Evaluation of Convolutional Neural Networks for Coffee Leaf Rust Classification

Thiago Vieira Machado[1], Leonardo Gabriel Ferreira Rodrigues[1], Bruno Augusto Nassif Travençolo[1], Cícero Lima Costa[4], Danielli Araújo Lima[4], Larissa Ferreira Rodrigues Moreira[2]

[1]School of Computer Science – Federal University of Uberlândia (UFU)
Uberlândia – MG – Brazil

[2]Institute of Exacts and Technological Sciences
Federal University of Viçosa (UFV)
Rio Paranaíba – MG – Brazil

[3]Laboratory of Computational Intelligence, Robotics and Optimization
Federal Institute of Education, Science and Technology of Triângulo Mineiro
IFTM Campus Patrocínio – MG – Brazil

{thiago.machado, leonardo.g.rodrigues, travencolo}@ufu.br

larissa.f.rodrigues@ufv.br

{cicero, danielli}@iftm.edu.br

Abstract. *Coffee is a beverage present in the lives of many people worldwide and is of great importance to the economies of various countries. Coffee leaf rust is a serious disease that affects crops worldwide, and identifying it quickly and accurately helps in its control. This paper uses a publicly available image dataset to evaluate the performance of Convolutional Neural Networks (CNNs) in the context of the automatic classification of rust on coffee leaves considering binary and multi-class classification. Among the evaluated networks, ResNet achieved the best results, with an accuracy of 95.19% for binary classification and 78.03% for multi-class classification. This study contributes to the application of deep learning as a tool for farmers, enabling the early detection of rust on coffee leaves and aiding in decision-making related to crop management.*

1. Introduction

Coffee is a beverage present in the lives of many people around the world. It is estimated that more than three billion cups are consumed every day. But coffee is not only important in the daily lives of the population; its cultivation is also of great importance to the economy of more than 50 countries [Silva et al. 2022]. Coffee-producing countries export most of what is produced, generating around 20 billion dollars annually from exports alone. However, when considering the entire coffee sector, revenues exceed 220 billion dollars [Organization 2019].

Considering the high value of the coffee industry, diseases affecting the crop also cause significant losses. Among the diseases that can affect coffee plants is leaf rust, caused by the fungus *Hemileia Vastatrix*, which causes premature leaf drop, branch dieback, and a considerable reduction in plant production [Santana et al. 2018].

Rust is a very severe disease that can affect the production of an entire crop. Its diagnosis is made by a specialist who analyzes the leaves to identify the disease. However, in many parts of the world, coffee cultivation is done by small producers who often cannot afford the costs of a specialist [Yebasse et al. 2021]. Thus, technology can be employed to identify diseases in coffee crops, aiding both large producers and the subsistence of small ones.

Among the existing technologies, Deep Learning, a subcategory of machine learning, has been gaining a lot of attention. It is widely used for speech recognition, Computer Vision, and natural language processing through the use of neural networks [Guo et al. 2016].

Within Deep Learning, Convolutional Neural Networks (CNNs) are probably the most well-known and used model for solving image classification tasks [Ponti et al. 2017][Rodrigues Moreira et al. 2025]. CNNs are a valuable alternative to assist in the process of identifying coffee rust, as they allow computers to be trained to understand and analyze images, enabling automated and accurate diagnosis [Yebasse et al. 2021][Ongsulee 2017]. Thus, this paper aims to analyze images of coffee leaves using deep learning techniques to identify rust on coffee leaves. The main contributions are: *(i)* comparing different CNNs in terms of accuracy, precision, recall, and F1-score; *(ii)* evaluating the performance of CNNs using k-fold cross-validation for training and validation sets; *(iii)* conducting binary classification to determine whether a leaf is rust-infected or not; *(iv)* conducting multi-class classification to distinguish between healthy leaves and four different rust severity levels.

2. Related Work

The use of image processing for detecting diseases in coffee and other plants has been studied by various researchers. In this regard, [Marcos et al. 2019] proposed a Convolutional Neural Network (CNN) for image segmentation, which contains two convolutional layers, followed by the ReLu filter and a max-pooling layer. Normalization was also applied to each output of the max-pooling layers. A proprietary image dataset was used to evaluate the CNN, composed of 159 images of coffee leaves, which were manually classified by a specialist who marked all points containing rust. The Dice coefficient was used to measure the CNN's performance, with results of 0.79 as the mean and 0.82 as the median.

[Yebasse et al. 2021] used the RoCoLe dataset [Parraga-Alava et al. 2019], composed of 1560 images of coffee leaves, with 791 healthy and 769 diseased. The diseased leaves were divided and classified as follows: 167 with red spider mite, 344 with rust level 1, 166 with rust level 2, 62 with rust level 3, and 30 with rust level 4. The study evaluated three image visualization techniques and two different approaches for detecting rust on coffee leaves. The first was the naive approach, which used the ResNet with some modifications, such as backbone and Grad-CAM as a visualization technique. In this approach, the model achieved 99% accuracy in training but only 77% in testing. In the guided approach, an image segmentation technique was used before training, and in this case, they achieved 98% accuracy in testing.

[Dutta and Rana 2021] used the BRACOL dataset [Krohling 2019], composed of 1747 images containing healthy leaves and leaves affected by major coffee diseases. The

dataset is divided into two parts: one containing images of whole leaves and another with cropped images showing only one of the diseases. The CNN MobileNet V2 was used for training and transfer learning. The proposed model achieved an accuracy of 98.51%.

[Suparyanto et al. 2022] evaluated the CNN ResNet-18 for classifying rust on coffee leaves, trained using Stochastic Gradient Descent (SGD). The study evaluated 100 images of coffee leaves, obtained exclusively for the project, which were assessed by disease and pest specialists who divided the leaves into diseased and healthy. Due to the limited dataset used, the CNN had significant overfitting, and the accuracy achieved after training was only 59%.

[Pandian et al. 2022] proposed the creation of a 14-layer CNN for detecting plant diseases. For training, validation, and testing, images from five different public repositories were used, including Bracol [Krohling 2019] and RoCoLe [Parraga-Alava et al. 2019], resulting in 61,459 images of plant leaves. After training, real images of leaves from various plants were submitted, and the CNN was able to identify the plant species and the disease present. The model achieved an accuracy of 99.96%, precision of 99.79%, recall of 99.79%, and F1-score of 99.79%.

[Montalbo 2022] presented a model composed of three aggregated CNNs as a single unit: DenseNet-121, VGG-16, and EfficientNetB0. The dataset used contains 4,675 images of coffee leaves, divided into seven different categories depending on the level and type of disease. This dataset was constructed by combining three other public image datasets: Bracol [Krohling 2019], RoCoLe [Parraga-Alava et al. 2019]), and LiCoLe [Montalbo and Hernandez 2020]. The proposed model achieved an overall precision of 95.98%.

The present work differs from the previously mentioned studies by comparing different CNNs, exploring data augmentation techniques and k-fold cross-validation. Additionally, the RoCoLe dataset has been little explored for image classification tasks, and this work may provide research opportunities in this field, which can directly assist coffee growers suffering from rust in their crops.

3. Material and methods

The publicly available RoCoLe [Parraga-Alava et al. 2019] dataset [1] was used, containing 1,560 images of coffee leaves, with 791 healthy leaves and 769 diseased leaves. Figure 1 shows an example of a healthy leaf and another with rust, which are available in the RoCoLe image dataset.

(a) Healthy (b) Rust

Figure 1. Example images from the RoCoLe dataset.

[1] Available at: `https://doi.org/10.17632/c5yvn32dzg.2`

Data augmentation techniques were used on the dataset images to increase the dataset size and prevent overfitting. The images were resized to fit the input size required by the CNNs used in this study. In addition, the dataset was split using the k-fold cross-validation technique with $k = 5$. Figure 2 summarizes the steps of our proposed method.

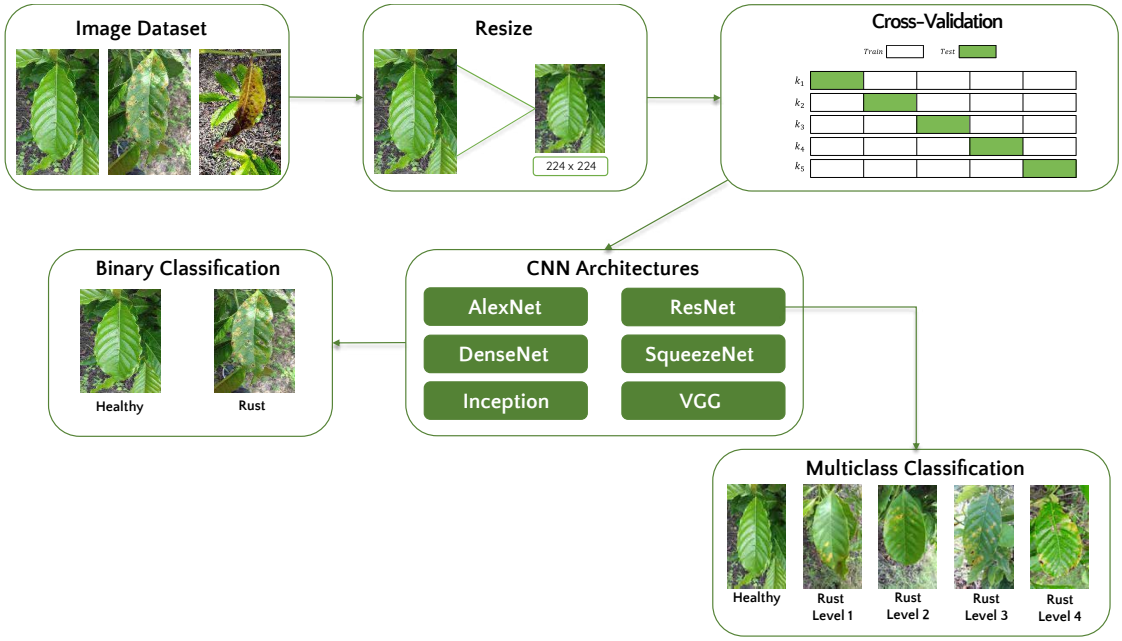


Figure 2. Steps of proposed method.

In this study we compared six CNN architectures: AlexNet [Krizhevsky et al. 2012], DenseNet [Huang et al. 2016], Inception [Szegedy et al. 2015], ResNet [He et al. 2016], SqueezeNet [Iandola et al. 2016], and VGG [Simonyan and Zisserman 2014].

AlexNet was proposed by [Krizhevsky et al. 2012] and was the first CNN to win the ILSVRC competition in 2012. AlexNet is composed of five convolutional layers, with three intercalated polling layers and three fully connected layers before the output layer. Even with existing modern CNN architectures, we chose to train AlexNet because of its historical importance and as a baseline for comparison with other architectures.

DenseNet was proposed by [Huang et al. 2016] and introduces dense connectivity, where each layer receives input from all previous layers and passes its feature maps to all subsequent layers. This architecture promotes feature reuse and mitigates the vanishing gradient problem, leading to improved parameter efficiency and model performance. DenseNet's innovative approach has made it a valuable addition to the deep learning community. We chose to train DenseNet to leverage its dense connections and assess its effectiveness compared to other state-of-the-art architectures.

Inception was proposed by [Szegedy et al. 2015] and introduced the concept of the Inception module, which allows for more efficient computation by capturing multi-scale features. Inception is known for its deep architecture and has achieved state-of-the-art performance on several benchmarks. Despite newer architectures, we chose to train Inception due to its innovative design and effectiveness in handling complex datasets.

ResNet, or Residual Network, introduced by [He et al. 2016] addresses the gradient vanishing problem in deep networks through residual connections. Each residual block contains a sequence of convolutional layers, and skip connections enable skipping some residual blocks and feeding the next block's input with the previous one's output. ResNet-50 is a specific variant that consists of fifty convolutional layers.

SqueezeNet, proposed by [Iandola et al. 2016], aims to achieve AlexNet-level accuracy with significantly fewer parameters. It introduces the Fire module, which uses squeeze and expand layers to reduce the model size without compromising performance. SqueezeNet's lightweight nature makes it suitable for deployment on devices with limited computational resources. We included SqueezeNet in our comparisons to evaluate its performance relative to larger models.

VGG was introduced by [Simonyan and Zisserman 2014] and is known for its simplicity and depth, consisting of 16 or 19 layers with small 3×3 convolutional filters. VGG has demonstrated excellent performance on various image recognition tasks and remains a popular choice for transfer learning. We chose to train VGG to benchmark its performance against other more recent and complex architectures.

We trained and evaluated the CNNs considering accuracy, precision, recall, and F1-score (as defined in Table 1). Two analyses were considered: (*i*) binary, where the images were classified into two classes (healthy and rust); and (*ii*) multiclass, where the images were classified into five classes (healthy and four different levels of rust severity).

Table 1. Evaluation Metrics

Metric	Definition	Formula
Accuracy	Hits of the classifier as a whole	$\frac{T_P+T_N}{T_P+T_N+F_P+F_N}$
Precision	Hit rate per class for positive cases	$\frac{T_P}{T_P+F_P}$
Recall	Rate of a relevant sample being correctly classified	$\frac{T_P}{T_P+F_N}$
F1-Score	Harmonic mean of Recall and Precision	$2 \times \frac{Recall \times Precision}{Recall+Precision}$

Where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative samples, respectively.

4. Results and discussion

All experiments were conducted on a computer with an AMD Ryzen 5 5600X processor, operating at a maximum frequency of 4.5 GHz, 32 GB of RAM, an NVIDIA RTX 3060 Ti GPU with 8 GB of memory, and Windows 10 operating system. All models were implemented using Python version 3.8, using PyTorch framework (version 2.0) [Paszke et al. 2019] with CUDA version 11.7 and cuDNN 8.0.

We evaluated the performance of six pre-trained CNNs selected for the image classification task to detect the presence of rust on coffee leaves. The CNNs were initially tested in binary classification to distinguish healthy from diseased leaves. Subsequently, the CNN with the best accuracy in binary classification was selected for a multiclass classification test, in which the leaves were divided into the following classes: healthy and rust levels ranging from one (mild symptoms) to four (severe symptoms). The results

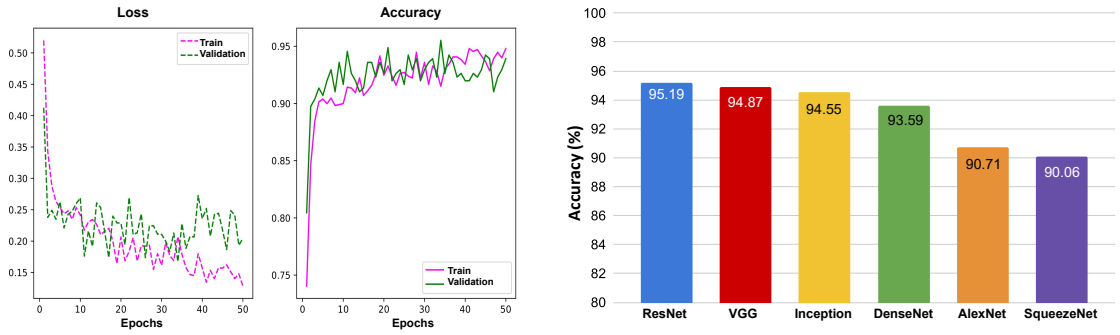
shows that CNNs are able to detect and differentiate rust on coffee leaves, contributing to the development of efficient disease detection techniques in coffee cultivation.

4.1. Binary Classification

For binary classification, images of coffee leaves were classified as either healthy or rusty. A total of 1,560 images present in the RoCoLe dataset [Parraga-Alava et al. 2019] were used, containing 791 images of healthy leaves and 769 images of rust. It should be noted that among the diseased leaves, there was a division by rust level, ranging from level one to level four, and a small number of images (only 167) of leaves attacked by red spider mites. Data augmentation was applied to the training images by random rotation (ranging from -30° to 30°) as well as vertical and horizontal flips.

Figure 3(a) presents, for example, the loss and accuracy graphs of ResNet generated during the training and validation of Fold 0. As can be seen, as the epochs progressed, the loss decreased, indicating that the model was improving. There is a rapid decrease at the beginning and stabilization at the end of the curve at a low level, suggesting that the model converges and achieves good performance. The accuracy increased over the epochs, indicating that the hit rate of the model increased and improved. At the end of the accuracy curve, it stabilizes at a high level, indicating that the model effectively learns the patterns in the data.

The average performance of the CNNs in terms of accuracy, which was calculated by dividing the number of correct predictions by the total number of coffee leaf images in the test set, is presented in Figure 3(b). It is interesting to note that all the CNNs achieved an average accuracy above 90%. Notably, ResNet achieved the best performance with an accuracy of 95.19%, followed by VGG with 94.87% and Inception with 94.55%.

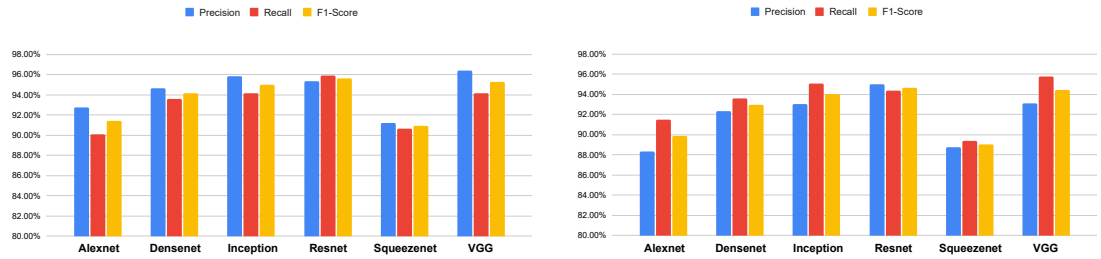


(a) Loss and Accuracy (b) Accuracy comparison

Figure 3. Experimental Results for Binary Classification.

The classification performance of the CNNs in terms of precision, recall, and F1-Score metrics is summarized in Figure 4, which illustrate the variations of each CNN. Among these metrics, recall is particularly noteworthy, because a high recall indicates that the model has a high capacity to correctly identify diseased leaves. In other words, a high recall indicates that the CNN has a lower false-negative rate, with a lower probability of failing to correctly identify a diseased leaf. For coffee growers, early identification of the affected leaves is crucial for control and prevention, allowing them to take appropriate measures more assertively. In this regard, all CNNs used in this study achieved a recall of above 90%. The three best performances were obtained from VGG, which achieved the

highest recall, correctly identifying 95.74% of the diseased samples, followed by Inception with 95.04%, and ResNet with 94.33%.



(a) Healthy leaves (b) Rust leaves

Figure 4. Performance variation of CNNs when evaluating healthy and rust leaves.

By analyzing Table 2, it can be noted from the confusion matrices from Fold 0 that the CNNs analyzed were able to accurately differentiate images of healthy leaves from diseased leaves. In particular, the DenseNet and VGG architectures were able to correctly classify a large number of healthy leaves, whereas for diseased leaves, DenseNet and ResNet stood out.

Table 2. Confusion matrices for each CNN model.

	Alexnet	
	Healthy	**Rust**
Healthy	151	20
Rust	11	130

	DenseNet	
	Healthy	**Rust**
Healthy	165	6
Rust	4	137

	Inception	
	Healthy	**Rust**
Healthy	161	10
Rust	9	132

	ResNet	
	Healthy	**Rust**
Healthy	159	12
Rust	7	134

	Squeezenet	
	Healthy	**Rust**
Healthy	160	11
Rust	11	130

	VGG	
	Healthy	**Rust**
Healthy	162	9
Rust	11	130

4.2. Multiclass Classification

For multiclass classification, ResNet achieved 78.03% accuracy, showing that the approach is promising for rust classification. However, it is evident that using the five classes is much more challenging for the CNN and needs to be further explored in future research.

The performance for the other metrics can be seen in Table 3, where it can be seen that for healthy leaves, the model achieved good results, but for rust level 2 and above, the performance dropped drastically.

This could be because of an imbalance in the dataset. Whereas healthy leaves represent 56.78% of the images, leaves with rust levels of three and four represent only 4.45% and 2.15%, respectively. Even when data augmentation was applied, it was still insufficient for the model to learn adequately.

We observed through the confusion matrix presented in Table 4 that ResNet correctly classified almost all healthy leaves and most leaves with rust level 1 – however, as the number of images per class decreases, the network's ability to correctly classify also decreases, causing it to misclassify the class to which each image belongs.

Table 3. Performance evaluation of the ResNet architecture in multiclass classification.

Class	Precision	Recall	F1-Score
Healthy	91.93%	94.74%	93.31%
Rust level 1	67.99%	69.81%	68.77%
Rust level 2	53.95%	45.62%	48.28%
Rust level 3	23.37%	20.69%	21.30%
Rust level 4	52.21%	35.52%	34.42%
Average	57.89%	53.28%	53.21%

Table 4. Confusion matrix for multiclass classification considering Fold 0 and ResNet architecture.

	Healthy	Rust level 1	Rust level 2	Rust level 3	Rust level 4
Healthy	150	8	0	0	0
Rust level 1	19	44	12	1	0
Rust level 2	0	8	11	2	1
Rust level 3	0	0	11	2	0
Rust level 4	0	1	2	1	6

5. Conclusion

Coffee is present in the lives of many families worldwide. Their importance in people's lives and daily routines is undeniable. Similarly, rust presents significant challenges and losses to coffee growers worldwide, whether they are large-scale producers or small family run farms. In this context, finding methods that can assist in rapid, accurate, and low-cost identification of the disease is crucial. In this regard, computer vision techniques are promising alternatives.

In this study, six CNNs were evaluated for classifying coffee leaf images with the goal of detecting the presence of diseased leaves in a dataset. Two different classification strategies were explored: binary (healthy or diseased leaves) and multi-class (healthy leaves or one of the four rust levels). Data augmentation techniques were applied to address imbalances in the dataset.

The architectures used in this study are AlexNet, DenseNet, Inception, ResNet, SqueezeNet, and VGG. Their performance was assessed based on accuracy, precision, recall, and F1-Score metrics. For the binary classification strategy, the best result was 95.19% accuracy achieved with ResNet CNN. Owing to its superior performance in binary classification, ResNet was selected for multiclass classification – however, this task proved quite challenging. Given the imbalanced dataset and the small number of images with severe rust levels, the CNN faced some difficulties, achieving an accuracy of 78.03%. The results suggest that the tested CNNs can assist farmers in identifying the disease, allowing them to control and prevent its spread to the rest of the coffee plantation.

Future work will aim to expand the dataset, explore new data augmentation techniques, test different CNN architectures, optimize hyperparameters, and develop a mobile application.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Bruno A. N. Travençolo is greatfull to CNPq for financial support (Grant #306436/2022-1). Larissa F. Rodrigues Moreira gratefully acknowledges the financial support of FAPEMIG (Grant #APQ00923-24).

References

Dutta, L. and Rana, A. K. (2021). Disease Detection Using Transfer Learning In Coffee Plants. In *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pages 1–4, Bangalore, India. IEEE.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48. Recent Developments on Deep Big Vision.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2016). Densely connected convolutional networks.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Krohling, R. A. (2019). BRACOL - A Brazilian Arabica Coffee Leaf images dataset to identification and quantification of coffee diseases and pests. Type: dataset.

Marcos, A. P., Silva Rodovalho, N. L., and Backes, A. R. (2019). Coffee Leaf Rust Detection Using Convolutional Neural Network. In *2019 XV Workshop de Visão Computacional (WVC)*, pages 38–42, São Bernardo do Campo, Brazil. IEEE.

Montalbo, F. J. and Hernandez, A. (2020). Classifying barako coffee leaf diseases using deep convolutional models. *International Journal of Advances in Intelligent Informatics*, 6(2):197–209.

Montalbo, F. J. P. (2022). Automated diagnosis of diverse coffee leaf images through a stage-wise aggregated triple deep convolutional neural network. *Machine Vision and Applications*, 33(1):19.

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. In *2017 15th international conference on ICT and knowledge engineering (ICT&KE)*, pages 1–6, Bangkok, Thailand. IEEE.

Organization, I. C. (2019). Coffee Development Report 2019 - Growing for Prosperity: Economic viability as the catalyst for a sustainable coffee sector. Technical report, International Coffee Organization, London.

Pandian, J. A., Kumar, V. D., Geman, O., Hnatiuc, M., Arif, M., and Kanchanadevi, K. (2022). Plant disease detection using deep convolutional neural network. *Applied Sciences*, 12(14).

Parraga-Alava, J., Cusme, K., Loor, A., and Santander, E. (2019). RoCoLe: A robusta coffee leaf images dataset for evaluation of machine learning based methods in plant diseases recognition. *Data in Brief*, 25:104414.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Ponti, M. A., Ribeiro, L. S. F., Nazare, T. S., Bui, T., and Collomosse, J. (2017). Everything You Wanted to Know about Deep Learning for Computer Vision but Were Afraid to Ask. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, pages 17–41, Niterói, Brazil. IEEE.

Rodrigues Moreira, L. F., Moreira, R., Travençolo, B. A. N., and Backes, A. R. (2025). Deep learning based image classification for embedded devices: A systematic review. *Neurocomputing*, 623:129402.

Santana, M. F., Zambolim, E. M., Caixeta, E. T., and Zambolim, L. (2018). Population genetic structure of the coffee pathogen hemileia vastatrix in minas gerais, brazil. *Tropical Plant Pathology*, 43(5):473–476.

Silva, M. d. C., Guerra-Guimarães, L., Diniz, I., Loureiro, A., Azinheira, H., Pereira, A. P., Tavares, S., Batista, D., and Várzea, V. (2022). An overview of the mechanisms involved in coffee-hemileia vastatrix interactions: Plant and pathogen perspectives. *Agronomy*, 12(2).

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

Suparyanto, T., Firmansyah, E., Wawan Cenggoro, T., Sudigyo, D., and Pardamean, B. (2022). Detecting Hemileia vastatrix using Vision AI as Supporting to Food Security for Smallholder Coffee Commodities. *IOP Conference Series: Earth and Environmental Science*, 998(1):012044.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, MA, USA. IEEE.

Yebasse, M., Shimelis, B., Warku, H., Ko, J., and Cheoi, K. J. (2021). Coffee disease visualization and classification. *Plants*, 10(6):1257.