, vol. 13, no. 8, p. 2139, 2023, doi: 10.3390/agronomy13082139.
- [56] N. S. Ouf, "Leguminous seeds detection based on convolutional neural networks: Comparison of Faster R-CNN and YOLOv4 on a small custom dataset," *Artificial Intelligence in Agriculture*, vol. 8, pp. 30-45, 2023, doi: 10.1016/j.aiia.2023.03.002.
- [57] StatCounter Global Stats, "Mobile operating system market share Brazil." [Online]. Available: <https://gs.statcounter.com/os-market-share/mobile/brazil>. [Accessed: Jun. 20, 2026].

APPENDIX A – Paper I



AgriEngineering



Article

Model Development for Identifying Aromatic Herbs Using Object Detection Algorithm

Samira Nascimento Antunes ¹, Marcelo Tsugio Okano ^{1,*}, Irenilza de Alencar Nääs ¹, William Aparecido Celestino Lopes ¹, Fernanda Pereira Leite Aguiar ¹, Oduvaldo Vendrametto ¹, João Carlos Lopes Fernandes ¹ and Marcelo Eloy Fernandes ²

¹ Graduate Program in Production Engineering, Universidade Paulista, R. Dr. Bacelar 1212, São Paulo 04026-002, Brazil; samira.antunes@aluno.unip.br (S.N.A.); irenilza.naas@docente.unip.br (I.d.A.N.); william.lopes12@aluno.unip.br (W.A.C.L.); fernanda.aguiar11@aluno.unip.br (F.P.L.A.); oduvaldo.vendrametto@docente.unip.br (O.V.); joao.fernandes1@docente.unip.br (J.C.L.F.)

² Faculdade de Tecnologia de Barueri, CEETEPS, Barueri 06401-136, Brazil; marcelo.fernandes3@fatec.sp.gov.br

* Correspondence: marcelo.okano1@docente.unip.br

Abstract: The rapid evolution of digital technology and the increasing integration of artificial intelligence in agriculture have paved the way for groundbreaking solutions in plant identification. This research pioneers the development and training of a deep learning model to identify three aromatic plants—rosemary, mint, and bay leaf—using advanced computer-aided detection within the You Only Look Once (YOLO) framework. Employing the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, the study meticulously covers data understanding, preparation, modeling, evaluation, and deployment phases. The dataset, consisting of images from diverse devices and annotated with bounding boxes, was instrumental in the training process. The model's performance was evaluated using the mean average precision at a 50% intersection over union (mAP50), a metric that combines precision and recall. The results demonstrated that the model achieved a precision of 0.7 or higher for each herb, though recall values indicated potential over-detection, suggesting the need for database expansion and methodological enhancements. This research underscores the innovative potential of deep learning in aromatic plant identification and addresses both the challenges and advantages of this technique. The findings significantly advance the integration of artificial intelligence in agriculture, promoting greater efficiency and accuracy in plant identification.

Keywords: aromatic herb; convolutional neural network; deep learning; computer vision; YOLO v8



Citation: Antunes, S.N.; Okano, M.T.; Nääs, I.d.A.; Lopes, W.A.C.; Aguiar, F.P.L.; Vendrametto, O.; Fernandes, J.C.L.; Fernandes, M.E. Model Development for Identifying Aromatic Herbs Using Object Detection Algorithm. *AgriEngineering* **2024**, *6*, 1924–1936. <https://doi.org/10.3390/agriengineering6030112>

Academic Editors: Yee Siang Gan and Sze-Teng Liong

Received: 28 April 2024

Revised: 18 June 2024

Accepted: 19 June 2024

Published: 21 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Since ancient times, aromatic herbs such as mint, basil, and rosemary have enhanced food flavor and aroma and prolonged shelf life due to their antiseptic properties [1]. These herbs play an integral role in daily life, with culinary and medicinal applications. Aromatic herbs are typically small plants whose leaves emit distinct aromas, occasionally leading to confusion with medicinal herbs. Although there are similarities, medicinal herbs have a broader range of uses involving various plant parts, whereas aromatic herbs are primarily valued for their leaves.

Aromatic plants have long been utilized in medicine, food preservation, seasoning, and religious ceremonies. The diversity among aromatic herbs is extensive, and their visual differentiation poses significant challenges. The enhanced visual characterization of these herbs can substantially benefit wholesalers, retailers, farmers, cooperatives, importers, exporters, and the agro-industry. This improved understanding facilitates accurately identifying desired herb types for purchase, sale, consumption, and registration purposes for research endeavors [2,3].

With the increasing prevalence of digital technologies and digital transformation, efforts have been made to observe how these technologies can help produce aromatic herbs.

In the digital transformation scenario, it is observed that new digital technologies facilitate improvements in business processes, simplify operations, and create new business models. Object detection and categorization in images is one of the most significant problems in computer vision and related fields [4]. Additionally, the process becomes more intricate due to the various perspectives, sizes, angles, perspectives, occlusions, and illumination.

Previous investigations [2,5,6] have attempted different approaches to improve the accuracy and efficiency of herb identification from images. However, there have been limitations regarding detecting the shape of herbs, a critical characteristic for identifying the leaf family and removing the background. These issues were investigated previously, but have not been solved.

One way to ensure better accuracy in image detection is with convolutional neural networks (CNNs), which, like a neural network in the human brain, is a circuit with connections of neurons with weights between nodes, where a positive weight represents a stimulating connection, while negative values represent inhibitory connections. With the development of technology, deep learning enables the development of an image object detection model. CNNs are amongst the most widely utilized deep learning models for image detection and classification, owing to their superior accuracy to other machine learning algorithms [7].

Owing to the substantial increase in the utilization of digital images, deep learning, an artificial intelligence (AI) method, through an image object detection scheme, can enhance customer experience by identifying diverse types of horticultural products in the supply chain, such as aromatic herbs. The research question is as follows: to what extent can a deep learning model employing the You Only Look Once (YOLO) v8 framework effectively differentiate and accurately identify various aromatic herbs, including rosemary, mint, and bay leaf?

The current study aimed to develop and present a procedure for identifying three types of aromatic herbs using object detection algorithms through YOLO v8 architecture. The present study is organized as follows: In addition to this introduction, Section 2 presents the background, Section 3 describes the methodology used, and Section 4 presents the results and summarizes the discussion. Finally, Section 5 presents the conclusion and future research.

2. Background

Object detection is a component of computer vision that involves recognizing and positioning objects in videos or images [8,9]. Object class detection is usually based on a set of features, meaning each object of a specific class has certain characteristics based on which it is classified into different classes. Deep learning is a machine learning method that allows machines to recognize patterns in data [9]. Given a large dataset of labeled examples, it enables the learning of the most predictive features directly from images. Within the machine learning scenario, CNNs have become extremely popular and effective [10]. A CNN comprises a series of layers, and each layer receives the previous layer's output as its input. The input plane receives images of characters with a standardized size and centeredness, and each layer's units receive input from a set of smaller units located in the previous layer [11]. Two unique attributes of the architecture of CNNs are their sparse connections and shared weights [12]. CNNs exploit spatially local correlation by imposing a pattern of local connectivity between neurons of adjacent layers, where units in layer m are connected to three spatially adjacent units in layer $m - 1$. Every convolutional filter in the layer is repeated across the entire layer, sharing the same weights and biases. Such an approach will help CNNs to generalize the identification of problems more effectively while also increasing the learning efficiency by reducing the number of free parameters to be learned.

Object detection is a fundamental and long-standing problem in computer vision. Its task is to generate bounding boxes (in 2D pixel coordinates) for detected objects in an image that belongs to pre-specified object classes and assign classification scores to them [13]. It has been observed that object detection has improved significantly due to deep

neural networks, with convolutional neural networks (CNNs) being the most prominent architectures. The window that slips between classifiers and single-shot convolutional neural networks generates similar high-density windows around a particular object's correct location. Non-maximum suppression (NMS) chooses the single most significant candidate within the cluster of detection results for each object in each image. Therefore, non-maximum suppression (NMS) is analogous to a traditional clustering problem and generally depends on two primary operations: determining the cluster assignment for each detection and identifying a representative for each cluster [14]. Sun et al. [11] explain how NMS eliminates redundant boxes and selects the candidate box that best represents the object. Meanwhile, the remaining candidate box is called the object representative. Among the various algorithms widely used in various object detector models, YOLO stands out.

Through a procedure called backpropagation, a CNN learns the weights and biases of the kernel from a collection of input images [15]. These values serve as parameters that encapsulate essential characteristics of the images, independent of their location. These kernel weights slide through an input image performing element-wise dot products, producing intermediate results that are subsequently summed with the learned bias value. Then, each neuron obtains an output based on the input image. These outputs are also called activation maps. To reduce the number of parameters and prevent overfitting, CNNs decrease the resolution of inputs using a different layer called pooling. Activation functions are employed in a CNN to augment the model with nonlinearity. This enables the model to learn more intricate patterns in data.

2.1. You Only Look Once (YOLO)

YOLO is an object detection approach that aims to predict which objects are present and where they are by looking at the image only once [16]. In recent years, there has been a rise in research articles addressing object identification using the YOLO algorithm. However, we found a few articles about using YOLO to detect herbs. There has been a theoretical basis for recognizing the rhizome and main root in plants through "internal content difference" and "external morphological difference" [17]. Weed detection was previously performed to assist laser weed removal robots [18]. Previous studies [3,19] determined YOLO's adaptability and efficiency in plant and herb identification through image analysis, leveraging its speed and accuracy in object detection tasks. The technology is beneficial for enhancing agricultural practices and scientific research by automating the extraction of detailed trait data from images.

Recent advances in research have demonstrated the effectiveness of image recognition systems based on YOLO. A notable example is the study by [20], which presents the implementation of an apple detection system using YOLO v3, and which was later integrated into a harvesting manipulator robot. The developed system improved apple identification accuracy under various lighting and obstruction conditions. Furthermore, the trajectory of the manipulator robot was optimized using reinforcement learning techniques, resulting in a 92% increase in accuracy in laboratory tests. Under controlled conditions, the system achieved a 100% recognition rate. These improvements concur with the increasing popularity of utilizing convolutional neural networks (CNNs) in agriculture, proving effective in solving complex computer vision tasks.

Similarly, a previous study was conducted [21] by applying YOLO v3 with advanced pre- and post-processing techniques for apple identification, highlighting the evolution of computer vision algorithms to meet the demands of modern horticulture. The improvements implemented resulted in reduced error rates and average identification time, making the system more efficient and comprehensive than previous versions, such as the faster region-based convolutional neural network (RCNN) and dynamic activation sparsity network (DaSNet) v2. This study highlights the importance of computer vision in agricultural automation.

On the other hand, the study by [22] discusses aspects related to object detection in images, covering everything from likelihood relationships to using scalable and efficient

neural networks. Given the increasing relevance of object identification and recognition tasks in images, the authors explore the development of convolutional neural networks to solve recognition problems. The study includes the evaluation of several pre-trained architectures, considering metrics such as error probabilities, precision, detection recall, intersection over union, and average interpolated precision.

Mirzaei et al. [23] contribute a comprehensive review on small object detection and tracking, addressing challenges and methods in computer vision. The study analyzes small object detection techniques, including methods based on CNNs, image processing algorithms, and machine learning techniques. This review article offers a detailed analysis of the current literature, serving as a valuable resource for researchers and professionals involved in developing small object detection and tracking systems.

2.2. YOLO v8

The YOLO v8 architecture consists of three main components: a backbone, a neck, and a decoupled head. The backbone employs an EfficientRep block, a convolutional neural network (CNN) designed to extract features from the input image. The neck, called the C2f module (C2f module refers to an optimized version of the CSP—cross stage partial bottleneck—with two convolutions, known as the C2 module), integrates these features from various depths of the backbone to generate a more detailed image representation. Finally, the decoupled head processes this fused feature map to estimate the size of the boxes surrounding the object in the image and the probability of the object being classified as a specific type [24], see Figure 1.

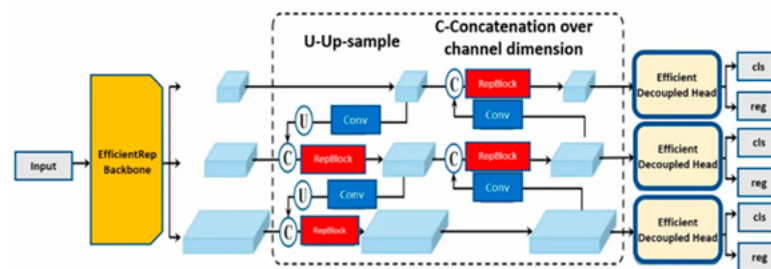


Figure 1. YOLO v8 architecture, adapted from [24].

The backbone is a lightweight CNN optimized for efficiency and accuracy. It comprises multiple convolutional layers, each designed to extract different levels of abstraction from the input image. Although the specific details of these convolutional layers are not depicted in the provided image, they are extensively documented in the YOLO v8 literature [24]. The neck, or C2f module, is a novel architectural feature introduced in YOLO v8. It aggregates feature maps from different depths of the backbone using a channel concatenation operation. This concatenation enables the network to learn relationships between features at various levels of abstraction, thereby enhancing the model's accuracy.

The decoupled head is tasked with generating the final predictions. It utilizes the fused feature map from the neck to predict bounding boxes and class probabilities for the objects in the image. The decoupled head includes several convolutional layers, culminating in one layer that predicts bounding boxes and another that predicts class probabilities.

Overall, the YOLO v8 architecture represents an advanced and efficient design for object detection. The EfficientRep backbone achieves an optimal balance between accuracy and speed, the C2f module facilitates the learning of richer feature representations, and the decoupled head ensures precise predictions of objects within the image. Research on aromatic herb identification has evolved significantly in recent years, addressing different approaches and methodologies. In this section, we review the main studies that form the

basis of our work, highlighting the most relevant contributions and the gaps that persist in the literature.

2.3. Related Works

Research on aromatic herb identification has evolved significantly in recent years, addressing different approaches and methodologies. In this section, we discuss the primary investigations that form the basis of our work, highlighting the most relevant contributions and the gaps that persist in the literature.

Che Soh et al. [25] investigated the process of herb identification, which is accomplished through organoleptic methods and relies heavily on botanical compounds; this makes it difficult to differentiate different species of herbs based on their aroma alone because of their shared physical attributes and aroma. Artificial technology is different from humans; unlike humans, it is thought to have the capacity to differentiate different species precisely. The authors developed an electronic nose that intended to recognize the scent of 12 species, including the Myrtaceae, Lauraceae, and Zingiberaceae families. The output received by electronic sensors for nasal gas was categorized using two different artificial intelligence techniques: ANN and ANFIS. From the outcome, ANFIS has a percentage of 94.8 compared to ANN's percentage of 91.7.

Previous research [26] studied the effectiveness of deep machine learning algorithms in classifying medicinal plant leaves from images. The investigation sought to create effective and accurate methods for differentiating between species of plants and diseases, using advanced technologies such as convolutional neural networks (CNNs) to improve the accuracy and reliability of plant recognition systems. The methodology included data collection and pre-processing steps, feature extraction, model development and training, and performance evaluation. The results indicated that using CNNs increased classification accuracy and efficiency, with high recognition rates and potential applications in environmental and agricultural monitoring.

Fernandes et al. [27] proposed a different approach, which was focused on creating an architecture that recognized 18 aromatic herbs using augmented reality and computer vision. The project and development of this architecture facilitated the overlay of virtual information onto the naturally identified environment, which contained an intuitive and simple-to-use mobile application. However, the primary obstacle was the creation of large databases for agricultural solutions with diverse and high-quality images. These images augmented the real world and provided solutions based on AI and neural networks for future endeavors.

Duth et al. [28] developed a classification system for medicinal herb leaves using digital image processing and machine learning techniques. The technologies associated with convolutional neural networks (CNNs), such as recurrent neural networks (RNNs), long short-term memory (LSTM), and object detection models like the region-based convolutional neural network (RCNN), Fast RCNN, and Faster RCNN, were employed. These technologies were employed to extract relevant features from leaf images and improve classification accuracy and speed. The proposed system achieved high accuracy and robustness in identifying different species of medicinal herb leaves. Compared to other existing techniques, the proposed method demonstrated effectiveness and accuracy in identifying the leaves of selected medicinal herbs.

In summary, although these studies provide a solid foundation for our research, there is a clear need for technologies to identify aromatic herbs. Our work aimed to develop and present a procedure for identifying three types of aromatic herbs using object detection algorithms through YOLO v8 architecture.

Training a network specifically for aromatic or medicinal herbs is crucial due to their visual similarity, which makes accurate identification challenging. Such a model enhances agricultural practices by improving crop management and harvesting, automating identification processes to reduce labor and increase efficiency, and enabling data-driven decision-making for better sustainability. Additionally, it ensures product authenticity for

consumers, fosters trust and satisfaction, and supports scientific research on plant characteristics and genetic diversity, leading to potential discoveries. These benefits highlight the importance of specialized deep learning models in agricultural and market operations.

3. Methods

The YOLO v8 model was employed for the real-time detection of three types of aromatic herbs (rosemary, mint, and bay), using a convolutional neural network architecture to process entire images and apply non-maximum suppression to refine predictions. Through training with a diverse set of annotated images, the model achieved mAP50 and precision measurements above 0.7, although it demonstrated recall values above 0.5, suggesting a trend towards over-detection. It was decided not to incorporate improvements to the YOLO v8 model, maintaining the original primary structure without changes. The decision was to assess the performance of the unmodified standard model in accurately identifying and categorizing aromatic herbs, solely utilizing the training dataset without any further adjustments to its architecture or hyperparameters.

The YOLO v8 model, with its advanced architecture and real-time processing capabilities, proved effective in identifying aromatic herbs. By following the CRISP-DM methodology, the study ensured a systematic approach to model development, from understanding the business requirements to deploying a robust solution in a practical setting.

The investigation intended to create a model for identifying some herbs commercialized at the largest Brazilian center for food distribution, the Company of Warehouses and General Warehouses of São Paulo (CEAGESP). Considering the proposed research problem, the study followed the Cross Industry Standard Process for Data Mining (CRISP-DM) model [29]. Table 1 presents the research stages.

Table 1. Summary of procedures performed using the CRISP-DM model.

CRISP-DM Stage	Activity	Tool/Method
Business Understanding	Identify experts within the organization. Gather and outline needs and expectations. Verify the existence of image databases within the organization.	Visits and meetings with representatives from CEAGESP.
Data Understanding	Understand the available data. Evaluate the quality of the available data. Verify if the volume of data meets business needs.	Analysis of the formats and characteristics of the three selected herbs.
Data Preparation	Select data for analysis. Cleanse the data. Format the data appropriately. Select the model training environment. Choose the tool for creating bounding boxes. Understand the model metrics.	Creation of the database using photos (in JPG format) obtained from different cell phone models, NIKON cameras, Samsung tablets, and websites. Use of Google Cloud via Colab Pro and Colab Pro+ products. Utilization of the CVAT [30].
Modeling	Select the appropriate technique for modeling.	Utilization of the YOLO V8 framework
Evaluation	Analyze performance metrics.	Analysis of the model's recognition accuracy.
Deployment	Prepare the analysis report.	Analyze the results.

The selected herbs for training were rosemary, mint, and bay leaf. According to the botanical description [31–33], rosemary can have opposite leaves, lack petioles, and be simple, linear, and leathery, with stellate hairs on the underside that are whitish in color and dark green on the upper surface. Mint leaves are described as having opposite leaves, provided with short petioles, oblong to oval in shape, with toothed margins. Bay leaf has persistent, petiolate, alternate leaves, elliptical or lanceolate, leathery with undulated, entire smooth edges, bright green on the upper surface, and pale green on the lower surface.

Table 2 shows the leaves of the aromatic herbs used in the study. Following this instruction, bounding boxes were created on the images used for training.

Table 2. Reference of the trained aromatic herbs.

Aromatic Herb	Leaf Reference
Rosemary	
Mint	
Bay Leaf	

After consolidating the images, a cleaning process was conducted where images with significant blur were discarded. Finally, for each database, 10% of the images were selected for the validation process and 90% of the images were selected for the training process, resulting in the following:

- Rosemary leaf: 27 images for validation and 234 images for training.
- Mint leaf: 35 images for validation and 302 images for training.
- Bay leaf: 31 images for validation and 262 images for training.

The storage of the image database on the Google Cloud provider was utilized due to its connection with the Collaboratory platform [30], which is also from Google. Individual databases and a single database containing all three aromatic herbs were created for each herb. The classes for the single database containing the three aromatic herbs were classified as bay leaf (0), rosemary (1), and mint (2). The bounding boxes were created using the computer vision annotation tool [30].

The selection between the validation database and the test database was made randomly. In summary, the image database was analyzed and, taking care to prevent very similar images from being trained and validated, to avoid overfitting, the images were manually separated.

The images obtained during a visit to CEAGESP are herbs that have suppliers from different cities in São Paulo, Brazil. Regarding the maturity of the herbs, there was no specific selection concerning this requirement, as the focus has always been on identifying the herb regardless of its maturity. Consumers will face this condition daily; therefore, the database was not limited to maturity. The challenges regarding the identification of mint start from the definition of the bounding box; that is, as this herb has small leaves, it is necessary to define it by leaf or by the set of leaves, which are characteristics of the herb. When benchmarking with models available in Roboflow, there are these two situations.

The bay leaf is used fresh or dehydrated; therefore, like the other herbs in the study, they were not separated by maturity levels.

Furthermore, it has a variety of shapes as it originates from several tree species, with Loureiro and Umbellularia being the most common. Rosemary is a small evergreen shrub with thick aromatic leaves and many varieties. Regarding the identification challenges, the bounding boxes for each shrub are created.

The adopted performance metrics by class highlight explanations which are critical for understanding the model's performance for each class, especially in datasets with multiple object categories. For each class in the dataset, the following were calculated. (1) Precision (P): the accuracy of detected objects, indicating how many detections were correct (Equation (1)); (2) recall (R): the model's ability to identify all instances of objects in the images (Equation (2)); (3) mAP50: average precision (AP) is calculated at an intersection over union (IoU) threshold of 0.50. This measures the model's accuracy by considering only "easy" detections (Equation (3)); and (4) mAP50–95: the average of the average precision calculated at various IoU thresholds, ranging from 0.50 to 0.95. This provides a comprehensive view of the model's performance at different detection difficulty levels.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

where TP (true positives) = number of correct positive predictions, and FP (false positives) = number of incorrect positive predictions made by the model.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

where TP (true positives = number of correct positive predictions, and FN (false negatives) is the number of positives the model incorrectly predicted as negative.

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (3)$$

where N = the number of classes and AP = average precision.

As for the generated graphical measures, the following were adopted: (1) the F1 score curve represents the F1 score at various thresholds. Interpreting this curve can provide insights into the model's balance between false negatives and false positives at different thresholds; (2) the precision–recall curve, which is a graphical representation of the precision values at different thresholds; and (3) the recall curve, which is a graph that illustrates how recall values change at different thresholds.

In the training quantity, epochs are defined, meaning one epoch is a complete iteration of the entire training dataset in one cycle to train the model. During the epoch process, the model processes each training sample in the dataset, and its weights and biases are updated according to the calculated loss or error. The number of epochs is a hyperparameter that defines how often the learning algorithm will run through the entire training dataset. The training database is privately owned, and the programming logic for using YOLO v8 followed the instructions in the model document.

The model detects herbs individually, not by mixing them in a different group. This means that each herb is identified and analyzed on its own merits rather than being combined with other herbs to form a composite group. This approach allows for the more precise identification and analysis of each herb's unique characteristics and properties.

Training a deep learning model for identifying aromatic herbs substantially benefits from a comprehensive and high-quality training database. Such a database, comprising annotated images, enhances the model's accuracy and precision, reducing over-detection issues by providing more representative samples. It ensures the model's robustness against variations in herb appearance due to environmental factors and different cultivation practices. Additionally, it improves efficiency in agricultural operations by ensuring correct herb identification, thus streamlining processes in wholesale markets and retail.

4. Results and Discussion

The input of a neural network has weights, and these numbers are initialized randomly. In the learning process, these numbers are updated every time the neural network is stimulated, meaning the weights are adjusted towards the ideal value. There is no definition of a standard number of epochs for a neural network, but the present research aimed for a mAP50 greater than or equal to 0.7. In this scenario, a separate model was generated for each aromatic herb, considering 300 epochs. After this process, a single model was generated for all three herbs, also considering 300 epochs. Table 3 presents the performance metrics for each trained aromatic herb.

Table 3. Summary of training for the aromatic herb model.

Aromatic Herb	Epoch Quantity	Time	mAP50	mAP50-95	P	R
Rosemary	300	13 h	0.739	0.369	0.724	0.786
Mint	300	15 h	0.738	0.408	0.739	0.714
Bay Leaf	300	16 h	0.7	0.387	0.611	0.727

The graphical measures of the F1 curve are shown in Figure 2a–c.

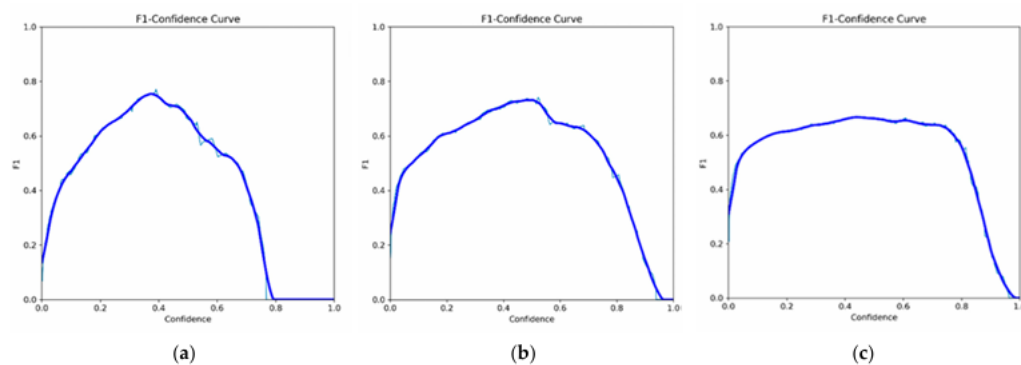


Figure 2. Graphical measures of the F1 curves for rosemary (a), mint (b), and bay leaf (c).

For the rosemary class, despite the mAP precision scores being above 0.7, it is noticeable that the recall measure exceeds the values of the previously mentioned measures, suggesting that negative objects are being incorrectly classified as positive. The F1 measure reinforces the need for improvement in detection, since it represents a balanced average of the model's ability to capture positive cases (recall) and to be precise with the cases it captures (precision). Additionally, it is noted that there is no need to change the number of epochs, since the mAP measure did not drop below 0.65 from 268 epochs onwards.

Although the recall measure for the mint class is below the precision and mAP measures, there is a slight difference. Therefore, it is understood that it follows the same behavior and improvements as the rosemary class. Additionally, the training occurred until 295 epochs, as there were no differences in the measures after this epoch.

Regarding the bay leaf class, the precision, mAP50, and F1 measures below the recall measure suggest that the object was not successfully identified with the trained sample. Concerning the variation in the mAP50 measure, after 192 epochs, the measure varied between greater than or equal to 0.6 and less than 0.7, meaning there were no measures above 0.7.

Google-enhanced computational resources (Colab Pro+) were used to train the model with the three aromatic herbs. Due to the machine's capacity and constant internet connection, the training occurred in six parts. Table 4 presents the performance metrics for the

trained consolidated model. According to Google, Colab Pro+ is a paid service offered by Google Colab, a cloud-based collaborative notebook platform. The central processing unit (CPU) is the general-purpose processing unit of a computer. While it is versatile and can handle a wide range of tasks, it may not be optimized for compute-intensive operations. On the other hand, the TPU (tensor processing unit) is custom-designed to accelerate machine learning workloads, particularly those that involve neural networks and large-scale data. Then, TPUs are highly specialized and can outperform GPUs in specific scenarios for training state-of-the-art deep learning models, especially when dealing with large datasets and complex neural networks in fields like natural language processing and computer vision.

Table 4. Summary of training for the consolidated model of aromatic herbs.

Step	Epoch Quantity	Time (h)	mAP50	mAP50–95	P	R
1	50	8	All: 0.480 Bay Leaf: 0.479 Rosemary: 0.516 Mint: 0.445	All: 0.221 Bay Leaf: 0.245 Rosemary: 0.203 Mint: 0.215	All: 0.531 Bay Leaf: 0.561 Rosemary: 0.513 Mint: 0.52	All: 0.507 Bay Leaf: 0.405 Rosemary: 0.750 Mint: 0.365
2	51	8	All: 0.691 Bay Leaf: 0.595 Rosemary: 0.823 Mint: 0.653	All: 0.321 Bay Leaf: 0.301 Rosemary: 0.328 Mint: 0.33	All: 0.672 Bay Leaf: 0.552 Rosemary: 0.862 Mint: 0.602	All: 0.724 Bay Leaf: 0.667 Rosemary: 0.821 Mint: 0.283
3	52	10	- All: 0.715 - Bay Leaf: 0.612 - Rosemary: 0.862 - Mint: 0.671	- All: 0.381 - Bay Leaf: 0.334 - Rosemary: 0.405 - Mint: 0.405	- All: 0.692 - Bay Leaf: 0.617 - Rosemary: 0.802 - Mint: 0.658	- All: 0.702 - Bay Leaf: 0.610 - Rosemary: 0.893 - Mint: 0.603
4	53	9	- All: 0.761 - Bay Leaf: 0.664 - Rosemary: 0.942 - Mint: 0.679	- All: 0.376 - Bay Leaf: 0.342 - Rosemary: 0.384 - Mint: 0.400	- All: 0.778 - Bay Leaf: 0.618 - Rosemary: 1 - Mint: 0.715	- All: 0.692 - Bay Leaf: 0.602 - Rosemary: 0.871 - Mint: 0.603
5	54	11	- All: 0.737 - Bay Leaf: 0.642 - Rosemary: 0.871 - Mint: 0.698	- All: 0.383 - Bay Leaf: 0.349 - Rosemary: 0.408 - Mint: 0.393	- All: 0.743 - Bay Leaf: 0.655 - Rosemary: 0.857 - Mint: 0.718	- All: 0.695 - Bay Leaf: 0.593 - Rosemary: 0.857 - Mint: 0.635
6	40	8	- All: 0.723 - Bay Leaf: 0.619 - Rosemary: 0.884 - Mint: 0.666	- All: 0.395 - Bay Leaf: 0.332 - Rosemary: 0.439 - Mint: 0.414	- All: 0.754 - Bay Leaf: 0.661 - Rosemary: 0.919 - Mint: 0.682	- All: 0.664 - Bay Leaf: 0.593 - Rosemary: 0.813 - Mint: 0.587

In the training of the three classes, it is noted that, concerning the measures presented, from the third training onwards, the variations in the measures were below 0.05. The recall and precision results showed the same characteristics as the individual training, i.e., values close together, indicating that the model is deficient in capturing positive cases (recall) and in being precise with the cases it captures (precision).

The class related to the aromatic herb rosemary presented mAP50, precision, and recall values above 0.8; thus, it was possible to recognize rosemary images during testing. However, when testing with an image of the herb chive, confusion in identification was noted, indicating an opportunity to add the chive class to the model, improve the database for the rosemary herb, and redo the bounding boxes, considering the herb, not the bunch.

The class related to the herb mint presented mAP50 and precision values close to 0.65, but the recall measure was also close to 0.6. The objective was to achieve a mAP50 value above 0.7. Considering the results of the other measures, one can understand the recognition confusion in the testing stage and the opportunities for improvement in the database for the mint herb and redo the bounding boxes. The class related to the bay leaf presented the worst performance measures. Despite the precision value of 0.66, the mAP50 and

recall values were close to 0.6. Therefore, there were various recognition confusions when considering different untrained aromatic herbs during the testing phase. Thus, there is an opportunity to improve the database for the bay leaf herb and redo the bounding boxes.

The challenges previously raised [8,18,34] were noted in the method stage, as image recognition became challenging due to variations in the aromatic herb's size, angle, and quality. These proportions or new or unusual configurations characteristic of aromatic herbs represent a challenge regarding the volume and diversity of the database to be used for training. Most difficulties we found in the present study, such as overlay detection and false positive detections, were also perceived by Weaver and Smith [32] when digitalizing an herbarium. We recommend increasing the number of images to overcome this issue.

After analyzing the results, we understand that the research's problem and objective have been achieved. Despite the opportunities for improvement in recognition through the improvement of the database and bounding boxes, it is noted that it is possible to train a deep learning model to identify diverse types of aromatic herbs.

Concerning the models' results for each herb and the model for the classes, there is an opportunity for improvement in the model training process for all aromatic herbs [17,18,35]. A model that does not produce false negatives, i.e., a model in which no undetected bounding boxes should be detected, has a recall of 1.0. Therefore, it is essential to classify negative samples. In the current approach, the mAP compares the ground truth bounding box with the detected bounding box. It returns a score, emphasizing the improvements to be implemented for a new training session.

Although this study was performed using the internet in real-time, there is an opportunity for training without internet dependency regarding the training environment when the training is conducted in an environment connected to the internet and from a cloud service provider. Thus, in addition to the cost of using this provider's environment, there were challenges regarding the constant connection of the environment. Internet speed does not influence training, but possible connection fluctuations do, because, as the training took place using resources from a public cloud, connection drops cause interruptions and continuity in training.

5. Conclusions

We found a model to automatically detect herbs (rosemary, mint, and bay leaf). The model with all classes showed mAP50 and precision measures greater than 0.7. However, it also showed values above 0.5 for the recall measure, indicating a possibility of over-detection. The recall measure indicates the ability of a model to find all the bounding boxes of the ground truth (positive samples); therefore, based on the results and discussions presented, the improvement of the database and bounding boxes stands out to better avoid noise in detection.

Concerning the database, as it was the first contact with the object under study and with the image identification process, the definition of the minimum number of images was presented in the study. Furthermore, no parameters were found to requalify the minimum quantity for training aromatic herbs in the literature. As mentioned, after analyzing the results, the need for database improvement is evident and is in progress.

Author Contributions: Conceptualization, S.N.A., W.A.C.L., J.C.L.F. and M.T.O.; methodology, S.N.A. and M.T.O.; software, S.N.A.; validation, W.A.C.L., J.C.L.F. and M.T.O.; formal analysis, M.E.F. and O.V.; investigation, S.N.A., W.A.C.L., J.C.L.F. and M.T.O.; resources, W.A.C.L., J.C.L.F., M.E.F., O.V. and M.T.O.; data curation, O.V.; writing—original draft preparation, S.N.A.; translation to English, F.P.L.A.; writing—review and editing, I.d.A.N.; visualization, M.T.O.; supervision, S.N.A.; project administration, M.T.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data will be available from the corresponding author.

Acknowledgments: We appreciate the support from CEAGESP (Companhia de Entrepósitos e Armazéns Gerais de São Paulo). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Fontes, V.; Pereira, D.C.; Pupin, B.; Sakane, K.K. Aplicação de Espectroscopia no Infravermelho: Como ferramenta para análise quantitativa de orégano. *Rev. Univap* **2020**, *26*, 15–25. [\[CrossRef\]](#)
- Carranza-Rojas, J.; Goeau, H.; Bonnet, P.; Mata-Montero, E.; Joly, A. Going deeper in the automated identification of Herbarium specimens. *BMC Evol. Biol.* **2017**, *17*, 1–14. [\[CrossRef\]](#)
- Triki, A.; Bouaziz, B.; Mahdi, W.; Gaikwad, J. Objects Detection from Digitized Herbarium Specimen based on Improved YOLO V3. In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Valletta, Malta, 27–29 February 2020; Volume 4, pp. 523–529. [\[CrossRef\]](#)
- Aamir, M.; Pu, Y.F.; Rahman, Z.; Abro, W.A.; Naeem, H.; Ullah, F.; Badr, A.M. A Hybrid Proposed Framework for Object Detection and Classification. *J. Inf. Process. Syst.* **2018**, *14*, 1176–1194. [\[CrossRef\]](#)
- Munisami, T.; Ramsurn, M.; Kishnah, S.; Pudaruth, S. Plant leaf recognition using shape features and colour histogram with k-nearest neighbour classifiers. *Procedia Comput. Sci.* **2015**, *58*, 740–747. [\[CrossRef\]](#)
- Agarwal, S.; Jalal, A.; Khan, M. Plant Identification Using Leaf Image Analysis. In Proceedings of the 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 2018, Malaviya National Institute of Technology, Jaipur, India, 26–27 March 2018.
- Véstias, M.P. A Survey of Convolutional Neural Networks on Edge with Reconfigurable Computing. *Algorithms* **2019**, *12*, 154. [\[CrossRef\]](#)
- Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and validation of a Deep Learning Algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **2016**, *316*, 2402–2410. [\[CrossRef\]](#)
- Johnston, T.; Young, S.R.; Hughes, D.; Patton, R.M.; White, D. Optimizing Convolutional Neural Networks for Cloud Detection. In Proceedings of the Machine Learning on HPC Environments—MLHPC’17, New York, NY, USA, 12 November 2017; Volume 4, pp. 1–9. [\[CrossRef\]](#)
- Lecun, Y.; Bengio, Y. *The Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, MA, USA, 1995.
- Sun, K.; Li, Z.; Zheng, Y.; Kuo, H.-W.; Lee, K.-P.; Tang, K.-T. An Area-Efficient Accelerator for Non-Maximum Suppression. *IEEE Trans. Circuits Syst. II Express Briefs* **2023**, *70*, 2251–2255. [\[CrossRef\]](#)
- Symeonidis, C.; Mademlis, I.; Pitas, I.; Nikolaidis, N. Neural Attention-Driven non-maximum suppression for person detection. *IEEE Trans. Image Process.* **2023**, *32*, 2454–2467. [\[CrossRef\]](#)
- Oro, D.; Fernández, C.; Martorell, X.; Hernandez, J. Work-Efficient Parallel Non-Maximum Suppression Kernels. *Comput. J.* **2022**, *65*, 773–787. [\[CrossRef\]](#)
- Wang, Z.J.; Turko, R.; Shaikh, O.; Park, H.; Das, N.; Hohman, F.; Kahng, M.; Chau, D.H.P. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 1396–1406. [\[CrossRef\]](#)
- Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 7–12 June 2015; pp. 779–788. Available online: <https://arxiv.org/abs/1506.02640> (accessed on 10 May 2024).
- Ultralytics YOLOv8 Docs. Available online: <https://docs.ultralytics.com/> (accessed on 12 March 2024).
- Zhang, H.; Pan, Y.; Liu, X.; Chen, Y.; Gong, X.; Zhu, J.; Yan, J.; Zhang, H. Recognition of the rhizome of red ginseng based on spectral-image dual-scale digital information combined with intelligent algorithms. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2023**, *297*, 122742. [\[CrossRef\]](#) [\[PubMed\]](#)
- Fatima, H.S.; Hassan, I.U.; Hasan, S.; Khurram, M.; Stricker, D.; Afzal, M.Z. Formation of a Lightweight, Deep Learning-Based weed detection system for a commercial autonomous laser weeding robot. *Appl. Sci.* **2023**, *13*, 3997. [\[CrossRef\]](#)
- Kholiya, D.; Mishra, A.; Pandey, N.; Tripathi, N. Plant Detection and Counting using Yolo based Technique. In Proceedings of the 3rd Asian Conference on Innovation in Technology (ASIANCON), Pune, India, 25 August 2023; pp. 1–5. [\[CrossRef\]](#)
- Andriyanov, N. Development of Apple Detection System and Reinforcement Learning for Apple Manipulator. *Electronics* **2023**, *12*, 727. [\[CrossRef\]](#)
- Kuznetsova, A.; Maleva, T.; Soloviev, V. Using YOLOv3 Algorithm with Pre- and Post-Processing for Apple Detection in Fruit-Harvesting Robot. *Agronomy* **2020**, *10*, 1016. [\[CrossRef\]](#)
- Andriyanov, N.A.; Dementiev, V.E.; Tashlinskii, A.G. Detection of objects in the images: From likelihood relationships towards scalable and efficient neural networks. *Comput. Opt.* **2022**, *46*, 139–159. [\[CrossRef\]](#)
- Mirzaei, B.; Nezamabadi-pour, H.; Raoof, A.; Derakhshani, R. Small Object Detection and Tracking: A Comprehensive Review. *Sensors* **2023**, *23*, 6887. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hussain, M. YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection. *Machines* **2023**, *11*, 677. [\[CrossRef\]](#)

25. Che Soh, A.; Yusof, U.K.; Radzi, N.F.M.; Ishak, A.J.; Hassan, M.K. Classification of aromatic herbs using artificial intelligent technique. *Pertanika J. Sci. Technol.* **2017**, *25*, 119.
26. Chanyal, H.; Yadav, R.K.; Saini, D.K.J. Classification of Medicinal Plants Leaves Using Deep Learning Technique: A Review. *Int. J. Intell. Syst. Appl. Eng.* **2022**, *10*, 78–87.
27. Fernandes, J.C.L.; Okano, M.T.; Lopes, W.A.C.; Antunes, S.N.; Vendrametto, O. An Architecture to Identify Aromatic Herbs using Augmented Reality (AR) and Mobile Application. *WSEAS Trans. Environ. Dev.* **2023**, *19*, 1459–1467. [CrossRef]
28. Duth, S.; Vedavathi, S.; Roshan, S. Herbal Leaf Classification using RCNN, Fast RCNN, Faster RCNN. In Proceedings of the 2023 7th International Conference on Computing, Communication, Control And Automation (ICCUBE), Pune, India, 18–19 August 2023; pp. 1–8.
29. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. *CRISP-DM 1.0: Step-By-Step Data Mining Guide*; Semantic Scholar Inc.: Seattle, DC, USA, 2000.
30. CVAT—Computer Vision Annotation Tool. Available online: <https://www.cvat.ai> (accessed on 12 November 2023).
31. Vaz, A.P.A.; Jorge, M.H.A. Alecrim. Corumbá, MS: Embrapa Pantanal, 2006. 1 p. (Plantas Medicinais, Codimentares e Aromáticas). Available online: <https://ainfo.cnptia.embrapa.br/digital/bitstream/CPAP/56607/1/FOL68.pdf> (accessed on 12 December 2023).
32. Vaz, A.P.A.; Jorge, M.H.A. Hortelã. Corumbá, MS: Embrapa Pantanal, 2006. Plantas Medicinais, Codimentares e Aromáticas. Available online: <https://ainfo.cnptia.embrapa.br/digital/bitstream/CPAP/56556/1/FOL90.pdf> (accessed on 12 December 2023).
33. Vaz, A.P.A.; Jorge, M.H.A. Louro. Corumbá, MS: Embrapa Pantanal, 2006. Plantas Medicinais, Codimentares e Aromáticas. Available online: <https://ainfo.cnptia.embrapa.br/digital/bitstream/CPAP/56112/1/FOL113.pdf> (accessed on 12 December 2023).
34. Cloud Google Colab. Available online: <https://colab.research.google.com/notebooks/intro.ipynb> (accessed on 12 December 2023).
35. Weaver, W.; Smith, S. From leaves to labels: Building modular machine learning networks for rapid herbarium specimen analysis with LeafMachine2. *Appl. Plant Sci.* **2023**, *11*, e11548. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

APPENDIX B – Paper II

IAES International Journal of Artificial Intelligence (IJ-AI)

Vol. 14, No. 5, October 2025, pp. 4123–4136

ISSN: 2252-8938, DOI: 10.11591/ijai.v14.i5.pp4123-4136

□ 4123

Enhanced classification of aromatic herbs using EfficientNet and transfer learning

Samira Nascimento Antunes, Madalena De Oliveira Barbosa Divino, Luana Dos Santos Cordeiro,
Fernanda Pereira Leite Aguiar, Marcelo Tsuguio Okano
Program of Production Engineering, Paulista University, São Paulo, Brazil

Article Info

Article history:

Received Sep 21, 2024

Revised Jun 20, 2025

Accepted Jul 10, 2025

Keywords:

Aromatic herbs
Classification
Convolutional neural network
Deep learning
EfficientNet
Image identification
Plant identification

ABSTRACT

Herbs have long been used for culinary and medicinal purposes, as well as in religious rituals, due to their essential oils and aromatic properties. However, distinguishing between aromatic and medicinal herbs based on visual characteristics alone can be challenging. With recent advances in computer vision, plant identification from images has seen significant growth, offering promising applications in several domains. This article aims to evaluate the classification of aromatic herbs using the EfficientNet convolutional neural network (CNN) technique with transfer learning. The methodology used is experimental research, systematically manipulating variables to observe their effects on the object of study. The researcher plays an active role in this process, rather than being a passive observer. Based on the results and the literature review, it is evident that the objective of this research was achieved, as despite the opportunities for improvement in training to achieve accuracy above 0.8, it was possible to evaluate the classification of aromatic herbs using EfficientNet CNN through the transfer learning technique.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Marcelo Tsuguio Okano
Program of Production Engineering, Paulista University (UNIP)
Rua Dr. Bacelar, No. 1212, Vila Clementino, São Paulo, Brazil
Email: marcelo.okano@unip.br

1. INTRODUCTION

The identification and classification of visually similar herbs present a significant challenge due to their resemblance in characteristics such as color, shape, and texture. Traditional methods often fail to distinguish these species with high accuracy, highlighting the need for more advanced technologies. Studies have shown that models based on convolutional neural networks (CNNs) can enhance this process, enabling automated and more precise identification. Herbs play a crucial role in both nutrition and phytotherapy, possessing historical and cultural significance. They have been widely used in rituals and religious ceremonies due to their aromatic and medicinal properties [1].

The wide variety of aromatic and medicinal herbs with similar shape, color, and geometry characteristics makes their visual differentiation difficult. However, recent advancements in computer vision have spurred rapid growth in plant identification research, with promising results that demonstrate high accuracy, precision, and practical real-world applications [2]. With digital technologies like artificial intelligence (AI), transfer learning, deep learning (DL), CNNs, and mobile devices, automating the classification process of aromatic and medicinal herbs has become feasible. Herb classification holds particular importance in medicine, plant science, and the food industry. While various parts of plants, such as leaves, flowers, fruits, seeds, and roots, can be used for species identification, leaves are especially advantageous due to their availability throughout much of the plant's lifecycle [3].

Journal homepage: <http://ijai.iaescore.com>

This study aims to evaluate specific classification techniques using the EfficientNet CNN and transfer learning to overcome these challenges. The research adopts an experimental methodology that allows for the controlled manipulation of study variables to analyze their impact on the object of investigation. This approach enables a comprehensive assessment of the system's behavior under different conditions, ensuring that the effects of each variable are systematically evaluated. The validation of the methodology is carried out through experiments that simulate various operational conditions, allowing the researcher to actively observe the influence of each variable [4].

The structure of this paper is as follows: section 2 presents the theoretical framework for the proposed topics including the methodology. Section 4 details the results and discussion. Finally, section 5 concludes with final observations and directions for future research.

2. THEORETICAL FRAMEWORK

2.1. Classification vs. image identification with EfficientNet convolutional neural network

The distinction between image classification and image identification is fundamental to understanding computer vision tasks. Image classification refers to assigning an image a single class or label. In contrast, image identification involves identifying one or more objects through bounding boxes associated with object classes. According to Russakovsky *et al.* [5], image classification is a problem where the goal is to assign a single class to an entire image. This process involves determining the general category to which the image belongs without considering the precise location of objects within the image. On the other hand, image identification, as defined by Bochkovskiy *et al.* [6], involves the localization and identification of specific objects or individuals within a complex image. In this context, the task is to recognize objects' presence and determine their locations and often their specific identities. The swift advancement of digital technology, combined with the growing integration of AI in agriculture, has opened new paths for innovative solutions in plant identification [7]. This process typically involves using object detection algorithms, such as you only look once (YOLO) or faster region-based convolutional neural network (Faster R-CNN), to identify regions of interest in the image. While image classification focuses on assigning labels to entire images, image identification involves detecting and identifying specific objects within the image. These distinctions are essential for understanding the different approaches and techniques used in machine learning.

This research focused on classification using the EfficientNet CNN. According to Tan and Le [8], the EfficientNet CNN proposes an innovative approach to designing scalable and computationally efficient CNNs. They introduce a scaling method consisting of depth, width, and resolution scales, allowing the model to be resource-efficient while maintaining or improving classification performance. Furthermore, Chenna [9] emphasize that the EfficientNet CNN represents a significant advancement in the design of CNN architectures. They note that the scaling method proposed by EfficientNet allows for effective adaptation to different dataset sizes and computational constraints, making it an attractive choice for various practical applications. The EfficientNet CNN utilizes conventional metrics for evaluating classification models, such as accuracy, to measure its effectiveness in image classification. According to Tan and Le [8], the image levels in EfficientNet refer to different variants of the EfficientNet architecture that are scaled in depth, width, and image resolution. These levels are represented by letters and numbers, such as B0, B1, B2, and B7, where B0 refers to the lowest resolution, and B7 refers to the highest resolution and requires greater complexity and computational capacity. Each image level has a specific combination of network depth, channel width, and input resolution. The higher the level number (e.g., B7 compared to B0), the greater the image's depth, width, and resolution, resulting in a more complex and powerful network capable of handling more demanding computer vision tasks. Table 1 summarizes this dynamic.

Table 1. EfficientNet model base according to image resolution

Model base	Resolution
EfficientNetB0	224
EfficientNetB1	240
EfficientNetB2	260
EfficientNetB3	300
EfficientNetB4	380
EfficientNetB5	456
EfficientNetB6	528
EfficientNetB7	600

This scalability approach allows EfficientNet CNNs to be adapted to different available computational resources and precision requirements, providing a flexible and effective solution for various DL applications in computer vision. According to Tan and Le [8], accuracy is an essential metric for evaluating the performance of the EfficientNet CNN in object classification tasks. They state that EfficientNet was

designed to maximize accuracy on reference datasets such as ImageNet, ensuring that the model can make accurate and reliable predictions across various object classes. Furthermore, Antunes *et al.* [7] note that accuracy is a crucial metric for determining EfficientNet's practical utility in different applications. They also emphasize that a prominent level of accuracy is essential to ensure that the model is dependable enough to be used in real-world environments, where prediction accuracy is paramount.

2.2. Deep learning, convolutional neural network, and deep transfer learning

The concepts of DL, CNNs, and deep transfer learning (DTL) have been crucial in the field of AI. They demonstrate the ability to manage complex data and increasingly sophisticated tasks, resulting in a positive impact due to their wide range of applications in various fields. The growing popularity of classification methods such as deep CNNs enhances AI's potential in multiple disciplines, particularly computer vision [10].

DL is a computer-implemented algorithm that utilizes a neural network to select data features automatically. AI does not require any extra enhancements. It combines low-level attributes to create high-level attributes that describe the distributed nature of the data and sample attributes. Compared to traditional image recognition and detection methods, DL is more effective in image recognition and target detection. Today, the main types of networks are multilayer perceptron's, CNNs, and recurrent neural networks (RNNs) [11]. The most common type of DL for spatial pattern analysis is a CNN (also called ConvNets). CNNs aim to acquire spatial skills, such as edges, corners, textures, or other shapes that more accurately describe the intended class or volume [11]. DL methods, especially those based on CNNs, are popular in the agricultural domain for detection and classification purposes [11]. The core to learning these features is a series of transformations of the input data that occur at different spatial scales (e.g., through pooling). This facilitates identifying and combining low-level properties and conceptual ideas [11].

According to Lu *et al.* [11], CNNs typically comprise convolutional, pooling, and fully connected layers. The convolutional layer considers the local association of image content to derive features. Like other typical neural network models, CNNs consist of neurons organized in layers and thus capable of learning complex representations, with neurons between layers connected through weights and biases. The first layer is the input layer, for example, remote sensing data, and the final layer is the output, such as a predicted classification of plant species. Between the layers are hidden transformations that alter how the input space is transformed into one that aligns with the output. CNNs have at least one convolutional layer as a hidden layer that explores patterns. Other non-convolutional layers can also be incorporated. The convolutional layers include various filters that can be optimized; these layers transform the input or previous hidden layer information. The number of filters determines the thickness of a convolutional layer. The resulting transformations aim to expose significant patterns that address the problem at hand [12].

DTL is a method that attempts to reduce dependence and costs by utilizing knowledge from a source task to train a target task [13]. Transfer learning relaxes the requirement that training and testing data be independent and identically distributed, which motivates the use of transfer learning to address the problem of insufficient training data. In transfer learning, the process of training and testing data does not need to be identically distributed, and the model in the target domain does not need to be trained from scratch, which can significantly reduce the demand for training data and the time required to train the model in the target domain [14], [15]. Transfer learning allows you to leverage what you have learned in one situation to learn more rapidly in another. It is commonly used for object recognition and image classification through pre-trained CNN models. Different methods of altering a pre-trained CNN and when to use each method are of great interest for research. The effect of each method on training time and test set accuracy affects the way transfer learning is employed. The herbs selected for analysis included *Laurus nobilis*, *Rosmarinus officinalis*, and *Mentha spicata*. The bay leaf (*Laurus nobilis*) is characterized by lanceolate leaves with an alternate arrangement and short petioles, measuring 5-8 cm in length and 3-4 cm in width. These leaves are leathery, punctate, with revolute and wavy entire margins. The upper surface is glabrous and glossy, ranging from olive green to brown, while the lower surface exhibits a dull olive green to brown coloration with a prominent midrib and veins. *Laurus nobilis* is widely used in the food and pharmaceutical industries due to its antioxidant and antimicrobial properties [16]. *Rosmarinus officinalis*, from the Lamiaceae family, is an aromatic perennial plant with erect stems and blue-white flowers. Commonly known as rosemary and native to the Mediterranean region, its fresh and dried leaves are widely used as a seasoning and for herbal tea preparation. It is particularly noted for its anti-inflammatory properties demonstrated in preclinical in vivo models [17]. Finally, spearmint (*Mentha spicata*) is distinguished by its opposite leaves with short petioles and oblong to oval shapes with serrated margins. This herb plays a crucial role due to its rich composition of phenolic compounds, which are effective in treating cardiovascular diseases. Additionally, it is frequently used as a seasoning and in infusions [18].

2.3. Methods

This research followed a structured experimental procedure to assess the classification of aromatic herbs using the EfficientNet CNN with transfer learning. The methodology is presented in two main phases

Enhanced classification of aromatic herbs using EfficientNet and transfer ... (Samira Nascimento Antunes)

to allow for replication by other researchers: i) literature review and dataset preparation, and ii) experimental model training and evaluation.

2.3.1. Literature review and dataset preparation

Initially, a comprehensive literature review was conducted to provide a solid theoretical foundation on relevant topics, including DL, CNN architectures, transfer learning, and EfficientNet's specific application to plant classification. This review aimed to identify knowledge gaps and validate EfficientNet's suitability for this task. The literature review is essential for understanding the evolution of CNNs, enabling the identification of advancements, challenges, and research opportunities. Recent studies emphasize their importance in analyzing architectures, optimizing models, and exploring practical applications [19]. Contemporary authors emphasize the relevance of literature reviews in constructing scientific knowledge. Literature reviews is a study that examines and synthesizes an existing body of literature by identifying, questioning, and developing the foundations of a theory through the analysis of a body (or multiple bodies) of prior work [20]. Additionally, Ebador and Ikhide [21] highlight the importance of including diverse sources, such as technical reports and specialized journals, to ensure a comprehensive approach in the literature review. The aim of this literature review was to seek the most recent contributions to understanding the topic. According to Barry *et al.* [22], regularly updating the literature review is crucial to contextualizing the study within the state of the art. The careful selection of source ensures the reliability and relevance of the information, which are essential for building a consistent theoretical framework [23].

Zoph *et al.* [24] introduced groundbreaking approach to scalable image recognition by developing transferable architectures through automated neural architecture search, significantly improving efficiency and accuracy in image recognition. This work laid foundation for further advancements in efficient model design, such as EfficientNet, particularly in resource-constrained environments. These improvements in network architectures have facilitated the development of more specialized techniques, such as semantic segmentation, which are essential for applications requiring high precision in visual detail. Consequently, the adoption of advanced semantic segmentation techniques to ensure the preservation of fine details in images is crucial for differentiating visually similar herbs. The use of methodologies like those presented by Shivaprasad and Wadhawan [25] highlights segmentation strategies that emphasize the preservation of subtle and essential features. This approach is vital for maintaining the integrity of visual information during image processing, enabling more accurate and effective classification by enhancing the critical details that distinguish each species. The dataset selected for model training consisted of images of three herb species: rosemary, mint, and bay leaf. The herbs were chosen based on their visual similarities in color, texture, and shape, which present challenges for image classification. The images were gathered from open sources and categorized based on their visual quality and relevance, containing individual leaf images as well as instances of leaves in bundles and mixed with other objects. This diversity was intended to replicate real-world conditions but also led to some classification challenges.

Given the challenges often imposed by the dataset size and the need to increase model robustness, the incorporation of self-supervised learning (SSL) techniques can be highly effective. Abdulrazzaq *et al.* [26] highlighted methods like missing part prediction and automatic colorization of black-and-white images offer promising approaches. These self-supervised methods enable the model to develop a deeper understanding of the intrinsic features of herbs without relying exclusively on labeled data. SSL has shown great potential in extracting discriminative representations, which are crucial for the precise distinction between visually similar aromatic herbs, allowing for more efficient learning and more accurate classification. In parallel, the precision in neural network selection and feature optimization plays a key role in enhancing classification performance. This is well illustrated in [27], which demonstrates the cross-disciplinary applicability of these techniques. The study shows how strategic adjustments in model selection and data treatment can significantly improve classification outcomes, offering valuable insights for addressing similar challenges in other technical domains.

Moreover, image quality is fundamental to the accuracy of the classification process. The integration of enhancement methods, such as deblurring and asymmetric spatial attention, as presented in the study [28], can significantly improve the sharpness and clarity of herb images. These enhancement techniques, applied during the preprocessing phase, are essential for addressing issues related to blurry or low-quality images, ensuring that the model receives higher-quality visual data. By improving data input quality, these methods enhance classification accuracy and reduce errors caused by inadequate image conditions.

2.3.2. Experimental model training and evaluation

The model training was conducted on Google Colab, using available graphics processing unit (GPU) resources to facilitate efficient processing. Python programming language was utilized, alongside key libraries such as TensorFlow, NumPy, Pandas, Matplotlib, and OpenCV. These tools allowed for precise manipulation and analysis of the image data.

- Transfer learning with EfficientNet: the EfficientNet CNN was pre-trained with ImageNet weights to leverage transfer learning, which accelerates training by adapting learned features from a large, general

- dataset to specific herb classification task. The model was fine-tuned for 50 epochs, and variants B0 through B7 were tested with resolutions ranging from 224 to 600 pixels. Each variant's configuration, including depth, width, and resolution scaling, was adjusted to achieve optimal accuracy and computational efficiency.
- Custom model training: in addition to transfer learning, a custom model was trained from scratch using only the dataset of rosemary, mint, and bay leaf images. This approach aimed to address the challenges encountered in transfer learning due to dataset limitations and verify the model's ability to learn exclusively from this specific dataset. For comparability, the custom model was also trained for 50 epochs, and variants B0 through B4 were tested.
- Metrics for evaluation: the primary evaluation metric was accuracy, defined as the proportion of correctly classified images out of the total classifications. Loss was also monitored to assess model convergence and detect potential overfitting. These metrics were recorded throughout training and on test sets to provide a comprehensive view of model performance.

The model's hyperparameters, such as learning rate, batch size, and number of epochs, were kept consistent across all training sessions to ensure comparability. The image data were split into training and test sets using an automated algorithm, maintaining balanced representation for each herb category. This setup is intended to enable other researchers to replicate the experiments and observe similar results under controlled conditions. This structured methodology allows for replication by detailing each experimental step, from dataset preparation to model evaluation, enabling future studies to build on these findings and refine herb classification techniques using CNNs and transfer learning.

2.4. Consolidation of AI techniques applied to research

The application of advanced AI techniques has become a robust and efficient approach for complex image classification tasks, especially in areas where subtle visual differences make recognition challenging. In this study, we used the EfficientNet architecture combined with transfer learning and data augmentation techniques to classify aromatic herbs, meeting the demand for an automated and precise solution. The EfficientNet architecture represents an innovation in the design of CNNs by introducing a compound scaling method that automatically adjusts the network's depth, width, and resolution based on the available computational resources and task complexity [8]. This approach is particularly valuable for plant classification problems, where the model needs to capture subtle variations in texture and color between different species. Recent studies have highlighted that compound scaling allows for high accuracy with optimized resource use, which is essential in large-scale image processing environments [25], [26].

Transfer learning was applied to leverage prior knowledge from models trained on large datasets, such as ImageNet, and adapt them for specific herb classification tasks. Transfer learning is a technique in which an algorithm learns image features within one domain and can then apply this acquired knowledge to a new domain, even with a smaller dataset [27]. The technique of transfer learning with fine-tuning enables starting with a model pre-trained for a specific task and then adjusting only certain layers of the neural network to adapt it for a similar yet distinct target task [28]. This technique is widely recognized for accelerating training and improving accuracy, especially when applied to limited or highly visually similar datasets [29].

Additionally, to improve generalization and reduce overfitting, data augmentation techniques were applied, including transformations such as rotation, brightness adjustment, and cropping. Data augmentation creates artificial variations of the training images, simulating a more robust dataset and helping the model handle variations in angle and lighting [30], [31]. Studies have shown that data augmentation is effective in plant classification domains, where the model needs to adapt to subtle changes in leaf appearance [32]. Finally, using accuracy and loss metrics was essential for evaluating and adjusting the model throughout the training process, helping to detect issues like overfitting or underfitting. In summary, combining EfficientNet, transfer learning, and data augmentation provides a solid framework for complex visual recognition tasks, such as the classification of aromatic herbs, demonstrating the potential of these techniques to enhance the accuracy and applicability of AI models in challenging contexts.

2.5. Comparative evaluation of convolutional architectures for aromatic herb classification

This study focused exclusively on using EfficientNet for the classification of aromatic herbs due to its compound scaling ability, which automatically adjusts the model's depth, width, and resolution according to the task's requirements. This characteristic makes EfficientNet computationally efficient and especially useful in resource-limited environments. However, a comparative analysis with other modern CNN architectures could provide a more comprehensive evaluation of the model's effectiveness and identify alternatives that may bring distinct advantages for herb classification. ResNet, for instance, stands out for its use of residual connections, which facilitate gradient flow in deep networks and mitigate the vanishing gradient problem. The model has one MaxPool layer, 48 convolutional layers, and a standard pool layer. In each identification block, as well as in each convolution block, there are three convolution layers [33]. Comparing ResNet with EfficientNet would help assess whether the generalization capacity of residual connections benefits the classification of classes with

high visual similarity, such as aromatic herbs. DenseNet, in turn, adopts a densely connected layer architecture, promoting feature reuse and improving information propagation throughout the network. This can be particularly advantageous in datasets where fine visual details, such as texture and edges, are crucial for classification. The ability to reuse intermediate information allows DenseNet to capture subtle visual nuances, offering an interesting alternative to EfficientNet in terms of differentiating visually similar species [34].

MobileNet, a lightweight architecture, was specifically designed for efficiency on mobile devices, making it ideal for applications requiring low resource consumption, such as field-based plant classification. In practical use situations, such as identifying herbs in remote locations, MobileNet could be a beneficial choice, balancing accuracy and computational efficiency on devices with lower processing power. Studies indicate that MobileNet performs well on mobile devices, maintaining competitive accuracy levels compared to more complex networks [35]. Architectures such as Inception, which combines convolutions of various sizes to capture a wide range of visual features, and Xception, which enhances this approach with separable convolutions for greater efficiency, could also enrich this analysis. Inception is effective with images exhibiting high pattern variability, while Xception combines accuracy and efficiency, making it suitable for contexts where balancing computational cost with feature extraction capacity is essential [36].

3. RESULTS AND DISCUSSION

3.1. Results

Table 2 presents the test set's results regarding the model training through the transfer learning process, considering the CNN EfficientNet pre-trained with the weights of the ImageNet dataset. Table 3 (see in Appendix) presents the evolution of accuracy and loss throughout the training process, providing a detailed view of the model's performance under different configurations. Analyzing these data is essential to understand how variations in network architecture and training parameters impact the final results. This analysis is crucial for optimizing the machine learning process, ensuring that the model is not only efficient but also capable of achieving the highest possible accuracy. Due to the models' accuracy being close to 0.5, the pronounced fluctuations of accuracy throughout the training, the average accuracy between 0.6 and 0.7 considering the test sets, and the confusion of results when testing with different images, it was decided to train the EfficientNet CNN without weights. Training a model considered customized, meaning the model learns only from the input dataset and with null input weight, aimed to identify the probable causes for the accuracy values under 0.8 and the different confusions during testing. Table 4 presents the test set's results concerning the model's training without the transfer learning process, considering the EfficientNet CNN without input weights and with only the selected aromatic herbs dataset. Notably, due to computational capacity, the models were trained to consider resolutions B0 through B4. Table 5 shows the evolution of accuracy and loss throughout the training sessions.

Regarding the training time, it is noted that it increased according to the resolution. Additionally, the time is considerably longer for a model learning without input weights. This dynamic is expected because, according to Tan and Le [8], the image levels in EfficientNet refer to different variants of the EfficientNet architecture, and the higher the level number, the greater the depth, width, and resolution of the image, resulting in a more complex and powerful network capable of handling more demanding computer vision tasks, hence the longer training time. When analyzing the loss, it is noticeable that both in the transfer learning process and in the custom process, it increases as the resolution increases. This is also expected due to the better resolution of the image pixels, which enables objects to be identified more accurately.

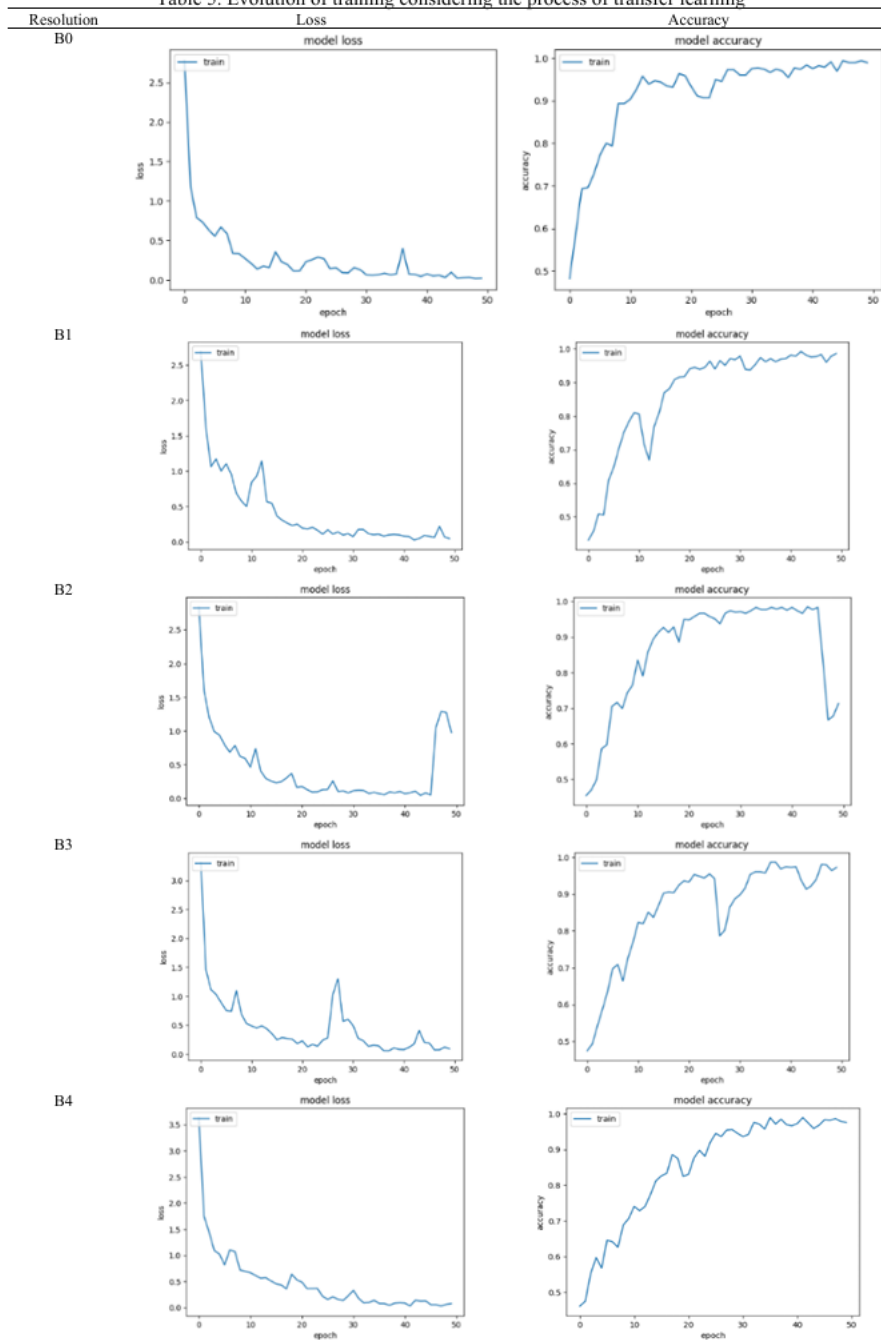
Table 2. Summary of EfficientNet model results using transfer learning

Model variant	Resolution	Epochs	Training time	Loss	Accuracy
B0	224	50	12m30s	0.79	0.71
B1	240	50	18m33s	0.77	0.51
B2	260	50	24m32s	0.80	0.77
B3	300	50	42m35s	0.72	0.77
B4	380	50	01h39m	0.73	0.80
B5	456	50	03h32m	0.84	0.57
B6	528	50	06h09m	0.69	0.74
B7	600	50	09h44m	0.77	0.65

Table 4. Custom model training results (without transfer learning)

Model variant	Resolution	Epochs	Time	Loss	Accuracy
B0	224	50	51 m 42 s	0.94	0.88
B1	240	50	01 h 11 m	1.09	0.80
B2	260	50	01 h 37 m	2.542	0.40
B3	300	50	02 h 50 m	0.96	0.80
B4	380	50	06 h 36 m	1.53	0.68

Table 5. Evolution of training considering the process of transfer learning



Enhanced classification of aromatic herbs using EfficientNet and transfer ... (Samira Nascimento Antunes)

The accuracy varied according to the resolution, meaning that the higher the resolution, the higher the accuracy. Additionally, the variation in accuracy in the transfer learning process had more significant oscillation than that in the custom model, which showed a growing learning trend over the epochs. This suggests that the custom model was able to learn more effectively from the available dataset, resulting in a more stable performance over time.

3.2. Discussion

This study addresses significant gaps observed in previous research on the classification of aromatic herbs using computer vision techniques. While prior studies have explored the use of CNNs and transfer learning for plant identification, they primarily focused on distinct plant species with visually apparent differences. These studies often emphasized general plant classification rather than the nuanced differentiation required for visually similar aromatic herbs, such as rosemary, mint, and bay leaf. Consequently, they did not explicitly address the influence of subtle variations in color, shape, and texture that are critical for accurately classifying these herbs.

This research specifically investigates the application of the EfficientNet CNN architecture to tackle these challenges, examining how transfer learning can be fine-tuned to improve classification accuracy for visually similar plant species, thereby filling a key gap in the existing literature. The use of transfer learning allowed the EfficientNet model, pre-trained on the ImageNet dataset, to be successfully adapted for specific aromatic herb classification. This technique proved efficient for accelerating training, reducing computational costs, and maintaining accuracy.

Notably, lower-resolution pre-trained models, such as EfficientNet B0, showed substantially shorter training times without significantly compromising accuracy. Higher-resolution models, including EfficientNet B3 and B4, demonstrated increased accuracy, achieving values over 0.8 in certain cases. However, the computational cost associated with higher resolutions, from EfficientNet B5 to B7, was considerably higher, with training times up to five times longer, which limits their practical feasibility for resource-constrained devices. This observation underscores the importance of balancing resolution with computational efficiency, especially in field applications where resources may be limited.

Although custom-trained models (trained from scratch) achieved competitive accuracy, they required significantly longer training times and displayed greater stability in accuracy over the course of training. Transfer learning models, in contrast, exhibited some fluctuations in accuracy across different resolutions, suggesting that the pre-trained model might be sensitive to the specific characteristics of the herb dataset. The variability observed in accuracy and loss over the epoch's points to potential overfitting, particularly with higher resolutions, highlighting the importance of employing data augmentation techniques to improve model robustness and enable better generalization to new images of herbs.

The study identified that model accuracy could be impacted by the presence of images of herb bunches mixed with images of individual leaves in the dataset. Thus, a key recommendation for future research is to refine the dataset to include only isolated leaf images, which could help reduce classification confusion. Additionally, further exploration of different resolution configurations and implementing fine-grained visual classification (FGVC) techniques are recommended to improve the model's ability to differentiate subtle variations between herb species.

In comparing transfer learning-based models with custom-trained models, each approach presents distinct advantages. Transfer learning significantly reduces training time and enables rapid results, whereas custom models offer greater stability in accuracy throughout training, especially at intermediate resolutions. Although more time-consuming, custom models demonstrated higher resilience to overfitting, showing more consistent performance on unseen data a valuable characteristic in real-world applications.

In terms of accuracy, mid-resolution models, particularly EfficientNet B2 to B4, seem to offer an optimal balance between computational efficiency and generalization capability, maintaining competitive accuracy without the high computational costs associated with higher-resolution models. This study provides a balanced perspective on the strengths and limitations of each approach, offering practical insights for the development of efficient models in aromatic herb classification and similar classification contexts. Table 6 summarizes the key findings and observations from the study, highlighting model efficiency, the impact of resolution, variability in performance metrics, and recommendations for future enhancements.

The effectiveness of the aromatic herb classification model depends significantly on the selection and optimization of the neural networks employed. The techniques discussed in [37], explore feature fusion from multiple neural network architectures. This approach aims to combine the strengths of different networks to capture a broader range of discriminative features. Network selection can enhance model

accuracy and optimize the learning process by leveraging information fusion to address the specific visual similarities of aromatic herbs, ensuring more reliable classification results.

Table 6. Key findings and observations

Aspect	Findings
Model efficiency	Transfer learning reduced training time but achieved comparable accuracy to custom models.
Resolution impact	Higher resolutions (B3, B4) improved accuracy, but increased computation time significantly.
Loss and accuracy variability	Transfer learning models showed greater variability in accuracy across resolutions.
Generalization issues	Custom model generalized well with lower accuracy oscillation but required extended training.
Future enhancements	Suggested refining the database with isolated leaf images and exploring different resolutions.

Accurate herb identification faces challenges due to the high visual similarity between different herb types. Studies have explored complementary strategies to address these challenges, including:

- SSL, this technique reduces reliance on labeled datasets by allowing models to learn more relevant representations of herbs from unlabeled data, enhancing classification by emphasizing the intrinsic features of the leaves [26].
- Network selection and information fusion, by optimizing the choice of neural network architectures and fusing information, this approach improves classification accuracy. According to the Zafar *et al.* [27], this strategy can be applied to determine the best-performing version of EfficientNet for herb identification.
- Feature preservation in images, advanced segmentation techniques help maintain subtle details in images, preventing the loss of crucial information necessary for distinguishing visually similar species [25].
- Image quality enhancement, methods for improving the sharpness and clarity of medical images can be adapted to ensure higher classification precision in herb analysis by providing higher-quality visual data before automated analysis [28].

Together, these techniques form a robust framework for enhancing the performance and reliability of automated herb classification models, addressing both data quality and model optimization.

4. CONCLUSION

This study successfully demonstrates the application of EfficientNet with transfer learning for aromatic herb classification, achieving notable accuracy while highlighting areas for database refinement and the importance of balancing resolution with computational efficiency for improved model generalization. This review has key limitations, including reliance on a specific image dataset of aromatic herbs, testing in a controlled environment, and a focus mainly on accuracy and computational efficiency, with limited attention to interpretability and real-time performance. Given these limitations, future research should aim to extend this investigation across diverse datasets that include a broader array of plant species and different environmental contexts to enhance the model's robustness and applicability. It is also recommended to explore higher levels of data diversity, including varying image qualities, lighting conditions, and even seasonal variations in plant appearance, to improve model generalization and adaptability. Furthermore, research should examine how FGVC and other specialized techniques can be integrated to enhance the classification accuracy for species with subtle visual distinctions. Expanding model interpretability to better understand the decision-making process of CNNs could also provide valuable insights, especially for practical applications in agriculture and botany. We propose exploring interpretability approaches such as gradient-weighted class activation mapping (Grad-CAM) and local interpretable model-agnostic explanations (LIME). These techniques would allow us to observe the regions of the image that most contribute to classification, providing transparency to the model and enabling researchers to validate whether the model is using valid botanical features, such as texture patterns, edges, and leaf veins. Expanding model interpretability to better understand the decision-making process of CNNs could provide valuable insights, especially for practical applications in agriculture and botany. The implications of this research are significant for fields that depend on accurate plant identification, such as agriculture, medicine, and the food industry. Automated herb classification could streamline processes, reduce human error, and increase efficiency in applications ranging from botanical research to commercial product development. As models continue to improve in accuracy and adaptability, they could be integrated into mobile or IoT devices, providing on-site plant identification and further democratizing access to AI-driven botanical classification tool.

FUNDING INFORMATION

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–Finance Code 001.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Samira Nascimento Antunes	✓	✓		✓	✓				✓					
Madalena De Oliveira Barbosa Divino			✓			✓		✓	✓					
Luana dos Santos Cordeiro						✓		✓	✓					
Fernanda Pereira Leite Aguiar			✓	✓	✓			✓	✓	✓	✓			
Marcelo Tsuguio Okano	✓			✓			✓		✓	✓		✓	✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

This study did not involve human participants or animals.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [MTO]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

- [1] A. Preeya V, S. P. Raja, B. K. Dhanalakshmi, H. Gururaj, V. Ravi, and P. Ravi, "Perceptual pigeon galvanized optimization of multi-objective CNN on the identification and classification of mango leaves disease," *The Open Agriculture Journal*, vol. 18, no. 1, Nov. 2024, doi: 10.2174/0118743315331264240917073431.
- [2] D. B. Valdez, C. J. G. Aliac, and L. S. Feliscuzo, "Medicinal plant classification using convolutional neural network and transfer learning," in *4th IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET 2022*, Sep. 2022, pp. 1–6, doi: 10.1109/IICAIET55139.2022.9936868.
- [3] S. Chaivivatrakul, J. Moonranta, and S. Chaiwivatrakul, "Convolutional neural networks for herb identification: plain background and natural environment," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 12, no. 3, pp. 1244–1252, Jun. 2022, doi: 10.18517/ijaseit.12.3.15348.
- [4] R. K. Domingues, J. A. Cordoli, and D. Silva, "Development of a methodology for diagnosing faults in bearings operating under variable operating conditions based on self-supervised learning," *Proceedings of the Annual Conference of the Prognostics and Health Management Society, PHM*, vol. 16, no. 1, 2024, doi: 10.36001/phmconf.2024.v16i1.4183.
- [5] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

- [6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOV4: optimal speed and accuracy of object detection," *arXiv-Computer Science*, pp. 1–17, 2020.
- [7] S. N. Antunes *et al.*, "Model development for identifying aromatic herbs using object detection algorithm," *AgriEngineering*, vol. 6, no. 3, pp. 1924–1936, Jun. 2024, doi: 10.3390/agriengineering6030112.
- [8] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *36th International Conference on Machine Learning, ICML 2019*, 2019, pp. 1–11.
- [9] D. Chenna, "Evolution of convolutional neural network (CNN): compute vs memory bandwidth for edge AI," *Feedforward*, vol. 2, no. 3, pp. 2–12, 2023.
- [10] S. Temsiririkkul, P. Siritanawan, and R. Temsiririkkul, "Which one is Kaphrao? identify Thai herbs with similar leaf structure using transfer learning of deep convolutional neural networks," in *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, Auckland, New Zealand, Dec. 2021, pp. 738–743, doi: 10.1109/TENCON54134.2021.9707324.
- [11] J. Lu, L. Tan, and H. Jiang, "Review on convolutional neural network (CNN) applied to plant leaf disease classification," *Agriculture*, vol. 11, no. 8, Jul. 2021, doi: 10.3390/agriculture11080707.
- [12] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 24–49, Mar. 2021, doi: 10.1016/j.isprsjprs.2020.12.010.
- [13] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," *Technologies*, vol. 11, no. 2, 2023, doi: 10.3390/technologies11020040.
- [14] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, Cham: Springer, 2018, pp. 270–279, doi: 10.1007/978-3-030-01424-7_27.
- [15] T. Alshalali and D. Josyula, "Fine-tuning of pre-trained deep learning models with extreme learning machine," in *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*, Dec. 2018, pp. 469–473, doi: 10.1109/CSCI46756.2018.00096.
- [16] A. Paparella, B. Nawade, L. S.-Harpaz, and M. Ibdah, "A review of the botany, volatile composition, biochemical and molecular aspects, and traditional uses of *Laurus nobilis*," *Plants*, vol. 11, no. 9, Apr. 2022, doi: 10.3390/plants11091209.
- [17] C. Gonçalves, D. Fernandes, I. Silva, and V. Mateus, "Potential anti-inflammatory effect of rosmarinus officinalis in preclinical in vivo models of inflammation," *Molecules*, vol. 27, no. 3, Jan. 2022, doi: 10.3390/molecules27030609.
- [18] S. Saqib *et al.*, "Mentha: nutritional and health attributes to treat various ailments including cardiovascular diseases," *Molecules*, vol. 27, no. 19, Oct. 2022, doi: 10.3390/molecules27196728.
- [19] Z. Du, "Review of convolutional neural network," *Science and Technology of Engineering, Chemistry and Environmental Protection*, vol. 1, no. 9, Oct. 2024, doi: 10.61173/1xfrez14.
- [20] C. Post, R. Sarala, C. Gatrell, and J. E. Prescott, "Advancing theory with review articles," *Journal of Management Studies*, vol. 57, no. 2, pp. 351–376, Mar. 2020, doi: 10.1111/joms.12549.
- [21] L.-L. Ebidor and I. G. Ikhide, "Literature review in scientific research: an overview," *East African Journal of Education Studies*, vol. 7, no. 2, pp. 179–186, 2024, doi: 10.37284/eajes.7.2.1909.
- [22] E. S. Barry, J. Merkebu, and L. Varpio, "State-of-the-art literature review methodology: a six-step approach for knowledge synthesis," *Perspectives on Medical Education*, vol. 11, no. 5, pp. 281–288, Sep. 2022, doi: 10.1007/s40037-022-00725-9.
- [23] G. de B. V. Junior *et al.*, "Métricas utilizadas para avaliar a eficiência de classificadores em algoritmos inteligentes," *Centro de Pesquisas Avançadas em Qualidade de Vida*, vol. 14, no. 2, pp. 1–14, 2022, doi: 10.36692/v14n2-01r.
- [24] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 8697–8710, doi: 10.1109/CVPR.2018.00907.
- [25] K. Shivaprasad and A. Wadhawan, "Deep learning-based plant leaf disease detection," *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 360–365, 2023, doi: 10.1109/ICICCS56967.2023.10142857.
- [26] M. M. Abdulrazzaq *et al.*, "Consequential advancements of self-supervised learning (SSL) in deep learning contexts," *Mathematics*, vol. 12, no. 5, 2024, doi: 10.3390/math12050758.
- [27] A. Zafar, J. Tanveer, M. U. Ali, and S. W. Lee, "BU-DLNet: breast ultrasonography-based cancer detection using deep-learning network selection and feature optimization," *Bioengineering*, vol. 10, no. 7, pp. 1–16, Jul. 2023, doi: 10.3390/bioengineering10070825.
- [28] S. M. A. Sharif, R. A. Naqvi, Z. Mehmood, J. Hussain, A. Ali, and S.-W. Lee, "MedDeblur: medical image deblurring with residual dense spatial-asymmetric attention," *Mathematics*, vol. 11, no. 1, 2022, doi: 10.3390/math11010115.
- [29] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021, doi: 10.1109/JPROC.2020.3004555.
- [30] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.
- [31] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 13001–13008, Apr. 2020, doi: 10.1609/aaai.v34i07.7000.
- [32] N. Saleem *et al.*, "Class-specific data augmentation for plant stress classification," *Plant Phenome Journal*, vol. 7, no. 1, 2024, doi: 10.1002/ppj2.20112.
- [33] V. Sharma, K. Chaurasia, and A. Bansal, "Optimizing species recognition in medicinal plants: a comprehensive evaluation of deep learning models," in *Proceedings of the 14th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2024*, Jan. 2024, pp. 738–743, doi: 10.1109/Confluence60223.2024.10463198.
- [34] E. Hassan, S. A. Ghazalah, N. El-Rashidy, T. A. El-Hafeez, and M. Y. Shams, "DenseNet model with attention mechanisms for robust date fruit image classification," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, 2025, doi: 10.1007/s44196-025-00809-4.
- [35] A. Howard *et al.*, "Searching for mobileNetV3," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.
- [36] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258, doi: 10.1109/CVPR.2017.195.
- [37] M. U. Ali *et al.*, "Deep learning network selection and optimized information fusion for enhanced COVID-19 detection," *International Journal of Imaging Systems and Technology*, vol. 34, no. 2, 2024, doi: 10.1002/ima.23001.

APPENDIX

Table 3. Evolution of training considering the transfer learning process

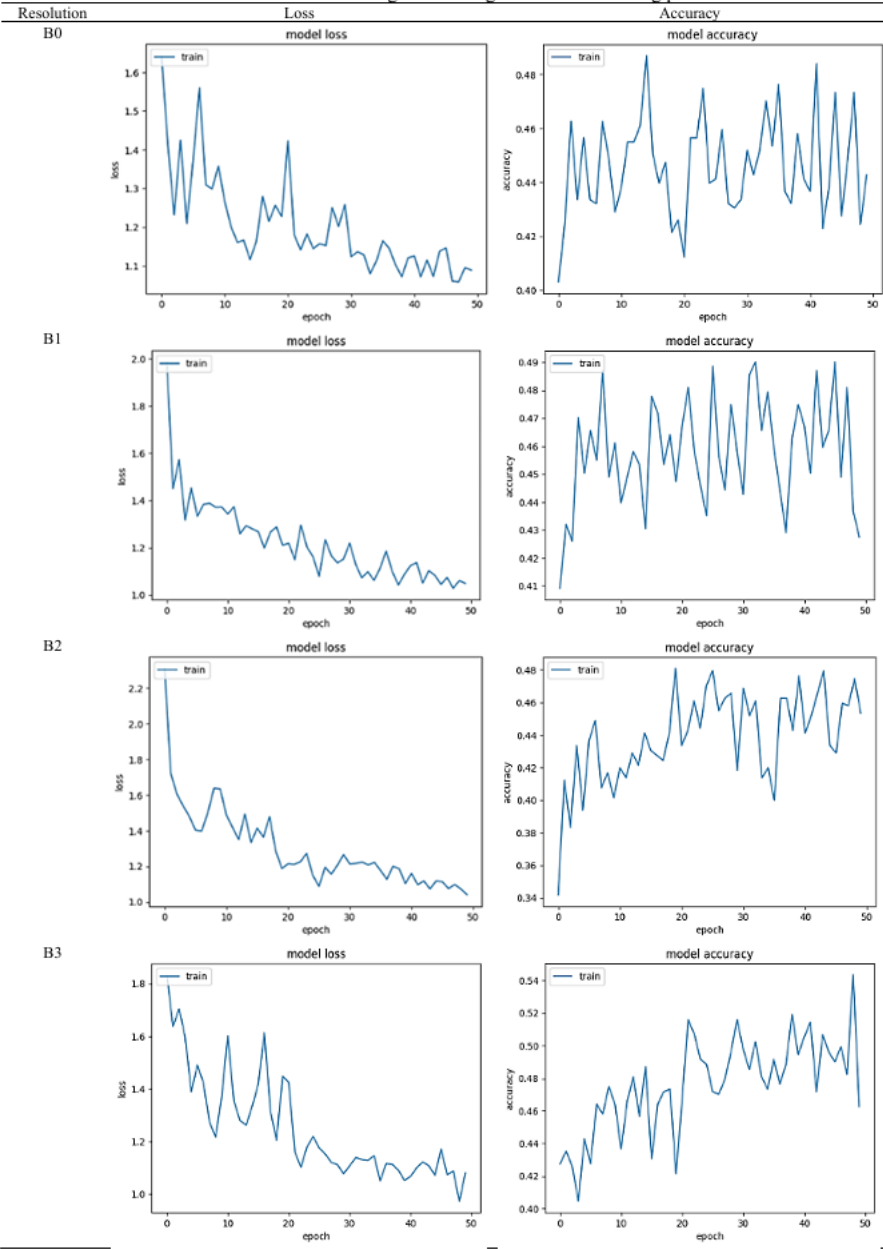
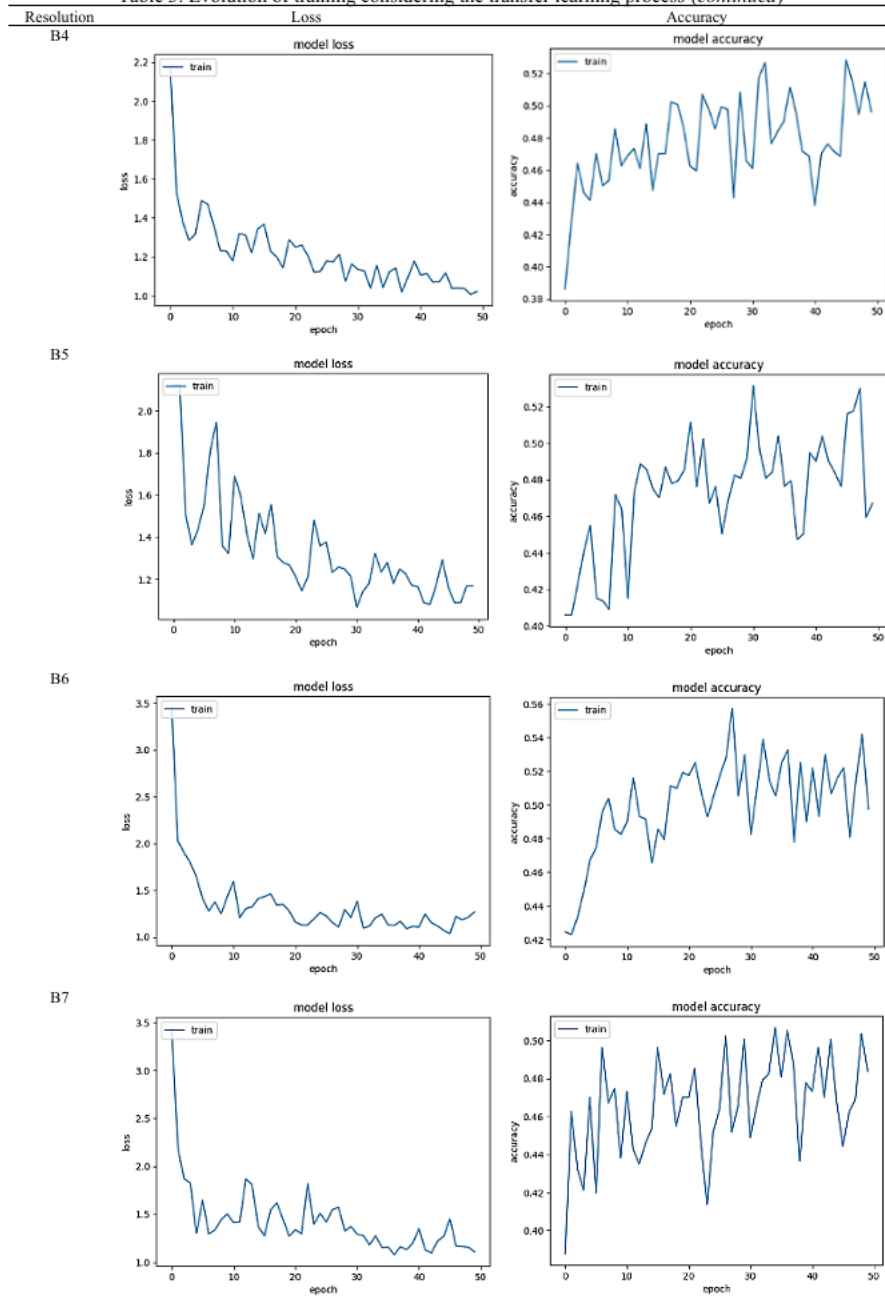
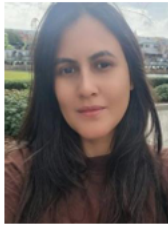
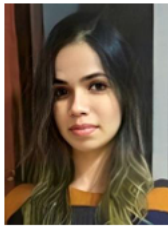


Table 3. Evolution of training considering the transfer learning process (*continued*)

BIOGRAPHIES OF AUTHORS



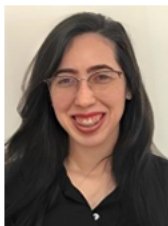
Samira Nascimento Antunes    is Ph.D. candidate in Digital Technologies and Technological Innovation in Operation Systems. This project is part of the research project Management of Production and Innovation in Agriculture, Industry and Services with a focus on fruit and herb recognition through artificial intelligence algorithms (CNN, YOLO framework) and development of a web and mobile interface as provision of product. She can be contacted at email: samira.antunes@aluno.unip.br.



Madalena De Oliveira Barbosa Divino    has a Bachelor's Degree in Systems Analysis and IT-FATEC Antônio Russo São Caetano do Sul (2012), a postgraduate degree in Teaching in Higher Education-SENAC (2015) and a degree in Pedagogical Training for Secondary Education from the State Center for Technological Education Paula Souza (2017). She currently works as a teacher-Etec Rio Grande da Serra in the technical area, since 2015. She also works in private education, as a teacher of the initial years and high school in the areas of technology, programming and robotics. She can be contacted at email: madalena.divino@etec.sp.gov.br.



Luana dos Santos Cordeiro    is Pedagogy, Computer Technologist with emphasis in Business Management from FATEC-Mauá, Master's student in Production Engineering. Currently, she is a High School and Technical School teacher-ETEC Professora Maria Cristina Medeiros and at the São Paulo State Department of Education (SEDUC). She has experience in the area of Education, in different teaching modalities, working for over 13 years in the area of education permeating between early childhood education, elementary school 1 and 2, high school, technical and vocational education, in addition to inclusive education. Enthusiastic, curious and interested in the area of project development in different segments. She can be contacted at email: lsc.luana@hotmail.com.



Fernanda Pereira Leite Aguiar    is Ph.D. Candidate in Production Engineering. This project seeks to develop an application, using artificial intelligence and computer vision to identify fear in beef cattle. Master in Productive Systems, at Centro Paula Souza (2023). Postgraduate MBA in Project Management from Centro Paula Souza (2018). Graduation in Information Technology Management from Faculdade de Tecnologia de Barueri (2016). Academic award for the undergraduate degree in Information Technology Management from Faculdade de Tecnologia de Barueri (2016). Participation in technology fairs (9th Technology Fair-Feteps) by the Barueri Faculty of Technology (2015). LIBRAS interpreter course by the InterLibras school (2019). Theology course by the Barueri Institute of Theology-ITEB (2019). She can be contacted at email: fernanda.p.aguiar@gmail.com.



Prof. Dr. Marcelo Tsugio Okano    is master's in administration (2006) from USCS, Ph.D. (2010) in Production Engineering from Universidade Paulista (UNIP), and Ph.D. (2020) in Administration from the School of Business Administration of São Paulo (EAESP) of Fundação Getúlio Vargas (FGV). He completed his Post-Doctorate (Production Engineering) in 2013 at FEG-UNESP and in 2015 he completed his Post-Doctorate (Production Engineering) in innovation and entrepreneurship at UNIP. Since 2022, he has been a professor in the Postgraduate Program in Production Engineering at UNIP, working in the concentration area of Operation Systems Management in projects related to the themes of information and communication technology, digital technologies, technological innovation, digital transformation, digital platforms, digital ecosystems and LGPD. Leader of the CNPq Research Group GEDITAD-Study and Development Group in Innovation, Technologies and Digital Artifacts at UNIP. Voluntary collaborating researcher at the Faculty of Technology of UNICAMP since 2020 working in the areas of IT and Telecommunications. Productivity Scholarship in Technological Development and Innovative Extension by CNPq level 2. He has several articles published in specialized journals and works presented in annals of scientific events, book chapters and published books, in addition to technological products and patented software. He can be contacted at email: marcelo.okano@unip.br.

APPENDIX C – Paper III

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Data-Centric Object Detection for Aromatic Herbs: Image Quality Filtering, Class Selection, and Model Comparison

Samira Nascimento Antunes¹, Won Suk Lee², Irenilza de Alencar Nääs¹ and Marcelo T Okano¹, Member, IEEE

¹Graduate Program in Production Engineering, Paulista University, Sao Paulo, SP 04026-002 Brazil

²Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL 32611-0570 USA

Corresponding author: Marcelo T Okano (e-mail: marcelo.okano1@docente.unip.br).

ABSTRACT The performance of object detection models depends on the quality and consistency of the training data. Most recent work targets model architecture. Dataset curation and class definition have received comparatively little attention. This work takes a different angle: instead of refining the model, we focused on the data itself, specifically on image quality and visual separability between herb classes. A dataset of 5,679 images across six herb classes was constructed from web-scraped and field-captured images, followed by quality-based curation and data augmentation to expand and balance the final dataset. A class selection strategy was adopted to reduce inter-class ambiguity by excluding visually indistinguishable species. Three detection architectures were trained and evaluated on the curated dataset: YOLOv11s, Faster R-CNN via Detectron2, and SSD300/VGG. On the independent test set, YOLOv11s achieved the best overall performance (mAP50 = 0.802; mAP50-95 = 0.593), followed by Faster R-CNN (AP50 = 0.798; AP = 0.541) and SSD300/VGG (mAP = 0.726). Although architectural differences influenced performance, the results suggest that dataset quality and class definition were critical factors for achieving stable detection performance across models. Across all three models, the quality of the training set and the class definition were stronger determinants of performance, lending empirical support to data-centric thinking in applied computer vision.

INDEX TERMS Aromatic Herb Recognition, Class Separability, Data-Centric AI, Deep Learning, Image Quality Assessment, Object Detection.

I. INTRODUCTION

Computer vision techniques have been increasingly applied in agricultural and food-distribution systems, enabling automated detection, localization, and classification of plant products from digital images. In this context, object detection has become a fundamental task in computer vision because it allows objects to be both identified and spatially localized in natural scenes [1]. Its applications are particularly relevant to precision agriculture and food supply chains, where visual detection can support decision-making, quality control, inventory monitoring, product sorting, and resource optimization [2]. Over the past two decades, object detection methods have evolved from handcrafted feature-based approaches to deep learning models capable of learning complex visual representations directly from data [1]. Architectures such as You Only

Look Once, region-based frameworks such as Detectron2, and single-shot detectors have expanded the practical use of computer vision in real-world agricultural environments. However, despite these advances, object detection remains challenging when images are acquired under non-standardized conditions, especially due to illumination variance, intra-class diversity, scale differences, occlusion, and localization accuracy [1].

Aromatic herbs represent a particularly difficult object-detection problem because many commercially relevant species exhibit strong morphological similarity and limited visual separability in RGB imagery. In several cases, herbs are distinguished by non-visual attributes, such as aroma, texture, or stem-base characteristics, which are not reliably captured by conventional images. This makes class definition and visual discrimination more complex than in

tasks involving fruits or vegetables with clearer color, shape, or size differences. In addition, images collected in real distribution or retail environments are affected by uncontrolled lighting, overlapping products, irregular object placement, heterogeneous backgrounds, blur, noise, and variable image quality. These factors increase both inter-class confusion and intra-class variability, reducing the reliability of detection models. In the specific context of CEAGESP, one of the largest food distribution centers in Brazil, aromatic herbs are handled and distributed in flexible and uncontrolled environments, where lighting conditions, product presentation, and image acquisition conditions vary substantially. For this reason, visually similar classes such as chives, chive blossoms, Chinese chives, dill, and fennel may introduce semantic noise when included without careful visual-separability criteria.

Recent research has emphasized that the performance of AI systems is not determined only by the model architecture, but also by the quality, consistency, representativeness, and semantic coherence of the training data [3], [4]. Nevertheless, most studies in object detection remain predominantly model-centric, focusing on improved architectures, optimization strategies, and computational efficiency, while treating datasets as fixed and sufficiently representative inputs. Works employing YOLO variants, prototypical networks, contrastive learning frameworks, and related deep learning approaches have reported strong results, yet the datasets underpinning these experiments are rarely scrutinized in terms of image quality, class coherence, or visual separability [5], [6]. Challenges such as inter-class similarity and intra-class variability are frequently acknowledged, but the usual response remains architectural, involving new loss functions, different backbones, refined augmentation strategies, or alternative training configurations [6]. A previous study [7] addressed aromatic herb detection using approximately 800 images across three classes, rosemary, mint, and bay leaf, and reported a consolidated mAP50 of 0.723. However, that work also identified important limitations, including over-detection tendencies, insufficient recall, the need for dataset expansion, bounding-box refinement, and the absence of systematic image-quality filtering, visual class-selection criteria, and an independent held-out test set. This reinforces the relevance of the data-centric perspective, particularly because real-world datasets are often affected by noise, heterogeneity, blur, uneven illumination, and cluttered backgrounds, confirming the principle that poor data quality leads models to perform worse [8]. In supply chain applications, such as food distribution centers, data are gathered under varied scenarios, with no visual constraints and substantial variation in image quality and object characteristics, making objects with irregular shape, orientation, or spacing particularly difficult to detect faithfully [9].

Therefore, this study investigates whether systematic dataset curation, including image-quality filtering and class selection based on visual separability, can improve and stabilize object-detection performance in a visually ambiguous multi-class aromatic-herb dataset across different detection architectures. The proposed approach expands the dataset to 5,679 images distributed across six visually distinguishable herb classes and introduces a structured pipeline for dataset construction, including quality assessment based on blur, illumination, and noise metrics, deduplication, class balancing, and explicit class-definition criteria. Three object detection architectures, YOLOv11s, Faster R-CNN implemented through Detectron2, and SSD300/VGG, are trained and evaluated under consistent experimental conditions using a common held-out test set of 856 images. The main contributions of this research are: a data-oriented pipeline for constructing and curating an aromatic herb dataset; a class-definition strategy grounded in visual separability to reduce semantic ambiguity; an empirical assessment of detection performance after applying image-quality filtering, deduplication, class balancing, and visual class-selection criteria; and a comparative analysis across different object detection architectures. By emphasizing data quality and class coherence rather than only architectural choice, this study supports the relevance of data-centric AI as a primary design principle for robust computer vision pipelines in agricultural and food-distribution applications.

II. RELATED WORK

A. Deep Learning-Based Object Detection

The introduction of convolutional neural networks (CNNs) as a dominant paradigm shifted object detection away from handcrafted feature engineering toward automatic extraction of hierarchical representations directly from image data [10]. That shift was not gradual. It was decisive. CNN-based pipelines quickly became the foundation for modern detectors across domains, and the architectural choices made in that period still shape how detection systems are built today.

Two broad families emerged from this transition. Two-stage approaches like R-CNN and Faster R-CNN first generate region proposals, then classify and refine them. Accurate, but computationally expensive. One-stage detectors like YOLO and SSD skip that intermediate step, predicting locations and classes in a single pass [11]. Speed gains are meaningful in practice, but not without tradeoffs in precision. Which approach works better depends entirely on the application: latency-sensitive systems lean toward one-stage models, tasks where missed detections are costly tend to stick with two-stage pipelines.

Small object detection is where both families struggle most. When targets occupy only a small fraction of the image, feature representations become sparse, spatial information is

lost during downsampling, and sensitivity to background clutter increases [12]. Detection accuracy for small objects remains significantly lower than for medium and large ones, even with state-of-the-art architectures. Multi-scale feature fusion and contextual aggregation strategies have been proposed to address this, with partial success [13]. The problem, however, is not purely architectural. In domains where objects are visually similar and data is collected under variable conditions, the quality and composition of the training set matter as much as the model design itself. Among one-stage detectors, YOLO has received particular attention in applied settings and warrants a closer look.

B. One-Stage and Two-Stage Detection Architectures

1) YOLO-Based Models

The shift from handcrafted features to deep learning transformed object detection. Within that transition, one-stage detectors emerged as a practical alternative to accuracy-focused but computationally expensive two-stage pipelines such as Faster R-CNN. YOLO (You Only Look Once) became the most representative method in this class, framing detection as a single regression problem across the entire image. Rather than generating region proposals and classifying them separately, YOLO predicts bounding boxes and class probabilities in a single forward pass, trading some precision for substantial speed gains [14], [15].

This efficiency has made YOLO the architecture of choice across a wide range of applied settings: aerial monitoring, UAV-based inspection, and agricultural surveillance, among others. Results in these domains are generally strong. Performance degrades noticeably, however, when models face occlusion, irregular backgrounds, or morphological similarity between target classes, factors that architectural refinement alone has not fully resolved [16], [17].

Small object detection compounds these difficulties further. When targets occupy only a fraction of the image, relevant features are easily lost during resizing or downsampling, a well-documented limitation in UAV-based applications where objects are frequently occluded or captured at varying scales [18]. Multi-scale detection has been proposed as a partial remedy, with encouraging results in controlled evaluations. What these solutions share, however, is that they remain architectural responses to what is, in part, a data problem.

Environmental variability such as inconsistent illumination, shifting viewpoints, and cluttered backgrounds affects generalization in ways that no single architectural choice fully resolves [6], [16]. Inter-class similarity and intra-class variability are consistently flagged as open challenges, yet the prevailing response is to adjust the model rather than examine the data [6], [17]. In fine-grained scenarios such as aromatic herb identification, where species may differ only in subtle morphological details, dataset quality and class separability are not secondary concerns. They are central to whether detection is reliable at all.

2) Faster R-CNN

Faster R-CNN operates differently from one-stage detectors. Rather than processing the entire image in a single pass, it first generates candidate regions using a Region Proposal Network (RPN). Then it performs classification and bounding-box regression on each proposed region. The two-stage design is slower, but it tends to pay off in scenarios involving small objects, partial occlusion, and high visual variability between classes [12].

A key architectural decision was integrating the RPN directly into the convolutional pipeline, allowing feature maps to be shared between the proposal and classification stages. This eliminated dependence on external proposal methods, such as Selective Search, and made end-to-end training both more efficient and more consistent [19]. The use of anchors across multiple scales and aspect ratios further enhances the model's ability to handle objects of varying sizes and shapes.

In practice, two-stage detectors generally outperform one-stage alternatives on small object benchmarks. Faster R-CNN has shown strong results in medical imaging, surveillance, and generic detection tasks, particularly in scenarios where localization accuracy matters more than inference speed [12], [19]. This advantage extends to low-light and thermal imaging contexts, where precise localization under challenging conditions is critical [20].

That said, small object detection remains a known limitation of deep CNN-based pipelines in general. Successive pooling and downsampling operations tend to reduce the spatial resolution of feature maps, which disproportionately affects small targets. Strategies such as multi-scale feature fusion, path augmentation, and customized anchor configurations have been proposed to mitigate this, with meaningful improvements reported across different application domains [12], [21].

For this work, Faster R-CNN serves as the two-stage baseline in the comparative evaluation and is implemented via the Detectron2 framework. Its inclusion allows for a direct assessment of how architectural differences interact with dataset quality under consistent experimental conditions.

3) Single Shot MultiBox Detector (SSD)

SSD was designed around a simple premise: eliminate the region proposal stage and perform localization and classification in a single forward pass [22]. Compared to two-stage detectors, this makes inference substantially faster. The tradeoff, as with YOLO, is some loss in precision, particularly for small objects.

What distinguishes SSD from other one-stage approaches is its multi-scale detection strategy. Rather than relying on a single feature representation, predictions are drawn from several convolutional layers at progressively decreasing spatial resolutions. This allows the model to handle objects of varying sizes without a dedicated proposal mechanism, and it

does so with computational costs low enough for real-time deployment [22].

That efficiency comes with known limitations. One-stage detectors depend heavily on anchor matching and feature map resolution, and objects that occupy only a small portion of the image tend to lose spatial detail during successive downsampling. This is especially problematic for thin or irregularly shaped objects, where anchor assignment strategies can produce suboptimal matches and reduce localization precision [23]. Improved localization metrics and adaptive anchor strategies have been proposed to address this, with some success.

Because of its reduced computational footprint, SSD has been widely adopted in embedded systems, industrial inspection pipelines, and edge computing scenarios where inference speed is a hard constraint [24]. More recent architectures have surpassed it in raw accuracy, particularly for small objects. Even so, SSD remains a relevant reference point in the evolution of one-stage detection and continues to influence how real-time frameworks are designed.

For this work, SSD serves as the lightweight baseline in the comparative evaluation, representing the efficiency-oriented end of the architectural spectrum alongside YOLO and Faster R-CNN.

Each of the three architectures reviewed here makes a different bet. Faster R-CNN prioritizes precision, accepting slower inference in exchange for more accurate localization. YOLO bets on speed, compressing the entire detection pipeline into a single pass. SSD sits somewhere in between, though closer to the speed end of the tradeoff [25]. For applications involving visually similar objects collected under variable conditions, such as the aromatic herb dataset examined in this work, the choice of architecture is only one piece of the puzzle. As the following sections discuss, the quality and composition of the training data may exert an equally significant influence on detection outcomes.

C. Object Detection in Agricultural Applications

Object detection in agricultural imagery remains challenging because field images often contain non-uniform illumination, occlusion, overlapping plant structures, scale variation, and visually complex backgrounds. Traditional pipelines based on hand-crafted features and classifiers like SVMs performed well under controlled conditions but quickly broke down when lighting shifted, plants overlapped, or the background was cluttered [26]. CNN-based models changed what was achievable in practice, mostly because they learned directly from data at a scale that older methods could not handle [26]. Single-stage detectors became the practical standard for field applications where response time matters. YOLO went through several generations, progressively improving the model's handling of scale variation, small targets, and cluttered scenes [27]. Two-stage detectors like Faster R-CNN localize with higher precision, but inference cost is high enough to be

a real constraint in many deployment scenarios [27],[28]. In agricultural practice, the gap between these families shows up less in benchmark tables and more in whether the system can run on the hardware available in the field.

The literature offers several instructive examples of where standard architectures hit their limits. In mango orchards, fruits at different growth stages appear beneath dense foliage with heavy overlap, conditions in which baseline YOLOv4 struggled with missed and false detections. Previous research [28] tackled this by incorporating DenseNet blocks into the backbone, improving how features propagated across layers. The result was a mAP of 96.20% compared to 91.47% for the unmodified model, with detection running at 44.2 FPS [28]. Inter-class similarity is a different kind of problem. A previous study [29] worked with 12 fruit blossom categories, in which species like apple and pear look nearly identical at the image level. Adding a contrastive learning component that explicitly separated class representations in the feature space reduced apple-to-pear misclassification by 4%, something that simply adding more parameters did not achieve [29]. Earlier authors [30] combined YOLOv5 with a regression head to estimate flea beetle damage per plant in oilseed rape. Plant detection was strong, but damage estimation degraded noticeably in severity ranges where training samples were scarce, even after augmentation [30].

Across these cases, the architecture was rarely the binding constraint. Occlusion, visual similarity between classes, and gaps in the training distribution each imposed ceilings that more complex models failed to overcome. Dataset composition, meaning which classes are included, how they are balanced, and what quality of images goes into training, appears to matter as much as model selection. That is the question this work is built around.

D. Data-Centric AI and Dataset Quality

The prevailing paradigm in machine learning research has long been model-centric: given a fixed dataset, researchers iteratively refine architectures, loss functions, and training procedures to achieve performance gains [8]. This approach treats data as a static artifact, overlooking a fundamental constraint: the quality and composition of the training set impose a ceiling that no architectural improvement can overcome. A growing body of work now advocates a data-centric perspective, in which the model is held stable while systematic effort is directed toward improving the data itself [31].

The distinction between these two paradigms carries practical consequences. A previous study [32] demonstrates this empirically in medical imaging: models for osteosarcoma detection trained on unannotated data plateaued at sensitivities between 60% and 70%, a ceiling that persisted regardless of architecture changes. Keeping the model fixed and investing instead in high-quality, expert-guided annotations raised sensitivity to 95.52% with an AUC of 0.989. The authors

conclude that for tasks where visual findings are subtle and training samples are scarce, the handling of training data matters more than the choice of model.

Dataset composition is equally consequential. Folkman et al. [33] expose this in underwater object detection, where mAP50 values exceeding 96% under random splitting collapsed to below 53% under domain-based splitting. Standard evaluation practices, they show, can mask severe distribution shift. Applying image enhancement preprocessing at the data level improved out-of-domain detection accuracy by up to 13.4 percentage points across two detector architectures, outperforming several model-centric domain generalization approaches. The same enhancements produced negligible gains under random splitting, confirming the improvements were attributable to the data rather than to incidental model effects.

A related concern is the structural quality of datasets themselves. A previous study [31] surveys dataset refinement methods in computer vision and identifies four recurring categories of dataset pathology: class imbalance, noisy labels, dataset bias, and high resource costs. A particularly important observation in their work is that augmenting a dataset that already contains problematic samples does not correct those problems but instead amplifies them. This challenges the common practice of treating augmentation as a general remedy for dataset deficiencies and argues for dataset refinement as a prerequisite to augmentation in any rigorous training pipeline.

Annotation quality represents a distinct but equally critical dimension. Earlier author [34] investigated this in aerial image segmentation through a controlled experiment in which five independent annotators labeled the same images. Inter-annotator agreement, measured via mean IoU, revealed strong class dependence: agreement reached 0.70 for visually unambiguous classes such as dormers, but fell to 0.15 for structurally ambiguous ones. The predictive performance of networks trained on this data tracked annotator agreement almost class by class, meaning classes that humans labeled inconsistently were precisely those the model struggled to predict.

Systematic annotation errors in large public datasets present an analogous problem. MS-COCO and the Open Images Dataset were analyzed, and three prevalent error types were identified [3]: missing annotations, incorrect bounding box sizes, and positional offsets. An automated correction pipeline they propose improves mAP by up to 32% across multiple detectors, including Faster R-CNN, EfficientDet, and RetinaNet, without modifying any model parameters. The authors note that most prior research kept these imperfect annotations fixed while investing effort in architectural modifications, leaving substantial performance gains on the table.

The translation of these principles to agricultural object detection is demonstrated by [2], who conduct a systematic data-centric analysis of on-tree fruit detection across seven

public datasets using a fixed YOLOv5s detector. Their results show that detection accuracy saturates beyond approximately 2,500 annotated objects for single-class tasks, and that training-validation similarity is strongly predictive of performance before any training is conducted. Object complexity inversely predicts accuracy, with morphologically irregular objects consistently yielding lower results. The most instructive case is the almond dataset, which underperformed relative to its complexity score due to approximately 15% mislabeled annotations, a direct consequence of high visual similarity between the fruit and surrounding foliage.

Whether the challenge is annotation noise [3], [34], distributional mismatch [33], structural dataset pathologies [31], label quality in specialized domains [32], or data-centric attributes in agricultural detection [2], the evidence consistently shows that targeted improvements to the data yield gains that are difficult to replicate through model refinement alone. This body of literature directly motivates the approach adopted in the present work: rather than focusing on architectural novelty, we concentrated on dataset construction quality, image filtering, and class selection as primary levers for improving detection performance.

Despite this evidence, few studies in agricultural object detection explicitly combine image-quality filtering, visual class-selection criteria, and multi-architecture benchmarking under a common held-out test set.

E. Image Quality Assessment (IQA)

Assessing the quality of images used to train object detection models is not a trivial task, as quality is multidimensional and context-dependent. In the literature, objective IQA methods are classified by whether a reference image is available for comparison. Full-reference approaches, such as SSIM, evaluate a distorted image against the original by jointly modeling luminance, contrast, and structural information. In many practical scenarios, however, no such reference exists, and blind or no-reference metrics must be used instead. A previous study [35] surveys this landscape in the context of remote sensing imagery, noting that pixel-level statistics such as MSE and PSNR are inadequate surrogates for perceptual quality and that methods aligned with the human visual system are necessary when images are destined for interpretation or automated analysis at scale.

The practical implications of this limitation become especially clear in agricultural image datasets, where data is collected under uncontrolled field conditions and no clean reference is available. Early research [36] addresses this challenge directly in a UAV-based study on potato plant stress detection, evaluating a battery of no-reference metrics including BRISQUE, NIQE, IL-NIQE, PIQUE, and PaQ-2-PiQ across multiple preprocessing and image fusion conditions. Their findings reinforce that different metrics capture different aspects of image degradation, and that relying on any single indicator risks overlooking relevant quality failures. The study also underscores a point directly applicable to web-scraped

datasets. When images originate from heterogeneous sources with no controlled acquisition protocol, no-reference evaluation is the only viable path.

Automation of this evaluation is equally important when working at the scale required by modern deep learning pipelines. Previous research [37] embedded a no-reference perceptual blur metric within a fully automated hyperspectral calibration workflow for high-throughput crop phenotyping, using it to trigger quality warnings and exclude compromised scans before they propagate through the processing chain. This positions IQA not as a post-hoc diagnostic but as an integral stage of data governance, a role it must play whenever dataset construction cannot be supervised sample by sample.

A related but distinct concern is whether an image that passes visual quality filters actually adds value to the dataset from a learning perspective. An earlier study [38] argues that conventional IQA, focused on degradation artifacts, says nothing about a sample's informational content relative to what the model has already seen. In a crop pest recognition study, they propose a distance-entropy indicator computed from the position of each candidate sample in the model's embedding space: images mapped to uncertain, boundary regions between classes carry high entropy. They are deemed informative, while those falling inside well-covered regions are treated as redundant. Selecting training samples by this criterion consistently improved classification accuracy over random selection, establishing informational value as a quality dimension that complements, rather than replaces, perceptual adequacy.

Visual adequacy alone is therefore an insufficient criterion for dataset curation in applied computer vision. Images must also contribute discriminative information not already well represented in the collection, and quality filtering must be feasible at scale without reference images. Both requirements shaped the curation methodology adopted in this work.

F. Class Separability and Visual Ambiguity

The visual separability between categories is a well-recognized factor in image classification performance, yet it receives comparatively little attention in dataset design. Standard loss functions such as cross-entropy treat all misclassifications equally, regardless of how visually similar two classes are. This assumption is fundamentally flawed [39]: the visual separability between object categories is highly uneven, and models systematically confuse categories whose visual features overlap. Through controlled experiments comparing datasets from visually similar and dissimilar categories, the authors showed a clear, measurable accuracy gap attributable solely to inter-class visual similarity, independent of model architecture. They further distinguished between semantic similarity and visual similarity, noting that machine learning models operate on visual feature space rather than semantic ontologies — a distinction with direct implications for class selection in applied recognition tasks.

This phenomenon is particularly relevant in fine-grained recognition of natural objects such as aromatic herbs, where multiple species share morphological characteristics including leaf shape, color, and texture. When visually indistinguishable classes are included in the same dataset, the model is exposed to irreducible ambiguity that no architectural improvement can fully resolve. Earlier author [40] quantified this effect directly in the latent space of object detectors, showing that low inter-class separability consistently predicts degraded detection performance across RetinaNet, FCOS, and YOLOv7. Their findings confirm that such degradation originates in the data rather than in model limitations. A principled response to this problem is to curate the class set based on visual separability criteria, excluding species whose appearance overlap would systematically degrade detection performance across all evaluated architectures. This data-centric strategy shifts the locus of intervention from the model to the dataset, consistent with the broader data-centric AI perspective discussed in Section II-D.

III. METHODOLOGY

A. Dataset Construction

1) Image Collection

Images were obtained from two complementary sources: original photographs captured at CEAGESP and controlled indoor settings, and web-scraped images retrieved using structured Portuguese and English keyword queries. After curation and dataset splitting, augmentation was applied only to the training subset. Collecting images encompasses three strategies. The first strategy consisted of original photographs taken directly at CEAGESP and in controlled indoor settings, capturing herbs under conditions that characterize the target deployment environment: variable lighting, overlapping products, and informal presentation. The second strategy relied on web scraping using the Bing Image Downloader library, with queries drawn from a structured keyword list prepared in both Portuguese and English. For each class, keyword combinations covered common names, cultivation contexts, commercialization settings, and visual presentation formats such as bundles, potted plants, and loose leaves in market environments. Queries were managed programmatically, with checkpoint-based resumption to handle timeouts during collection, and the third strategy was data augmentation.

2) Image Standardization and Deduplication

Following collection, all images for each class were consolidated into a single folder, and JPEG format was applied to ensure consistent input for later stages of processing. Next, duplicate and near-duplicate images were found and removed. This step was performed to reduce redundancy in the training set and minimize the risk of artificially inflated performance metrics due to repeated samples appearing in both the train and test splits.

3) Visual Curation Based on CEAGESP Context

After deduplication, a manual visual review was conducted with explicit reference to the CEAGESP distribution context. Images were assessed for compatibility with the appearance of herbs during handling and commercialization at the facility. Samples were removed when they presented low visual quality, excessive noise, inadequate botanical representation, or an appearance too distant from the commercial product, such as scientific illustrations, heavily processed photographs, or images depicting growth stages not representative of the point of sale. This curation step was intentionally conservative: when in doubt, images were excluded rather than retained.

4) Data Augmentation

Augmentation was applied using the Albumentations library. The goal was straightforward: exposing the model to more of the visual variation it would encounter in practice, without collecting additional images.

Five transformations were selected, each addressing a specific source of real-world variability. Gaussian blur mimics camera focus problems. Brightness adjustment covers the inconsistent lighting common in market and warehouse settings. Gaussian noise replicates sensor artifacts from lower-quality cameras. Random cropping introduces framing variation. A horizontal flip doubles the orientation diversity at no additional collection cost.

Per class, 100 images were randomly drawn from the curated pool and processed through each transformation independently, yielding up to 500 augmented candidates per class. These candidates were subsequently reviewed, and only those meeting the visual quality criteria established in Section 3.1.3 were retained. Approved samples were added exclusively to the training split, keeping validation and test sets free of artificially generated images.

5. Class Balancing and Dataset Split

Following curation and augmentation, the dataset was approximately balanced across the six classes. Four classes (basil, mint, bay leaf, and rosemary) were maintained at 950 images each. Leek reached 945 images and Sage 934, reflecting the volume available after quality filtering. The final dataset comprised 5,679 images in total.

The dataset was divided into training, validation, and test subsets using an approximately stratified split of 70%, 15%, and 15% per class. The test set was held out throughout all training runs and used exclusively for final evaluation, with no information from it influencing any training or hyperparameter decision.

B. Class Definition and Selection Criteria

Not every herb commercialized at CEAGESP was a viable target for detection. The fundamental constraint of this work is that all classification decisions must be made from RGB

images alone, with no access to aroma, texture under touch, or any other non-visual cue. That constraint drove the entire class selection process.

Species were evaluated against three questions: Can it be identified consistently from images collected under uncontrolled conditions? Is it visually distinct enough from the other candidate classes? Does it appear regularly enough in the CEAGESP context to justify its inclusion? Species that answered yes to all three were retained. Those that did not were dropped or folded into whichever candidate class best represented them visually.

Six species passed: basil, mint, sage, bay leaf, leek, and rosemary. Bay leaf is rigid and elongated. Sage has a grayish, velvety surface. Mint has serrated borders. Basil is broad and flat, glossy. Leek is linear and upright, structurally unlike any leafy class in the set. Rosemary has needle-like leaves on thin stems. All six hold up visually under variable lighting and partial occlusion.

The exclusions were equally principled. Chives, chive blossoms, and Chinese chives were indistinguishable in practice; their differences are sensory, not visual. Dill and fennel diverge mainly at the base, which was frequently missing from collected samples. Tarragon was too close to rosemary to separate reliably; rosemary stayed; tarragon did not. The same logic applies to marjoram relative to basil. Oregano and thyme have small, variable leaves that proved inconsistent across lighting conditions. Coriander and parsley share a fine leaf shape that market photography cannot reliably separate.

Celery deserves a separate note. It was considered alongside leek precisely because both are vertically oriented. The concern was that tall, upright herbs generate bounding boxes with extreme aspect ratios and frequent partial occlusion from neighboring products, conditions that complicate both annotation and detection. The leek results confirmed this: it was the weakest-performing class across all three models. Celery, sharing that structural profile, was excluded to avoid adding a second problematic class to an already difficult scenario.

The six retained classes form a set that is visually separable, contextually relevant, and annotatable with reasonable consistency. The impact of this selection on model performance is examined in Section V.

C. Annotation Pipeline

1) Auxiliary Model for Pre-annotation

Before annotating the full dataset, a lightweight auxiliary model was trained on an initial subset to accelerate labeling. The subset was small by design: 480 training images and 120 for validation, drawn automatically from the curated pool. A YOLOv11s initialized with pretrained weights was trained on it for 100 epochs. Best epoch results were mAP50 of 0.553 and mAP50-95 of 0.323 on validation — modest numbers, but sufficient for the task at hand. These results were sufficient for

the intended purpose of generating approximate bounding box proposals, but not for inclusion in the final model comparison.

2) Automatic Pre-annotation and Manual Review in CVAT

The best weights obtained from auxiliary training were imported into CVAT, an open-source annotation platform, and applied to generate automatic pre-annotations across the full dataset, including training, validation, and test images. All pre-annotations were subsequently reviewed manually, image by image. The review covered four things: wrong class labels, imprecise bounding boxes, false positives, and objects the auxiliary model missed entirely. All images were manually reviewed and corrected after automatic pre-annotation. The point was not to fix the auxiliary model; it was to make sure the final annotations did not inherit its mistakes.

3) Annotation Export and Validation

After manual review, annotations were exported from CVAT in YOLO format as text files containing class identifiers and normalized bounding box coordinates. The exported files were validated using custom Python scripts that verified image-annotation correspondence, the absence of empty annotation files, class identifier validity, and coordinate range consistency. Only images passing all validation checks were included in the final dataset.

D. Experimental Setup

1) YOLOv11s

Training ran on Windows using the Ultralytics framework in a Conda environment, with an NVIDIA GeForce RTX 4090 Laptop GPU as the hardware backend. The starting point was the YOLOv11s.pt pretrained checkpoint — fine-tuning made more sense than training from scratch given that the dataset has roughly 4,000 training images. One hundred epochs completed in about 3 hours. Inference on the test set averaged 21.1 ms per image when accounting for the full pipeline, including preprocessing, inference, and postprocessing. The inference-only component averaged approximately 1.2 ms per image on the same hardware, consistent with the latency profile reported for YOLOv11s in similar deployment scenarios.

2) Faster R-CNN via Detectron2

Faster R-CNN was set up under Ubuntu 24.04.2 LTS via WSL2, using Detectron2 with Python 3.12 in a dedicated virtual environment. Running it on Linux was not an arbitrary choice. Detectron2 has known compatibility issues on Windows, and WSL2 was the most stable path forward. The configuration used was faster_rcnn_R_50_FPN_3x, which pairs a ResNet-50 backbone with a Feature Pyramid Network. Since the original annotations were in YOLO format, they needed to be converted to COCO JSON before training could begin. Custom scripts handled this, producing instances_train.json, instances_val.json, and

instances_test.json. Fine-tuning from pretrained COCO weights ran for 20,000 iterations and finished in about 2 hours on the same RTX 4090 Laptop GPU used for the other models.

3) SSD300/VGG

SSD300 with a VGG backbone was trained using the ssd.pytorch repository on Windows with a dedicated Conda environment. Annotations were converted from YOLO format to Pascal VOC/XML format, and the repository was adapted to support six herb classes, including the background class. The model was initialized with VGG16 pretrained weights and trained for 20,000 iterations with a learning rate of 0.001, a batch size of 32, and learning rate decay at iterations 12,000 and 17,000. Training completed in approximately 6 hours and 24 minutes on the same GPU.

4) Evaluation of Metrics and Comparison Protocol

Due to differences in evaluation protocols across the three frameworks, a fully unified metric comparison is not achievable. AP50, defined as average precision at an IoU threshold of 0.5, is adopted as the common baseline for cross-model comparison, as all three evaluation pipelines natively support it. For YOLOv11s and Faster R-CNN, mAP50-95 is additionally reported as a stricter localization metric, reflecting average precision across IoU thresholds from 0.5 to 0.95 in steps of 0.05. This metric is not available for SSD300/VGG under the Pascal VOC evaluation protocol, which computes mAP only at IoU 0.5. This limitation is acknowledged in the comparative analysis presented in Section V. All models were evaluated on the same held-out test set of 856 images across six classes.

IV. EXPERIMENTS AND EVALUATION

A. Dataset and Experimental Setup

The final dataset comprised 5,679 images and 13,490 annotated herb instances distributed across six culinary herb classes: basil (2,350), mint (3,960), sage (2,410), bay leaf (1,960), leek (920), and rosemary (1,390). The difference between the number of images and annotated instances reflects the fact that several images contained more than one visible herb object or bundle. Bounding box annotations follow YOLO and VOC/COCO formats depending on the target framework. Spatial distribution analysis confirms that object centers are uniformly distributed across the image plane, while bounding box dimensions are predominantly small to medium, with normalized width and height concentrated below 0.3. This distribution reflects realistic conditions of partial occlusion and variable plant density in natural imagery. The dataset was partitioned into training (approximately 70%), validation (849 images), and held-out test (856 images) splits, with these splits held constant across all three models to ensure a fair comparison.

Three object detection architectures were evaluated: YOLOv11s, Faster R-CNN with a ResNet-50 backbone, and SSD300 with a VGG-16 backbone. YOLOv11s was trained for 100 epochs using SGD with cosine learning rate annealing (initial LR $\approx 9.9 \times 10^{-4}$, final LR $\approx 2.0 \times 10^{-5}$), with input images resized to 640×640 pixels, completing training in approximately 3 hours on an NVIDIA GeForce RTX 4090 (16 GB). Faster R-CNN was trained for 20,000 iterations using the Detectron2 framework with a stepped LR schedule (initial LR = 5.0×10^{-3} , decayed to 2.5×10^{-6}), and evaluated using the COCO API. SSD300/VGG was trained for 20,000 iterations with an initial LR of 1.0×10^{-3} that decayed at steps 12,000 and 17,000 ($\gamma = 0.1$), a batch size of 32, and evaluated using the VOC 2007 protocol, completing training in approximately 6 hours and 24 minutes. All models used pretrained backbone weights.

B. Training Convergence

YOLOv11s exhibited rapid and stable convergence across all three loss components (box regression, classification, and distribution focal loss). Training box loss decreased from 1.231 at epoch 1 to 0.603 at epoch 100, while classification loss dropped from 1.996 to 0.431, with no signs of overfitting. Validation mAP@0.5 reached 0.730 by epoch 20 and stabilized around 0.808 from epoch 40 onward. Faster R-CNN converged consistently over 20,000 iterations, with total loss decreasing from 2.434 at iteration 19 to 0.359 at iteration 19,999, a reduction of approximately 85%. Final foreground classification accuracy reached 94.8%, and the false negative rate stabilized at approximately 10%, confirming effective region proposal learning.

SSD300/VGG was also trained over 20,000 iterations, with the total loss (running average) decreasing from approximately 4.5 in the early iterations to 2.695 at completion, and learning-rate steps at iterations 12,000 and 17,000 produced visible reductions in loss. Training time was the longest among the three models at 6 hours and 24 minutes, reflecting the larger batch size and VOC-format data pipeline.

C. Validation Performance

Table I presents per-class and overall mAP results on the validation set for all three models.

TABLE I
Validation mAP@0.5 Per Class

Class	YOLOv11s	SSD300/VGG
Basil	0.809	0.730
Mint	0.811	0.762
Sage	0.823	0.749
Bay Leaf	0.865	0.790
Leek	0.839	0.756
Rosemary	0.765	0.723
mAP@0.5 (all)	0.819	0.752

† Faster R-CNN is not included in this validation table because per-class validation AP values were not extracted during the original Detectron2

evaluation. Faster R-CNN is included in the held-out test set comparison in Table II.

YOLOv11s achieved the highest validation mAP@0.5 of 0.819, outperforming SSD300/VGG (0.752) by 6.7 percentage points. Bay leaf consistently achieved the highest per-class AP across both models. At the same time, rosemary remained the most challenging class in all experiments, consistent with its fine-grained, needle-like morphology, which increases susceptibility to background confusion.

D. Test Set Performance

Table II summarizes the detection performance on the held-out test set across all three models.

TABLE II
Test Set Performance Comparison

Model	mAP@0.5	mAP@0.50:0.95	Basil	Mint	Sage	Bay Leaf	Leek	Rosemary
YOLOv11s	0.802	0.593	0.792	0.816	0.799	0.857	0.839	0.713
Faster R-CNN	0.798	0.541	0.740	0.790	0.780	0.840	0.878	0.759
SSD300/VGG	0.726	—	0.714	0.759	0.715	0.776	0.751	0.638

Note: Faster R-CNN per-class AP values reported as AP@0.5 ($\times 1$) derived from the COCO API. SSD300/VGG evaluated using VOC 2007 protocol (11-point interpolation); mAP@0.50:0.95 not available. YOLOv11s is evaluated using COCO-style interpolation.

YOLOv11s achieved the highest overall mAP@0.5 on the test set, 0.802, followed closely by Faster R-CNN at 0.798, a difference of only 0.4 percentage points. SSD300/VGG trailed both with a test mAP of 0.726. When evaluated under the stricter mAP@0.50:0.95 criterion, YOLOv11s (0.593) demonstrated a clear advantage over Faster R-CNN (0.541), indicating superior bounding box localization quality at higher IoU thresholds.

Notably, despite the near-equivalent mAP@0.5 between YOLOv11s and Faster R-CNN, their per-class profiles differ substantially. YOLOv11s achieves considerably higher AP on leek (0.839 vs. 0.418) and rosemary (0.713 vs. 0.388), while Faster R-CNN's bay_leaf AP (0.653) remains competitive but lower than YOLOv11s (0.857). This suggests that the single-stage architecture of YOLOv11s generalizes more uniformly across the dataset's visually heterogeneous classes.

SSD300/VGG, despite its lower overall mAP, shows more balanced per-class performance than Faster R-CNN, with all six classes above 0.63, suggesting that the multi-scale prior-box design of SSD provides reasonable coverage of the diversity of herb morphology.

E) Inference Speed

YOLOv11s was the only model explicitly profiled for inference latency, achieving approximately 1.2 ms per image for model inference alone and 21.1 ms per image for the full preprocessing-inference-postprocessing pipeline on the RTX 4090. SSD300/VGG processed the validation set of 849 images in approximately 13 seconds under its evaluation routine, corresponding to an approximate throughput estimate of 6.87 images per second at batch size 10. Faster R-CNN inference time was not explicitly logged in this experiment. Therefore, speed-related conclusions are restricted to measured or directly observed values.

From a deployment perspective, YOLOv11s remains the most practical candidate among the evaluated models because it combines the highest test-set $\text{mAP}@0.5$ and $\text{mAP}@0.50:0.95$ with explicitly measured low-latency inference. SSD300/VGG provides a lightweight single-stage alternative, although with lower test accuracy. Faster R-CNN is retained as a two-stage accuracy-oriented baseline; however, without measured wall-clock inference time in this experiment, its deployment cost is discussed qualitatively rather than treated as a quantified comparison.

F) Summary

Among the three architectures evaluated, YOLOv11s stood out as the most competitive option. It recorded the highest $\text{mAP}@0.5$ (0.802) and $\text{mAP}@0.50:0.95$ (0.593), while also showing more consistent detection across all six herb classes and considerably lower inference latency. Faster R-CNN came close on $\text{mAP}@0.5$, yet at a much steeper computational cost; moreover, its per-class results were notably uneven, with leek and rosemary proving particularly problematic. SSD300/VGG, by contrast, distributed its errors more evenly across classes, though its overall accuracy remained below the other two models. Taken together, these results indicate that YOLOv11s is the most promising candidate among the tested models for real-time aromatic-herb detection, although deployment-level validation under operational CEAGESP conditions remains necessary.

V. CONCLUSION

This work examined whether systematic data curation, specifically image quality filtering and class selection grounded in visual separability, could drive object detection performance in a fine-grained aromatic herb dataset without relying on architectural modifications.

The answer emerging from the experiments is affirmative. YOLOv11s, Faster R-CNN, and SSD300/VGG were each trained on the same curated set of 5,679 images and evaluated on a common held-out test set. Their $\text{mAP}@0.5$ scores ranged from 0.726 to 0.802. What is notable is not the gap between them, but how narrow it is across three models with fundamentally different designs. A dataset with substantial noise or semantic ambiguity would tend to

produce more divergent outcomes. The relative consistency observed here points to the dataset itself as a stabilizing factor.

Within the curation pipeline, two decisions most directly shaped the outcome. Filtering images by blur, illumination, and noise severity likely reduced sources of visual inconsistency that could impair model learning, particularly in a dataset composed of web-scraped and field-captured images. Class selection was equally consequential. Species such as dill and fennel, or the members of the chive family, were excluded not for convenience but because their visual overlap with retained classes could not be resolved from RGB images under real acquisition conditions. Keeping them would have embedded ambiguity into the training set at the class level, a problem that lies outside the reach of any detection architecture.

YOLOv11s produced the strongest results overall. Its $\text{mAP}@0.5$ of 0.802 and $\text{mAP}@0.50:0.95$ of 0.593 on the test set, combined with an inference time around 1.2 ms per image, make it the most practical option for deployment in operational settings. Faster R-CNN came close with an $\text{mAP}@0.5$ of 0.798, though its per-class results were uneven and its computational demands were considerably higher. SSD300/VGG reached $\text{mAP}@0.5$ of 0.726 with a more consistent per-class distribution. Rosemary was the hardest class for all three models, a finding consistent with its needle-like structure and tendency to blend into cluttered backgrounds. Bay leaf was the most reliably detected, favored by its distinct elliptical shape.

Compared to the earlier study reported in [7], which used a single YOLOv8 model, roughly 800 training images, and three classes, reaching a consolidated mAP50 of 0.723 while flagging over-detection and low recall as open problems, the present results show clear progress. Expanding the dataset, applying structured quality filters, and defining classes by visual separability raised mAP50 to 0.802 and mAP50-95 to 0.593. That the improvement required no change to the model architecture is the more telling result: it suggests that the earlier performance ceiling was a data problem, not a modeling one.

For researchers facing similar recognition tasks in agricultural or supply chain contexts, the practical implication is that effort directed at dataset construction and class definition is likely to produce more reliable gains than effort directed at model selection. Future studies should compare curated and noncurated datasets, evaluate each curation stage via ablation experiments, test external generalization with newly collected CEAGESP images, and assess deployment feasibility on mobile or edge computing devices.

REFERENCES

- [1] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023, doi: 10.1109/JPROC.2023.3238524.

- [2] A. Wang, J. Tang, and M. Whitty, "Data-centric analysis of on-tree fruit detection: Experiments with deep learning," *Comput. Electron. Agric.*, vol. 194, p. 106748, Mar. 2022, doi: 10.1016/j.compag.2022.106748.
- [3] S. Nou, J.-S. Lee, N. Ohyama, and T. Obi, "The improvement of ground truth annotation in public datasets for human detection," *Mach. Vis. Appl.*, vol. 35, p. 49, Apr. 2024, doi: 10.1007/s00138-024-01527-1.
- [4] D. Malerba and V. Pasquidibisceglie, "Data-centric AI," *J. Intell. Inf. Syst.*, vol. 62, pp. 1493–1502, Oct. 2024, doi: 10.1007/s10844-024-00901-9.
- [5] R. Avanzato and F. Beritelli, "Automatic ECG diagnosis using convolutional neural network," *Electronics*, vol. 9, no. 6, p. 951, Jun. 2020, doi: 10.3390/electronics9060951.
- [6] J. Zhu, K. Yang, N. Guan, X. Yi, and C. Qiu, "HCPNet: Learning discriminative prototypes for few-shot remote sensing image scene classification," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 123, p. 103447, Aug. 2023, doi: 10.1016/j.jag.2023.103447.
- [7] S. N. Antunes, M. T. Okano, I. A. Nääs, W. A. C. Lopes, F. P. L. Aguiar, O. Vendrametto, J. C. L. Fernandes, and M. E. Fernandes, "Model development for identifying aromatic herbs using object detection algorithm," *AgriEngineering*, vol. 6, no. 3, pp. 1924–1936, Jun. 2024, doi: 10.3390/agriengineering6030112.
- [8] M. H. Jarrahi, A. Memariani, and S. Guha, "The principles of data-centric AI," *Commun. ACM*, vol. 66, no. 8, pp. 84–92, Aug. 2023, doi: 10.1145/3571724.
- [9] D. Im, S. Lee, H. Lee, B. Yoon, F. So, and J. Jeong, "A data-centric approach to design and analysis of a surface-inspection system based on deep learning in the plastic injection molding industry," *Processes*, vol. 9, no. 11, p. 1895, Oct. 2021, doi: 10.3390/pr9111895.
- [10] T. Hoerster and C. Kuenzer, "Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends," *Remote Sensing*, vol. 12, no. 10, p. 1667, May 2020, doi: 10.3390/rs12101667.
- [11] I. Diaz-Cano, A. Morgado-Estevéz, J. M. Rodríguez Corral, P. Medina-Coello, B. Salvador-Dominguez, and M. Alvarez-Alcon, "Automated fillet weld inspection based on deep learning from 2D images," *Appl. Sci.*, vol. 15, no. 2, p. 899, Jan. 2025, doi: 10.3390/app15020899.
- [12] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep learning methods for small object detection," *Expert Syst. Appl.*, vol. 172, p. 114602, Jun. 2021, doi: 10.1016/j.eswa.2021.114602.
- [13] Y. Huang, Y. Qian, H. Wei, Y. Lu, B. Ling, and Y. Qin, "A survey of deep learning-based object detection methods in crop counting," *Comput. Electron. Agric.*, vol. 215, p. 108425, Dec. 2023, doi: 10.1016/j.compag.2023.108425.
- [14] L. Jiao and M. I. Abdullah, "YOLO series algorithms in object detection of unmanned aerial vehicles: a survey," *Serv. Oriented Comput. Appl.*, vol. 18, pp. 269–298, Mar. 2024, doi: 10.1007/s11761-024-00388-w.
- [15] R. Sharma, K. Patel, S. Shah, and M. Aibin, "Aerial footage analysis using computer vision for efficient detection of points of interest near railway tracks," *Aerospace*, vol. 9, no. 7, p. 370, Jul. 2022, doi: 10.3390/aerospace9070370.
- [16] B. Deng, Y. Lu, and J. Xu, "Weed database development: An updated survey of public weed datasets and cross-season weed detection adaptation," *Ecol. Inform.*, vol. 81, p. 102546, Jun. 2024, doi: 10.1016/j.ecoinf.2024.102546.
- [17] J. Zhang, Y. Li, Z. Tong, L. He, M. Zhang, Z. Niu, and H. He, "GLCANet: Global-local context aggregation network for cropland segmentation from multi-source remote sensing images," *Remote Sens.*, vol. 16, no. 24, p. 4627, Dec. 2024, doi: 10.3390/rs16244627.
- [18] B. Lawrence, E. de Lemmus, and H. Cho, "UAS-based real-time detection of red-cockaded woodpecker cavities in heterogeneous landscapes using YOLO object detection algorithms," *Remote Sens.*, vol. 15, no. 4, p. 883, Feb. 2023, doi: 10.3390/rs15040883.
- [19] L. Aziz, M. S. B. H. Salam, U. U. Sheikh, and S. Ayub, "Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review," *IEEE Access*, vol. 8, pp. 170461–170495, Sep. 2020, doi: 10.1109/ACCESS.2020.3021508.
- [20] A. Sai Charan, M. Jitesh, M. Chowdhury, and H. Venkataraman, "ABIFN: Attention-based bi-modal fusion network for object detection at night time," *Electron. Lett.*, vol. 56, no. 24, pp. 1309–1311, Nov. 2020, doi: 10.1049/el.2020.1952.
- [21] J. Xu, H. Ren, S. Cai, and X. Zhang, "An improved faster R-CNN algorithm for assisted detection of lung nodules," *Comput. Biol. Med.*, vol. 153, p. 106470, Feb. 2023, doi: 10.1016/j.combiomed.2022.106470.
- [22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherlands, 2016, pp. 21–37.
- [23] L. Gou, S. Wu, J. Yang, H. Yu, and X. Li, "Gaussian guided IoU: A better metric for balanced learning on object detection," *IET Comput. Vis.*, vol. 16, no. 6, pp. 556–566, 2022, doi: 10.1049/cvi2.12113.
- [24] A.-A. Tulbure, A.-A. Tulbure, and E.-H. Dulf, "A review on modern defect detection models using DCNNs – Deep convolutional neural networks," *J. Adv. Res.*, vol. 35, pp. 33–48, 2022, doi: 10.1016/j.jare.2021.03.015.
- [25] N. Bilous, V. Malko, M. Frohme, and A. Nepochenko, "Comparison of CNN-based architectures for detection of different object classes," *AI*, vol. 5, no. 4, pp. 2300–2320, Nov. 2024, doi: 10.3390/ai5040113.
- [26] A. J., J. Eunice, D. E. Popescu, M. K. Chowdhury, and J. Hemanth, "Deep learning-based leaf disease detection in crops using images for agricultural applications," *Agronomy*, vol. 12, no. 10, p. 2395, Oct. 2022, doi: 10.3390/agronomy12102395.
- [27] S. Kamalesh Kanna, R. Kumaraperumal, P. Pazhanivelan, R. Jagadeeswaran, and P. C. Prabu, "YOLO deep learning algorithm for object detection in agriculture: a review," *J. Agric. Eng.*, vol. 55, no. 4, p. 1641, 2024, doi: 10.4081/jae.2024.1641.
- [28] A. M. Roy and J. Bhaduri, "Real-time growth stage detection model for high degree of occultation using DenseNet-fused YOLOv4," *Comput. Electron. Agric.*, vol. 193, p. 106694, Feb. 2022, doi: 10.1016/j.compag.2022.106694.
- [29] W. Zhou, Y. Cui, H. Huang, H. Huang, and C. Wang, "A fast and data-efficient deep learning framework for multi-class fruit blossom detection," *Comput. Electron. Agric.*, vol. 217, p. 108592, Feb. 2024, doi: 10.1016/j.compag.2023.108592.
- [30] A. Bereciartua-Pérez, M. Monzón, D. Múgica, G. De Both, J. Baert, B. Hedges, N. Fox, J. Echazarra, and R. Navarra-Mestre, "Estimation of flea beetle damage in the field using a multistage deep learning-based solution," *Artif. Intell. Agric.*, vol. 13, pp. 18–31, 2024, doi: 10.1016/j.iaia.2024.06.001.
- [31] Z. Wan, Z. Wang, C. Chung, and Z. Wang, "A survey of dataset refinement for problems in computer vision datasets," *ACM Comput. Surv.*, vol. 56, no. 7, Art. no. 172, Apr. 2024, doi: 10.1145/3627157.
- [32] J. Hasei et al., "High-quality expert annotations enhance artificial intelligence model accuracy for osteosarcoma X-ray diagnosis," *Cancer Sci.*, vol. 115, no. 11, pp. 3695–3704, Nov. 2024, doi: 10.1111/cas.16330.
- [33] L. Folkman, K. A. Pitt, and B. Stantic, "A data-centric framework for combating domain shift in underwater object detection with image enhancement," *Appl. Intell.*, vol. 55, Art. no. 272, Jan. 2025, doi: 10.1007/s10489-024-06224-0.
- [34] S. Krapf, L. Bogenrieder, F. Netzler, G. Balke, and M. Lienkamp, "RID—Roof information dataset for computer vision-based photovoltaic potential assessment," *Remote Sens.*, vol. 14, no. 10, Art. no. 2299, May 2022, doi: 10.3390/rs14102299.
- [35] G. Chen, Q. Pei, and M. M. Kamruzzaman, "Remote sensing image quality evaluation based on deep support value learning networks," *Signal Process. Image Commun.*, vol. 83, p. 115783, Apr. 2020, doi: 10.1016/j.image.2020.115783.
- [36] S. Modak, J. Heil, and A. Stein, "Pansharpening low-altitude multispectral images of potato plants using a generative adversarial network," *Remote Sens.*, vol. 16, no. 5, p. 874, Mar. 2024, doi: 10.3390/rs16050874.
- [37] V. Sagan et al., "Data-driven artificial intelligence for calibration of hyperspectral big data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, p. 5510320, 2022, doi: 10.1109/TGRS.2021.3091409.
- [38] Y. Li and X. Chao, "Distance-entropy: An effective indicator for selecting informative data," *Front. Plant Sci.*, vol. 12, p. 818895, Jan. 2022, doi: 10.3389/fpls.2021.818895.

-
- [39] X. Yu, Y. Chen, G. Zhou, Y. Liu, F. Li, and Z. Wang, "Error Refactor loss based on error analysis in image classification," *IET Computer Vision*, vol. 16, no. 2, pp. 192–203, 2022. doi: 10.1049/cvi2.12079
- [40] J. Ammeling, J. Ganz, F. Wilm, K. Breininger, and M. Aubreville, "Investigation of class separability within object detection models in histopathology," *IEEE Transactions on Medical Imaging*, vol. 44, no. 8, pp. 3162–3174, Aug. 2025. doi: 10.1109/TMI.2025.3560134