FRANCISCO ERON CORDEIRO CARVALHO

Em direção a uma abordagem holística para o monitoramento de fazendas de café por meio de inteligência artificial

LAVRAS - MG

2023

FRANCISCO ERON CORDEIRO CARVALHO

Em direção a uma abordagem holística para o monitoramento de fazendas de café por meio de inteligência artificial

Monografia apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Ciências Biológicas, para a obtenção do título de Bacharel.

Prof.Dr. Antonio Chalfun Junior
Orientador

Dr. Raphael Ricon de Oliveira

Coorientador

Dr. Muhammad Noman

Coorientador

LAVRAS - MG

2023

FRANCISCO ERON CORDEIRO CARVALHO

Em direção a uma abordagem holística para o monitoramento de fazendas de café por meio de aplicativos de inteligência artificial

Towards a Holistic Approach to Coffee Farm Monitoring by Artificial Intelligence Applications

Monografia apresentada à Universidade Federal de Lavras, como parte das exigências do Curso de Ciências Biológicas, para a obtenção do título de Bacharel.

APROVADA em 14 de Julho de 2023.

Dr. Antonio Chalfun-Junior UFLA

Dr. Raphael Ricon de Oliveira UFLA

Dr. Pedro Luiz Lima Bertarini UFU

Dr. Cleverson Carlos Matiolli ITQB-UNL

Prof. Dr. Antonio Chalfun Junior
Orientador

Dr. Raphael Ricon de Oliveira

Coorientador

Dr. Muhammad Noman

Coorientador

LAVRAS - MG

2023

RESUMO

Neste trabalho, busca-se utilizar metodologias de inteligência artificial para um monitoramento holístico do cafeeiro, incluindo a quantificação dos frutos e seus estágios de maturação, a detecção de doenças foliares do café e a análise da florada. Estudos anteriores já haviam utilizado a detecção automática de frutos com base em visão computacional para estimar o rendimento do café durante a colheita, porém, há poucas pesquisas sobre a quantificação dos frutos de café na própria planta. Para preencher essa lacuna, o estudo utiliza pela primeira vez a versão mais recente do algoritmo de ponta chamado YOLOv8 (You Only Look Once). O YOLOv8 foi treinado em três conjuntos de dados diferentes: frutos de café, doenças foliares do café e detecção de árvores de café. Além disso, foram utilizados modelos de K-means para gerar classes de cores por meio de anotação semi supervisionada, possibilitando a identificação de diferentes estágios de maturação dos frutos de café com base na cor da epiderme. Essa abordagem inovadora de anotação de dados permite um processamento eficiente das imagens e tem potencial para revolucionar a avaliação da maturação dos frutos de café em termos de precisão e escalabilidade. Após o treinamento bem-sucedido dos modelos de detecção de objetos, a saída desses modelos é utilizada como entrada para um modelo de fundação denominado SAM (*Segment Anything Model*), que extrai informações detalhadas de segmentação de instâncias. Essa segmentação permite a implementação de metodologias de quantificação, possibilitando a obtenção de informações precisas sobre a quantidade e a distribuição dos frutos de café na planta. Além disso, a combinação das saídas dos modelos de detecção de objetos e do SAM é transformada em texto e alimentada em um *Large Language Model*. Esse modelo de linguagem inteligente atua como um assistente virtual, permitindo a interação e a tomada de decisões informadas no contexto da produção de café. Por meio dessa abordagem, é possível monitorar de forma eficiente os campos de café remotamente e tomar decisões baseadas em informações sobre irrigação, aplicação de fertilizantes e outras medidas de manejo oportuno do campo, promovendo assim a agricultura de precisão. Essa tecnologia baseada em inteligência artificial não se restringe apenas à produção de café, mas também pode ser adaptada para outras culturas de frutas. A integração dessa abordagem com outras ferramentas possibilita um monitoramento mais abrangente e eficiente das plantações, fornecendo informações valiosas para os produtores agrícolas e contribuindo para a otimização da produtividade e da qualidade dos cultivos.

Palavras-chaves: Café, agricultura de precisão, aprendizado de máquina, inteligência artificial, fenotipagem por imagens.

ABSTRACT

In this work, the aim is to use artificial intelligence methodologies for a holistic monitoring of coffee plants, including the quantification of fruits and their maturation stages, detection of coffee leaf diseases, and analysis of flowering. Previous studies have used computer vision-based automatic fruit detection to estimate coffee yield during harvesting, but there is limited research on quantifying coffee fruits directly on the plant. To fill this gap, the study employs the latest version of the state-of-the-art YOLOv8 (You Only Look Once) algorithm. YOLOv8 is trained on three different datasets: coffee fruits, coffee leaf diseases, and coffee tree detection. Additionally, K-means models are utilized to generate color classes through semi-supervised annotation, enabling the identification of different maturation stages of coffee fruits based on their epidermis color. This innovative data annotation approach allows for efficient image processing and has the potential to revolutionize the assessment of coffee fruit maturation in terms of accuracy and scalability. After successful training of the object detection models, their output is used as input for a foundational model called SAM (Segment Anything Model), which extracts detailed instance segmentation information. This segmentation facilitates the implementation of quantification methodologies, providing precise information on the quantity and distribution of coffee fruits on the plant. Furthermore, the combination of the object detection and SAM outputs is transformed into text and fed into a Large Language Model. This intelligent language model acts as a virtual assistant, enabling interaction and informed decision-making in the context of coffee production. Through this approach, it is possible to efficiently monitor coffee fields remotely and make data-driven decisions on irrigation, fertilizer application, and other timely field management measures, thus promoting precision agriculture. Importantly, this AI-based technology is not limited to coffee production and can be adapted for other fruit crops. Integration with other tools allows for comprehensive and efficient monitoring of plantations, providing valuable information for agricultural producers and contributing to optimizing productivity and crop quality.

Keywords: Coffee, precision agriculture, machine learning, artificial intelligence, image-based phenotyping.

SUMÁRIO

PRIMEIRA PARTE……1

INTRODUÇÃO GERAL……1

SEGUNDA PARTE - ARTIGO……...2

ARTIGO - TOWARDS A HOLISTIC APPROACH TO COFFEE FARM MONITORING BY ARTIFICIAL INTELLIGENCE……2

1.INTRODUCTION..……..3

1.1. CONTEXTUALIZATION AND BACKGROUND……...3

1.2. GENERAL ARCHITECTURE……...7

2. MATERIALS AND METHODS……9

2.1. DATA COLLECTION……9

2.2. DATA SPLITTING……10

2.3. DATA ANNOTATION……10

2.4. DEVELOPING A SEMI-SUPERVISED ANNOTATION SYSTEM FOR COFFEE FRUIT MATURATION STAGES……..11

2.5. TRAINING DETAILS……12

2.6. VALIDATION OF THE OBJECT DETECTION MODELS……..13

2.7. IMPLEMENTATION OF FOUNDATIONAL MODELS TO ENHANCE OBJECT DETECTION CAPABILITIES……14

3. RESULTS……15

3.1. COFFEE FRUITS……...15

3.2. LEAF DISEASES……29

3.3. COFFEE TREE……...33

DISCUSSION……..35

CONCLUSION……...38

Introdução Geral

O café é uma das *commodities* mais negociadas em todo o mundo e desempenha um papel significativo no desenvolvimento socioeconômico de muitos países tropicais. Além de contribuir para a geração de renda e emprego, a produção de café é importante para a redução da pobreza e para o alcance dos Objetivos de Desenvolvimento Sustentável. No entanto, a indústria cafeeira enfrenta desafios, como desequilíbrios na oferta e demanda e os impactos das mudanças climáticas.

Para enfrentar esses desafios, a utilização de tecnologias de inteligência artificial tem se mostrado promissora, permitindo a aquisição de dados de maneira não invasiva. No artigo científico apresentado neste documento foi empregada uma metodologia baseada em visão computacional e algoritmos de inteligência artificial para monitorar e avaliar de forma holística as plantações de café. O objetivo do trabalho baseia-se na quantificação de frutos e seus estágios de maturação, detecção e quantificação de doenças foliares e análise de florada. O artigo foi produzido seguindo as normas da revista Scientia Horticulturae para publicação.

A metodologia utilizada envolveu o treinamento de modelos de detecção de objetos, que foram alimentados com conjuntos de dados específicos de frutos de café, doenças foliares e detecção de árvores de café. Além disso,desenvolvemos uma nova metodologia para classificação automática de diferentes cores dos frutos de café, facilitando a identificação dos diferentes estágios de maturação dos frutos com base na cor da epiderme.

Após o treinamento dos modelos de detecção de objetos, a saída desses modelos foi combinada com um modelo de segmentação de instâncias, permitindo a obtenção de informações detalhadas sobre frutos, folhas e doenças, e florada, permitindo o desenvolvimento de métricas confiáveis. Essas informações foram transformadas em texto e alimentadas em um modelo de linguagem de grande porte, possibilitando a interação e a tomada de decisões informadas no contexto da produção de café.

Original research article *Scientia Horticulturae*

Towards a Holistic Approach to Coffee Farm Monitoring by Artificial Intelligence Applications

Francisco Eron[a#], Muhammad Noman[a#], Raphael Ricon de Oliveira[a#], Antonio Chalfun-Junior[a]*

Abstract

Earlier, researchers have employed computer vision-based automatic fruit detection to estimate coffee yield at the time of harvest, however, studies on the on-plant quantification of the coffee fruit are scarce. In this study, the latest version of the state-of-the-art algorithm YOLOv8 (You Only Look Once) was used for the first time. YOLOv8 was trained in 3 different datasets: coffee fruits, coffee leaf diseases and coffee tree detection. Meanwhile, the K-means models were trained which led to machine-generated color classes of coffee fruit for semi-supervised image annotation for different stages in coffee fruit maturation. After successfully training object detection models, we used the output from detection to prompt SAM, extracting instance segmentation, output from these 2 models combined were then transformed into text and fed to a Large Language Model. This AI-based technology when integrated with other tools such as UAV would efficiently remotely monitor coffee fields for informed decisions about irrigation, fertilizer application and other measures of timely field management, hence advancing precision agriculture. Moreover, this machine learning intelligent model can be tailored for various other fruit farming.

Keywords: Coffee, precision agriculture, machine learning, artificial intelligence, digital phenotyping

1. Introduction

1.1. Contextualization and Background

Coffee is a highly traded commodity globally ranking second only to oil in terms of traded value (FAO, 2023). The crop is a major contributor to the socio-economic development of tropical developing countries, with millions of people globally depending on it for their livelihoods. Aside from its contribution to agricultural GDP, coffee production is a significant employer and supports poverty alleviation (Chemura et al., 2016; Läderach et al., 2017). Thus, coffee cultivation is considered an avenue for realizing several of the Sustainable Development Goals (SDGs), such as generating income, creating rural employment, and poverty alleviation (FAO, 2023). Coffee cultivation takes place in over 60 countries, primarily in tropical regions that are conducive to its growth. Brazil, Vietnam, and Colombia are the leading producers worldwide, with Brazil alone accounting for 36% of global coffee production (USDA), while the U.S, Brazil and Europe are its top consumers. Additionally, coffee plantations, especially shaded farms, provide crucial ecosystem services such as biodiversity conservation (Jha et al., 2014), carbon sequestration (van Rikxoort et al., 2014), and soil protection (Meylan et al., 2017).

The coffee market is subject to recurrent supply-demand imbalances and uneven income distribution along the value chain. The global exports of coffee were recorded to be 10.88 million bags by December 2022 (ICO, 2023). Per data provided by the International Coffee Organization (ICO) in 2023, global coffee production was estimated to have reached 169.34 million bags, with each bag weighing 60 kg, signifying a decline of 2.2% compared to the previous year. In 2021, Brazil suffered a 21.7% drop in coffee production, which amounted to an estimated 67.2 million bags due to weather-associated factors such as drought and frost. The sustainability of coffee bean production and the impact of climate change are key sources of uncertainty for the coffee industry. Climatic conditions, especially during the vegetative and reproductive phases of the coffee plant, significantly influence coffee yield (Tavares et al., 2018). Rising temperatures and precipitation shortages affect flowering, fruiting, and bean

quality. Furthermore, climate variability is a key factor influencing the incidence of severe pests and diseases such as coffee leaf rust and coffee berry borer, which can decrease coffee yield and quality and increase production costs (Krishnan, 2017).

Globally, *Coffea arabica* and *Coffea canephora*, commonly referred to as Arabica and Robusta coffees respectively, constitute approximately 99% of the coffee production (Jayakumar et al., 2017). The quality of beans and yield of both species declines when outside these optimal temperature ranges (18-22°C for Arabica, while 22-28°C for Robusta), suggesting significant sensitivity to climatic changes (Magrach & Ghazoul, 2015). Therefore, from a socio-economic standpoint, it is crucial to comprehend the degree of climate-driven impacts on coffee production and the advantages of potential adaptation strategies to maintain and enhance coffee productivity and profitability while sustaining the livelihoods of smallholder producers globally. To protect coffee farms from adverse climatic conditions, keep a sustainable production and even enhance coffee yield and productivity, coffee farms demand continuous monitoring of every aspect.

Nonetheless, the phenomenon of asynchronous flowering poses a significant challenge for coffee growers, leading to irregular fruit ripening (López et al., 2021). Consequently, this causes problems during the harvesting process, as careful attention must be paid to ensure optimal timing. Oftentimes, for quality coffee production, coffee farmers resort to the practice of lapsed harvesting, wherein they must wait for the next batch of cherries to ripen before harvesting. This approach is not only time-consuming but also labor-intensive, requiring frequent visits and manual screening of the fruits. It is important to note that the quality of coffee is largely dependent on the ripeness of the fruits (Thompson et al., 2012). Coffee fruits, commonly referred to as red cherries, undergo a color transformation during the ripening process (Haile & Kang, 2019). The term "red cherry" is used to describe the fruit's epidermis when it achieves a uniform and intense red color at full ripeness, having progressed through various shades of green, orange, and pink. Overripe cherries turn dark violet, while the presence of green, overripe, or dry cherries in the harvested mass negatively impacts the quality of the beverage

and subsequently, its value in the international market (Velásquez et al., 2019). In particular, the proportion of green cherries in the harvested mass can significantly affect the beverage's acidity. To maintain high-quality standards and command a premium price, it is crucial to ensure that at least 98% of the harvested cherries are fully ripe (Leroy et al., 2006).

In the current landscape, the adoption of new technologies and innovation is imperative for the beverage industry to increase productivity and competitiveness. To this end, the scientific community is making significant efforts to develop automatic systems that can enhance the inspection process. Numerous studies have already been conducted, resulting in the development of various applications that have improved for example sorting processes for different fruits and vegetables (Hameed et al., 2018). Technological advancements in precision agriculture play a vital role in obtaining accurate and reliable measurements for crop monitoring. Precision agricultural practices, aimed at achieving high levels of productivity while promoting sustainability, can maximize the potential of each region, resulting in increased crop productivity and quality and reduced cost. Remote sensing has emerged as a promising technology for coffee management, with studies demonstrating its efficacy in evaluating coffee leaf rust levels through the use of Sentinel 2 sensor and Random Forest (RF) algorithms combined with vegetation indices, as described earlier (Chemura et al., 2017).

Computer vision has enabled the implementation of non-destructive techniques for detecting and identifying vegetative structures in the field using images. These techniques have been successfully applied to a wide range of crops including corn (Guerrero et al., 2013) tomatoes (Verma et al., 2014), and oranges (Patel et al., 2011). Besides, these techniques have also been implemented with grapes (Dey et al., 2012), pineapples (Moonrinta et al., 2010), and vegetable crops (Jay et al., 2015). Efficient decision-making on the appropriate harvesting period for coffee fruits can be facilitated by tracking their maturation stages through digital phenotyping. Ramos (2018) suggests that this can be achieved by determining the percentage of mature fruits on tree branches (Ramos et al., 2018). While previous studies have relied on destructive sampling, mainly post-harvest, to quantify and classify fruit for yield estimation (Carrillo &

Penaloza, 2009; de Oliveira et al., 2016), only a limited number of studies have explored the classification of coffee fruits before harvest, which can significantly benefit coffee farmers decision-making. Earlier, Avendano et al. (2017)) developed a system that constructs a 3D representation of coffee branches and classifies their vegetative structures (Avendano et al., 2017). In this pursuit, another group came with a brilliant idea of developing a CV-based non-destructive method of fruit counting and classification similar to ours, with few shortcomings (Ramos et al., 2017).

Few advancements were recently seen in this field. For example, a study developed a vegetation index (VI) for coffee ripeness based on the imaging data obtained from coffee fields through an RGB and a five-band multi-spectral cameras, each fixed on a separate UAV (Nogueira Martins et al., 2021). Similarly, Rodriguez et al. earlier came with a classic computer vision approach, however, it involved many instruments for image acquisition, a complex image processing system with precision values (Rodríguez et al., 2020). Although this technique requires the extraction of various features and their input into a classification algorithm, recent advancements in computer vision systems based on deep learning allow for the automatic extraction of multiple features. These techniques have gained popularity due to their speed and accuracy. Some recent studies devised yield mapping techniques during harvest based on imaging from a camera mounted over the harvesting machine, using YOLOv4 (Bazame et al., 2021, 2022; Martello et al., 2022).

The current study aimed at implementing the state-of-the-art CNN-based computer vision algorithms (Wang et al., 2022) to detect and classify coffee fruits on tree branches at different maturation stages, detect and classify different occurrences of leaf diseases and detect coffee trees. In order to further enhance the capabilities of our implemented model, we leverage the output from trained object detection models to automatically prompt state-of-the-art attention-based foundational models. This approach allows us to achieve instance segmentation and facilitates interaction with a Large Language Model through conversational context, without the need for additional training.

The object detection algorithm was initially trained on separate datasets for coffee fruits, coffee leaf diseases, and tree detection. The training data consisted of 80% of each dataset, while the remaining 20% was used for evaluation. Additionally, we aim to introduce a novel semi-supervised method for annotating coffee fruit maturation stages, which offers time-saving benefits and the ability to handle large datasets with varying numbers of fruit maturation categories based on the color of the epidermis. This innovative approach to annotation allows for efficient processing of data and holds the potential to revolutionize coffee fruit maturation assessment in terms of both accuracy and scalability.

By integrating this AI-based technology with other tools such as UAVs, coffee fields can be efficiently monitored remotely. This integration enables informed decisions regarding irrigation, fertilizer application, and other timely field management measures. The implementation of precision agriculture in sustainable quality coffee production chains allows for better decision-making, and our Large Language Model plays a crucial role in facilitating these informed decisions through conversation. Furthermore, the adaptable nature of this machine learning model makes it suitable for various other fruit farming applications.

1.2. *General Architecture*

Different object detection models were trained in down-stream tasks to locate and classify different stages of coffee fruit maturation, coffee leaf diseases and to locate coffee trees. Trained object detection model was used to prompt SAM (Kirillov et al., 2023) (Segment Anything Model, meta), a novel foundation (Bommasani et al., 2021) computer vision model for segmentation, where foundational models are trained in a broad range of images and are adaptable to generalize to a wide range of unseen data without further training through zero-shot-generalization (Oh et al., 2017) .The data extracted from the combination of these 2 computer vision model was then transformed into text and fed as a prompt to a Large Language Model API. The overall architecture (Figure 1.) is as follows: Input image enters into the encoder portion, where a d-dimensional embedding is extracted from the input image through

the encoder, ViT (Virtual Transformer architecture) (Bommasaniet al., 2021) in this case, the embedding, representation of features in the image to a lower-dimension space, is then fed to the decoder portion of the SAM architecture.

In the decoder portion of the architecture, the prompted object localization (in our case, detected objects from YOLO) (**1B**) is summed with the embedding extracted from the image, and upscale to the output mask (extracted pixels from detected object). Through acquiring masks, we proceed into generating background removal, flower density estimation and disease leaf severity.

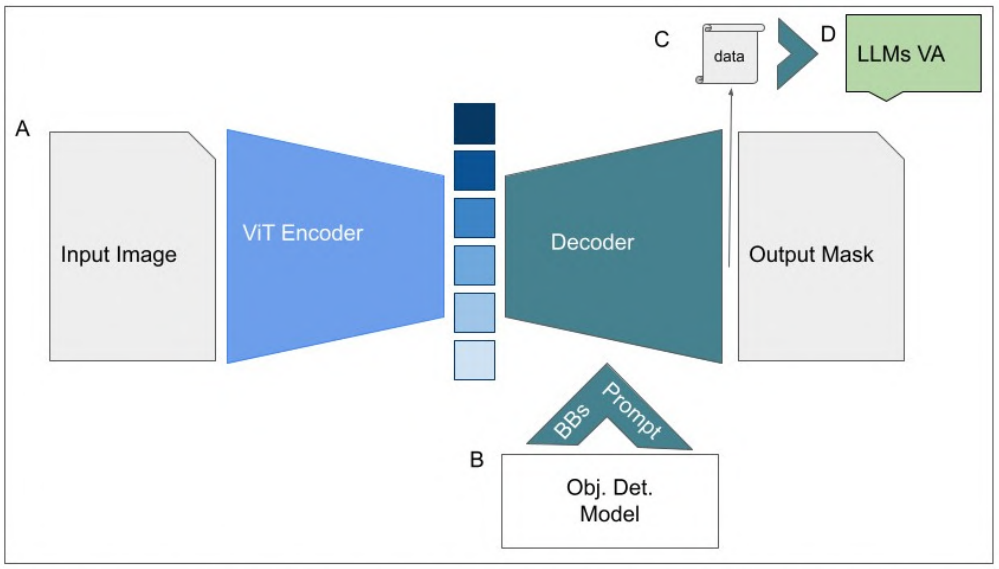


Figure 1. Architecture of SAM+YOLO+LLM model. Input image is fed onto the SAM and YOLO, the output from YOLO is used as a prompt (bounding-boxes, object location) to SAM decoder, being capable of extracting the exact pixels that comprises the detected object.

We used automatic prompt-engineering to propose bounding-boxes to SAM through the use of fine-tuned object detection models in fruits, tree, leaf and diseases in coffee. From the bounding boxes prompted to the decoder portion, associated with the embedding state extracted from input image, SAM is capable of creating an output mask (instance segmentation), enhancing the capacities in precision agriculture.

The data extracted from the combination of these 2 computer vision models (SAM and YOLO) is then transformed into text (**1C)** and fed to a Large Language Model API (**1D)**, to act as a virtual assistant. Through textual prompt engineering (LIU, 2023) and contextualization (providing information to the model, such as contacts, disease scientific description, and so on), we can extract valuable information without further training of the LLM (being trained in a large array of text and languages from the internet, such as GPT or BERT). The curation of the responses are done through selecting the positive-feedback answers and saving them into a d-dimensional embedding space, the correct answers thus can be used as contextualization to similar questions in the future, by searching and extracting similar (close) questions and answers in the embedding space, allowing the model to answer the individual question using good and curated information contextualized as reference.

2. Materials and Methods

2.1. Data Collection

In the creation of the coffee fruit and tree detection datasets, we initially gathered images from diverse coffee farms located in Lavras and surrounding areas of Minas Gerais, the highest coffee-producing region in Brazil. To ensure broad applicability, the collected images encompassed coffee fruits (branches) and flowering trees. To ensure comprehensive data representation, photographs were captured at various stages, spanning from unripe green fruits to ripened cherries and raisins, covering the entire maturation process. For optimal data diversity, fruit-bearing branches and flowering coffee trees were photographed using multiple smartphone cameras from different angles, ensuring the inclusion of representative data.

For the creation of the leaf disease dataset, we utilized curated and published datasets, as diseases necessitate a higher level of expertise and accurately annotated data. This approach ensured the inclusion of professionally validated information and enabled us to capture the diverse range of leaf diseases prevalent in coffee plants.

2.2. Data Splitting

All datasets were split into training and validation sets, with a training split of 80% and a validation split of 20%. This random division of the dataset ensures that the models are trained on a diverse set of images and can generalize well to new data.

2.3. *Data Annotation*

The images were manually annotated using Label Studio (Label Studio), an open-source platform for creating labeled datasets, to accurately identify the coffee cherries on the tree canopy, coffee trees and coffee leaf diseases. In the process of annotation of the coffee fruit dataset, the scale presented previously (Ságio, 2009), was used as reference.

To facilitate model training, all images were resized to 640 x 640 pixels. To further improve the model ability to generalize, default data augmentation techniques specific to the implemented models were used. In particular, we used mosaic augmentation, as described earlier (Bochkovskiy et al., 2020), to randomly combine multiple images into a single training sample.

YOLO is a part of a family of one-stage object detectors and is popular for its speed and accuracy (Wu et al., 2020). Here we used YOLO as the object detection portion of our general architecture (Figure 1), however, we first evaluated and compared the efficiency of YOLOv5 (Jocher et al., 2022), YOLOv5m6 (Li et al., 2023), YOLOv6 (Li et al., 2022, YOLOv7 (Wang et al., 2022) and YOLOv8 (Jocher G et al., 2023), which are the latest and have not been employed before for this purpose. An ideal state-of-the-art model should have (1) a faster and stronger network architecture; (2) a more effective feature integration method; (3) a more accurate detection method; (4) a more robust loss function; (5) a more efficient label assignment method; and (6) a more efficient training method. As compared to YOLOv4, YOLOv7 has been proved to be more efficient even with 75% less parameters and 36% less computation (Wang et al., 2022). And the later version of YOLO, YOLOv8, was demonstrated to be more efficient than YOLOv7 (Jocher G et al., 2023).

A diverse set of images in the training data helped the models learn to better handle

occlusions and other challenging conditions. Notably, the collected images contained a certain level of noise, reflecting the reality of field data collection and further challenged the models ability to generalize learning. Equations 1 and 2 were used in comparing the different object models efficiency at each dataset.

$$IoU = \frac{A \cap B}{A \cup B} \quad (1)$$

whereas; IoU - Intersection over Union, A - Ground Truth Boxes, B - Predicted Boxes,

$$APi = \int_0^1 P(R)dR \quad (2)$$

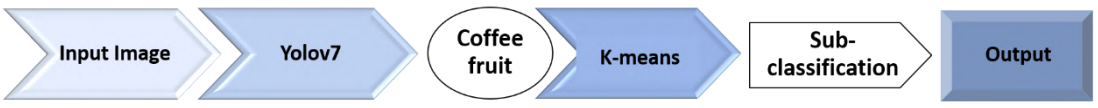
whereas; AP - Average Precision, P – Precision, R – Recall

“Precision” measures the number of correct positive predictions.

“Recall” measures the number of correctly identified positive class samples present in the dataset.

2.4. Developing a semi-supervised annotation system for coffee fruit maturation stages

Using a trained object detection model, we performed automatic fruit localization in the dataset. The located fruits were then cropped and resized to dimensions of 28x28x3. To account for potential lighting variations, we converted the resized RGB (Red, Green, Blue) images into the LAB color space. We focused on the A and B color channels for further analysis, while excluding the L channel to minimize bias from shadow and light variations. The AB color space images were represented as vectors in a multidimensional space. To create distinct color classes for the fruits, we trained K-means models with various k-sizes (ranging from 2 to 7). Approximately 36,000 fruits were randomly selected from the dataset for this purpose. The outlined strategy is depicted below.



To enable semi-supervised learning, we curated annotations consisting of detected bounding

boxes from object detection model (object location), YOLOv, but unsupervised sub-categories of fruits (classification performed by K-means algorithm). By leveraging these annotations, we performed semi-supervised learning in object detection tasks, which is crucial for real-world applications. By allowing the creation of in-demand complex subcategories of objects, the selected model was trained in the semi-supervised learned sub-categories of fruits and contrasted with the performance from the model of supervised learning from the same number of categories and hyperparameters. Semi-supervised learning categories can provide a more accurate representation of the mathematical process of categorization in AI. This approach helps prevent human errors from propagating through the machine learning metrics by avoiding the imposition of categories or scales, especially in subjective areas such as assessing coffee fruit maturation based on epidermis color. Additionally, unsupervised learned categories expedite the annotation process and can be utilized to create mathematically optimized models and scales.

2.5. Training Details

All object detection models were trained using the default hyperparameters specified in the respective papers or repositories, except for the batch size and number of epochs. For this study, a batch size of 16 and 100 epochs were used for all models. The training and evaluation were conducted on a Tesla T4 GPU available in Google Colab (Bisong & Bisong, 2019). The evaluation metrics used in this study included Precision (Equation 3), Recall (Equation 4), and mAP (Equation 5).

$$P = \frac{TP}{TP+FP} \tag{3}$$

whereas; P - Precision, TP – Total Positives, FP – False Positives,

$$R = \frac{TP}{TP+FN} \tag{4}$$

whereas; R - Recall, TP – Total Positives, FN - False Negatives

$$mAP@.5 = \frac{1}{n} \sum_{i=0}^{n} AP_i^{0.5} \quad (5)$$

whereas; mAP - mean Average Precision, AP – Average Precision

2.6. Validation of the object detection models (YOLO)

The objective of our study was to develop a model capable of quantifying and categorizing different aspects of coffee crops. We utilized three distinct datasets: coffee fruits, coffee leaf diseases, and coffee tree detection.

For the coffee fruit dataset, we employed a trained model to quantify the fruits and classify them based on their maturity level, distinguishing between unripe and ripe fruits. The model was trained on a large dataset of labeled coffee fruit images, enabling it to learn features and patterns associated with different maturation stages. This trained model was then applied to the entire dataset, predicting the class of each fruit and providing fruit count per image. To evaluate the model's performance, we compared its predictions with ground truth labels obtained through manual annotation.

Simultaneously, we also worked with the coffee leaf diseases dataset, where our goal was to identify and classify diseases affecting coffee leaves. The model was trained on a comprehensive dataset of labeled images capturing various leaf diseases. By learning from these examples, the model gained the ability to detect and categorize different types of coffee leaf diseases.

Additionally, we tackled the coffee tree detection dataset, aiming to develop a model capable of detecting and localizing coffee trees within images. The model was trained on a diverse dataset of labeled images containing coffee tree instances. Through this training process, the model learned to identify and accurately locate coffee trees.

Individual performance object detection model was evaluated by comparing its predictions with the manual annotations in each dataset, using metrics described at Section 2.5.

2.7. *Implementation of foundational models to enhance object detection capabilities*

Unlike the training and validation process specific to the object detection model demonstrated in sections 2.5 and 2.6, foundational models are trained in advance on a diverse range of data. They acquire knowledge during the training phase, which can then be transferred to specific tasks without requiring further training (Bommasani et al., 2021). These models typically operate with short prompts, such as overall object location (Kirillov et al., 2023) or question-answering conversations (Kumar et al., 2022) . This transfer of knowledge allows the foundational models to effectively contribute to various tasks within the architecture, providing sophisticated results without the need for task-specific training.

To enhance the capabilities of our object detection model, we incorporated two foundational models into our general architecture (Section 1.2). One of these models is SAM (Segment Anything Model), which is based on autoencoders (Bank et al., 2020). SAM has the ability to perform pixel-wise classification without requiring additional training. This is accomplished by leveraging the object location and classification information obtained from the trained object detection model, specifically for tasks such as coffee fruit detection (Section 2.5., Section 2.6).

By incorporating prompts related to object location (bounding box) and object classification from trained object detection models, integrated with the d-dimensional space embedding of the input image through encoder portion processing based on Visual Transformers (Liu et al., 2023), SAM's light-weight decoder portion (Kirillov et al., 2023) enables us to generate an output mask. This mask provides pixel-level categorization, differentiating between various classifications of fruits, diseases, and coffee tree instances, along with the background. This integration not only enhances the segmentation results within the object detection framework but also enables more precise estimations, such as coffee leaf disease severity and coffee tree flower density. With this approach, we can achieve more accurate and detailed insights into the coffee crop's health and productivity.

Additionally, we included a second foundation model, a Large Language Model (Zhao et al., 2023) in our architecture. With the scalability in parameters and training-data in these large models scaling to specific tasks (Kaplan et al., 2020). Due to the large number of parameters, LLMs are hard to fine-tune, since they demand a high computational power generally available only for large-companies (Bommasaniet al., 2021). To extract conversational and knowledge transmission through the implementation of LLMs in our specific task (coffee farm parameters), we use contextualization (providing curated and good information to the machine) and prompt-engineering (Brown et al., 2020) through the model API (GPT).

3. Results

3.1. Coffee Fruits

3.1.1. YOLOv8 processed the images with the highest mean average precision in coffee fruit dataset

In the dataset of total 406 images, 324 were used as training data while 82 as validation data. After training the algorithm, the selected models (YOLOv5, YOLOv5m6, YOLOv6, YOLOv7 and YOLOv8), chosen for their high mean average precision at 50% intersection over union (mAP@.5) and real-time object detection capabilities in COCO dataset, were trained on our training dataset (324 images). While comparing the object detection efficiency of five different YOLO versions, the results showed that YOLOv8 achieved the highest mAP@.5 values in all modes, followed by YOLOv7, YOLOv5, Yolov5m6, and YOLOv6 (Table 1). YOLOv8 showed mAP@.5 value of 0.920 for mono class (only fruit), 0.899 for binary class (unripe/ripe) while 0.705 for multiclass (green, yellow-green, cherry, raisin, dry).

Table 1. **Comparison of object detection performance of five different YOLO versions in five different modes (Mono, Binary, and Multiclass)**. The values of precision (P), recall (R) and mAP@.5 are calculated using the validation data. The parameters values indicate the complexity of the models.

Model	*P*	*R*	mAP@.5$_{val}$	Parameters
Yolov8 (Mono)	0.858	*0.865*	0.920	25.9M
Yolov8 (Binary)	0.838	0.844	0.899	25.9M
Yolov8 (Multiclass)	0.632	0.723	0.705	25.9M
Yolov7 (Mono)	0.852	0.871	0.904	36.9M
Yolov7 (Binary)	0.845	0.852	0.892	36.9M
Yolov7 (Multiclass)	0.627	0.682	0.605	36.9M
Yolov5 (Mono)	0.875	0.819	0.885	21.2M
Yolov5 (Binary)	0.844	0.821	0.866	21.2M
Yolov5 (Multiclass)	0.64	0.562	0.555	21.2M
Yolov5m6 (Mono)	0.873	0.833	0.898	35.7M
Yolov5m6 (Binary)	0.848	0.821	0.875	35.7M
Yolov5m5 (Multiclass)	0.721	0.547	0.556	35.7M
Yolov6 (Mono)	0.650	0.700	0.580	34.9M
Yolov6 (Binary)	0.625	0.650	0.647	34.9M
Yolov6 (Multiclass)	0.450	0.650	0.418	34.9M

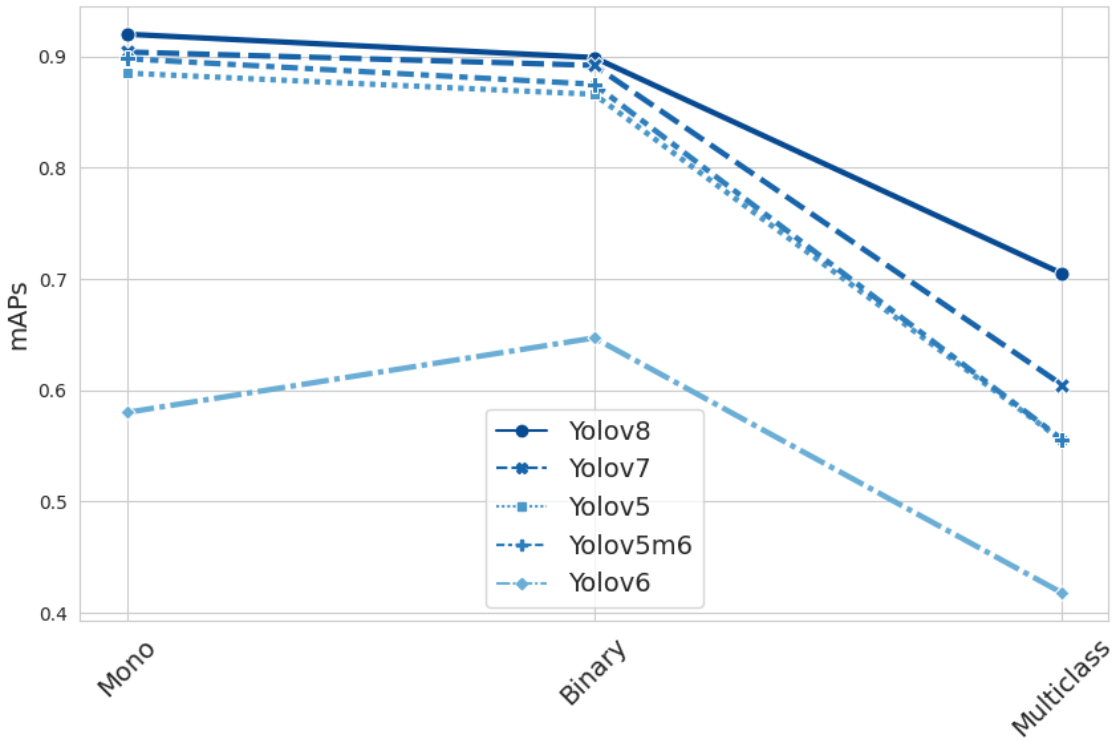


Figure 2. Comparison of the performance of five different YOLO versions. For the mean average precision at 50% intersection over union (mAP@.5) in three modes of the dataset: Mono (only fruits), Binary (unripe and ripe fruits), and Multiclass (continuous classification scale - unripe, yellow, cherry, raisin and dry), YOLOv8 outperformed YOLOv7, YOLOv5, YOLOv5m6, and YOLOv6.

3.1.2. A novel semi-supervised annotation system was developed for automatic labeling

After the model was trained with the training data, and YOLOv8 selected to proceed with, instead of manual annotation (object classification), which is obviously time-consuming and error-prone, we attempted to automate the annotation of images, which was termed as semi-supervised annotation. In this approach, the bounding boxes for training data were located by the object detection model, however, the classification was automatically performed by machine learning, hence named as semi-supervised. In other words, an object detection identified the location of the fruits in the image, and a second machine learning (K-means) algorithm labeled them with their respective categories. This successfully led to the development of a semi-supervised annotation system for training data. Here, K-means clustering was used to create categories of coffee fruits based on their color. We trained K-means models with different k-sizes ranging from 2 to 7 and evaluated their performance based on their ability

to identify distinct color clusters. To our interest, the K-means model efficiently identified distinct color clusters within the high-dimensional (28*28*2 axis) and created categories of the various stages of coffee fruits that were visually distinguishable. In this novel approach, the categories were composed of coffee fruits with similar color representing similar ripening stage, which can be useful for further analysis and classification. After categorization, we performed PCA in the high-dimensional vector to produce visualization of the boundaries among clusters (Figure 3).

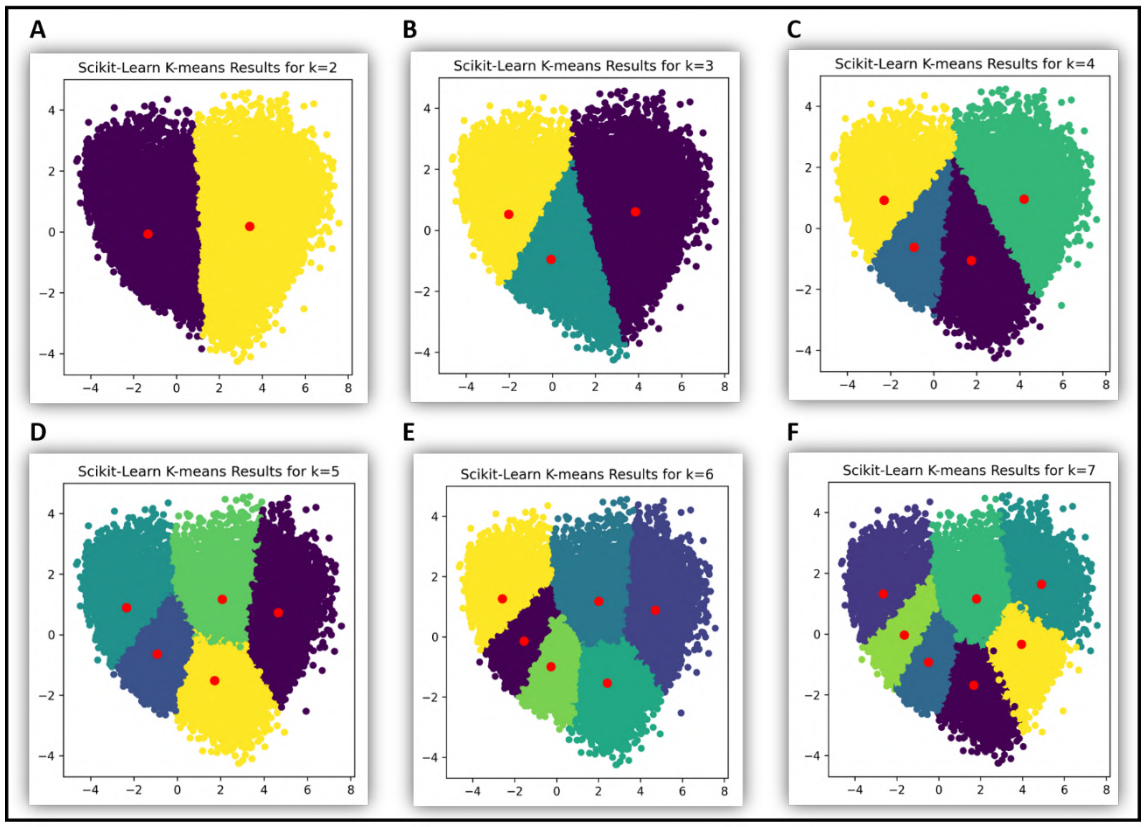


Figure 3. Principal Component analysis of dataset annotated through semi-supervised approach. Various clusters in each class are visible in different colors. As the number of classes increased from two to seven classes, the mAP@.5 value also increased, however 4 classes appeared to be optimum.

The seven various color classes of the coffee fruit ripening stage created through k-means are shown in figure 4. Employing semi-supervised approach, we attempted to annotate the same training and validation datasets which were earlier annotated using supervised method. Interestingly, the precision of detection increased with the increase in number of classes,

whereas the optimal number of classes was determined to be 4, which was also visualized through the Elbow method.

However, it is important to note that the categories created by K-means were based solely on color and did not necessarily correspond to different types or varieties of coffee fruits, such as maturity. Therefore, further analysis and classification was required to accurately identify the different types of coffee fruits being represented in each category. The further analysis to acknowledge this problem was performed by correlating the output categories with the Moraes categories used in the annotation process. By doing this, we defined categories in a crescent order of maturity.

Figure 3. K-means-based machine–generated color classes as semi-supervised method. There were up to 7 classes generated keeping k = 7, however the optimal number of classes was determined as 4.

3.1.3. Semi-supervised system proved faster and more accurate

To validate the efficiency as well as consistency of the novel machine learning method, we compared the performance of the semi-supervised model with the unsupervised one. Figure 5 shows the comparative performance of supervised and semi-supervised annotations in binary as well as multiclass modes. To our interest, the semi-supervised model performed faster and more

accurate annotation than the supervised one. Figure 5A depicts output of the supervised (5A) and semi-supervised (5B) method for binary class annotation. The number of ripe (R) and unripe (UR) fruit through ML-based annotation is written above each image. The output of comparison of supervised and semi-supervised methods in multi-class mode is shown in figure 5C and 5D, respectively. The number as well as class of the fruit detected are mentioned in the figure. The images in both the binary and multiclass represent four different time points of fruit ripening (from earlier (A, B) to later stages (C, D)). Comparing both the cases (Figure 5A, B, C, and D), it is clear that semi-supervised annotation surpassed the supervised annotation in terms of speed and accuracy. This is further simplified graphically in Figure 6. For binary class, the supervised and semi-supervised training models had an equal mAP@.5 of .89 (Figure 6A), showing similar performance for both methods. However, for multi-class detection (Figure 6B), the mAP@.5 was 0.77 in case of semi-supervised model, which was only 0.6 with the supervised method, keeping the number of categories the 4 in both cases. It proves the high resolving power of the semi-supervised annotation. Moreover, its faster and more accurate annotation feature will aid in machine learning of large dataset, in less time. This is a novel and rigorous approach to analyze large-scale coffee-fruits datasets, which can have significant implications for various fields such as computer vision, image processing, and machine learning.

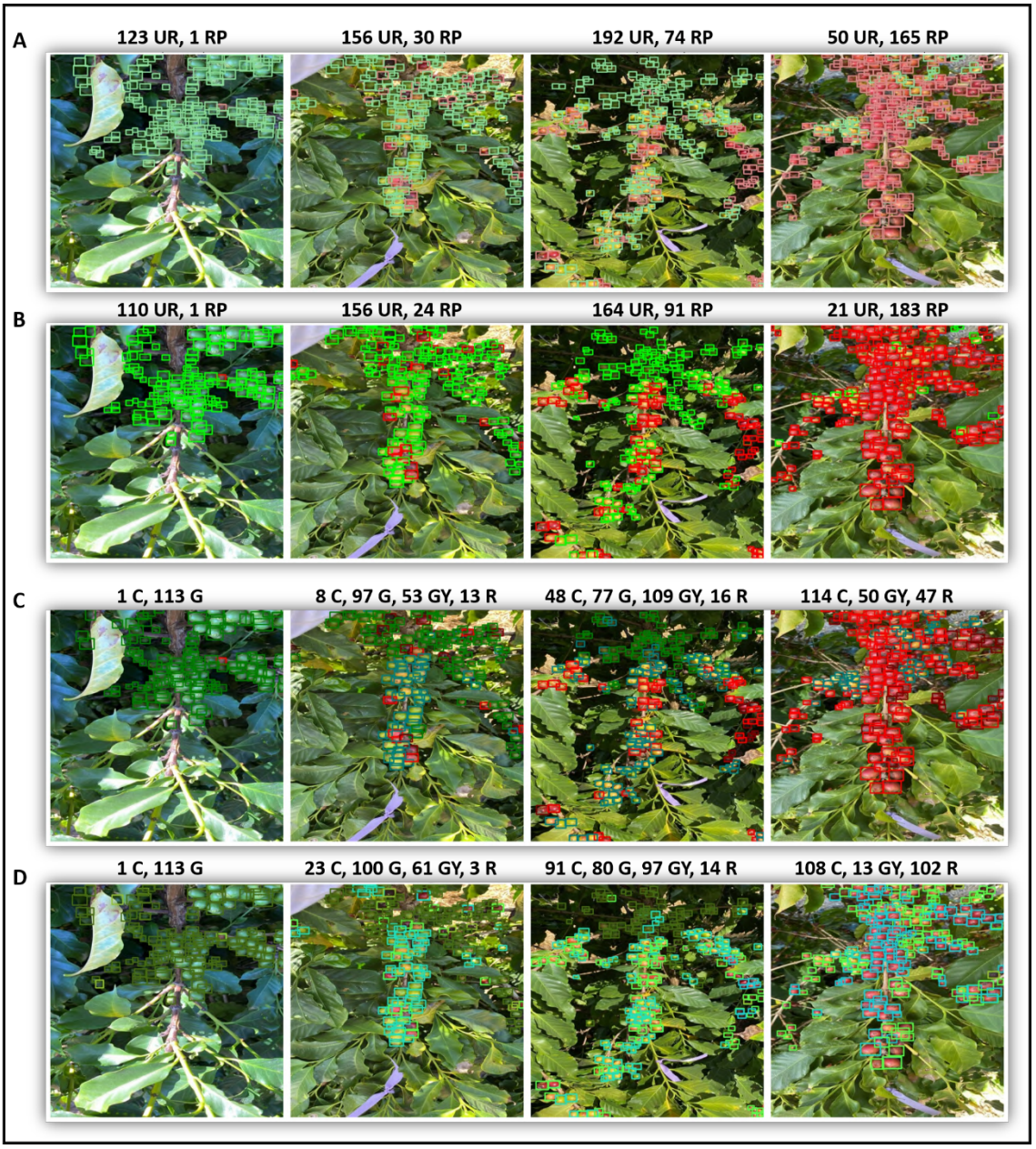


Figure 5. Comparative performance of supervised and semi-supervised methods in coffee fruits dataset. The novel method of semi-supervised annotation was compared with the supervised for binary (Supervised – **1A** Semi-supervised – **1B**) and multi-class (Supervised – **2A**, Semi-supervised –**2B)** modes. The numerals show fruit counts while letters denote fruit type as *UR – Unripe, RP – Ripe, C – Cherry G – Green, GY – Green-yellow, R – Raisin.*

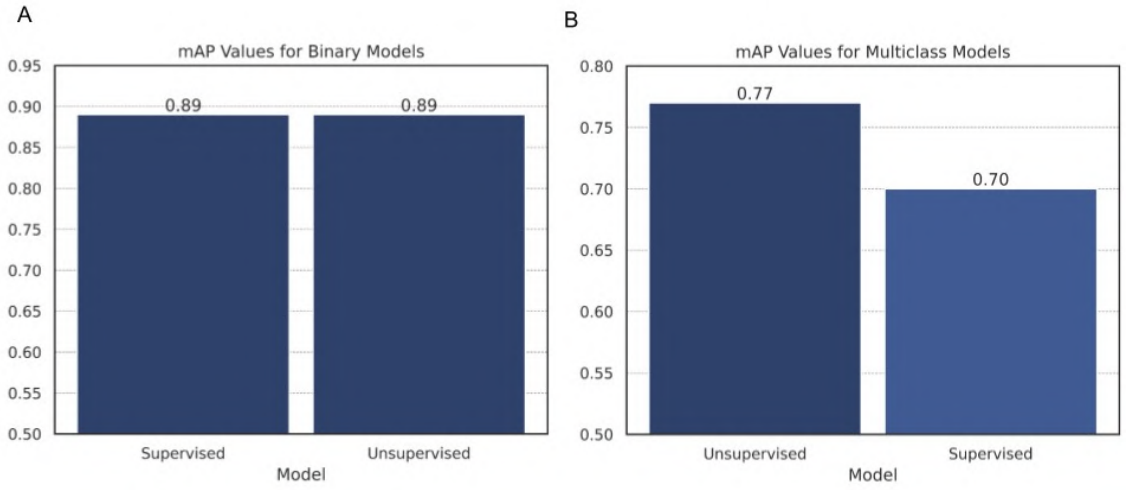


Figure 6. Graphical representation of the comparative performance of supervised and semi-supervised methods in coffee fruits dataset. For binary class, the supervised and semi-supervised training models had an equal mAP@.5 of .89, showing similar performance for both methods (**A**). However, for multi-class annotation, the semi-supervised method displayed a higher mAP@.5 value of 0.77 as compared to 0.70 of the supervised method (**B**), showing better performance.

3.1.4. The established model was validated using test images outside our dataset

To check the efficiency of our trained model, we initially tested it by feeding raw images, not included in our initial dataset. Afterwards, we also tracked the ripening of coffee fruits in real time. Both the approaches proved the higher image processing efficiency of our established model. Figure 7 depicts raw images not originally included in our dataset. The raw images from the field were analyzed with the model whereas; figure 7A shows the binary class detection counting only ripe and unripe fruits. However, multi-class fruit detection and quantification, classifying them into green, green-yellow, cherry and raisin is also shown in figure 7B. The number and category of the fruit are written above each image. This proved the model was successful in image processing. A collection of data like this will provide a broad picture of the fruit ripening pattern, estimate yield and harvesting time. The big data will eventually aid in informed decision on coffee crop management specially plans for harvest and post-harvest measures.

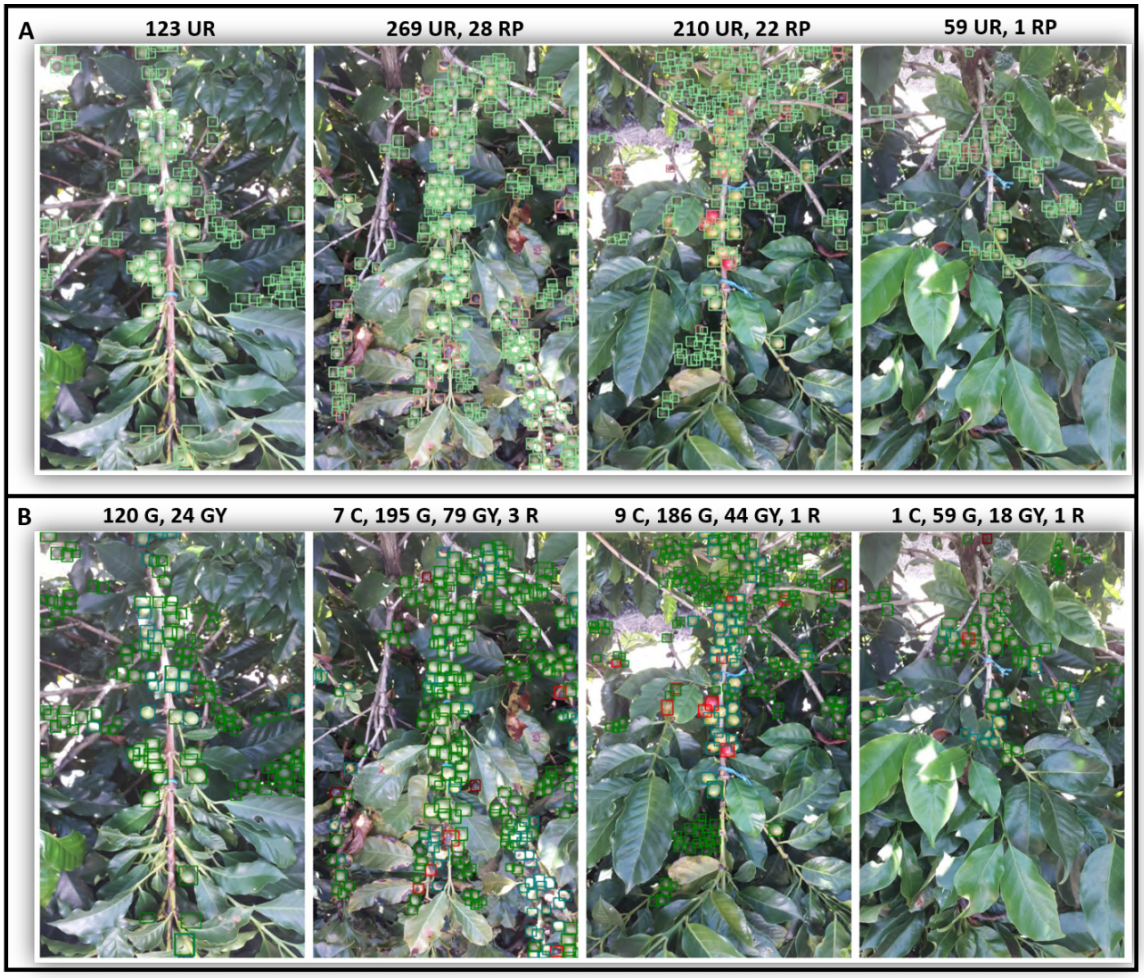


Figure 7. Evaluation the established model using test images outside of dataset. Raw images from coffee field were analyzed by binary (**A**) and multi-class (**B**) fruit detection. The numerals show fruit counts while letters denote fruit type as *UR – Unripe, RP – Ripe, C – Cherry G – Green, GY – Green-yellow, R – Raisin.*

For further validation, we tracked and analyzed the fruit ripening in a coffee field in real time for 90 days. To estimate the percentage of ripeness and unripeness in the plant, equation 6 and 7 were used . Figure 8A shows the ripening in binary mode (unripe and ripe) over the said time duration. Figure 8B depicts the ripening information of the same data in multi-class mode over the 3-month period. It is obvious from this example that for about the initial 40 days there is higher percentage of unripe fruit which turn ripe after this period. We additionally demonstrated the potential of this algorithm to be used in crop analysis and estimation, utilizing regression and ridgeline plots (Figure S2) to illustrate the variation in mature fruit proportions over the growing season.

In addition to its yield estimation capabilities, the object detection model, Yolov8, can also extract valuable information about the ripening process of crops over time. By calculating the percentage of ripe fruits over months, we can create plots that visualize the progression of ripeness levels as the crops mature, as well as the categorization of the pattern of maturation present in the farm. These plots provide farmers and researchers with valuable insights into the development of the crop, enabling them to plan harvesting schedules, optimize yields, and better understand the underlying biological processes. By quantifying ripeness in this way, we can improve our ability to predict and manage crop yields, ultimately leading to more efficient and sustainable agricultural practices. Furthermore, the data collected from the model allowed for an analysis of the distribution of ripe and unripe fruits throughout the growing season.

$$\text{Ripeness (\%)} = \frac{N(ripe\ fruits)}{N(total\ fruits)} \times 100 \quad (6)$$

$$\text{Unripeness (\%)} = 100 - Ripeness\ (\%) \quad (7)$$

Figure 8. Tracking coffee fruit ripening with the developed model over a period of 3 months. Plots show the percentage of ripe and unripe fruits over time in binary (**A**) and multi-class (**B**) modes.

3.1.5. YOLO + SAM + LLM Architecture Capabilities in Fruit Phenotyping

The workflow of image processing through the overall architecture for detecting and segmenting instances of fruit in coffee trees is shown in Figure 9. The input image (A) is first processed by YOLOv8, a trained object detection model, to identify the location and classification of fruits (B). The output of YOLOv8 is then used to prompt a second deep learning model, SAM, to generate instance segmentation for each fruit detected (C). The output mask from SAM prompted by YOLOv8 is then combined with the detection boxes from YOLOv8 to produce a final image of the fruits with background removal (C).

Figure 10 shows an example of a conversation between a user and an LLM VA (Large Language Model Virtual Assistant). The conversation is in two colors: blue for the LLM VA's responses and gray for the user's input.

The user is a coffee farmer who is using object detection and segmentation to extract (automatically used as textual prompt to the Virtual Assistant) information about their coffee farm. They are interested in learning more about coffee productivity. The LLM VA is able to provide the user with detailed information about these factors. The LLM VA provides the user with some tips on how to improve their coffee cultivation practices, such as watering their trees regularly, fertilizing them every 6 months, and pruning them regularly. The conversation between the user and the LLM VA is an example of how this technology can be used to help coffee farmers improve their productivity and quality of their coffee beans. By using object detection and segmentation to extract information about their farms, coffee farmers can better understand the factors that are affecting the growth and production of their coffee trees, combinating with the capacity of conversation and knowledge transmission from the Virtual Assistant. This conversation is an example on how the entire model can be used to make informed decisions about how to improve their farming practices.

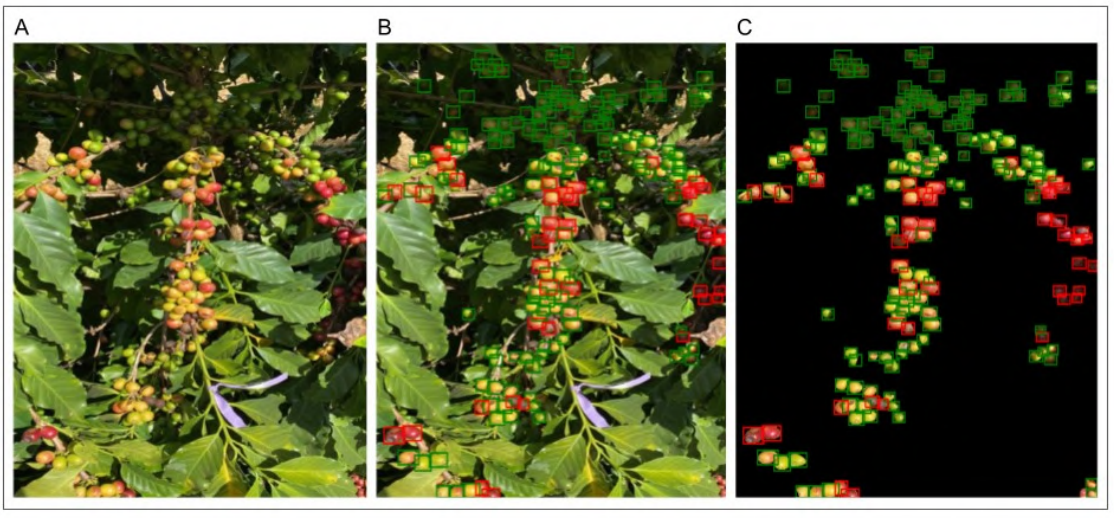


Figure 9. Detecting and Segmenting Instances of Fruit. Plots show the workflow of image processing through the overall architecture (**A**) Input Image (**B**) Fruit location and classification (detection) by YOLOv8 (**C**) Background removal through SAM mask output (prompted by YOLOLv8 detection) + YOLOv8 detection (boxes).

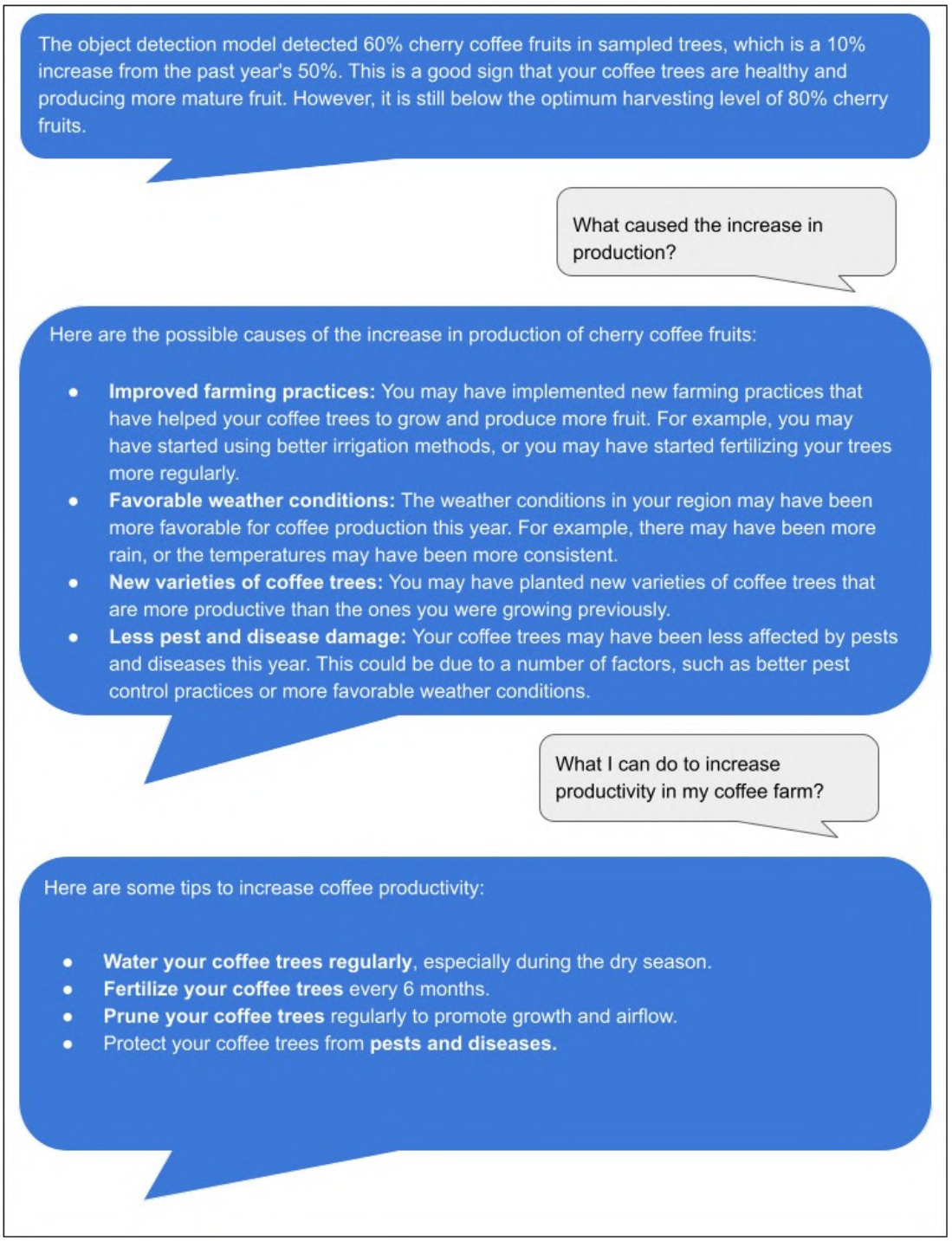


Figure 10. Example of conversation between user and LLM VA (Large Language Model Virtual Assistant) about coffee fruits and coffee farm productivity. In blue are shown answers from the Large Language Model API and in gray user input.

3.2. Leaf Disease

3.2.1. YOLOv8 processed the images with the highest mean average precision in coffee leaf disease dataset

In order to train a computer vision-based machine, such as an algorithm designed to identify target objects in images, annotated images must be provided as the initial input. For the leaf disease dataset, we obtained data by sampling various curated and publicly available coffee leaf diseases datasets (Brito Silva et al.,2020; Madhukar et al., 2020; Krohling et al., 2019). This dataset consisted of a total of 6,427 images, which were divided into 80% (5,304) for training and 20% (1,123) for validation purposes. Different detection object models were trained using this dataset, as illustrated in Figure 11.

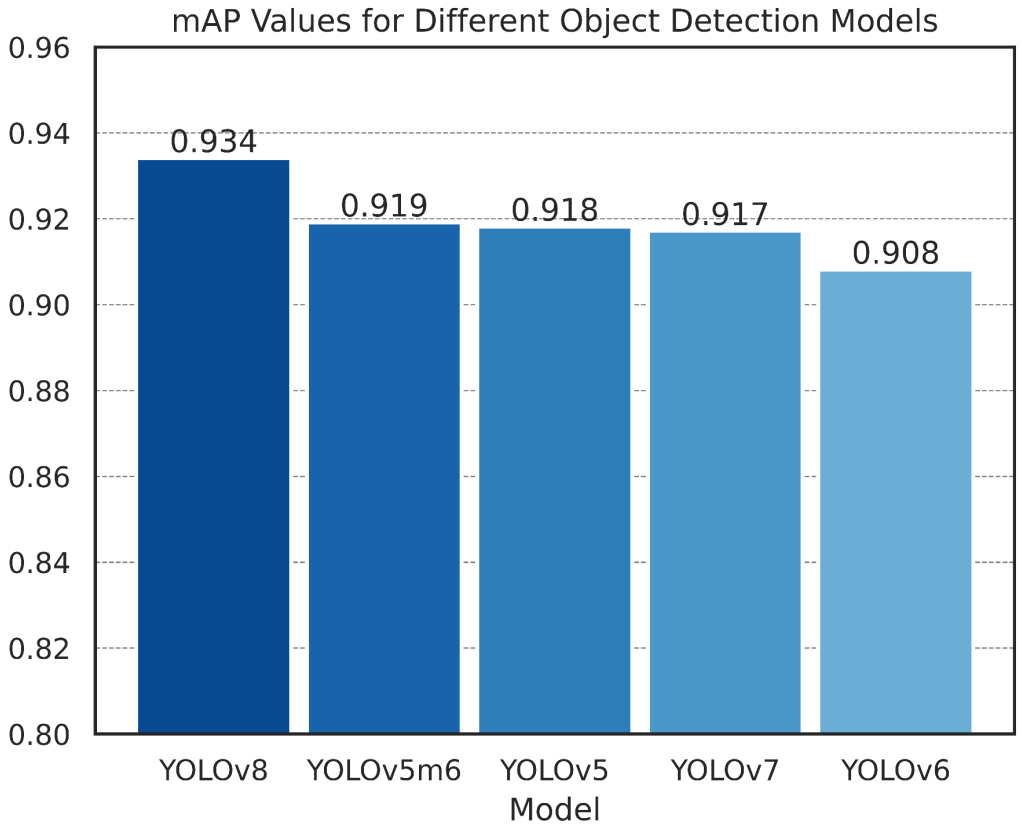


Figure 11: Comparison of the performance of five different YOLO versions in coffee leaf disease detection. For the mean average precision at 50% intersection over union (mAP@.5) in coffee leaf disease dataset (rust, miner, phoma,and cercospora) YOLOv8 outperformed YOLOv5m6, YOLOv5, YOLOv7, and YOLOv6.

While comparing the object detection efficiency of the different YOLO versions, the results showed that YOLOv8 achieved the highest mAP@.5 values followed by YOLOv5m6, YOLOv5, YOLOv7, and YOLOv6 (Table 2). YOLOv8 showed mAP@.5 value of 0.934.

Table 2. **Comparison of object detection performance of five different YOLO versions in coffee leaf disease dataset**. The values of precision (P), recall (R) and mAP@.5 are calculated using the validation data. The parameters values indicate the complexity of the models.

Model	*P*	*R*	mAP@.5$_{val}$	Parameters
YOLOv8	*0.915*	*0.883*	*0.934*	25.9M
YOLOv5m6	0.942	0.876	0.919	35.7M
YOLOv5	0.93	0.869	0.918	21.2M
YOLOv7	*0.929*	*0.879*	*0.917*	36.9M
YOLOv6	0.908	0.488	0.907	34.9M

3.2.3. YOLO+SAM+LLM architecture potentials in leaf disease phenotyping

The leaf input image is first processed by YOLOv8, a trained object detection model, to identify the location and classification of coffee leaf diseases. The output of YOLOv8 is then used to prompt a second deep learning model, SAM, to generate instance segmentation for each disease detected. The output mask from SAM prompted by YOLOv8 is then combined with the detection boxes from YOLOv8 to produce a final image of the leaf and disease with background removal. To assess the severity of a particular disease, we can utilize the output mask obtained from SAM. By calculating the ratio of disease pixels to leaf pixels, we can quantify the severity as a percentage using Equation 8. This calculation indicates the proportion of the coffee leaf that has been affected by the specific disease (Figure 12).

$$\text{Severity (\%)} = \frac{Disease_{pixels}}{Leaf_{pixels}} \times 100 \quad (8)$$

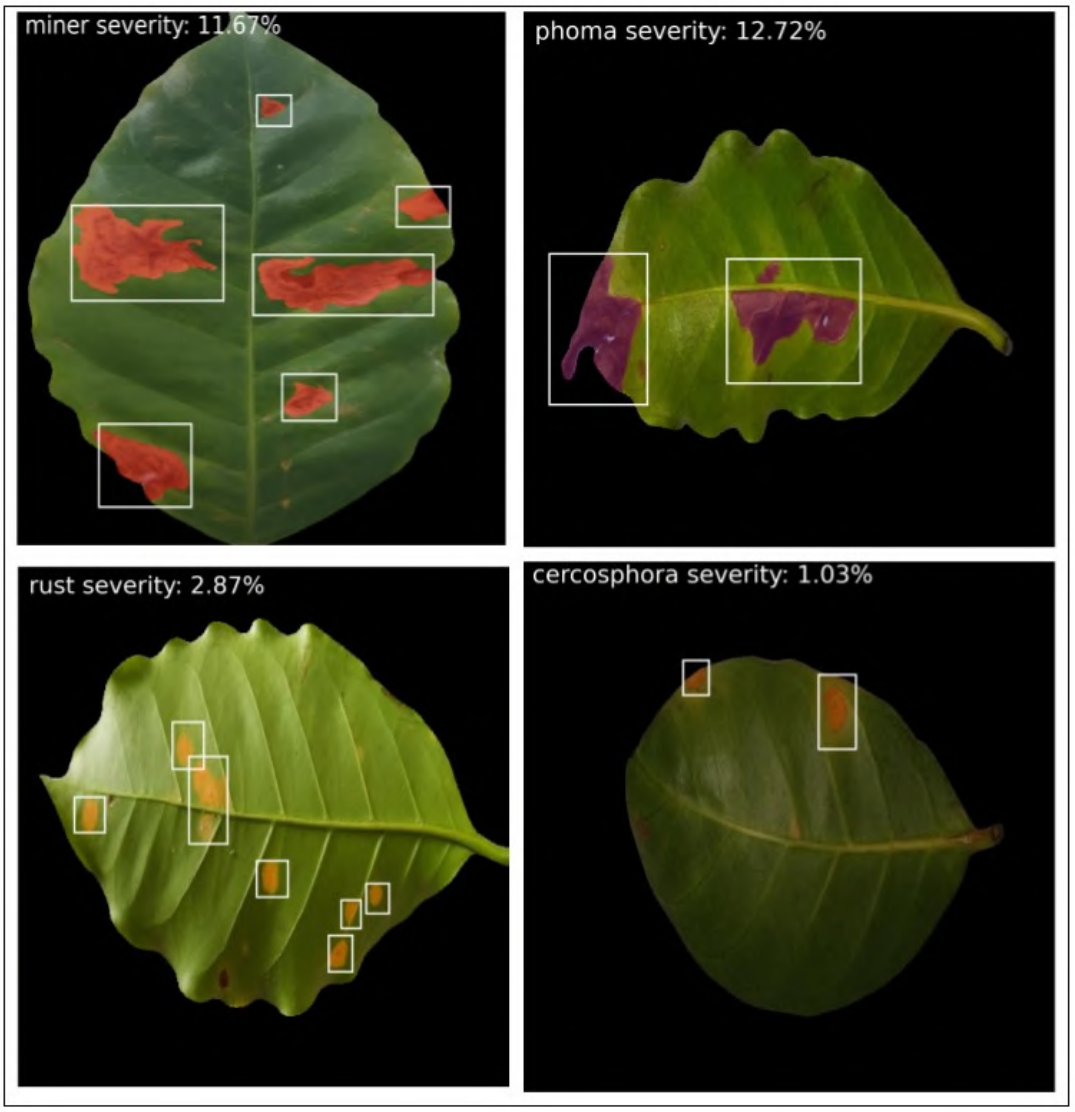


Figure 12 Different output instances of coffee leaf disease dataset processed by YOLO+SAM. The detection capacity of YOLOv8 is demonstrated by the presence of white bounding boxes and disease names, indicating its ability to identify and classify coffee leaf diseases. SAM, prompted by YOLOv8's leaf and disease detections, showcases its capabilities in leaf segmentation, background removal, and disease severity estimation.

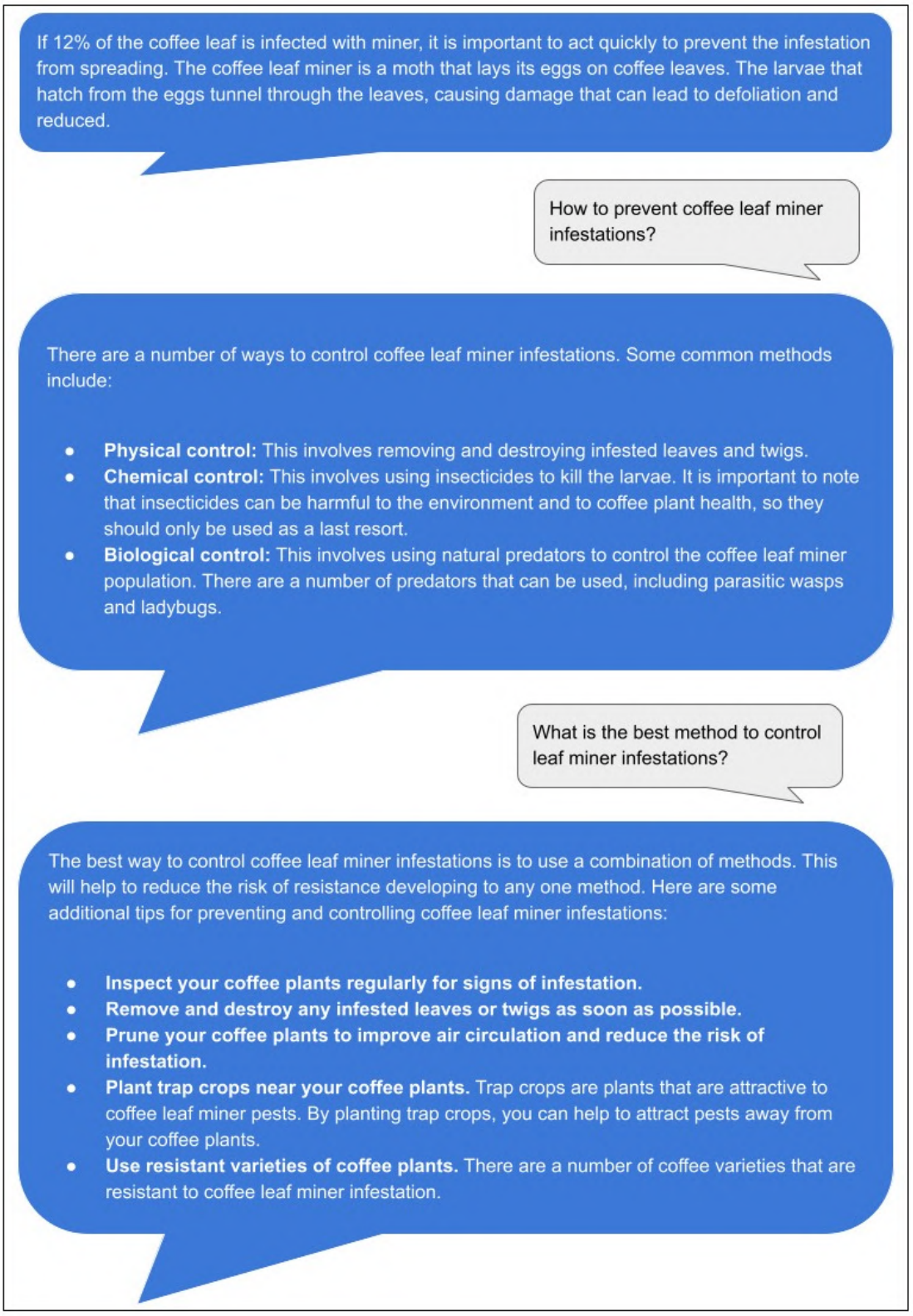


Figure 13. Example of conversation between user and LLM VA (Large Language Model Virtual Assistant). User and Large Language Model conversation about coffee leaf diseases. In blue are shown answers from the Large Language Model API and in gray user input.

3.3. Coffee Tree

3.3.1. YOLOv8 processed the images with the highest mean average precision in coffee tree detection dataset

The tree dataset consists of 174 images splitted in 80% (139) training and 20% validation (35). A similar methodology to the disease severity (section 3.2.2.) was applied to segment the tree and white pixels (flowers) and calculate the density (relation of white pixels to tree). To do this, we trained an object detection model to detect trees in complex environments, fed the output as a prompt to SAM and performed quantification of pixels in the output mask. The results for trained object detection models are shown in figure 14.

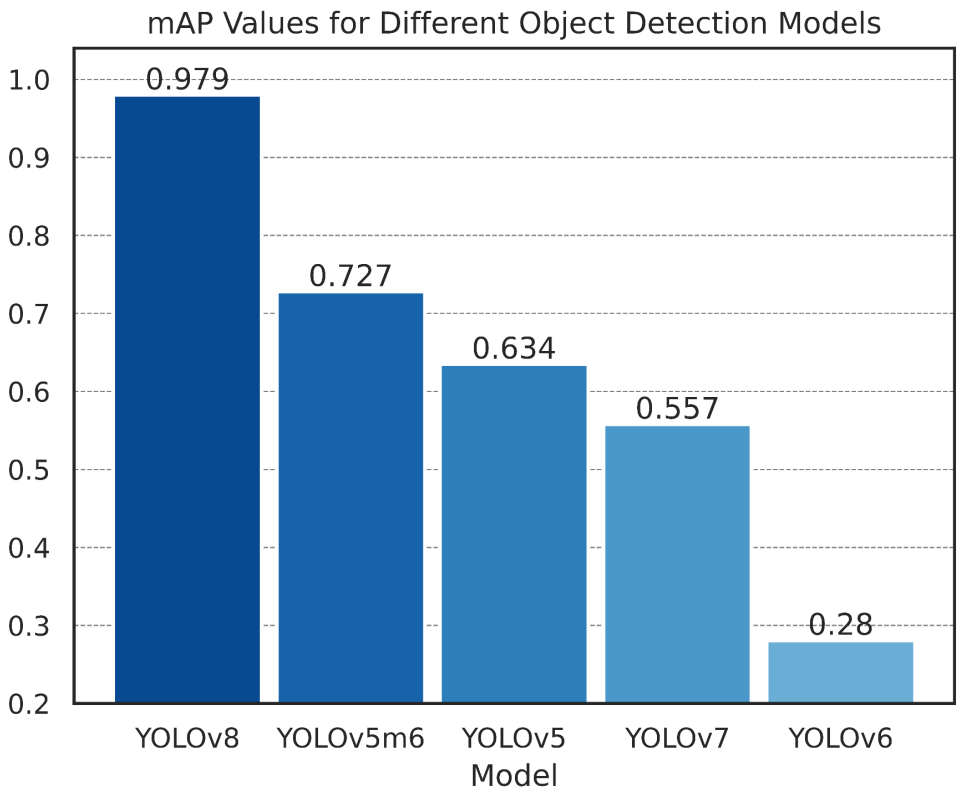


Figure 14: Comparison of the performance of five different YOLO versions in coffee tree detection. For the mean average precision at 50% intersection over union (mAP@.5) in coffee tree detection dataset. YOLOv8 outperformed YOLOv5m6, YOLOv5, YOLOv7, and YOLOv6.

While comparing the object detection efficiency of the different YOLO versions, the results showed that YOLOv8 achieved the highest mAP@.5 values followed by YOLOv5m6, YOLOv5, YOLOv7, and YOLOv6 (Table 2). YOLOv8 showed mAP@.5 value of 0.979.

Table 3. **Comparison of object detection performance of five different YOLO versions in coffee tree detection**. The values of precision (P), recall (R) and mAP@.5 are calculated using the validation data. The parameters values indicate the complexity of the models.

Model	*P*	*R*	mAP@.5$_{val}$	Parameters
YOLOv8	*0.949*	*0.952*	*0.979*	25.9M
YOLOv5m6	0.697	0.924	0.727	35.7M
YOLOv5	0.595	0.929	0.634	21.2M
YOLOv7	*0.579*	*0.646*	*0.557*	36.9M
YOLOv6	0.207	0.637	0.280	34.9M

3.3.3. YOLO+SAM+LLM architecture potentials in coffee tree flower density estimation

The process to estimate tree flower density in coffee tree (Figure 15) began by detecting the tree in the image, through our trained object detection, YOLOv8 (Section 3.3.2.). To accomplish a better estimation of coffee flower density, we used output from object detection (15A) as prompting to SAM, which performed segmentation (15B).

Once the segmentation was complete, the resulting tree segmentation mask for each tree was obtained (15C). This mask provided a visual representation of individual trees and their regions occupied by flowers in the tree in white. Filters were applied to create binary masks, where instances of flower were distinctly from tree, while the background was removed in the segmentation proccess (15D).

To extract the white pixels corresponding to the tree from the segmentation mask, we applied specific filters. These filters enhanced the visibility of the white areas, making them more pronounced and distinguishable from the tree pixels. Consequently, the image transformed into a binary representation, where the background pixels were assigned one value (e.g., black or zero), and the white pixels corresponding to the tree were assigned another value (e.g., white or one). This binary representation simplified subsequent analysis and processing tasks involving the tree, as it clearly differentiated between the tree and the surrounding environment. By acquiring the flower and tree pixels, we used equation 9 to estimate the flower density.

$$\text{Flower Density (\%)} = \frac{Flower_{pixels}}{Tree_{pixels}} \times 100 \quad (9)$$

Figure 15: Estimation of flower density in coffee tree process. (**A**) Different tree detect by YOLOv8 (object detection model), (**B**) Individual tree is sampled from multi object detection output, (**C**) Individual tree is segmented by prompting SAM with bounding boxes from YOLOv8 output, (D) Segmented tree is then passed through white pixel extraction and flower density is estimated (28.5%).

Discussion

Biotic and abiotic stresses are becoming more pronounced with the ongoing climate change. Unusual weathers such as intense rain, temperature fluctuations and prolonged droughts are now damaging our crops more than ever before. The fruit crop such as coffee is temperature-sensitive, and is highly vulnerable to temperature ups and downs. The optimal temperature for the main growing species (Arabica and Robusta) is between 18-28 ℃ (Magrach & Ghazoul, 2015). Temperature beyond this range for certain period causes serious damages to coffee, leading to significant yield losses. The common symptom of such adversity is the dried fruit (black). Coffee fruit typically has a long and non-uniform ripening window (Kazama et al., 2021), and is thus relatively more prone to environmental adversities during ripening period. Another issue associated with coffee is the asynchronous ripening which is due to asynchronous flowering (Cardon et al., 2022). While for the quality coffee, the ripe cherries should make the maximum percentage of harvested fruit, farmers must take care at the harvesting stage.

Traditional methods of yield estimation and decision about harvest time for crops like coffee involve manual counting of fruits (usually through destructive sampling), which is laborious, time-consuming, and often error-prone. Federación Nacional de Cafeteros (FNC) reported a significant portion of coffee fruit is wasted by the destructive sampling which includes 60 coffee trees per hectare in an area of 2000 hectares, for coffee yield estimation l (Ramos, et al., 2016). To rapidly monitor coffee crop specially during fruit ripening time which is usually from January to May and June, estimate the yield and harvest time, forecasting market supply, and reducing production costs, precision agricultural approaches present an efficient and reliable solution. These take help from artificial intelligence technology, especially computer vision.

In this study, we present a novel computer vision-based approach for estimating coffee crop yields as well as harvest time. Our approach involves training of a model for which the image dataset was annotated using online tools. Our dataset included images of fruit bearing branches of coffee plants. While addressing the above-mentioned issues, we trained the state-of-the-art object detection model YOLOv8 (Wang et al., 2022) to count and classify coffee fruit in images. YOLO is CNN-based object detection model used in numerous areas such including traffic. Convolutional Neural Networks (CNNs) (Sultana et al., 2020) are a type of deep learning model that are inspired by the human visual system and consist of multiple convolutional and pooling layers followed by fully connected layers. The input layer of a CNN takes in the input data, which could be an image, and passes it through a series of convolutional and pooling layers.

The convolutional layers apply convolution operations to the input data to extract features such as edges, corners, and textures, which are important for object recognition. The pooling layers reduce the spatial dimensions of the feature maps obtained from the convolutional layers, which helps in reducing computational complexity and improving the model's ability to generalize. The output of the last pooling layer is then flattened into a vector, which is passed through fully connected layers. These fully connected layers learn complex patterns and relationships between the features extracted from the convolutional layers, and produce the final prediction output (Bellocchio et al., 2019).

During the training process, the forward computation is performed to make predictions, and the backward computation is performed to compute the gradients of the model parameters based on the prediction output and the labeled ground-truth. The gradients are then used to update the parameters of the model, typically using optimization algorithms such as gradient descent, which iteratively adjust the parameters to minimize the loss or cost function. The training process continues for a determined number of iterations of forward and backward stages, also known as epochs, until a stopping criterion is met, such as reaching a certain level of accuracy or a maximum number of epochs. This helps the model learn the optimal parameters for making accurate predictions on the training data (Liu et al., 2015).

Comparing YOLOv8 with the two other versions, YOLOv7, YOLOv6, YOLOv5 and YOLOv5m6 (Table 1,2,3), we obtained highest mAP value with YOLOv8 in all datasets. mAP is the average precision of accurately identifying the target object in images and thus larger values means better performance of the model.

One of the challenges of annotation using an adapted scale in the field is the increase of confusion between classes due to the overlapping of features present in the classes. This can be seen by comparing the unsupervised and supervised methods Figure 6. This confusion can affect the reliability and validity of the annotation process and compromise the quality of the data. Therefore, it is important to create mathematically optimized scales that can reduce ambiguity and increase consistency among annotators. Such scales can also facilitate the analysis and interpretation of the data and enhance the scientific contribution of the research. Although YOLOv8 displayed higher values even in multiclass mode, yet it still needs further improvement to achieve the best results. Enlarging and further diversifying the training dataset could also further improve its performance.

The semi supervised model we devised is a novel method for annotating coffee images in training as well as validation data. It holds enormous potential in handling a dataset despite its size. One thing to note is the number of classes at which the color classification is optimum. In our case, we determined the optimal number of color classes to be 4. Our semi-supervised

method outperformed the supervised method as it gave a mAP@.5 of 0.77 compared to 0.70 of the supervised method in multiclass detection mode. The performance could even be enhanced in future as newer improved versions of YOLO become available. However, increasing this value should be addressed in future CV studies regarding coffee.

In addition to implementing trained object detection models for various aspects of coffee farms, such as coffee leaf and diseases, coffee fruits, and coffee tree detection, we further improved the capabilities of these models. This enhancement was achieved by utilizing foundation models like SAM (Segmentation Algorithm Model) and LLMs (Large Language Models), which enabled us to extract more precise information from the initial outputs.

Using SAM, we were able to obtain more accurate estimations of flower density by segmenting the images and extracting specific information related to the presence and distribution of flowers. This allowed us to understand the flowering patterns of coffee trees in greater detail.

Furthermore, by employing LLMs, we incorporated a virtual assistant into our system. This assistant could provide valuable insights and support in analyzing the collected data, allowing for a more comprehensive understanding of the coffee farm and its production processes on a larger scale.

By quantifying and analyzing the three parameters of coffee, diseases, and flowers, we adopted a holistic approach towards comprehending coffee farm dynamics. This approach facilitated a deeper understanding of coffee production, enabling farmers and researchers to make informed decisions and take appropriate actions to optimize their farming practices at large scales.

Conclusions

As a temperature sensitive tropical plant, weather changes can greatly affect coffee yield. Thus, the continuous monitoring of coffee, especially during fruit ripening is indispensable for its production in a sustainable manner. Under these scenarios, computer vision-aided coffee fruit quantification and yield estimation studies have been carried out in the recent past, however

some of them used expensive machinery and complex image processing while some used outdated machine learning models. Using the latest state-of-the-art YOLOv7, we obtained an mAP@.5 of 0.89, the highest ever so far. We also devised a semi-supervised method of annotating coffee images (training data), which would greatly aid in handling large datasets and save time. The algorithm is efficient in CV-aided coffee fruit counting, through which the yield and harvest time can be estimated. With the integration of UAV and other value addition, the developed system holds enormous potential to be used in monitoring coffee farms for informed decision on timely field management, harvest time and post-harvest measures, which will ultimately enhance coffee yield and contribute to sustainable coffee production.

References

Avendano, J., Ramos, P. J., & Prieto, F. A. (2017). A system for classifying vegetative structures on coffee branches based on videos recorded in the field by a mobile device. *Expert Systems with Applications*, *88*, 178–192. https://doi.org/https://doi.org/10.1016/j.eswa.2017.06.044

Bank, D., Koenigstein, N., & Giryes, R. (2020). Autoencoders. *arXiv preprint arXiv:2003.05991*.

Bazame, H. C., Molin, J. P., Althoff, D., & Martello, M. (2021). Detection, classification, and mapping of coffee fruits during harvest with computer vision. *Computers and Electronics in Agriculture*, *183*, 106066. https://doi.org/https://doi.org/10.1016/j.compag.2021.106066

Bazame, H. C., Molin, J. P., Althoff, D., Martello, M., & Corrêdo, L. D. P. (2022). Mapping coffee yield with computer vision. *Precision Agriculture*, *23*(6), 2372–2387. https://doi.org/10.1007/s11119-022-09924-0

Bellocchio, E., Ciarfuglia, T. A., Costante, G., & Valigi, P. (2019). Weakly Supervised Fruit Counting for Yield Estimation Using Spatial Consistency. *IEEE Robotics and Automation*

Letters, *4*(3), 2348–2355. https://doi.org/10.1109/LRA.2019.2903260

Bisong, E., & Bisong, E. (2019). Google colaboratory. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, 59–64.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *ArXiv Preprint ArXiv:2004.10934*.

Brito Silva, L., Cavalcante Carneiro, A. L., & Silveira Almeida Renaud Faulin, M. R. (2020). and Leaf Miner (Leucoptera coffeella) in Coffee Crop (Coffea arabica). *Mendeley Data*, *4*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Cao, L., Zheng, X., & Fang, L. (2023). The Semantic Segmentation of Standing Tree Images Based on the Yolo V7 Deep Learning Algorithm. In *Electronics* (Vol. 12, Issue 4). https://doi.org/10.3390/electronics12040929

Cardon, C. H., de Oliveira, R. R., Lesy, V., Ribeiro, T. H. C., Fust, C., Pereira, L. P., Colasanti, J., & Chalfun-Junior, A. (2022). Expression of coffee florigen CaFT1 reveals a sustained floral induction window associated with asynchronous flowering in tropical perennials. *Plant Science*, *325*, 111479.

Carrillo, E., & Penaloza, A. A. (2009). Artificial vision to assure coffee-Excelso beans quality. *Proceedings of the 2009 Euro American Conference on Telematics and Information Systems: New Opportunities to Increase Digital Citizenship*, 1–8.

Chemura, A., Kutywayo, D., Chidoko, P., & Mahoya, C. (2016). Bioclimatic modelling of current and projected climatic suitability of coffee (Coffea arabica) production in Zimbabwe. *Regional Environmental Change*, *16*(2), 473–485. https://doi.org/10.1007/s10113-015-0762-9

Chemura, A., Mutanga, O., & Dube, T. (2017). Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions. *Precision Agriculture*, *18*, 859–881.

de Oliveira, E. M., Leme, D. S., Barbosa, B. H. G., Rodarte, M. P., & Pereira, R. G. F. A. (2016). A computer vision system for coffee beans classification based on computational intelligence techniques. *Journal of Food Engineering*, *171*, 22–27. https://doi.org/https://doi.org/10.1016/j.jfoodeng.2015.10.009

Dey, D., Mummert, L., & Sukthankar, R. (2012). Classification of plant structures from uncalibrated image sequences. *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, 329–336.

FAO. (2023). *Food and Agriculture Organization*. https://www.fao.org/markets-and-trade/commodities/coffee/en/

Guerrero, J. M., Guijarro, M., Montalvo, M., Romeo, J., Emmi, L., Ribeiro, A., & Pajares, G. (2013). Automatic expert system based on images for accuracy crop row detection in maize fields. *Expert Systems with Applications*, *40*(2), 656–664.

Haile, M., & Kang, W. H. (2019). The harvest and post-harvest management practices' impact on coffee quality. *Coffee-Production and Research*, 1–18.

Hameed, K., Chai, D., & Rassau, A. (2018). A comprehensive review of fruit and vegetable classification techniques. *Image and Vision Computing*, *80*, 24–44.

ICO. (2023). *International Coffee Organization*. 2023. https://www.ico.org/

Janandi, R., & Cenggoro, T. W. (2020). An Implementation of Convolutional Neural Network for Coffee Beans Quality Classification in a Mobile Information System. *2020 International Conference on Information Management and Technology (ICIMTech)*, 218–222. https://doi.org/10.1109/ICIMTech50083.2020.9211257

Jay, S., Rabatel, G., Hadoux, X., Moura, D., & Gorretta, N. (2015). In-field crop row

phenotyping from 3D modeling performed using Structure from Motion. *Computers and Electronics in Agriculture*, *110*, 70–77.

Jayakumar, M., Rajavel, M., Surendran, U., Gopinath, G., & Ramamoorthy, K. (2017). Impact of climate variability on coffee yield in India—with a micro-level case study using long-term coffee yield data of humid tropical Kerala. *Climatic Change*, *145*(3), 335–349. https://doi.org/10.1007/s10584-017-2101-2

Jha, S., Bacon, C. M., Philpott, S. M., Ernesto Méndez, V., Läderach, P., & Rice, R. A. (2014). Shade Coffee: Update on a Disappearing Refuge for Biodiversity. *BioScience*, *64*(5), 416–428. https://doi.org/10.1093/biosci/biu038

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., & Kwon, Y. (2022). ultralytics/yolov5: V6. 1-TensorRT TensorFlow edge TPU and OpenVINO export and inference. *Zenodo*, *2*, 2.

Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics. Version 8.0.0. https://github.com/ultralytics/ultralytics.

Kazama, E. H., da Silva, R. P., Tavares, T. de O., Correa, L. N., de Lima Estevam, F. N., Nicolau, F. E. de A., & Maldonado Júnior, W. (2021). Methodology for selective coffee harvesting in management zones of yield and maturation. *Precision Agriculture*, *22*, 711–733.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

Krishnan, S. (2017). *Sustainable Coffee Production*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199389414.013.224

Krohling, Renato A.; Esgario, José; Ventura, José A. BRACOL–a Brazilian Arabica Coffee Leaf images dataset to identification and quantification of coffee diseases and pests. **Mendeley**

Data, v. 1, 2019.

Kumar, H., Musabirov, I., Shi, J., Lauzon, A., Choy, K. K., Gross, O., ... & Williams, J. J. (2022). Exploring the design of prompts for applying gpt-3 based chatbots: A mental wellbeing case study on mechanical turk. *arXiv preprint arXiv:2209.11344*.

LabelStudio. (n.d.). *Open Source Data Labeling*. 2021. Retrieved March 13, 2023, from https://labelstud.io/

Läderach, P., Ramirez–Villegas, J., Navarro-Racines, C., Zelaya, C., Martinez–Valle, A., & Jarvis, A. (2017). Climate change adaptation of coffee production in space and time. *Climatic Change*, *141*(1), 47–62. https://doi.org/10.1007/s10584-016-1788-9

Leroy, T., Ribeyre, F., Bertrand, B., Charmetant, P., Dufour, M., Montagnon, C., Marraccini, P., & Pot, D. (2006). Genetics of coffee quality. *Brazilian Journal of Plant Physiology*, *18*, 229–242.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., ... & Wei, X. (2022). YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*.

Li, C., Li, L., Geng, Y., Jiang, H., Cheng, M., Zhang, B., Ke, Z., Xu, X., & Chu, X. (2023). YOLOv6 v3. 0: A Full-Scale Reloading. *ArXiv Preprint ArXiv:2301.05586*.

Liu, T., Fang, S., Zhao, Y., Wang, P., & Zhang, J. (2015). Implementation of training convolutional neural networks. *ArXiv Preprint ArXiv:1506.01195*.

Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., ... & He, Z. (2023). A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*.

López, M. E., Santos, I. S., de Oliveira, R. R., Lima, A. A., Cardon, C. H., & Chalfun-Junior, A. (n.d.). An overview of the endogenous and environmental factors related to the Coffea arabica flowering process. *Beverage Plant Research*, *1*(1), 1–16. https://doi.org/10.48130/BPR-2021-0013

Magrach, A., & Ghazoul, J. (2015). Climate and Pest-Driven Geographic Shifts in Global Coffee Production: Implications for Forest Cover, Biodiversity and Carbon Storage. *PLOS ONE*, *10*(7), e0133071. https://doi.org/10.1371/journal.pone.0133071

Martello, M., Molin, J. P., & Bazame, H. C. (2022). Obtaining and Validating High-Density Coffee Yield Data. In *Horticulturae* (Vol. 8, Issue 5). https://doi.org/10.3390/horticulturae8050421

Madhukar, R. K., Chaurasiya, A., & Chaturvedi, P. (2022, October). A Systematized Chronicity based Disease Classification in Coffee Leaves using Deep Learning. In *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 1336-1342). IEEE.

Meylan, L., Gary, C., Allinne, C., Ortiz, J., Jackson, L., & Rapidel, B. (2017). Evaluating the effect of shade trees on provision of ecosystem services in intensively managed coffee plantations. *Agriculture, Ecosystems & Environment*, *245*, 32–42. https://doi.org/https://doi.org/10.1016/j.agee.2017.05.005

Moonrinta, J., Chaivivatrakul, S., Dailey, M. N., & Ekpanyapong, M. (2010). Fruit detection, tracking, and 3D reconstruction for crop mapping and yield estimation. *2010 11th International Conference on Control Automation Robotics & Vision*, 1181–1186.

Muñoz Pérez, C. (2017). *Development of an Android application that allows processing and storing records of coffee branches* [Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/bitstream/handle/unal/62027/1018417862.2017.pdf?sequence=1&isAllowed=y

Nogueira Martins, R., de Carvalho Pinto, F. D., Marçal de Queiroz, D., Magalhães Valente, D. S., & Fim Rosas, J. T. (2021). A Novel Vegetation Index for Coffee Ripeness Monitoring Using Aerial Imagery. In *Remote Sensing* (Vol. 13, Issue 2). https://doi.org/10.3390/rs13020263

Oh, J., Singh, S., Lee, H., & Kohli, P. (2017, July). Zero-shot task generalization with multi-task deep reinforcement learning. In *International Conference on Machine Learning* (pp. 2661-2670). PMLR.

P. Ramos, F. Prieto, C. Oliveros, N. F. Aleixos, F. Albert, B. J. (2016). Medición del porcentaje de madurez en ramas de café mediante dispositivos móviles y visión por computador. *VIII Congreso Ibérico de Agroingeniería. Retos de La Nueva Agricultura Mediterránea*, 917–925. http://hdl.handle.net/20.500.11939/6828

Patel, H. N., Jain, R. K., & Joshi, M. V. (2011). Fruit detection using improved multiple features based algorithm. *International Journal of Computer Applications*, *13*(2), 1–5.

Ramos, P. J., Avendaño, J., & Prieto, F. A. (2018). Measurement of the ripening rate on coffee branches by using 3D images in outdoor environments. *Computers in Industry*, *99*, 83–95. https://doi.org/https://doi.org/10.1016/j.compind.2018.03.024

Ramos, P. J., Prieto, F. A., Montoya, E. C., & Oliveros, C. E. (2017). Automatic fruit count on coffee branches using computer vision. *Computers and Electronics in Agriculture*, *137*, 9–22. https://doi.org/https://doi.org/10.1016/j.compag.2017.03.010

Rodríguez, J. P., Corrales, D. C., Aubertot, J.-N., & Corrales, J. C. (2020). A computer vision system for automatic cherry beans detection on coffee trees. *Pattern Recognition Letters*, *136*, 142–153. https://doi.org/https://doi.org/10.1016/j.patrec.2020.05.034

Ságio, S. A. (2009). *Características fisiológicas e bioquímicas de frutos de duas cultivares de café de ciclos de maturação precoce e tardio*.

Sultana, F., Sufian, A., & Dutta, P. (2020). *A Review of Object Detection Models Based on Convolutional Neural Network BT - Intelligent Computing: Image Processing Based Applications* (J. K. Mandal & S. Banerjee (eds.); pp. 1–16). Springer Singapore. https://doi.org/10.1007/978-981-15-4288-6_1

Tavares, P. da S., Giarolla, A., Chou, S. C., Silva, A. J. de P., & Lyra, A. de A. (2018). Climate

change impact on the potential yield of Arabica coffee in southeast Brazil. *Regional Environmental Change*, *18*(3), 873–883. https://doi.org/10.1007/s10113-017-1236-z

Thompson, S. S., Miller, K. B., Lopez, A. S., & Camu, N. (2012). Cocoa and coffee. *Food Microbiology: Fundamentals and Frontiers*, 881–899.

van Rikxoort, H., Schroth, G., Läderach, P., & Rodríguez-Sánchez, B. (2014). Carbon footprints and carbon stocks reveal climate-friendly coffee production. *Agronomy for Sustainable Development*, *34*(4), 887–897. https://doi.org/10.1007/s13593-014-0223-8

Velásquez, S., Peña, N., Bohórquez, J. C., Gutierrez, N., & Sacks, G. L. (2019). Volatile and sensory characterization of roast coffees–Effects of cherry maturity. *Food Chemistry*, *274*, 137–145.

Verma, U., Rossant, F., Bloch, I., Orensanz, J., & Boisgontier, D. (2014). Shape-based segmentation of tomatoes for agriculture monitoring. *ICPRAM*.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *ArXiv Preprint ArXiv:2207.02696*.

Wu, D., Lv, S., Jiang, M., & Song, H. (2020). Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Computers and Electronics in Agriculture*, *178*, 105742.

Yuan, W. (2023). Accuracy Comparison of YOLOv7 and YOLOv4 Regarding Image Annotation Quality for Apple Flower Bud Classification. In *AgriEngineering* (Vol. 5, Issue 1, pp. 413–424). https://doi.org/10.3390/agriengineering5010027

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.