Evaluation and Optimization of an AI Model for European Canker Detection in Apple Trees

Camile Coelho Arruda[1], **Jonatam Sturcio Corrêa**[1],
Wilson Castello Branco Neto[1], **Robson Costa**[1]

[1]Instituto Federal de Santa Catarina (IFSC) - Campus Lages
Rua Heitor Villa Lobos, 225 - 88.506-400 - Lages - SC - Brasil

camilecoelho23@gmail.com, jonatamsturcio.c@gmail.com
wilson.castello@ifsc.edu.br, robson.costa@ifsc.edu.br

Abstract. *This paper presents a study on the use of Convolutional Neural Networks (CNNs) for the detection of European Canker (Neonectria ditissima) in apple tree leaves. Several CNN architectures were experimentally evaluated using Data Augmentation, ensemble strategies, and threshold-based decision methods, each tested over ten independent replications to ensure robustness and reproducibility. The experiments demonstrated the effectiveness of these approaches for reliable and accurate disease detection, highlighting their potential to support early diagnosis and management decisions in apple orchards. The best-performing models achieved test-set precision values ranging from 0.801138 to 0.871632 and accuracy values from 0.76542 to 0.842679, which are comparable to those obtained by two agronomists who evaluated the same image set, with precision scores of 0.792207 and 0.885496 and accuracy values ranging from 0.838006 to 0.872274.*

1. Introduction

Plant diseases represent one of the major challenges facing Brazilian agriculture, directly impacting crop productivity, food quality, and farmers' profitability. Identifying causal agents, understanding disease transmission mechanisms, and implementing effective management strategies are critical steps to protect agricultural systems and ensure food security. In this context, the adoption of technological tools that support accurate and timely disease diagnosis plays a fundamental role in mitigating losses and promoting sustainable agricultural practices.

Figure 1. Symptoms of European canker (*Neonectria ditissima*) on an apple tree branch, showing necrotic tissue, bark cracking, and canker formation.

"European canker is one of the most devastating fungal diseases of apple in most temperate regions. The causal agent, *Neonectria ditissima*, infects trees through wounds in the bark forming cankers that girdle the stem and eventually cause tree death." (Harteveld et al., 2023, p. 1). In Brazil, particularly in the southern regions, European canker is a growing concern for apple producers. The disease primarily affects the woody parts of the plant, as shown in Figure 1, and can also cause fruit rot, leading to substantial economic losses. Effective management strategies are crucial to mitigate these impacts (Gelain and De Mio, 2019).

The economic impact of European canker is particularly severe in Brazil, where the states of *Santa Catarina* and *Rio Grande do Sul* account for approximately 90% of national apple production (EPAGRI, 2024). The disease leads to seedling mortality and a decline in plant vigor, directly affecting yield and fruit quality. Economic studies indicate that financial losses associated with the disease are substantial. For instance, a technical report from Brazilian Agricultural Research Corporation (EMBRAPA – *Empresa Brasileira de Pesquisa Agropecuária*) highlights the economic and financial burden on apple production systems in *Vacaria, Rio Grande do Sul*, due to additional control costs and reduced profitability (Lazzarotto and Alves, 2015). Beyond direct expenses related to management and disease control, producers face productivity losses and the need to replant affected areas, increasing operational costs and reducing profit margins. These factors underscore the necessity for effective monitoring and management strategies to mitigate both the economic and environmental consequences of European canker.

Therefore, early detection of European canker is essential to minimize damage and prevent the spread of the disease. However, visual diagnosis can be challenging and requires expert knowledge, emphasizing the need for more accurate and accessible diagnostic methods.

The Cancontrol platform[1] was developed in 2021 through a partnership between the Federal Institute of Santa Catarina (IFSC – *Instituto Federal de Santa Catarina*) and the *Santa Catarina* Agricultural Research and Rural Extension Company (Epagri – *Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina*). Since its release, it became one of the main resources available for monitoring European canker in apple orchards. Cancontrol allows producers to submit images of their apple trees for analysis by *phytopathologists*[2]. These experts assess the submitted photos and provide diagnostic results on the presence of the disease. In addition, Cancontrol offers educational materials, including news articles, images, and videos, to raise awareness of European canker (Branco Neto et al., 2021).

However, this process still relies on human evaluation, which can result in diagnostic delays. The integration of computer vision models has the potential to enhance disease identification by enabling an initial automated screening, thus reducing the workload of specialists. With advances in artificial intelligence, Convolutional Neural Networks (CNNs) have become promising tools to assist in plant disease identification. These models have already been successfully applied to the detection of various plant diseases across multiple species, demonstrating their potential for agricultural diagnostics (Mohanty et al., 2016). Incorporating this model into Cancontrol could improve diagnostic

[1]Official website of the Cancontrol platform: `http://www.cancroeuropeu.com.br`.

[2]Specialized professional in the study of plant diseases.

efficiency by providing faster results and supporting decision-making in apple orchard management.

This work presents the development and evaluation of a Convolutional Neural Network (CNN)-based models for the identification of European canker in apple orchards. Although the proposed solution is designed to be compatible with future integration into the Cancontrol system, this work focuses on studying and identifying the most effective model for disease detection. The main goal is to improve detection accuracy and support decision-making by farmers and experts through reliable and efficient image-based diagnosis. To guide the development of this work, the following specific objectives have been established:

- Investigate previous studies on the application of CNNs for European canker detection, with emphasis on the theoretical and practical foundations that support their use.
- Conduct experiments using the technique that showed the best performance in previous research, applying it to a new dataset to determine whether improvements in accuracy are due to a more robust model or to the quality and diversity of the images.
- Analyze strategies, such as *Data Augmentation*[3] techniques and the use of *Ensembles methods*[4] to mitigate *overfitting*[5] and improve model performance.
- Evaluate the impact of different threshold values on image classification, aiming to determine the value that provides the best balance between sensitivity and precision.
- Design and run a comprehensive series of tests that integrate the aforementioned techniques to validate the improvements achieved.

This integrated approach ensures a systematic progression from theoretical exploration to experimental validation and, ultimately, to practical application, contributing to the development of an effective tool for the early and accurate detection of European canker in apple orchards.

To achieve the proposed objectives, this study will begin with a review of the existing literature on the use of CNNs for plant disease detection, with an emphasis on European canker. This review will provide the theoretical foundation to guide model selection and training strategies. Subsequently, the CNN architecture that demonstrated the best performance in previous research will be implemented and trained using a new dataset of apple tree images affected by the disease.

To improve the model's generalization and mitigate overfitting, techniques such as Data Augmentation (DA) and Ensemble methods will be applied. Throughout the development process, performance metrics, including accuracy, precision, recall, and F1-score, will be employed to evaluate the model.

[3]Data Augmentation is a technique used to artificially expand a dataset by generating modified versions of real images through transformations including rotation, flipping, and scaling.

[4]Ensembles are machine learning techniques in which multiple models are combined to generate predictions that are more accurate and robust than those of a single model.

[5]Overfitting occurs when a machine learning model learns the training data too closely, including its noise or anomalies, resulting in poor generalization to unseen data.

This paper is organized into four main sections. Section 2 presents the theoretical background, including the state of the art in the application of artificial intelligence for plant disease detection, with a particular focus on Convolutional Neural Networks (CNNs). This section will also review related studies, investigating whether CNNs have previously been applied to the detection of European canker or other diseases. Section 3 describes the methodology employed, detailing the development of the proposed system, from data collection and preprocessing to model training and evaluation. Section 4 presents the results of the experiments conducted. Finally, Section 5 presents the final considerations, the contributions of this work, its limitations, and potential directions for future research.

2. Literature Review

This section presents the theoretical foundation supporting the development of a deep learning-based system for detecting European canker in apple orchards. The review is structured to encompass both agricultural and computational perspectives. Section 2.1 introduces key concepts in plant pathology, highlighting the importance of disease management in apple cultivation and the specific challenges posed by European canker, particularly the absence of image-based diagnostic tools. Section 2.2 provides an overview of CNNs, detailing their architecture, the use of transfer learning, and prominent model families. Section 2.3 discusses DA techniques used to enhance model generalization by transforming training data. Section 2.4 explores Ensemble learning strategies, describing their categories and relevance to plant disease classification. Finally, Section 2.5 reviews relevant prior studies that contextualize and justify the proposed research.

2.1. Phytopathology

Plant pathology is the scientific discipline dedicated to understanding plant diseases caused by pathogens such as fungi, bacteria, and viruses, as well as by adverse environmental conditions. These diseases have a significant impact on crop health, yield, and quality, making early detection and effective management essential to ensure sustainable agricultural practices (Ahmed and Yadav, 2023; Jafar et al., 2024). In apple cultivation, phytosanitary measures are crucial to ensure productivity and fruit quality. Apple orchards are susceptible to various diseases, including apple scab, powdery mildew, and fire blight, which can lead to substantial economic losses if not properly controlled (Assis, 2023; Singh et al., 2024).

Among these, European canker (*Neonectria ditissima*) is particularly severe, affecting branches and trunks and leading to necrotic lesions, dieback, and, in advanced stages, tree mortality. The disease typically presents subtle symptoms during its early stages, rendering timely detection a challenge. Manual inspections — currently the standard practice — are time-consuming, labor-intensive, and prone to subjectivity (Lazzarotto and Alves, 2015). Furthermore, visual inspections often depend on the availability of trained specialists, whose presence may be scarce in rural or large-scale orchard settings.

To illustrate the contrast between leaf-infecting diseases and those that affect woody tissues, Figure 2 presents a side-by-side comparison between apple scab, a foliar disease with clearly distinguishable lesions, and European canker, whose symptoms

on the trunk are far more subtle. Leaf diseases typically exhibit characteristic spots, making visual diagnosis easier. In contrast, European canker produces lesions that frequently merge with the natural bark texture, thereby significantly increasing the difficulty of field detection.

Image-based diagnosis has emerged as a promising alternative for plant disease detection, allowing symptom identification through computational analysis of visual patterns on leaves, stems, and trunks. Deep Learning (DL) models, particularly Convolutional Neural Networks (CNNs), have achieved high effectiveness in detecting foliar diseases such as apple scab, rust, and black rot, with precision levels exceeding 85% in some studies (Assis, 2023). These approaches has been increasingly enhanced through techniques such as Data Augmentation, Transfer Learning, and Ensemble modeling.

(a) European canker on the trunk, showing subtle lesions that are difficult to visually detect.

(b) Apple scab symptoms on leaves, which are visually distinct and easier to identify.

Figure 2. Comparison between foliar symptoms (apple scab) and woody tissue symptoms (European canker).

Despite such advancements, deploying these systems in real-world scenarios still poses significant challenges. Image datasets used for training are often captured under controlled lighting and environmental conditions, which differ significantly from real-world field conditions, resulting in reduced performance under natural settings (Xu et al., 2023). In addition, symptoms of various plant diseases may overlap, and environmental factors can obscure key visual cues. Deep learning models remain sensitive to dataset quality, balance, and variability, often exhibiting reduced performance when confronted with underrepresented or noisy classes in the dataset (Shafay et al., 2025).

Research efforts on European canker remain limited. An initial attempt by Mattos and Ribeiro (2022) employed CNNs on a limited dataset comprising infected apple tree images. Although some promising outcomes were reported, the model's performance was compromised by overfitting and insufficient data diversity. A subsequent study by Salvi and Camargo Jr. (2023) built upon these findings by applying DA strategies and evaluating additional architectures such as *ResNet* and *InceptionV3* in an attempt to enhance classification accuracy. Although some modest improvements were observed, the fundamental challenge of dataset scarcity persisted, particularly due to the discrete and trunk-localized nature of canker symptoms, which are visually subtle and more difficult to capture than

foliar lesions.

These findings highlight a critical gap in the current literature. Although several public datasets exist for apple leaf disease detection, such as PlantVillage (Hughes and Salathé, 2015), the Plant Pathology 2020 Challenge dataset (Thapa et al., 2020), and AppleLeaf (Yang et al., 2022), all of them focus exclusively on foliar symptoms. None of these datasets include images of European canker infections, which are essential for European canker identification. This lack of publicly available and specific datasets limits reproducibility and restricts the development of scalable and field-ready diagnostic systems.

They focus exclusively on foliar symptoms and do not include any canker-related diseases. The absence of images depicting canker infections limits reproducibility and constrains the development of scalable and field-ready diagnostic systems tailored to stem and bark diseases.

To address these limitations, the present work proposes a more robust diagnostic pipeline that combines an expanded and curated image dataset with advanced deep learning methodologies. The goal is to improve the accuracy, robustness, and generalizability with the aim of enabling earlier detection and more effective control.

2.2. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a class of deep learning models optimized to process grid-like data structures, such as images. Their architecture typically includes convolutional layers that apply filters to extract spatial features, pooling layers to reduce dimensionality, and fully connected layers for final classification (Rana et al., 2022). Due to their ability to automatically learn hierarchical patterns directly from raw pixel data, CNNs have become the state-of-the-art for visual recognition tasks, including plant disease detection.

A key enabler of CNN-based solutions, particularly in domains with limited labeled data, is *transfer learning*. This technique involves reusing models pre-trained using large-scale datasets such as ImageNet and fine-tuning them on more specific, often smaller, datasets. Transfer learning not only accelerates training but also improves generalization by leveraging features learned from a broader visual domain.

Several well-known CNN architectures have successfully been used in transfer learning scenarios. Each architecture presents distinct characteristics that make it more or less suitable depending on task requirements, hardware constraints, and dataset complexity:

- *VGG* (e.g., *VGG16, VGG19*): characterized by a simple and uniform structure that uses 3×3 convolutional filters and max pooling. *VGG* models are easy to implement and perform well in general classification tasks. However, they contain a large number of parameters, which increases training time and memory usage (Simonyan and Zisserman, 2015).
- *ResNet* (e.g., *ResNet18, ResNet50, ResNet101*): introduces residual connections that facilitate the training of very deep networks by allowing gradients to propagate directly through shortcut connections, thereby mitigating the vanishing gradient problem. This architecture improves accuracy and stability, especially in complex tasks (He et al., 2016).

- *MobileNet* (e.g., *MobileNetV1, MobileNetV2*): designed for efficiency in mobile and embedded applications. These models use depthwise separable convolutions and, in the case of *MobileNetV2*, incorporate inverted residuals and linear bottlenecks, which significantly reduce computational demands and model size while maintaining competitive accuracy (Howard et al., 2017; Sandler et al., 2018).
- *EfficientNet* (B0–B7): employs compound scaling to systematically scale depth, width, and resolution, achieving state-of-the-art performance with fewer parameters. *EfficientNet* offers an excellent balance between accuracy and efficiency, making it suitable for both cloud and edge-based deployment (Tan and Le, 2019).
- *AlexNet*: a pioneering architecture in deep learning, *AlexNet* helped popularize CNNs after its success in the 2012 *ImageNet* competition. Although relatively shallow by contemporary standards, it remains relevant for educational purposes and tasks with low complexity (Krizhevsky et al., 2012).
- *DenseNet* (e.g., *DenseNet121, DenseNet201*): employs dense connections in which each layer receives input from all preceding layers, promoting feature reuse and efficient gradient flow. *DenseNet* architectures achieve strong performance with fewer parameters, particularly in deep networks (Huang et al., 2017).

When choosing a CNN architecture, several factors must be considered: the size and complexity of the dataset, computational resources, target hardware (e.g., mobile vs. server), latency constraints, and the required balance between accuracy and efficiency. Lightweight models like *MobileNet* are well-suited for real-time applications on resource-constrained devices, while deeper models like *ResNet* and *EfficientNet* are preferable for high-accuracy tasks where computational resources are more abundant.

To illustrate the typical structure of a CNN, Figure 3 presents the *VGG19* model as a representative example. Although this model is not the focus of this study, it is used here to illustrate the general structure of CNNs, which typically include convolutional layers, pooling layers and fully connected layers. Other architectures, such as *ResNet*, *EfficientNet* and *MobileNet*, differ mainly in the number of layers, type of connections and convolutional filters used.

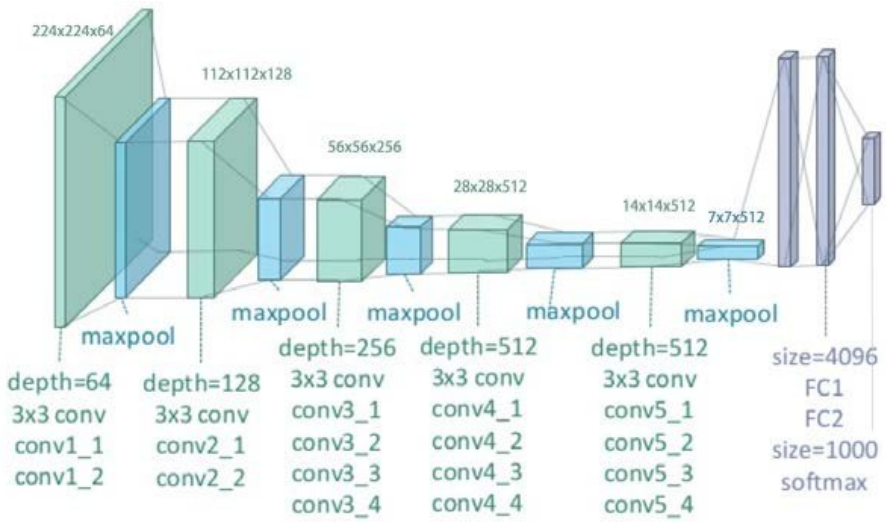


Figure 3. Example of a Convolutional Neural Network (CNN) architecture based on *VGG19* (Simonyan and Zisserman, 2014).

2.3. Data Augmentation

Data Augmentation (DA) is a strategy employed in machine learning to artificially expand the size and diversity of training datasets by applying various transformations to existing data. This technique is particularly important in domains such as agriculture, where collecting large, annotated image datasets is labor-intensive and time-consuming (Min et al., 2023).

The primary objectives of DA are to:

- **Enhance generalization**: by introducing variability, models can better generalize to unseen data.
- **Reduce overfitting**: augmented data prevents models from memorizing training data, promoting better performance on test sets.
- **Simulate real-world variations**: reflects possible changes in real-world scenarios, such as lighting conditions, orientations, and occlusions.

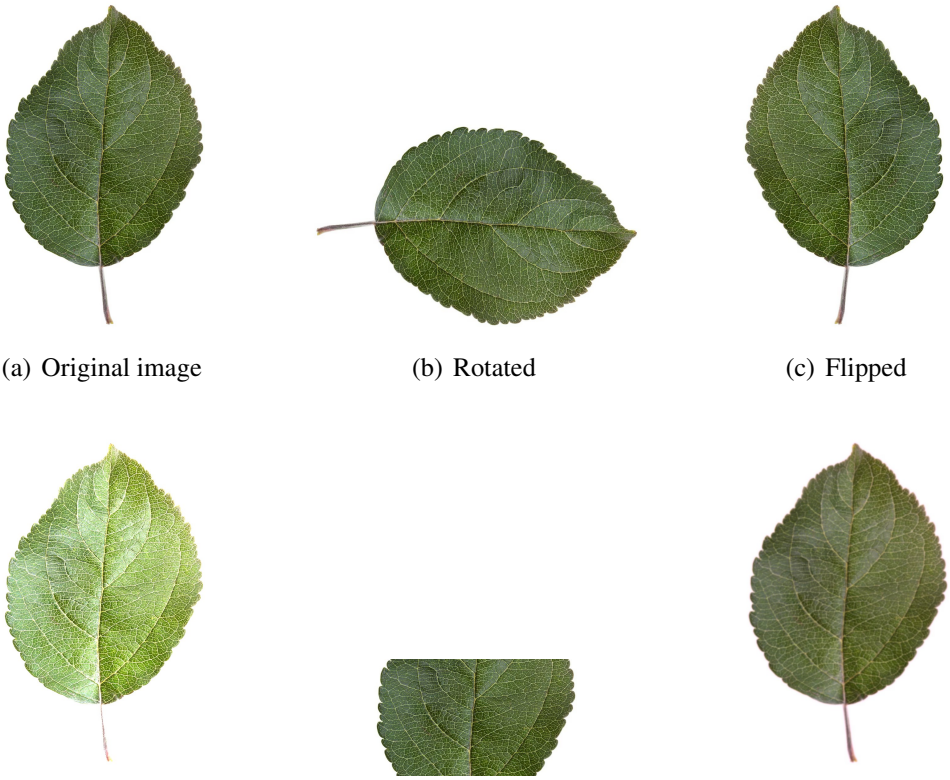
Data Augmentation techniques can be broadly categorized into several groups (Perez and Wang, 2017; Shorten and Khoshgoftaar, 2019):

1. *Geometric transformations*:
    - *Rotation*: rotating images by a certain angle to help models become invariant to orientation changes as illustrated in Figure 4(b).
    - *Flipping*: horizontally or vertically flipping images to introduce mirror images, as seen in Figure 4(c).
    - *Scaling*: zooming in or out to simulate different distances.
    - *Translation*: shifting images along the X or Y axis.
2. *Color space transformations*:
    - *Brightness adjustment*: modifying the brightness to simulate various lighting conditions, as illustrated in Figure 4(d).
    - *Contrast adjustment*: changing the contrast to highlight or diminish features.
    - *Saturation and hue alteration*: adjusting color intensity and tone.
3. *Cropping and padding*:
    - *Random cropping*: extracting random portions of images to focus on different areas, such as in Figure 4(e).
    - *Padding*: adding borders to images, which can help in preserving aspect ratios after transformations.
4. *Noise injection*:
    - *Gaussian noise*: adding random noise to simulate sensor imperfections.
    - *Salt and pepper noise*: introducing random black and white pixels to mimic data corruption, as shown in Figure 4(f).

Figure 4 demonstrates several augmentation techniques applied to plant leaf images. The original sample (a) is modified through rotation (b), horizontal flipping (c), brightness adjustment (d), cropping (e), and gaussian noise injection (f). These transformations simulate real-world variability, which improves the model's robustness during inference.

In plant pathology, DA plays a pivotal role. Studies have shown that integrating *MobileNetV2* with background-removed images improves classification performance for

apple leaf diseases (Ferdi, 2024). Such targeted augmentations are especially effective for conditions like European canker, where datasets are often limited and heterogeneous, underscoring the need for robust training methodologies.



(a) Original image (b) Rotated (c) Flipped

(d) Brightness adjusted (e) Cropped (f) Noisy

Figure 4. Examples of DA strategies applied to leaf images.

Furthermore, advanced strategies such as those using generative models like CycleGAN can synthesize realistic images of diseased leaves, effectively expanding the dataset and improving model generalization as outlined by Li and Zhang (2023).

A comprehensive DA pipeline is crucial for building accurate and resilient deep learning models, particularly in data-scarce domains. By enriching the dataset and simulating diverse input conditions, augmentation techniques substantially enhance the model's generalization and overall performance.

2.4. Ensembles

Ensemble learning is a machine learning framework that combines multiple models, *known as base learners*, to generate predictions that are more accurate, stable, and generalizable than those produced by any individual model. In plant disease detection, particularly in complex cases such as European canker, Ensembles are valuable tools for capturing diverse visual patterns across infection stages and host variability.

The primary objective of Ensemble methods is to reduce bias and variance present in standalone models, thereby enhancing robustness, precision, and sensitivity in classification tasks. This is especially relevant in agricultural scenarios, where images captured

under varying lighting conditions, viewing angles, or camera qualities can introduce noise and inconsistencies. Studies from the last few years (2022–2025) have demonstrated that Ensemble methods, particularly those employing lightweight CNN architectures, can significantly improve classification performance while maintaining low computational overhead, a critical factor for deployment on mobile and edge devices (Al-Gaashani et al., 2022; Ahmed et al., 2025).

Machine learning typically employs two main Ensemble approaches (Breiman, 1996):

Bagging (Bootstrap aggregating) involves training multiple versions of a model on different random subsets of the training data (with replacement), and aggregating their outputs through majority voting for classification or averaging for regression. This approach primarily reduces variance and is effective with high-variance models such as decision trees. In CNN-based applications, bagging can be approximated by training multiple instances on different data partitions and aggregating their predictions.

Boosting adopts a sequential strategy, where each new model is trained to correct the errors made by its predecessor. This method focuses on reducing bias by placing greater emphasis on misclassified instances in successive rounds. While classical methods such as *AdaBoost* and *Gradient Boosting* are widely used in traditional machine learning, deep learning adaptations often employ gradient-based meta-learners. For instance, the Ensemble proposed by Al-Gaashani et al. (2022) incorporates stacking with *XGBoost*[6] as a meta-learner, thereby inheriting characteristics of boosting.

Although bagging and boosting are rarely used in their classical forms, stacking has been widely adopted in studies from 2022 onward as an effective approach to enhance model performance in plant disease classification tasks. This method combines the outputs of multiple base learners (level-0) using a meta-learner (level-1), often achieving classification accuracies above 98%. Building on this paradigm numerous studies have successfully employed stacking by integrating lightweight Convolutional Neural Networks, such as *MobileNetV2*, *NASNetMobile*, and custom-designed CNNs, demonstrating substantial improvements over single-model baselines. In some cases, these models have been combined with meta-learners or weighted voting schemes to further increase accuracy in tasks such as cucumber or tomato leaf disease classification (Al-Gaashani et al., 2022).

Therefore, Ensemble learning emerges as a promising and effective strategy for the development of robust plant disease recognition systems. Given its demonstrated success in accurately classifying various leaf diseases, even under diverse conditions and using lightweight architectures, this approach has strong potential for application in the automated detection of European canker. Although no current studies directly apply Ensemble methods to this specific disease, the methodology's demonstrated capacity to generalize across complex classification tasks, as evidenced by the works presented in the following section.

[6] *XGBoost (Extreme Gradient Boosting)* is a machine learning algorithm that builds strong prediction models by combining several simpler models in sequence. It is known for being fast, accurate, and effective in classification tasks.

2.5. Related Work

The increasing use of Machine Learning (ML) and Deep Learning (DL) techniques to identify plant diseases has demonstrated promising results. Kumar et al. (2023) conducted a systematic review highlighting the prevalence of Convolutional Neural Networks (CNN) in classifying diseases in crops such as tomato, grapevine, and apple. Specifically, in the apple crop domain, studies such as Zhang et al. (2021) and Khan et al. (2024) demonstrated the ability to achieve high classification accuracy using CNN-based models.

Jafar et al. (2024) reviewed AI-driven diagnostic frameworks that use ML and DL models, including Support Vector Machines (SVM), CNN, and Deep Neural Networks (DNN), to detect diseases in crops like tomato, potato, cucumber, and chili. Their study presents a structured pipeline involving image acquisition, segmentation, feature extraction, and classification, and discusses how AI integration with IoT technologies such as drones and sensors supports real-time field diagnostics. Reported accuracies range from 85% to 97% depending on the crop and model used.

Similarly, Jackulin and Murugavalli (2022) explored a wide range of ML and DL applications for plant disease detection using computer vision. Their review illustrates the use of algorithms in traditional leaf classification tasks as well as in advanced forecasting scenarios involving environmental variables and climate data. CNNs are highlighted for their autonomous feature extraction capabilities. In contrast to single-model strategies, hybrid approaches, combining classifiers such as KNN, Naive Bayes, and Random Forest, achieved accuracies ranging from 88% to 94%, especially in noisy or uncertain data environments.

In addition to classification strategies, recent studies have explored the integration of AI with complementary technologies to enhance plant disease management. For instance, Gawande et al. (2023) discuss the role of optical sensors, autonomous robots, and AI algorithms in enhancing plant disease management, from early detection to preventive interventions. These systems enable real-time monitoring and data-driven decisions, reducing the need for manual inspection and improving crop health outcomes with model accuracies reported above 90% in controlled deployments.

Khan et al. (2023) offer an in-depth review of AI-based disease detection techniques, addressing both applications and limitations. While automation improves diagnostic efficiency, challenges remain regarding generalization across environmental conditions, dataset quality, and computational constraints in field configurations. The authors advocate for robust model validation, sensor fusion, and the development of lightweight frameworks deployable on edge devices, with typical model accuracy reaching up to 95%.

Focusing on comparative performance, Ahmed and Yadav (2023) evaluated the performance of ML and DL models on 17 plant diseases using the PlantVillage dataset. Their pipeline included preprocessing steps such as color space conversion and K-means clustering, followed by texture feature extraction and classification. The study found that CNNs achieved an accuracy of 98.2%, significantly outperforming SVMs (90.4%) and Random Forests (92.7%) in multiclass classification.

In the Brazilian context, Assis (2023) and Almeida (2016) applied CNNs to detect leaf disorders in apple trees, by adapting models to local agricultural conditions. Reported accuracies reached 97.3% and 94.3%, respectively. Li and Zhang (2023) introduced *Cy-*

cleGAN for visual enhancement, increasing classification accuracy from 91.8% (without enhancement) to 95.4%. Meanwhile, Bedi and Gole (2021) explored a CNN-autoencoder hybrid model[7] that achieved 98.38% accuracy, improving feature extraction compared to standard CNNs (baseline accuracy: 94.2%).

Assis (2023) proposed a CNN-based model trained on 7,767 images of apple leaves categorized into four classes: healthy, scab, black rot, and cedar rust. Their model achieved 97.3% accuracy. The architecture consisted of two convolutional layers with ReLU activation, followed by max-pooling and dropout layers to prevent overfitting. In comparison to other studies, their work emphasized preprocessing techniques, such as normalization and rotation, which enhanced training efficiency and model generalization.

A notable contribution in apple disease detection is presented by Zhang et al. (2021), which achieved over 96% accuracy using a custom CNN trained on four apple leaf disease classes. This study emphasized the role of color and shape features and the importance of balanced datasets. In contrast to Assis (2023), who used a simpler CNN structure, their model was designed to handle high variability in leaf textures and symptoms, leading to superior generalization.

Complementing this approach, Almeida (2016) addressed the classification of foliar disorders in apple trees using CNNs trained on images collected under real field conditions in southern Brazil. Their model achieved 94.3% accuracy and focused on practical deployment in heterogeneous environments, considering challenges such as background noise, lighting variability, and leaf occlusion. The findings demonstrate the feasibility of applying CNNs to support phytopathological monitoring in regional agricultural settings, reinforcing the model's adaptability to local contexts.

Khan et al. (2024) proposed an intelligent system for identifying apple leaf diseases using CNNs, achieving 97.8% accuracy. The study used a pre-trained CNN architecture enhanced via transfer learning, achieving robust classification results while maintaining computational efficiency. Notably, the authors integrated explainable AI (XAI) techniques to visualize feature maps and to interpret how the model distinguishes between disease categories. This approach enhances confidence in AI-driven decisions and supports agronomists in verifying the diagnostic process, particularly in large-scale orchard monitoring scenarios.

In the context of DA, studies from 2023 and 2024 explored strategies to improve model robustness and generalization in plant disease classification. Min et al. (2023) proposed an augmentation pipeline including geometric transformations (rotation, flipping, scaling), brightness adjustment, and noise injection, which significantly enhanced CNN performance on leaf datasets. Their experiments showed accuracy improvements from 91.3% (without DA) to 96.7% (with DA), demonstrating that such synthetic data variations reduce overfitting and enhance robustness against real-world variability in image capture conditions. The study also emphasized the necessity of augmentation in situations where disease classes are imbalanced or underrepresented.

Focusing specifically on apple leaf disease detection, Ferdi (2024) investigated a DA method based on background removal in conjunction with the lightweight *Mo-*

[7] Autoencoders are neural networks used for unsupervised feature learning and dimensionality reduction. In plant disease detection, they improve feature extraction before classification.

bileNetV2 model. By isolating the leaf from complex backgrounds and generating a variety of synthetic samples, the approach not only improved classification accuracy, from 93.2% with the base model to 97.6% with augmentation, but also reducing training time. The combination of DA and background segmentation allowed the model to focus on disease-related features without distraction from irrelevant image elements. This study highlights how augmentation strategies can significantly influence the success of mobile and embedded deep learning systems in agricultural diagnostics.

To address the limitations of single CNN models, Al-Gaashani et al. (2022) proposed a stacked Ensemble integrating *MobileNetV2*, *NasNetMobile*, and a custom CNN with an XGBoost meta-learner. The base models individually achieved accuracies between 94%–96%, while the Ensemble boosted performance to 99% accuracy on the PlantVillage dataset and remained efficient enough for deployment on low-power devices. By fine-tuning pre-trained models and applying strategies such as dropout and early stopping, the authors demonstrated an effective balance between performance and model complexity, enabling practical use in precision agriculture environments.

In a complementary approach, Bedi and Gole (2021) presented a hybrid model combining a Convolutional AutoEncoder (CAE) with a CNN for disease detection in peach plants. This model was specifically designed to reduce the number of training parameters without compromising accuracy. By leveraging the dimensionality reduction strengths of CAEs, the model achieved 98.38% accuracy using only 9,914 parameters, outperforming a conventional CNN baseline that reached 94.2%. The reduced computational load and training time make the proposed system particularly suitable for real-time agricultural applications in resource-constrained environments.

Pushing the complexity even further, Ahmed et al. (2025) introduced an Ensemble architecture tailored for cucumber leaf disease classification. Their system incorporated custom CNN models (CucuNet-CNN1 and CucuNet-CNN2), pre-trained deep learning models, and a biologically inspired Spiking Neural Network (SNN). By combining these components in a five-model Ensemble and optimization via grid search, the model achieved an accuracy of 98.73%, outperforming individual model accuracies that ranged from 92% to 96.5%. The SNN component further improved the performance to 98.91% while maintaining energy efficiency, indicating its potential for real-time agricultural diagnostics.

Table 1 summarizes the key studies discussed in Section 2.5, outlining their techniques, focus on apple crops, DA, Ensemble usage, and reported accuracy.

These studies collectively demonstrate the evolution of plant disease detection from traditional machine learning techniques to increasingly sophisticated deep learning and hybrid models. While Convolutional Neural Networks remain central to most high-performing approaches, emerging trends, such as Ensemble learning, lightweight architectures, and model interpretability, reflect a shift toward more adaptable and field-ready solutions. However, despite significant progress in classifying diseases in crops like tomato and apple, no work has specifically addressed the detection of European canker, reinforcing the novelty and relevance of the present research. By leveraging CNNs, DA, and Ensemble methods, this research aims to deliver a robust, automated diagnostic solution for a disease that remains underexplored in the current state-of-the-art.

Table 1: Comparison of related works based on technique, apple as a target crop, use of Data Augmentation, Ensemble methods, and accuracy.

Reference	**Technique**	**Apple as Target**	**DA**	**Ensembles**	**Accuracy**
Kumar et al. (2023)	CNN	Yes	No	No	99.0–99.9%
Jafar et al. (2024)	SVM, CNN, DNN	No	No	No	85–97%
Jackulin and Murugavalli (2022)	KNN, Naive Bayes, RF, CNN	No	No	Yes	88–94%
Gawande et al. (2023)	CNN	No	No	No	>90%
Khan et al. (2023)	CNN, XAI	No	No	No	95%
Ahmed and Yadav (2023)	CNN, SVM, RF	No	Yes	No	90.4–98.2%
Almeida (2016)	CNN	Yes	No	No	94.3%
Li and Zhang (2023)	CNN	Yes	Yes	No	91.8–95.4%
Bedi and Gole (2021)	CNN	No	No	Yes	98.38%
Assis (2023)	CNN	Apple	Yes	No	97.3%
Zhang et al. (2021)	CNN	Yes	No	No	>96%
Khan et al. (2024)	CNN (TL + XAI)	Yes	No	No	97.8%
Min et al. (2023)	CNN	No	Yes	No	91.3–96.7%
Ferdi (2024)	*MobileNetV2*	Yes	Yes	No	93.2–97.6%
Al-Gaashani et al. (2022)	*MobileNetV2, NasNetMobile*, CNN	No	Yes	Yes	99%
Ahmed et al. (2025)	Custom CNNs, SNN	No	No	Yes	98.91%

3. Materials and Methods

This study is classified as applied research with a quantitative approach and an exploratory objective. The technical procedures adopted were literature review and experimental research.

The research was carried out in three main stages. In the first stage, images of apple trees were collected from multiple sources. These sources include: Epagri, CIDASC (Integrated Agricultural Development Company of the State of Santa Catarina), SENAR (National Rural Learning Service), and the image database of the Cancontrol platform, which includes photos uploaded directly by apple growers.

In the second stage, a Convolutional Neural Network (CNN) was trained to classify whether the images showed symptoms of European canker. The model was implemented in Python using the PyTorch library. Statistical analysis and performance evaluation using the scikit-learn library.

In the third and final stage, DA techniques were applied to expand the dataset's diversity and size. These techniques included geometric transformations such as rotation, flipping, scaling, and brightness adjustments. Additionally, Ensemble learning methods were employed to combine the outputs of multiple models, enhancing robustness and reducing classification variance. The models were retrained with the augmented dataset and reevaluated using the same performance metrics.

3.1. Image Acquisition and Dataset Curation

A total of 1,652 images were collected from diverse and trusted institutional sources, including EPAGRI, CIDASC, SENAR, and the image database of the Cancontrol platform. All images were manually inspected to remove inadequate samples, such as those that were blurry, poorly lit, duplicated, or contained irrelevant content. Images displaying only fruits, rather than trunks or branches where the disease typically manifests, were also excluded, resulting in the elimination of 39 images. The final dataset consisted of 909 healthy (negative) and 704 diseased (positive) samples. To maintain class balance, the number of negative samples was not significantly increased through augmentation. The images were then preprocessed through resizing, pixel intensity normalization, and other enhancement techniques to ensure high-quality inputs for model training, validation, and testing, aiming to simulate real-world conditions and support robust model selection.

The training images are used by the model to calculate gradients and update the network's weights. Validation images are employed to evaluate the model by calculating the error and other metrics that help assess whether its performance is improving or deteriorating. The objective is to provide the model with data different from the data used during training, in order to obtain results that more accurately reflect real-world conditions. The outcomes obtained during validation serve as the basis for selecting the most appropriate model according to the specific characteristics of the problem. Once the model is selected, the test data are used to analyze its performance in a potential production environment.

3.2. Model Induction and Evaluation

We evaluated twenty-two different pre-trained models available in the *Torchvision framework 2.2.2*. This framework was chosen due to its greater flexibility in constructing and

maintaining architectures compared to alternative frameworks. Additionally, Torchvision benefits from a large and active community, which facilitates access to documentation and solutions for potential issues.

The models used include: *MobileNetV2*; *MobileNetV3* in both Large and Small versions; *ResNet* with 18, 50, and 101 layers; *VGG* with 11, 16, and 19 layers; *AlexNet*; *EfficientNet* from B0 to B7; and *DenseNet* with 121, 161, 169, and 201 layers. These pre-trained models generally achieve good results and are trained to classify a wide range of image types. These models are then fine-tuned for the present task.

As illustrated in Figure 5, the training pipeline comprises the complete process of training a machine learning model, with an emphasis on optimizing the F-Beta metric. The optimizer is initialized, and a tolerance value is defined as the stopping criterion, acting as a patience counter: if the model fails to improve for a number of consecutive epochs equal to this value, training is terminated early. The process begins by reading the essential hyperparameters, including the number of epochs, learning rate, weight decay, number of replications, and the models to be used, followed by loading the selected models and the dataset for training and validation.

The core of the training process takes place within a loop that iterates through the specified number of epochs. In each iteration, the models are trained and evaluated, and the F-score obtained is compared to the best result achieved thus far. If the current F-score surpasses the previous best, the model is saved as the new best model, and the tolerance counter is reset. If there is no improvement, the tolerance counter is decremented, indicating that the model may be approaching a plateau. Regardless of the outcome, epoch statistics are recorded for future analysis. Once the maximum number of epochs is reached or the tolerance is exhausted, the loop terminates, concluding the training process.

To assess and select the most effective models, several metrics are recorded during the training phase, including accuracy, precision, recall, F1-score, and F-beta.

Accuracy measures the ratio of correct predictions to the total number of samples. However, in many cases, a single metric is insufficient to adequately evaluate model performance. Datasets may be imbalanced, with an unequal number of samples across classes, or the problem may require greater emphasis on a specific class.

Precision calculates the proportion of true positive predictions relative to all instances classified as positive. Recall measures the proportion of true positives out of all actual positive cases, including those incorrectly classified as negative. Precision reflects how accurate the model is when it predicts a positive instance, whereas recall indicates how many actual positive cases were successfully identified by the model.

Precision and recall are often inversely related: depending on the problem, a model may exhibit high precision but low recall, or vice versa. A model may achieve perfect recall by classifying all inputs as positive, but this would drastically reduce precision.

The F-beta score is a weighted harmonic mean of precision and recall, defined in (1). The parameter β allows for adjustment of the emphasis placed on each metric: values of $\beta > 1$ give more weight to recall, while values of $\beta < 1$ favor precision. In the case of the standard F1 score, $\beta = 1$, that is, precision and recall are equally weighted.

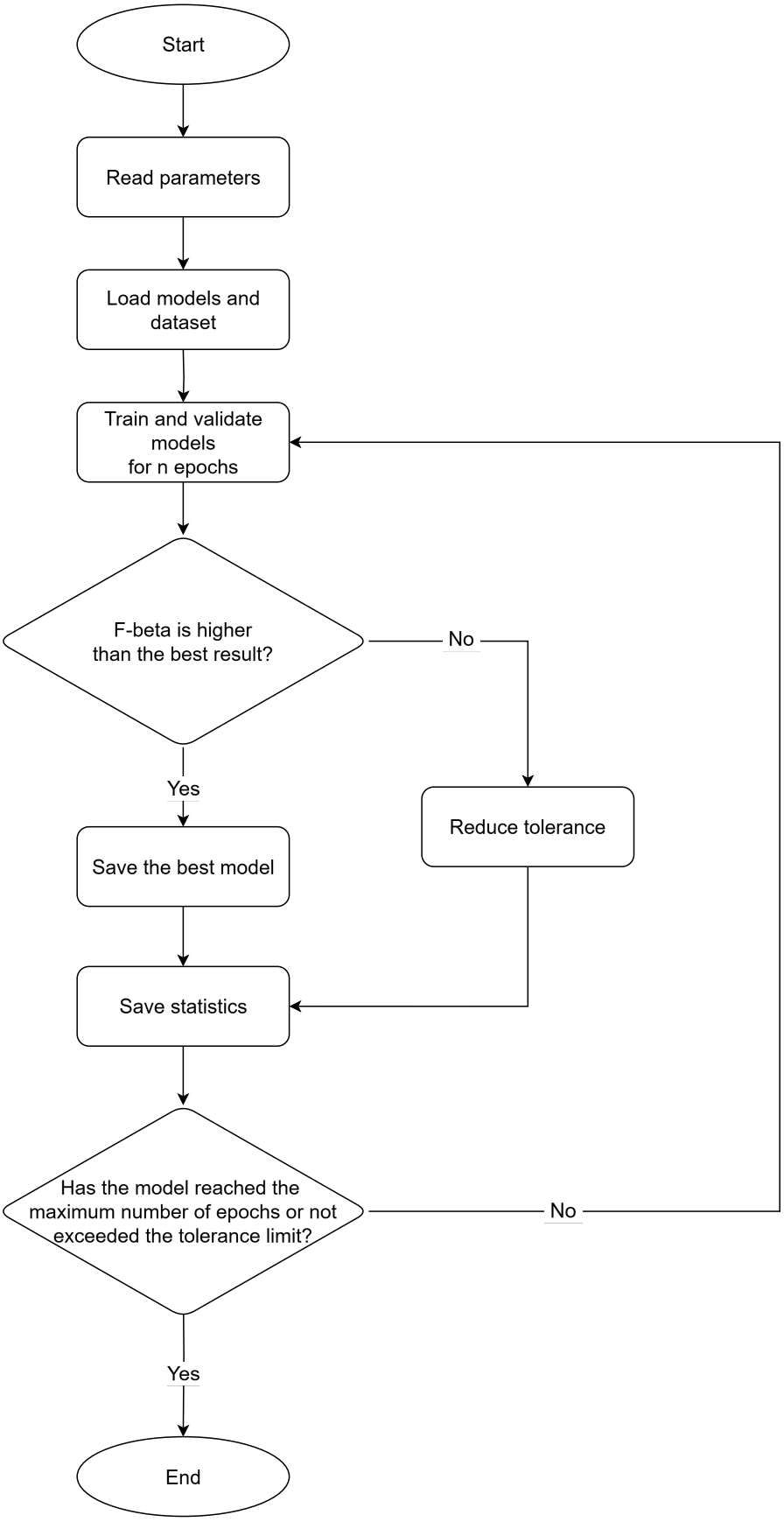


Figure 5. Flowchart of the machine learning model training and validation process.

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \tag{1}$$

The F-beta metric was chosen as the main evaluation criterion, with $\beta = 0.5$, since a value of $\beta < 1$ places more emphasis on precision than on recall. This emphasis is justified in the context of disease detection, where a false positive — the incorrect classification of a healthy tree as infected — may lead to unnecessary and costly actions, such as the removal and destruction of healthy crops. In contrast, false negatives are less critical, as these cases are subject to further examination by phytopathologists before any definitive measures are taken.

In addition to these metrics, the cross-entropy loss function was used as the optimization objective, measuring the divergence between predicted probabilities and true labels. The validation loss also served as the early stopping criterion; training was halted when it failed to decrease for a predefined number of consecutive epochs, preventing overfitting and reducing computation.

3.3. 1st Experiment: Baseline Model Selection

The objective of Experiment 1 is to identify the most suitable CNN model architecture for the task of detecting European canker in apple trees, using only the original dataset, without DA or Ensemble methods.

A total of 22 widely used convolutional models were evaluated, including: *MobileNetV2*, *MobileNetV3* (Large and Small), *ResNet* (18, 50, and 101 layers), *VGG* (11, 16, and 19 layers), *AlexNet*, *EfficientNet* (B0–B7), and *DenseNet* (121, 161, 169, and 201 layers). These models were selected due to their prevalence in the literature and their implementation in the *Torchvision 2.2.2*.

All models were initialized with default hyperparameters provided by *Torchvision*. This includes the Learning Rate (LR), Weight Decay (WD), and optimizer settings, which were not modified to ensure a fair and standardized baseline comparison across models. Each model was trained and evaluated across ten independent runs to account for performance variability due to the inherent randomness in model initialization and data shuffling.

Initially, each model was trained for up to 50 epochs. The cross-entropy loss function served as the primary control metric for early stopping and training stability. Specifically, the training was halted if no improvement in loss was observed over a number of epochs equal to the defined tolerance threshold. For models that consistently reached the maximum epoch limit without triggering early stopping in the majority of replications, the training was extended to 80 epochs and subsequently 100 epochs. No models were trained beyond 100 epochs, as the reduction in training loss became negligible.

3.4. 2nd Experiment: Evaluation of Data Augmentation Techniques

Experiment 2 was designed to evaluate the impact of DA techniques on the performance of the CNN architectures previously selected. This experiment aimed to investigate whether these additional techniques could enhance the accuracy, precision, recall, and F-beta score when compared to the baseline models established in Experiment 1. The primary goal was

to assess the contribution of these techniques to improve the robustness and generalization capacity of the models in the context of European canker detection.

A wide range of transformations was tested in this stage of the research. The DA techniques evaluated in this study were selected based on common practices and comprehensive surveys in the literature (Shorten and Khoshgoftaar, 2019; Perez and Wang, 2017):

- Random rotation;
- Random horizontal flip;
- Random vertical flip;
- Random resized crop;
- Color jitter (brightness, contrast, saturation, hue);
- Random grayscale;
- Gaussian blur;
- Random perspective;
- Random affine transformations;
- Random erasing.

Initially, all techniques were applied in a preliminary evaluation using three runs for each model. This pre-selection phase allowed us to identify transformations that consistently yielded improvements across the three runs. Only the techniques that demonstrated potential benefits were advanced to a complete experimental phase consisting of ten repetitions. This procedure was necessary given the high computational cost of training multiple configurations. Therefore, by filtering out methods that did not provide measurable improvements, the evaluation process was made more efficient without compromising result reliability.

Similar to Experiment 1, the models were trained with the default hyperparameter of the *Torchvision* framework. The number of epochs allocated to each model remained the same as in Experiment 1, ensuring a fair comparison between experiments. This consistency in training parameters allowed the effect of DA techniques to be isolated and properly analyzed in relation to the previously established baseline.

3.5. 3rd Experiment: Bagging Ensemble of Baseline Models

Experiment 3 aimed to evaluate the performance of an Ensemble of the baseline CNN models trained in Experiment 1, without applying any DA. The Ensemble method employed was bagging, which seeks to improve robustness and reduce variance by aggregating the predictions of multiple models. The models were not retrained. Instead, the pre-trained models from Experiment 1 were directly used. Each image in the validation set was passed through all selected models, and their prediction of each model was recorded as either correct, if it matched the true label or incorrect otherwise.

A total of 26.334 model combinations were evaluated, considering each Ensemble consisting of five models. For each Ensemble combination, the predicted label for a given image was determined using a majority-voting scheme: if at least three models predict the image label correctly, the Ensemble prediction was considered correct; otherwise, it was considered incorrect. This procedure was applied across all images in the validation set, and standard performance metrics such as accuracy, precision, recall, and F-beta were calculated for each Ensemble.

The results of this experiment enabled the identification of the most effective model combinations for European canker detection, providing insight into the possible performance gains achievable through bagging Ensembles of pre-trained CNNs.

3.6. 4th Experiment: Threshold Optimization

Experiment 4 was designed to analyze the impact of varying the classification threshold on the performance of the selected models from previous experiments. In standard binary classification using CNNs, the final class is assigned based on a predefined threshold, typically 0.50. However, depending on the application context, adjusting this threshold can improve specific performance metrics.

In the present study, threshold values were tested from 0.50 to 0.70, with increments of 0.02. Since the primary objective was to prioritize precision, values below 0.50 were not considered. A lower threshold would likely result in an increased recall and potentially higher accuracy and F-beta values. However, it would also increase the rate of false positives, which is undesirable in disease detection, as misclassifying healthy tree as infected may lead to unnecessary economic losses.

Conversely, the threshold was limited to a maximum of 0.70. Although thresholds above 0.70 could further improve precision, they would substantially reduce recall and negatively affect accuracy and F-beta. Thus, thresholds higher than 0.70 were excluded to avoid overly conservative models that fail to detect a significant portion of diseased samples.

For each selected threshold value, the predictions generated by the best-performing models from previous experiments were re-evaluated. Metrics including accuracy, precision, recall, and F-beta were computed to identify the threshold that best balances the trade-off between false positives and false negatives in the specific context of European canker detection.

3.7. Final Test

To conclude the experimental phase, a final test was conducted using images unseen by any model during training, validation, or threshold optimization. This dataset was used to evaluate the generalization ability of the best-performing models from the previous experiments.

In this experiment, the models that achieved the highest precision in each of the previous experiments were tested, based on the same metrics adopted throughout the study: accuracy, precision, recall, and F-beta. This comparison enabled to evaluate the strengths and limitations of each approach: baseline models, Data Augmentation, Ensemble learning, and threshold adjustment.

The same test images were also evaluated by two agronomists. The first was a researcher specialized in plant pathology with expertise in European canker; The second was an agronomist working in field extension services who provides technical assistance to apple growers and has practical experience with the disease, although he is not a specialist. This comparison aimed to compare the performance of the proposed models with the evaluations made by professionals with different levels of expertise.

4. Results

As detailed in Section 3, the results presented in this section were obtained through the training of 22 selected CNN architectures: *MobileNetV2*, *MobileNetV3* (Large and Small), *ResNet* (18, 50, and 101 layers), *VGG* (11, 16, and 19 layers), *AlexNet*, *EfficientNet* (B0–B7), and *DenseNet* (121, 161, 169, and 201 layers). The images used for training were divided into three distinct sets: 60% for training, 20% for validation, and 20% for testing. The results aim to identify the most suitable CNN architecture for European canker detection and determine their optimal combination for Ensemble methods.

4.1. 1st Experiment: Baseline Model

The analysis of variance (ANOVA) performed for accuracy, precision, and recall revealed extremely low p-values of the F-test in all cases:

- Accuracy: 2.62×10^{-74};
- Precision: 3.05×10^{-73};
- Recall: 6.46×10^{-73}.

All p-values are far below the conventional significance threshold $(p < 0.05)$, indicating that the observed differences in performance among the 22 models are statistically significant. This indicates that, for any of the three metrics considered, the variation in results is not attributable to random chance but rather to substantial differences among model architectures.

Tukey's Honest Significant Difference (HSD) test was applied to compare the mean precision values among all evaluated models. The "Tukey Group" column in Table 2 presents the groupings obtained exclusively based on the precision metric, without considering accuracy or recall for the formation of these groups. Models sharing the same letter do not present statistically significant differences in precision, whereas models assigned to different letters exhibit significant differences at a predefined confidence level.

The results show that *EfficientNetB2*, present exclusively in group A, achieved the highest mean precision (0.859540), significantly outperforming all other models. In contrast, group L includes two models — *EfficientNetB4* and *VGG16* — which recorded the lowest mean precision values (0.775388 and 0.767762, respectively).

Table 2: Average accuracy, precision, recall, and Tukey grouping for all evaluated models.

Model	Avg. Accuracy	Avg. Precision	Avg. Recall	Tukey Group
EfficientNetB2	0.861994	0.859540	0.861040	A
DenseNet121	0.843302	0.848175	0.839207	B
EfficientNetB0	0.841433	0.840302	0.836821	B
EfficientNetB1	0.837383	0.835223	0.833635	B C
DenseNet201	0.836137	0.834101	0.832530	B C
AlexNet	0.833022	0.833129	0.827907	B C
ResNet18	0.828349	0.831911	0.819071	B C D
DenseNet161	0.830841	0.831509	0.823465	B C D

Continued on next page

Table 2 – *Continued from previous page*

Model	Avg. Accuracy	Avg. Precision	Avg. Recall	Tukey Group
DenseNet169	0.830841	0.830063	0.825002	B C D
ResNet101	0.823676	0.822190	0.819538	C D E
EfficientNetB7	0.823053	0.821535	0.826509	C D E F
EfficientNetB3	0.819315	0.817526	0.814053	D E F G
ResNet50	0.816511	0.814280	0.816421	E F G H
MobileNetV2	0.814019	0.812641	0.808872	E F G H
EfficientNetB6	0.805919	0.807103	0.811801	E F G H I
MobileNetV3 Small	0.809657	0.806571	0.806379	F G H I
VGG11	0.799377	0.804553	0.786989	G H I
VGG19	0.797819	0.800987	0.787145	H I J
EfficientNetB5	0.793770	0.792257	0.796821	I J
MobileNetV3 Large	0.790343	0.788259	0.783427	J K
EfficientNetB4	0.777259	0.775388	0.769317	K L
VGG16	0.765109	0.767762	0.752395	L

4.2. 2nd Experiment: Data Augmentation

The second experiment investigated the effects of DA techniques on the performance of the CNN models. To evaluate the statistical significance of these effects, an analysis of variance (ANOVA) was performed on accuracy, precision, and recall. The ANOVA model considered three factors: the model, the DA technique, and their interaction (Model x DA). Table 3 summarizes the p-values obtained for each factor, presenting the results for accuracy, precision, and recall.

Table 3: ANOVA p-values for accuracy, precision and recall considering Model, DA Technique, and their interaction.

	Accuracy	Precision	Recall
Model	4.23×10^{-132}	2.82×10^{-114}	6.27×10^{-139}
DA Technique	7.04×10^{-6}	6.96×10^{-5}	2.74×10^{-8}
Model x DA	1.01×10^{-19}	1.35×10^{-19}	9.59×10^{-17}

Although the p-values for DA Technique were statistically significant, they were substantially lower for the interaction term (Model x DA). This indicates that the interaction between model and Data Augmentation influenced the apparent value of DA Technique (increasing the F-value and reducing the p-value). Consequently, the improvement cannot be attributed exclusively to the DA techniques. For this reason, Tukey's HSD test was applied to further investigate the influence of DA.

The average precision values obtained for each DA technique are presented in Table 4. These results were subjected to Tukey's HSD test, which revealed that none of the pairwise comparisons reached statistical significance. This outcome is represented by assigning the same letter (A) to all techniques in the Tukey Group column, indicating that the observed differences are not statistically meaningful when considered in isolation.

Table 4: Mean precision of DA techniques and Tukey grouping (all differences non-significant).

DA Technique	Avg. Precision	Tukey Group
All	0.842896	A
Rotation	0.841392	A
Vertical Flip	0.840950	A
Erasing	0.840441	A
Perspective	0.839828	A
Color Jitter	0.838859	A
Grayscale	0.837550	A
None	0.837498	A
Horizontal Flip	0.836031	A

To further explore the interaction between each model and each DA technique, precision values were aggregated by model and technique. Table 5 presents the 15 highest mean precision values from this aggregation. Three main insights emerge, which corroborate the results presented in the previous tables, can be drawn:

1. The influence of the model is considerably greater than that of the DA technique. For instance, *EfficientNetB2* consistently achieved the highest performance, independent of the applied technique.
2. No single DA technique stands out among most models. Instead, each model exhibits a distinct response, benefiting from a specific transformation. Therefore, no single DA technique can be considered universally superior.
3. All models shown improvement when the results obtained without any DA technique were compared to those achieved with their best-performing augmentation strategy. This finding reinforces the effectiveness of DA in enhancing model generalization and robustness, regardless of the architecture used.

Table 5: Top 15 results of Model–DA interaction by mean precision.

Model	DA Technique	Avg. Precision
EfficientNetB2	All	0.875571
EfficientNetB2	Erasing	0.874805
EfficientNetB2	Rotation	0.867818
EfficientNetB2	Horizontal Flip	0.866998
EfficientNetB2	Grayscale	0.861560
EfficientNetB2	Vertical Flip	0.860059
EfficientNetB2	None	0.859540
EfficientNetB2	Perspective	0.858202
EfficientNetB0	All	0.856580
EfficientNetB0	Color Jitter	0.856131
EfficientNetB2	Color Jitter	0.854548
DenseNet201	Rotation	0.850235
DenseNet121	All	0.849750
DenseNet201	Perspective	0.849136
EfficientNetB2	Vertical Flip	0.848455

Finally, Table 6 highlights the DA technique that yields the highest precision for each model. As shown in Table 6, the *EfficientNetB2* model once again stands out as the best performing architecture. However, the Tukey grouping indicates that *EfficientNetB0* may perform at a level comparable to *EfficientNetB2*, and in some cases, its performance

overlaps with that of several other models, suggesting no statistically significant difference among some of them.

Notably, for every model analyzed, the application of the best-performing DA technique resulted in a higher precision compared to the scenario without any augmentation. This consistent improvement across all architectures reinforces the positive impact of DA on model precision and generalization capability.

Table 6: Best DA technique per model based on mean precision, including precision without DA.

Model	Best DA Technique	Precision with DA	Precision without DA	Tukey Group
EfficientNetB2	All	0.875571	0.8595395	A
EfficientNetB0	All	0.856580	0.8403019	A B
DenseNet201	Rotation	0.850235	0.8341013	B C
DenseNet121	All	0.849750	0.8417055	B C
DenseNet169	Rotation	0.846639	0.8300633	B C
EfficientNetB1	Vertical Flip	0.840302	0.8352225	B C
DenseNet161	Color Jitter	0.841560	0.8315090	B C
AlexNet	All	0.840586	0.8331290	B C
ResNet18	Erasing	0.834840	0.8319106	C

4.3. 3rd Experiment: Ensembles

The third experiment focused on evaluating the performance of Ensemble configurations derived from the best-performing individual models identified in the previous experiments. Each Ensemble was constructed by combining the output probabilities of selected CNN architectures through averaging strategies, aiming to investigate whether Ensemble integration could further improve the detection accuracy of European canker.

Unlike the first and second experiments, it was not possible to perform an analysis of variance (ANOVA) due to the large number of Ensemble configurations and replications involved. The resulting dataset contained a high number of rows (Ensembles × replications), which made the computation of ANOVA unfeasible and would not yield reliable results due to the heterogeneous sample sizes among the Ensembles.

For this reason, the statistical analysis in this section relies exclusively on Tukey's Honest Significant Difference (HSD) test, which allows direct pairwise comparison of the mean precision values between Ensembles. This approach enables the identification of statistically significant differences among Ensemble configurations without requiring the homogeneity of variances or balanced sample sizes typical of ANOVA.

Table 7 presents the ten best-performing Ensemble configurations according to their mean precision values. The results indicate that there are no statistically significant differences among these top Ensembles based on Tukey's HSD test. Furthermore, it is noteworthy that more than one hundred additional Ensemble configurations demonstrated equivalent statistical performance to those listed in Table 7.

Table 7: Best Ensembles combination based on mean precision.

Combination	Precision	Tukey Group
AlexNet-EfficientNetB1-EfficientNetB2-ResNet101-ResNet18	0.852679	A

Continued on next page

Table 7 – *Continued from previous page*

Combination	Precision	Tukey Group
AlexNet-EfficientNetB2-EfficientNetB7-ResNet101-ResNet18	0.851859	A
AlexNet-DenseNet161-EfficientNetB1-EfficientNetB4-ResNet101	0.849112	A
AlexNet-DenseNet161-EfficientNetB7-ResNet101-ResNet18	0.848990	A
AlexNet-DenseNet121-DenseNet161-EfficientNetB1-ResNet18	0.848268	A
AlexNet-DenseNet161-EfficientNetB1-EfficientNetB2-ResNet101	0.847780	A
AlexNet-DenseNet121-EfficientNetB1-EfficientNetB2-ResNet18	0.847614	A
AlexNet-DenseNet161-EfficientNetB2-ResNet101-ResNet18	0.847510	A
AlexNet-EfficientNetB1-EfficientNetB2-EfficientNetB4-ResNet101	0.847281	A
AlexNet-DenseNet161-EfficientNetB1-ResNet101-ResNet18	0.847073	A

4.4. 4th Experiment: Threshold Optimization

This experiment was conducted using the best-performing Ensemble and individual CNNs identified in the previous experiments. The objective was to investigate how adjusting the classification threshold could influence performance metrics. Unlike earlier experiments, which used a fixed threshold of 0.50, this experiment explored threshold values ranging from 0.50 to 0.70, in increments of 0.02, to evaluate their impact on model performance.

The analysis of variance (ANOVA) indicated statistically significant effects for both the model and threshold factors, as well as for their interaction ($p < 0.05$). This result demonstrates that the classification threshold influences the models' performance, and that the magnitude of this effect differs among the evaluated CNNs.

To further investigate the significant effects detected by ANOVA, Tukey's Honest Significant Difference (HSD) test was applied to perform pairwise comparisons of mean precision across both threshold levels and models. This analysis aimed to determine whether specific threshold adjustments produced statistically significant improvements, providing a more detailed understanding of how precision varied with changes in the decision boundaries.

Table 8: Best thresholds optimizations based on mean precision.

Model	Threshold	Precision	Tukey Group
EfficientNetB0 - All	0.70	0.907143	A
EfficientNetB0 - All	0.68	0.900597	A
ResNet18 - random_erasing	0.70	0.900317	A
DenseNet161 - color_jitter	0.70	0.899991	A
Ensemble - none	0.70	0.896068	A
EfficientNetB0 - All	0.66	0.894842	A
EfficientNetB2 - All	0.70	0.894388	A
EfficientNetB2 - All	0.68	0.891689	A
ResNet18 - random_erasing	0.68	0.891616	A
DenseNet161 - color_jitter	0.68	0.889375	A

Table 8 presents the threshold values that achieved the highest mean precision. The Model column specifies each CNN architecture along with the Data Augmentation (DA) technique used during training. The entry labeled *Ensemble – none* represents the best ensemble configuration identified in Experiment 3, which combines the models *ResNet18*, *ResNet101*, *EfficientNetB1*, *EfficientNetB2*, and *AlexNet*.

The Tukey HSD results indicate that higher thresholds (around 0.68–0.70) tended to produce higher precision values across all evaluated models. However, no statistically significant differences were found among the top-performing configurations, all grouped under the same Tukey category (A). Consistent with theoretical expectations, increasing the threshold improved precision but reduced recall and, in some cases, also decreased accuracy.

4.5. 5th Experiment: Final Test

This final experiment evaluated the generalization capability of the best-performing configurations identified in the previous stages. Specifically, it compared the average precision obtained on the validation set (Sections 4.1 to 4.4) with that achieved on the test set, which contained images not previously seen by the models. The evaluated models were:

- 1st Experiment: *EfficientNetB2* without Data Augmentation (DA);
- 2nd Experiment: *EfficientNetB2* with all DA techniques applied together;
- 3rd Experiment: an ensemble composed of *AlexNet*, *EfficientNetB1*, *EfficientNetB2*, *ResNet101*, and *ResNet18*.
- 4th Experiment: *EfficientNetB0* with all DA techniques applied together and a threshold set to 0.70.

Furthermore, the evaluations from two human experts were included as a reference point for assessing model performance. Both experts evaluated only the test set. Expert 1 is a plant pathologist specialized in European canker, whereas Expert 2 is an agronomist working in field extension services.

In Table 9, the Evaluator column lists all CNN models and the human experts considered in this experiment. The model entries follow the pattern *CNN – DA technique*. The entry labeled *Ensemble – none* refers to the best ensemble configuration identified in Experiment 3, which combines *ResNet18*, *ResNet101*, *EfficientNetB1*, *EfficientNetB2*, and *AlexNet*. The column prefixes $V\_$ and $T\_$ denote the validation and test sets, respectively.

As expected, precision on the test set was slightly lower than on the validation set. The reduction was approximately 5% for *EfficientNetB2* (with or without Data Augmentation) and just over 1% for the ensemble, indicating good generalization across models. *EfficientNetB0*, which employed all Data Augmentation techniques and a threshold of 0.70, achieved the highest precision on the validation set and retained a relatively high value on the test set. However, this result came at the cost of a substantial decrease in recall, reflecting a stronger tendency toward conservative predictions. In contrast, the ensemble and *EfficientNetB2* with full Data Augmentation demonstrated a more balanced performance when both accuracy and F-beta were taken into account, indicating a more stable compromise between predictive precision and error tolerance.

When compared with human experts, all CNN models achieved results within a similar range. The best-performing models, particularly *EfficientNetB2* (with DA) and the ensemble, showed precision and recall values between those of Expert 1 and Expert 2. This finding indicates that, under the evaluated conditions, the models reached performance levels comparable to those of human specialists in identifying European canker.

Table 9: Comparative results between the evaluated models and human experts.

Evaluator	V_Acc	V_Prec	V_Recall	V_F-beta	T_Acc	T_Prec	T_Recall	T_F-beta
EfficientNetB2 - None	0.861994	0.85954	0.86104	0.860024	0.813707	0.801138	0.761429	0.792965
EfficientNetB2 - All	0.877881	0.875571	0.877312	0.876185	0.842679	0.839762	0.788571	0.82753
Ensemble - None	0.87757	0.852679	0.87	0.855965	0.841121	0.84117	0.783571	0.829231
EfficientNetB0 - All	0.822741	0.907143	0.661429	0.844030	0.76542	0.871632	0.542143	0.776759
Expert 1	-	-	-	-	0.872274	0.885496	0.816901	0.871071
Expert 2	-	-	-	-	0.838006	0.792207	0.859154	0.804861

5. Conclusion

This study demonstrates the potential of deep learning approaches to support the diagnosis of European canker in apple orchards. The experimental results indicate that pre-trained CNN architectures can identify disease with competitive levels of accuracy, precision, and recall. Furthermore, the performance differences observed among the twenty-two evaluated models, confirmed through statistical analysis, highlight the importance of the comprehensive comparative evaluation conducted in this work.

EfficientNetB2 emerged as the top-performing model, although its accuracy remains lower than that reported for CNN-based detection of leaf diseases in other crops. This difference is partly explained by the fact that European canker affects branches and stems, where symptoms are less distinct and more difficult to capture than in leaves. In addition, the limited availability of images for this disease and the use of real, unprocessed images for both training and testing make achieving higher accuracies more challenging. Nonetheless, this approach suggests that the results are likely more reproducible under real-world conditions than those derived from models trained on pre-processed images.

Data Augmentation (DA) techniques improved the performance of all models, even though no single technique consistently enhanced performance across all architectures. This suggests that applying DA is beneficial, but the most suitable technique must be selected for each specific model.

The ensemble model, composed of *AlexNet*, *EfficientNetB1*, *EfficientNetB2*, *ResNet101*, and *ResNet18*, demonstrated strong generalization across both validation and test sets. Although its performance on the validation set did not surpass that of *EfficientNetB2* with DA, the ensemble achieved higher precision on the test set than any individual models. These findings demonstrate the value of combining multiple architectures, as ensemble approaches can mitigate the limitations of single models and yield more robust results in practical applications.

Adjusting the decision threshold to 0.70 was useful for analyzing the balance between precision and recall. For *EfficientNetB0*, this adjustment increased precision but reduced recall, leading to more conservative predictions. Although this configuration did not outperform *EfficientNetB2* or the ensemble overall, this finding underscores the importance of threshold calibration for tailoring model behavior to the practical requirements of disease detection.

It is important to note that, in this study, both experts evaluated only photographic images, without access to the standard complementary diagnostic procedures normally

employed in practice. In routine diagnosis, in case of uncertainty, samples can be incubated to promote fungal growth and confirm the presence of *Neonectria ditissima*. Consequently, experts' errors do not reflect a lack of knowledge, but rather the limitations inherent to visual inspection alone, which reinforces the complexity of diagnosing European canker.

Overall, these findings highlight the progress made and the remaining challenges. Although the results are promising, further improvements are still necessary. Future work will involve integrating the trained models into the Cancontrol mobile application to enable real-world use. In this context, the emphasis on precision becomes essential: the system could automatically validate highly confident detections while forwarding uncertain cases to human experts for additional evaluation. Another promising direction involves developing a model specifically tailored for European canker detection. Certain architectures, such as *AlexNet*, achieved competitive results despite their simplicity, suggesting that a specialized model trained from scratch may be able to capture distinctive features of this disease more effectively than generic pre-trained networks.

References

Ahmed, I. e Yadav, P. K. (2023). A systematic analysis of machine learning and deep learning based approaches for identifying and diagnosing plant diseases. *Sustainable Operations and Computers*, 4:96–104.

Ahmed, S. et al. (2025). Cucunetcnns: application of novel ensemble deep neural networks for classification of cucumber leaf disease. *Ain Shams Engineering Journal*, 16.

Al-Gaashani, M. S. A. M., Shang, F., e El-Latif, A. A. A. (2022). Ensemble learning of lightweight deep convolutional neural networks for crop disease image detection. *World Scientific Journals*, 32(5).

Almeida, B. (2016). Classificação de distúrbios em folhas de macieiras utilizando redes neurais convolucionais. Dissertação de Mestrado – Universidade Federal de Pelotas (UFPel).

Assis, M. O. D. (2023). Classificação de doenças em folhas de macieira utilizando redes neurais convolucionais. Trabalho de Conclusão de Curso – Universidade Federal da Paraíba, UFPB.

Bedi, P. e Gole, P. (2021). Plant disease detection using hybrid model based on convolutional autoencoder and convolutional neural network. *Miscellaneous*, 5.

Branco Neto, W. C., Araujo, L., Pinto, F. A. M. F., Machado, R. A., Ribeiro, Y. F. B., Cordova Junior, W. F., e Mattos, K. M. (2021). Título do trabalho. In: *Anais do XIII Congresso Brasileiro de Agroinformática*, pages 44–52, Porto Alegre, RS, Brasil. SBC.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

EPAGRI (2024). Análise da evolução da produção de maçã em santa catarina. *Revista de Agricultura Catarinense*, 37(3). Accessed: 2025-11-04.

Ferdi, Y. (2024). Data augmentation through background removal for apple leaf disease classification using the MobileNetV2 model. *arXiv preprint arXiv:2412.01854*.

Gawande, V., Patil, A., Ingle, P., e Patil, R. (2023). From detection to protection: the role of optical sensors, robots, and artificial intelligence in modern plant disease management. *Artificial Intelligence in Agriculture*, 7:41–54.

Gelain, J. e De Mio, L. L. M. (2019). Podridão de *Neonectria ditissima* em frutos. In: Alves, S. A. M. e Czermainski, A. B. C., editors, *O cancro europeu no Brasil*, chapter 12, pages 169–180. Embrapa.

Harteveld, D. O. C., Goedhart, P., Houwers, I., Kohl, J., de Jong, P., e Wenneker, M. (2023). Detecting the asymptomatic colonization of apple branches by neonectria ditissima, causing european canker of apple. *SpringerNature Complete Journals*, 166(3):291–301.

He, K., Zhang, X., Ren, S., e Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., e Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G., Liu, Z., Van Der Maaten, L., e Weinberger, K. Q. (2017). Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708.

Hughes, D. e Salathé, M. (2015). Plantvillage dataset. Figshare. Includes apple (Malus x domestica) leaf images labelled "Healthy", "Scab", "Black Rot", "Cedar Rust" under controlled conditions.

Jackulin, C. e Murugavalli, S. (2022). A comprehensive review on detection of plant disease using machine learning and deep learning approaches. *Measurement: Sensors*, 24:100441.

Jafar, A., Bibi, N., Naqvi, R. A., Sadeghi-Niaraki, A., e Jeong, D. (2024). Revolutionizing agriculture with artificial intelligence: plant disease detection methods, applications, and their limitations. *Frontiers in Plant Science*, Volume 15 - 2024.

Khan, S. et al. (2024). Smart agriculture: an intelligent approach for apple leaf disease identification based on convolutional neural network. *International Journal of Intelligent Systems*, 172(4).

Khan, S., Rathore, M. M., Alhussein, M., Shah, M. A., e Kim, B.-K. (2023). Revolutionizing agriculture with artificial intelligence: plant disease detection methods, applications, and their limitations. *Computer Standards & Interfaces*, 88:103662.

Krizhevsky, A., Sutskever, I., e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25.

Kumar, P., Gupta, G., et al. (2023). A systematic analysis of machine learning and deep learning based approaches for identifying and diagnosing plant diseases. *Sustainable Operations and Computers*.

Lazzarotto, J. J. e Alves, S. A. M. (2015). Prejuízos econômicos e financeiros associados ao cancro europeu em sistemas de produção de maçã de vacaria, RS. Available at: http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1016470. Accessed on: January 24, 2025.

Li, H. e Zhang, W. (2023). Apple leaf disease identification via improved cyclegan and convolutional neural network. *Neural Computing and Applications*, 27(14).

Mattos, K. M. e Ribeiro, Y. F. B. (2022). Diagnóstico do cancro europeu da macieira com redes neurais convolucionais. Trabalho de Conclusão de Curso – Instituto Federal de Santa Catarina, IFSC Lages.

Min, B., Kim, T., Shin, D., e Shin, D. (2023). Data augmentation method for plant leaf disease recognition. *Applied Sciences*, 13(3):1465.

Mohanty, S. P., Hughes, D. P., e Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419.

Perez, L. e Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Rana, A. et al. (2022). A comprehensive review on detection of plant disease using machine learning and deep learning approaches. *Computers and Electronics in Agriculture*, 24.

Salvi, A. d. A. e Camargo Jr., P. C. (2023). Aplicação de data augmentation em redes convolucionais para detecção do cancro europeu das pomáceas. Trabalho de Conclusão de Curso – Instituto Federal de Santa Catarina, IFSC Lages.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., e Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520.

Shafay, M., Hassan, T., Owais, M., Hussain, I., Khawaja, S. G., Seneviratne, L., e Werghi, N. (2025). Recent advances in plant disease detection: challenges and opportunities. *Plant Methods*, 21:140.

Shorten, C. e Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Simonyan, K. e Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K. e Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*. arXiv:1409.1556.

Singh, A. et al. (2024). Revolutionizing agriculture with artificial intelligence: plant disease detection methods, applications, and their limitations. *Frontiers in Plant Science*, 15.

Tan, M. e Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6105–6114.

Thapa, R., Snavely, N., Belongie, S., e Khan, A. (2020). The plant pathology 2020 challenge dataset to classify foliar disease of apples. *Applications in Plant Sciences*, 8(9):e11390.

Xu, M., Kim, H., Yang, J., Fuentes, A., Meng, Y., Yoon, S., Kim, T., e Park, D. S. (2023). Embracing limited and imperfect training datasets: opportunities and challenges in plant disease recognition using deep learning. *Frontiers in Plant Science*, 14:1225409.

Yang, Q., Duan, S., e Wang, L. (2022). Efficient identification of apple leaf diseases in the wild using convolutional neural networks. *Agronomy*, 12(11):2784.

Zhang, L. et al. (2021). Disease detection in apple leaves using deep convolutional neural network. *Agriculture*, 11(7):617.