GUSTAVO PUCCI BOTEGA

VISÃO COMPUTACIONAL APLICADA A ANÁLISE DE FRUTOS DE *C. Arabica*

LAVRAS – MG

2023

GUSTAVO PUCCI BOTEGA

VISÃO COMPUTACIONAL APLICADA A ANÁLISE DE FRUTOS DE *C. Arabica*

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para obtenção do título de Doutor.

Orientadora
Profa. Flávia Maria Avelar Gonçalves

LAVRAS – MG

2023

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Botega, Gustavo Pucci.
Visão computacional aplicada a análise de frutos de Coffea arabica / Gustavo Pucci Botega. - 2023.
86 p. : il.

Orientador(a): Flavia Maria Avelar Gonçalves.

Tese (doutorado) - Universidade Federal de Lavras, 2023.
Bibliografia.

1. Maturação. 2. Redes neurais convolucionais. 3. Inteligência artificial. I. Gonçalves, Flavia Maria Avelar. II. Título.

GUSTAVO PUCCI BOTEGA

VISÃO COMPUTACIONAL APLICADA A ANÁLISE DE FRUTOS DE *C. ARABICA*

COMPUTER VISION APPLIED TO THE ANALYSIS OF *C. ARABICA* FRUITS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para obtenção do título de Doutor.

APROVADA em 16 de março de 2023.

Dra. Flavia Maria Avelar Goncalves	UFLA
Dr. César Elias Botelho	EPAMIG
Dr. Vinícius Quintão Carneiro	UFLA
Dr. Tiago Teruel Rezende	UFLA
Dr. Tiago de Souza Março	UFLA

Orientadora
Profa. Flávia Maria Avelar Gonçalves

LAVRAS – MG

2023

Aos meus pais e irmãos.
A minha namorada
Aos meus familiares
A Bella

DEDICO

AGRADECIMENTOS

Primeiramente agradeço à Deus por sempre se fazer presente, iluminando o meu caminho.

Aos meus pais, Luiz e Andrea, pelo amor, apoio e incentivo em todos os momentos da minha vida.

Tatiane, pelo amor, companheirismo e ser a base de meu sucesso, me apoiando e motivando em qualquer situação.

Aos meus tios, pela torcida e incentivo.

Aos meus irmãos da república Galo Bravo, por toda amizade e companheirismo.

À republica Forasteiras minha segunda família.

À professora Flávia pela orientação, pelos ensinamentos, pela paciência e confiança e por ter sido uma amiga e um guia acima de tudo.

À Universidade Federal de Lavras e ao programa de pós-graduação em Genética e Melhoramento de Plantas, pela oportunidade concedida.

A todos os professores do programa de pós-graduação em Genética e Melhoramento de Plantas, pelos conhecimentos transmitidos.

Aos amigos do grupo Melhoramento de Plantas Perenes, por toda ajuda e amizade.

Aos amigos do GEN pela amizade e convivência.

À todos que contribuíram direta e indiretamente para que este sonho se tornasse realidade.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior –CAPES pela concessão de bolsa de estudos.

Meu muito obrigado!

“*Conheça todas as teorias, domine todas as técnicas, mas, ao tocar uma alma humana, seja apenas outra alma humana.”*

Carl Gustav Jung

RESUMO GERAL

Nos centros de pesquisa de café, diversos caracteres são fenotipados pelos pesquisadores. Alguns deles são determinados diretamente pela fenotipagem dos frutos, como a maturação. A maturação dos frutos é um caráter importante de ser mensurado pois permite o lançamento de cultivares que apresentem diferentes ciclos de maturação, o que é fundamental para os produtores, uma vez que permite escalonar a produção e maximizar a eficiência e lucratividade. Contudo, nos programas de melhoramento a mensuração desse caráter possui diversas dificuldades. Este estudo foi dividido em três capítulos os quais apresentam uma avaliação ampla e profunda de frutos de café e de aspectos associados à seleção e avaliação da maturação em *Coffea arabica*. No capítulo um, frutos de café obtidos de uma plataforma de fenotipagem foram avaliados detalhadamente, examinando suas características morfológicas e de coloração com a utilização de visão computacional, para tanto foi necessário criar um modelo de classificação baseado em redes neurais convolucionais para classificação dos diferentes estágios de maturação. No capítulo dois, a partir do banco de imagens gerados, foi sintetizados imagens para o treinamento de um modelo de visão computacional baseado na arquitetura de redes neurais YOLO, para classificação e detecção de frutos de café em diversos cenários e ambientes. No capítulo três, o objetivo foi estabelecer o tamanho de amostra ideal na avaliação do caráter e verificar os erros associados em adotar cada tamanho, como também demonstrar que o método de agrupamento por meio de *K-means* pode ser uma alternativa para auxiliar os pesquisadores na tomada de decisão acerca dos genótipos constituintes da população de melhoramento. Analisou-se detalhadamente frutos de 21 cultivares, fornecendo valiosas informações para os pesquisadores sobre suas características morfológicas e de coloração, criou-se 36.879 imagens de frutos de café nos diferentes estágios de maturação. A utilização do modelo YOLO permite a avaliação de frutos de café em diferentes cenários e ambientes, reduzindo e facilitando o processo de fenotipagem do caráter. Verificou-se que amostras superiores a 500 ml de frutos demostra ser um excelente tamanho amostral e o uso do técnica de *K-means* para agrupar os dados nos diferentes ciclos de maturação pode ser uma excelente alternativa para os pesquisadores, permitindo uma análise e tomada de decisão precisa e eficiente.

Palavras-chave: Rede neurais convolucionais, YOLO, café, melhoramento genético, fenotipagem

GENERAL ABSTRACT

At coffee research centers, various traits are phenotyped by researchers. Some of these traits are directly determined by fruit phenotyping, such as ripening. Fruit ripening is an important trait to measure, because it allows for cultivars release with different ripening cycles, which is essential for farmers as it allows for the scaling of production and maximization of efficiency and profitability. However, measuring this trait in breeding programs presents several challenges. This study was divided into three chapters that present a comprehensive evaluation of coffee fruit and aspects associated with the selection and evaluation of ripening in *Coffea arabica*. In the first chapter, coffee fruits obtained from a phenotyping platform were thoroughly evaluated, by examining their morphological and color characteristics using computer vision. To achieve it, a classification model based on convolutional neural networks was created to classify the different stages of ripening. In the second chapter, images were synthesized from the generated image dataset to train a computer vision model based on the YOLO neural network architecture for direct classification and detection of coffee fruits in numerous scenarios and environments. In chapter 3, the objective was to establish the ideal sample size for ripening fruit evaluation and to verify the associated errors in adopting each sample size, as well as to demonstrate that the K-means clustering method can be an alternative to assist researchers in making decisions about the constituent genotypes of the breeding population. Detailed analysis was conducted on fruits from 21 cultivars, providing valuable information to researchers about their morphological and color characteristics. A total of 36.879 images of coffee fruits at different ripening stages were created. The use of the YOLO architecture allows for the direct evaluation of coffee fruits in different scenarios and environments, reducing and facilitating the process of phenotyping the trait. It was found that samples larger than 500 ml of fruits demonstrate an excellent sample size, and the use of the K-means technique to group data into different ripening cycles can be an excellent alternative for researchers, allowing for precise and efficient analysis.

Keywords: Convolutional neural network, YOLO, Coffea, Breeding plants, phenotyping

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO GERAL

O café é indiscutivelmente uma das bebidas mais populares do mundo, com um papel fundamental, tanto no âmbito comercial quanto social. O Brasil, em particular, desempenha um papel crucial na produção e exportação de café, sendo o maior produtor e exportador do mundo, além de ser o segundo maior consumidor da bebida. Embora existam mais de 100 espécies do gênero *Coffea*, apenas duas são relevantes comercialmente: *Coffea arabica* (arabica) e *Coffea canephora* (robusta). O *C.arabica*, responde por aproximadamente 59% da produção mundial, enquanto o robusta representa cerca de 41% (CONAB, 2022; DAVIS *et al.*, 2006).

No Brasil, a pesquisa cafeeira é em grande parte realizada por instituições públicas e no âmbito das mesmas, diversas são as características avaliadas. Dentre essas, algumas estão diretamente ou indiretamente relacionada a fenotipagem dos frutos, como maturação, uniformidade, doença e produção de grãos (BERTRAND *et al.*, 2011; CARVALHO, 2008; SILVA *et al.*, 2022; SOUSA *et al.*, 2019). Neste contexto, a avaliação quanto a maturação dos frutos é de extrema importância, pois a planta do cafeeiro apresenta um florescimento desuniforme, resultando no momento da colheita, em frutos nos diferentes estágios de maturação. Essa variação tem influência tanto ambiental quanto genética (CAMAYO *et al.*, 2003; MAJEROWICZ e SÖNDAHL, 2005)

Segundo Pezzopane *et al.* (2003) a maturação dos frutos é determinada da seguinte maneira: após a fecundação é iniciada a formação dos frutos (fase chumbinho), sem crescimento visível. Depois os frutos se expandem rapidamente (fase expansão dos frutos) até atingir seu tamanho máximo. Após a máxima expansão do fruto a cor verde do fruto é acentuada (fase grão verde). A partir dessa fase é iniciada a maturação dos frutos, processo que consiste na degradação da clorofila e a síntese de carotenoides, fazendo com que a cor verde perca gradativamente sua intensidade (fase verde cana), evoluindo até o estádio de amarelo ou vermelho (fase cereja). Posteriormente os frutos começam a secar (fase passa) até estar completamente seco (fase seco).

Algumas estratégias de avaliação da maturação dos frutos têm sido empregada nos centros de pesquisa, tais como, a contagem de uma amostra de frutos da parcela determinando os estádios fenológicos dos frutos, como também a atribuição de notas (COSTA *et al.*, 2013; NOGUEIRA *et al.*, 2005; PETEK *et al.*, 2006; SOUSA *et al.*, 2019). Contudo, não há estudos

comprovando a eficácia dessas estratégias, nem mesmo qual é o tamanho de amostra ideal para se ter uma boa representatividade da parcela. Outro ponto, é que estas avaliações exigem uma grande quantidade de mão de obra e tempo para serem realizadas, e por muitas vezes possuem altos erros associados, por se tratar de julgamentos subjetivos (NOGUEIRA *et al.*, 2005; SANTORO *et al.*, 2019).

Uma opção para avaliação do caráter é o emprego da visão computacional. Ela tem se destacado como uma das técnicas mais eficazes para a fenotipagem, permitindo a quantificação de caracteres complexos tanto em condições experimentais controladas (LI *et al.*, 2018) quanto diretamente no campo (DYRMANN *et al.*, 2016), sendo capaz de avaliar caracteres agronômicos e morfofisiológicos de centenas ou milhares de amostras, em um pequeno intervalo de tempo, sem demandar uma grande quantidade de mão-de-obra. A visão computacional é uma área da ciência que se concentra em permitir que computadores possam interpretar e compreender informações visuais a partir de imagens ou vídeos digitais. Ela utiliza de técnicas de processamento de imagem, aprendizado de máquina e inteligência artificial para analisar, interpretar e compreender o conteúdo visual (PRATT, 2013).

As técnicas de processamento de imagem permitem extrair informações relevantes de imagens. Essas técnicas podem ser usadas para melhorar a qualidade de uma imagem, remover ruídos e artefatos, detectar características específicas, segmentar regiões de interesse e extrair informações quantitativas. Alguns exemplos de técnicas de processamento de imagem incluem, a transformada de Fourier, a convolução, a segmentação por limiarização, a detecção de bordas e as operações morfológicas. Com a evolução da tecnologia, as técnicas de processamento de imagem tornaram-se cada vez mais avançadas e eficientes, permitindo uma ampla variedade de aplicações em diferentes áreas do conhecimento (PRATT, 2013).

Dentro da área de visão computacional a aprendizagem profunda com redes neurais convolucionais, também conhecida como *Convolutional Neural Networks* (CNNs), é uma técnica amplamente aplicada, tendo sido utilizada em diversos estudos, incluindo classificação (KRIZHEVSKY; SUTSKEVER; HINTON, 2017), detecção de objetos (GIRSHICK *et al.*, 2014) e segmentação por instância/semântica (HE *et al.*, 2017; LONG; SHELHAMER; DARRELL, 2015). Essa técnica também tem tido um impacto significativo na fenotipagem de plantas, como na detecção de plantas daninhas (MILIOTO; LOTTES; STACHNISS, 2017), avaliação de doenças (GHOSAL *et al.*, 2018), detecção de frutos (BRESILLA *et al.*, 2019) entre outras aplicações descritas em duas revisões (KAMILARIS; PRENAFETA-BOLDÚ,

2018; LI *et al.*, 2020). As CNNs são compostas por várias camadas de "neurônios" que realizam operações de convolução, ativação, *pooling* e classificação, permitindo que a rede aprenda automaticamente as características importantes de uma imagem durante o treinamento. Uma das principais vantagens das CNNs é sua capacidade de processar grandes quantidades de dados de forma eficiente, tornando possível treinar modelos complexos em conjuntos de dados muito grandes.

A grande quantidade de dados é fundamental para o treinamento de modelos via CNNs, pois esses modelos são altamente dependentes de dados de treinamento para aprender as características importantes das imagens. Quanto mais dados de treinamento uma CNN tem disponível, mais preciso e eficiente será seu modelo. Isso ocorre porque permite que a rede neural aprenda e generalize melhor os padrões presentes nas imagens, o que pode levar a resultados mais confiáveis e precisos (KRIZHEVSKY; SUTSKEVER; HINTON, 2017). Além disso, a grande quantidade de dados também ajuda a evitar o problema de *overfitting*, que ocorre quando o modelo aprende a identificar apenas os dados de treinamento específicos, mas não consegue generalizar para novos dados (ZHANG; ZHANG; JIANG, 2019).

Diversos são os problemas envolvidos na criação de um conjunto de dados para treinamento, os principais são o processo de aquisição e anotação das imagens. A anotação da imagem é um processo de marcação de informações específicas de cada imagem, com o objetivo de rotular os dados para serem usados no treinamento. A anotação de imagem pode incluir a marcação de objetos específicos dentro de uma imagem, como por exemplo demarcar todos os frutos de café dentro da imagem, a segmentação de regiões de interesse, a determinação da classe de uma imagem, entre outras formas.

Uma alternativa para a resolução dos desafios relacionados à anotação e aquisição de imagem é a utilização de imagens sintéticas. Estas imagens podem ser geradas de diversas maneiras e podem ser usadas para simular situações e ambientes específicos, o que permite a criação de conjuntos de dados rotulados para o treinamento dos modelos (TODA *et al.*, 2020; YANG *et al.*, 2021). A geração de imagens sintéticas pode ser mais eficiente e escalável do que a coleta de imagens reais, uma vez que as imagens sintéticas podem ser criadas rapidamente em grande quantidade. Além disso, a utilização de imagens sintéticas permite a criação de conjuntos de dados rotulados para cenários específicos, que podem ser difíceis ou impossíveis de serem capturados por meio de imagens reais.

Diante do exposto, o objetivo desta tese foi solucionar os problemas relacionados a fenotipagem para os estágios de maturação de cafeeiros arabica. A tese constitui de três capítulos, nos quais fornecem informações valiosas quanto a morfologia dos frutos e suas diferenças, classificação dos frutos nos estágios de maturação em diferentes cenários com a utilização da visão computacional. Assim como, verificar qual é o tamanho de ideal da amostra para avaliação do caráter e a melhor maneira de lidar com esse tipo de dado.

O primeiro capítulo desta tese, intitulado " Fenotipagem de frutos de café arabica com uso de visão computacional", aborda a avaliação detalhada de frutos de café, examinando suas características morfológicas e de coloração com a utilização de visão computacional, verificou-se a possibilidade de distinção de cultivares baseado na morfologia de seus frutos e seleção de grãos graúdos por meio da avaliação dos frutos.

O segundo capítulo desta tese, intitulado "Detecção e identificação de frutos de café utilizando imagens sintéticas, YOLO e inferência por fatia" é proposta a utilização de imagens sintéticas para o treinamento de um modelo de detecção de objetos em diversos cenários e ambientes.

O terceiro e último capítulo desta tese, intitulado “Tamanho amostral e análise de dados de maturação”, teve como objetivo estudar os tamanhos de amostra e os erros associados para cada tamanho, na avaliação dos estágios de maturação em cafeeiros arabica. Como também, sugere uma abordagem por meio de *K-means* a fim de selecionar os cafeeiros baseados em nos diferentes ciclos de maturação.

REFERÊNCIAS

BERTRAND, B.; ALPIZAR, E.; LARA, L.; SANTACREO, R.; HIDALGO, M.; QUIJANO, J. M.; MONTAGNON, C.; GEORGET, F.; ETIENNE, H. Performance of Coffea arabica F1 hybrids in agroforestry and full-sun cropping systems in comparison with American pure line cultivars. **Euphytica**, v. 181, n. 2, p. 147–158, 2011. Disponível em: <https://doi.org/10.1007/s10681-011-0372-7>.

BRESILLA, K.; PERULLI, G. D.; BOINI, A.; MORANDI, B.; CORELLI GRAPPADELLI, L.; MANFRINI, L. Single-shot convolution neural networks for real-time fruit detection within the tree. **Frontiers in plant science**, v. 10, p. 611, 2019.

CAMAYO, G. C.; CHAVES, B.; ARCILA, J.; JARAMILLO, A. Desarrollo floral del cafeto y su relación con las condiciones climáticas de Chinchiná Caldas. 2003.

CARVALHO, C. H. S. de. Cultivares de café: origem, características e recomendações. **Brasília: Embrapa Café**, v. 334, 2008.

CONAB. Acomp. safra brasileira de café. **COMPANHIA NACIONAL DE ABASTECIMENTO – CONAB**, v. 9, n. 4- Quarto levantamento, p. 52, 2022.

COSTA, J. C.; CARVALHO, C. H. S.; MATIELLO, J. B.; ALMEIDA, S. R.; CARVALHO, S. P.; BALIZA, D. P. Comportamento agronômico de progênies e cultivares de cafeeiro com resistência específica à ferrugem. **Coffee Science-ISSN 1984-3909**, v. 8, n. 2, p. 183–191, 2013.

DAVIS, A. P.; GOVAERTS, R.; BRIDSON, D. M.; STOFFELEN, P. An annotated taxonomic conspectus of the genus Coffea (Rubiaceae). **Botanical Journal of the Linnean Society**, v. 152, n. 4, p. 465–512, 2006.

DYRMANN, M.; MORTENSEN, A. K.; MIDTIBY, H. S.; JØRGENSEN, R. N. Pixel-wise classification of weeds and crops in images by using a fully convolutional neural network. In: Proceedings of the International Conference on Agricultural Engineering, Aarhus, Denmark, 2016, [...]. 2016. p. 26–29.

GHOSAL, S.; BLYSTONE, D.; SINGH, A. K.; GANAPATHYSUBRAMANIAN, B.; SINGH, A.; SARKAR, S. An explainable deep machine vision framework for plant stress phenotyping. **Proceedings of the National Academy of Sciences**, v. 115, n. 18, p. 4613–4618, 2018.

GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; MALIK, J. Rich feature hierarchies for

accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, [...]. 2014. p. 580–587.

HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, 2017, [...]. 2017. p. 2961–2969.

KAMILARIS, A.; PRENAFETA-BOLDÚ, F. X. Deep learning in agriculture: A survey. **Computers and electronics in agriculture**, v. 147, p. 70–90, 2018.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Communications of the ACM**, v. 60, n. 6, p. 84–90, 2017.

LI, D.; CAO, Y.; TANG, X.; YAN, S.; CAI, X. Leaf segmentation on dense plant point clouds with facet region growing. **Sensors**, v. 18, n. 11, p. 3625, 2018.

LI, Z.; GUO, R.; LI, M.; CHEN, Y.; LI, G. A review of computer vision technologies for plant phenotyping. **Computers and Electronics in Agriculture**, v. 176, p. 105672, 2020.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, [...]. 2015. p. 3431–3440.

MAJEROWICZ, N.; SÖNDAHL, M. R. Induction and differentiation of reproductive buds in Coffea arabica L. **Brazilian Journal of Plant Physiology**, v. 17, p. 247–254, 2005.

MILIOTO, A.; LOTTES, P.; STACHNISS, C. REAL-TIME BLOB-WISE SUGAR BEETS VS WEEDS CLASSIFICATION FOR MONITORING FIELDS USING CONVOLUTIONAL NEURAL NETWORKS. **ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences**, v. 4, 2017.

NOGUEIRA, Â. M.; CARVALHO, S. P. de; BARTHOLO, G. F.; MENDES, A. N. G. Fruit ripening evaluations of Catuai Amarelo and Vermelho (Coffea arabica L.) lineages of coffee cultivar, planted isolated and in combinations. **Ciência e Agrotecnologia**, v. 29, p. 18–26, 2005b.

PETEK, M. R.; SERA, T.; SERA, G. H.; FONSECA, I. C. de B.; ITO, D. S. Seleção de progênies de Coffea arabica com resistência simultânea à mancha aureolada e à ferrugem alaranjada. **Bragantia**, v. 65, p. 65–73, 2006.

PEZZOPANE, J. R. M.; PEDRO JÚNIOR, M. J.; THOMAZIELLO, R. A.; CAMARGO, M.

B. P. de. Escala para avaliação de estádios fenológicos do cafeeiro arábica. **Bragantia**, v. 62, p. 499–505, 2003.

PRATT, W. K. **Introduction to digital image processing**. [s.l.] CRC press, 2013.

SANTORO, P. H.; FANTIN, D.; MACHADO, A. H. R.; MENEZES, K. C. de; OLIVEIRA, J. T. de; MARIOTO, R. F.; SILVA, A. C. R. da. Incidência de Hypothenemus hampei e maturação de frutos de café arábica em sistema agroflorestal. 2019.

SILVA, V. A.; ABRAHÃO, J. C. de R.; REIS, A. M.; SANTOS, M. de O.; PEREIRA, A. A.; BOTELHO, C. E.; CARVALHO, G. R.; CASTRO, E. M. de; BARBOSA, J. P. R. A. D.; BOTEGA, G. P. Strategy for Selection of Drought-Tolerant Arabica Coffee Genotypes in Brazil. **Agronomy**, v. 12, n. 9, p. 2167, 2022.

SOUSA, T. V.; CAIXETA, E. T.; ALKIMIM, E. R.; OLIVEIRA, A. C. B.; PEREIRA, A. A.; SAKIYAMA, N. S.; ZAMBOLIM, L.; RESENDE, M. D. V. Early selection enabled by the implementation of genomic selection in Coffea arabica breeding. **Frontiers in Plant Science**, v. 9, p. 1934, 2019.

TODA, Y.; OKURA, F.; ITO, J.; OKADA, S.; KINOSHITA, T.; TSUJI, H.; SAISHO, D. Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. **Communications biology**, v. 3, n. 1, p. 1–12, 2020.

YANG, S.; ZHENG, L.; HE, P.; WU, T.; SUN, S.; WANG, M. High-throughput soybean seeds phenotyping with convolutional neural networks and transfer learning. **Plant Methods**, v. 17, n. 1, p. 50, 2021.

ZHANG, H.; ZHANG, L.; JIANG, Y. Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems. In: 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), 2019, [...]. 2019. p. 1–6.

SEGUNDA PARTE – ARTIGOS

ARTIGO 1 – THOROUGH EVALUATION OF FRUIT OF ARABICA COFFEE CULTIVARS USING COMPUTER VISION

Artigo redigido conforme a normas da revista Frontiers in Plant Science

(VERSÃO PRELIMINAR)

ABSTRACT

In research, many traits are evaluated that are directly or indirectly related to fruit phenotyping. The aim of this study was to thoroughly examine the morphological, color, and ripening characteristics of 21 Arabica coffee cultivars. To do so, computer vision was used to analyze images of coffee fruit acquired by a specially developed phenotyping platform. Individual fruit segmentation was accurately obtained through the platform and the pipeline that was developed, allowing for creation of 36,876 individual fruit images, classified according to their ripening stage. For that purpose, a classification model was developed using the ResNet101 architecture, and the classification from the model was compared with the classification performed by three different evaluators, revealing differences in judgments for each stage. The possibility of distinguishing cultivars based on differences of their fruit was also examined, as well as indirect selection for bean size based on fruit morphological traits. This study is the first to thoroughly evaluate Arabica coffee fruit, providing valuable information for researchers and serving as a reference for future research in the area.

Keywords: coffee maturation, convolutional neural network, deep learning, bean classification, visual evaluation, elliptic Fourier descriptors, *C. arabica*.

1 INTRODUCTION

In coffee research centers, numerous traits are phenotyped. Some of them are directly determined by phenotyping fruit, such as ripening, uniformity, and diseases in fruit (CARVALHO, 2008). Others use fruit in indirect determinations, such as for coffee bean yield, which in some studies is calculated based on the volume or weight of the fruit (BERTRAND et al., 2011; SILVA et al., 2022; SOUSA et al., 2019). In addition to these traits, others have potential for indirect measurement, such as bean size; several researchers report that genotypes with large fruit produce large beans; however, corroborative studies have yet to be performed.

Computer vision has emerged as one of the most efficient phenotyping techniques. It allows for the quantification of complex traits under both controlled (LI et al., 2018) and field (DYRMANN et al., 2016) conditions. It stands out as a fast, efficient, and non-destructive process, using advanced image processing techniques and machine learning algorithms to analyze and understand visual information. Computer vision is a constantly evolving field, with

new algorithms being developed every year, allowing for increasingly advanced solutions for a wide range of applications (LI et al., 2020).

Deep learning is widely applied in the field of computer vision, having been used in various studies, including classification (KRIZHEVSKY; SUTSKEVER; HINTON, 2017), object detection (GIRSHICK et al., 2014), and instance/semantic segmentation (HE et al., 2017; LONG; SHELHAMER; DARRELL, 2015). This technique has also had a significant impact on plant phenotyping, such as in weed detection (MILIOTO; LOTTES; STACHNISS, 2017), disease assessment (GHOSAL et al., 2018), and fruit detection (BRESILLA et al., 2019), among other applications described in two reviews (KAMILARIS; PRENAFETA-BOLDÚ, 2018 and LI et al., 2020).

To train deep learning models, the number of images is fundamental. Images are the "fuel", and without an adequate image dataset, it is not possible to evaluate and improve the accuracy of the models. Moreover, a high-quality image dataset is essential to perform experiments and analyses in different contexts and applications, enabling the development of advanced solutions in the fields of computer vision and artificial intelligence (PARIKH; ZITNICK, 2010).

Although advances have been made in computer vision, few studies have used this technique in the phenotyping of *C. arabica* traits. For example, Bazame et al. (2022), Ramos et al. (2017), and Ramos et al. (2018) used computer vision to identify and classify fruit directly on the plants. Others developed a system for classifying vegetative structures of coffee branches based on 2D and 3D features obtained from field videos (AVENDANO; RAMOS; PRIETO, 2017). Additionally, other authors implemented a computer vision model to detect and classify coffee fruit and map the maturity stage during harvest by harvesters (BAZAME et al., 2021).

This study is the first to comprehensively evaluate Arabica coffee fruit, examining its morphological, color, and ripening characteristics using computer vision. It showed inconsistencies in evaluation of the ripening stage by evaluators, and computer vision could be an alternative to mitigate this problem. Various maturity stages were represented in the 36,876 fruit images made. We examined if morphological traits are correlated with bean size and if this allowed indirect selection, as well as if Arabica cultivars could be distinguished based on the differences in their fruit.

2 MATERIALS AND METHODS

2.1 Experimental description

The *C. arabica* fruit used was from an experiment with 21 cultivars: Catiguá MG-1, Catiguá MG-2, Catiguá MG-3, Oeiras, Paraíso MG H 419-1, Araponga, Pau Brasil, Acauã Novo, Siriema, Clone 312, Saíra II, IPR 100, Catuaí Vermelho IAC 62, IPR 102, Catuaí Amarelo IAC 99, IPR 103, Rubi MG 1192, Topázio MG 1190, Travessia, Mundo Novo IAC 379-19, and Guará. The experiment was conducted at the Universidade Federal de Lavras, Lavras, MG, Brazil in a randomized complete block design, with three replicates, eight plants per plot, and plant spacing of 3.6 × 0.7 m. With the exception of Clone 312, additional information on the cultivars can be found in Carvalho (2008), Sera et al. (2017a), and Sera et al. (2017b).

2.2 Fruit sample acquisition

At the time of harvest, which occurred in 2021, one liter samples of fruit were collected at random from each plot. The samples were divided into two 500-milliliter sub-samples. Each cultivar was represented by six 500-milliliter samples, except for the Araponga cultivar, which had only four samples due to the absence of production in one of the replicates. Thus, 124 samples were evaluated.

2.3 Image acquisition

The images were captured using a phenotyping platform constructed as a "box" measuring 80 cm width by 80 cm length by 60 cm height made of wood and lighted by four 18W fluorescent lamps with a color temperature of 6500 K. The lamps were positioned at the top of the “box” in a square arrangement (Figure 1). At the center top, an opening was made to insert a camera. The platform is an adaptation of Mendoza and Aguilera (2004).

Figure 1. Phenotyping platform

To highlight the fruit, white ethylene-vinyl acetate (EVA) with a rectangular shape was used, while light blue EVA was used as a background for the platform. Additional information such as treatment, replication, sample number, sample size, and date of analysis was also included. Printed numbers were used to obtain this information, and a 10-cent coin was placed beside the samples to serve as a real reference. The fruit was uniformly arranged to obtain the images (Figure 2).

Figure 2. Image obtained from the phenotyping platform.

The images were taken with a Canon EOS 60D camera with the following settings: shutter speed of 1/13, no flash, aperture of F10, aspect ratio of 3:2, ISO 200, with a resolution of 5184 × 3456 pixels, and the images were stored in JPG format. The camera was connected to a laptop computer and controlled by it.

2.4 Image processing from the phenotyping platform

To segment the fruit, the OpenCV (BRADSKI; KAEHLER, 2008) and scikit-image (VAN DER WALT et al., 2014) libraries were used. To analyze the images obtained from the

phenotyping platform, first, the red channel of the RGB color space was extracted. Then, a threshold of 180 pixels was applied to segment the fruit and information.

This information was extracted using the EasyOCR library (https://github.com/JaidedAI/EasyOCR) which enables high-precision optical character recognition (OCR) using a deep learning framework, ensuring better recognition of letters and numbers.

Segmentation was performed using a series of image processing techniques. First, a median blur with a kernel of five was applied to the image to smooth the edges and reduce noise. Then, two morphological operations of erosion and dilation were performed with a kernel of 21 and two iterations. These operations helped remove unwanted pixels. This process allowed a segmented fruit mask to be obtained with high quality and precision. Additionally, other techniques such as color-based segmentation were also used to help refine the segmentation.

For efficient segmentation of coffee fruit that was touching, the following steps were performed: the exact Euclidean distance transformation of the fruit mask was obtained; next, using a minimum distance of 50 pixels, the image peaks were obtained as coordinates; and then the watershed transformation was applied to obtain the segmentation of each piece of fruit individually (MEYER, 1994). In post-processing, contours with less than 200 pixels were eliminated, as these pixels were considered noise in the image.

The combination of the exact Euclidean distance and the watershed algorithm is a robust technique for segmenting objects that are touching. This approach uses the geometric features of the objects to perform segmentation, resulting in more accurate and high-quality segmentation.

2.5 Ripeness evaluation by evaluators

In evaluation of fruit ripeness, one 500-milliliter sample from each plot was classified by three different evaluators based on their visual perceptions as being in the unripe, ripe, or overripe stage. After classification, the fruit was placed on a white cardboard and an image was captured. These images also contained important information, such as treatment, replication, sample number (evaluator), sample size, and evaluation date (Figure 3).

Figure 3. Fruit sample separated by one evaluator for the ripening stages.

To accurately count the fruit at each ripening stage, we used an application specifically designed for this task. The application allows counting through sequential mouse clicks and has additional functions, such as the option to reset the count and enter large numbers (Figure 4). After counting how many pieces of fruit are in each image, the information is stored, and at the end of the process, a spreadsheet is generated with all the information acquired. This procedure was adopted due to the large number of errors found in fruit counting when performed by evaluators, as previously found in undisclosed studies.

Figure 4. Counting a sample of fruit using the specially developed application.

2.6 Morphological traits

One of the main applications of fruit segmentation is morphological measurement for comparison and analysis among cultivars. Morphological variables such as length, width, area, perimeter, and eccentricity were calculated for each piece of fruit using the measure.regionprops module of the scikit-image library (VAN DER WALT *et al.*, 2014).

Different ripening stages have different probability distributions; thus, for comparative purposes, the variables were normalized for each ripening stage. Then, in order to analyze the morphological and color characteristics of mature fruit from different cultivars, principal component analysis was performed considering only that fruit and the morphological variables that were measured, along with the medians of each channel in the RGB color space.

2.7 Elliptic Fourier descriptors

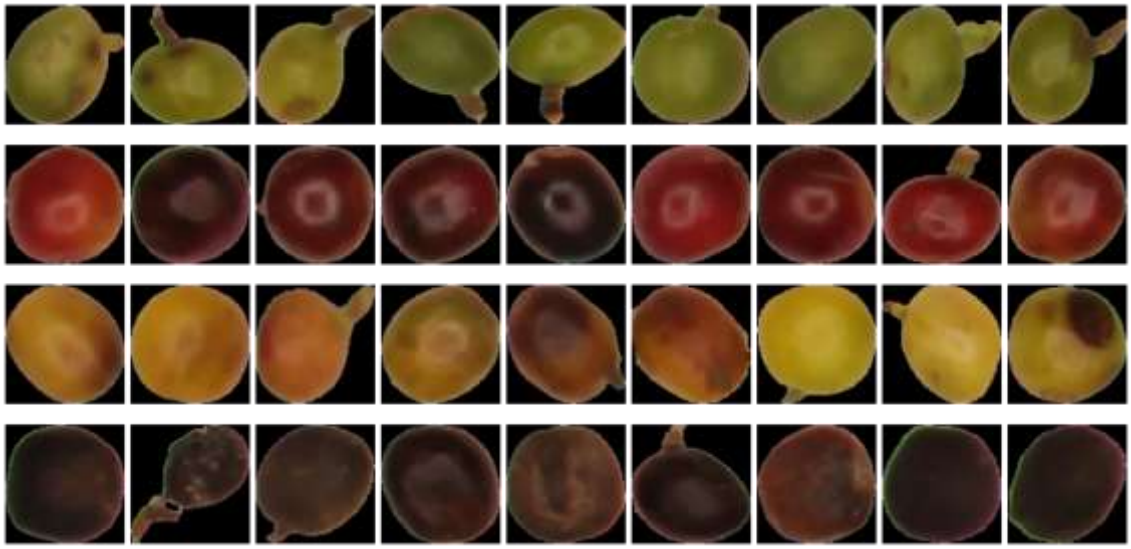
To analyze the elliptic Fourier descriptors (EFD) (KUHL; GIARDINA, 1982), fruit from the three replicates and two samples of each cultivar were considered. The EFD has been used to quantify the shape of fruit (MAEDA; AKAGI; TAO, 2018; OSAKO et al., 2020) and seeds (TODA et al., 2020; WILLIAMS; MUNKVOLD; SORRELLS, 2013). For the analysis, the segmented fruit images were converted into a binary mask where the background pixel was set to zero, and the fruit area pixel was set to one. Then, the contours were detected using the find_contours function of the OpenCV library (BRADSKI; KAEHLER, 2008). The detected contours were converted into EFD coefficients using the elliptic_fourier_descriptors function of the pyefd library (https://github.com/hbldh/pyefd) using a harmonic number of 20, considered sufficient to describe the morphological shape variation (GENTALLAN et al., 2019; SAYINCI et al., 2015; YOSHIOKA et al., 2004), and the coefficients were normalized to be invariant in size and orientation. The matrix was scaled, converting it from 4 × 20 to 80. In

principal component analysis, 77 of the 80 variables of the EFD coefficients were used, as the first three showed no variation due to normalization.

2.8 Classification model

To train the classification model for ripening stages, the fruit from 10 images was segmented, and each segmented piece of fruit was saved as an individual image with a black background. In all, 2886 images were generated; an example of these images is shown in Figure 5. These images were labeled as unripe, red ripe, yellow ripe, and overripe based on the consensus of three evaluators. For the classes, 879 images of unripe, 696 of red ripe, 645 of yellow ripe, and 666 of overripe fruit were obtained.

Figure 5. Examples of fruit used to train the classification model.



To classify the fruit, deep learning models were used. Several architectures were trained for 50 epochs and batch size of 32. The Adagrad optimization algorithm was used to optimize the parameters of the model (DUCHI; HAZAN; SINGER, 2011). The best architecture was selected based on the accuracy obtained in the test images, and the ResNet101 architecture (HE et al., 2016) obtained the best estimate. The images were resized to 120 × 120 pixels before training and divided into 80% for training and 20% for testing. The process was carried out using the Keras library in Python 3.6 and Tensorflow (ABADI et al., 2016).

The ResNet101 architecture is a variation of the ResNet (Residual Network) architecture, which has 101 convolutional layers. The ResNet architecture is known to be able to handle the vanishing and exploding gradient problems, which are common problems in deep neural

networks. It does this through the use of residual connections, which allow the gradient to flow directly to the deeper layers of the network (HE et al., 2016).

2.9 Statistical analysis

The statistical analysis included analysis of variance (ANOVA) when necessary, taking the experimental design into account. The significance of factors was assessed by *p* values. Heritability was calculated using the formula: $H^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$, where σ_g^2 is the genotypic variance and σ_e^2 is the residual variance. Additionally, the residual coefficient of variation was calculated using the following formula: $CVe(\%) = \frac{\sigma_r}{\underline{X}} * 100$, where σ_r is the residual standard deviation and $\underline{X}$ is the mean of the variable being analyzed.

2.10 Bean size classification

The fruit samples were dried on a coffee drying patio until reaching a moisture content of about 11%. They were then processed, and the beans were classified according to size. Flat beans were classified based on the percentage retained on circular sieves (18, 17, 16, 15, 14, and 13), while peaberries beans were classified based on oblong sieves (13, 12, 11, 10, and 9).

2.11 Color palette

In order to elucidate the process of color change in fruit during maturation for the cultivars studied, color palettes were created for each cultivar using the following procedures: (1) Fruit was selected from all samples for each cultivar. (2) For each piece of fruit, the median of the RGB color space was calculated. (3) Pieces of fruit were divided into different maturity stages, and for each stage, 500 pieces of fruit were randomly sampled. (4) For each stage, fruit was sorted in the R, G, and B color spaces, respectively. (5) Finally, the three sorting results were combined, and then a median blur was applied to smooth the images.

3 RESULTS

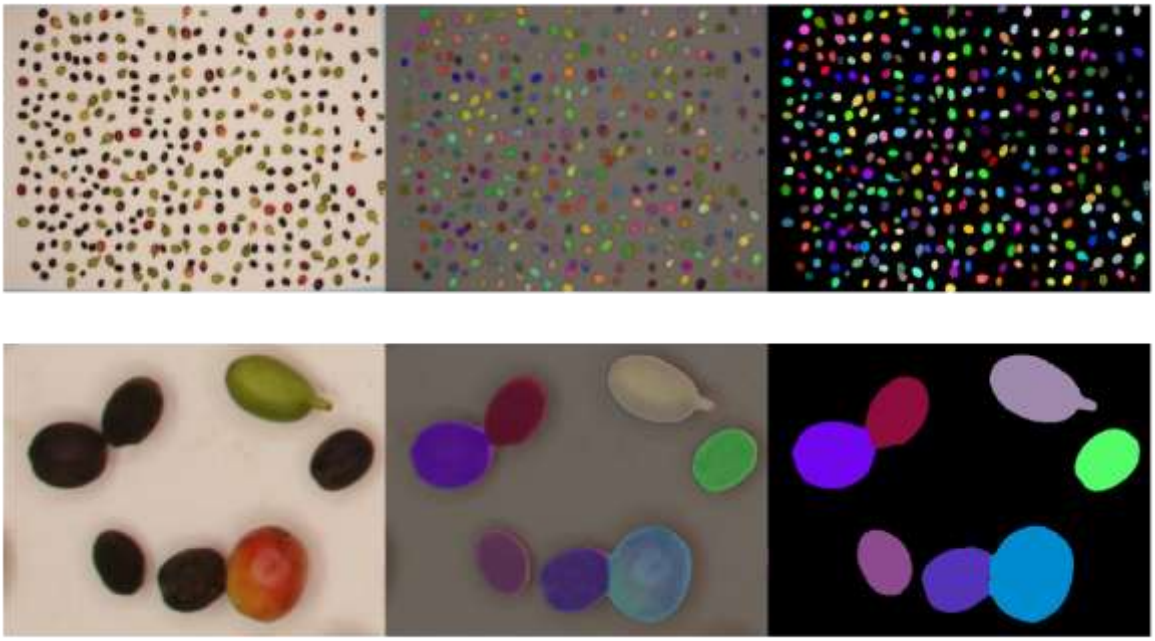
3.1 Fruit segmentation

The fruit segmentation pipeline proved to be efficient (Figure 6). Comparative analysis between the count performed by the specially developed application and the count obtained by the fruit segmented from each image revealed no significant differences (*p* value = 0.99), through analysis of variance (ANOVA).

Figure 6. Fruit segmentation by the specially developed pipeline: (A) segmented fruit, each color represents a piece of fruit, and (B) demonstration of the fruit in contact and its segmentation.

(A)

(B)



It can be seen that there was good separation between the fruit and the background of the image, which indicates that the segmentation was quite successful. During image acquisition, an attempt was made to spread the fruit out uniformly without any contact between two pieces of fruit, but even so, most of the images had fruit that was in contact. The watershed algorithm allowed the fruit to be accurately segmented when in contact, as shown in Figure 6B.

3.2 Classification model

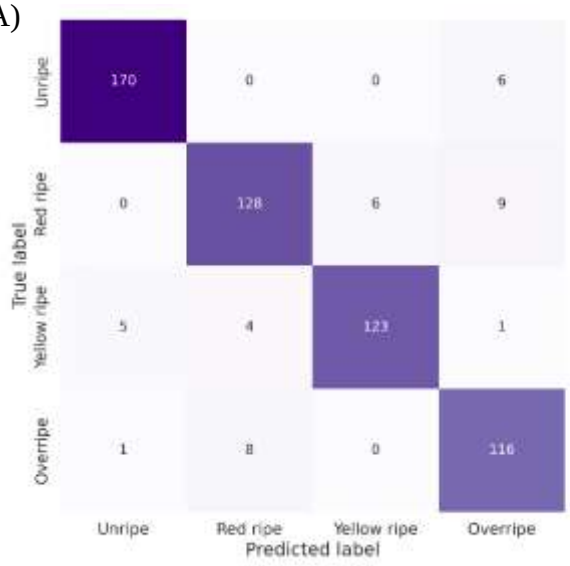
The model showed an accuracy of 98% in the training images and 93% in the test images. Figure 7A shows the confusion matrix for the test images. The confusion matrix is crucial for evaluating the efficiency of the model, because it allows precise information to be obtained regarding its performance and the number of correctly classified and misclassified images to be checked, providing an overall view of the precision and accuracy of the model. Moreover, it is useful for identifying weak points in the model and can be used to improve the model.

Analysis of the confusion matrix shows that the model had greater difficulty in classifying red ripe fruit (90% accuracy), followed by yellow ripe (92%), overripe (93%) and unripe (97%) fruit. Although the images were evaluated through consensus among three evaluators, errors may still have occurred during the annotations by assigning fruit to the wrong class.

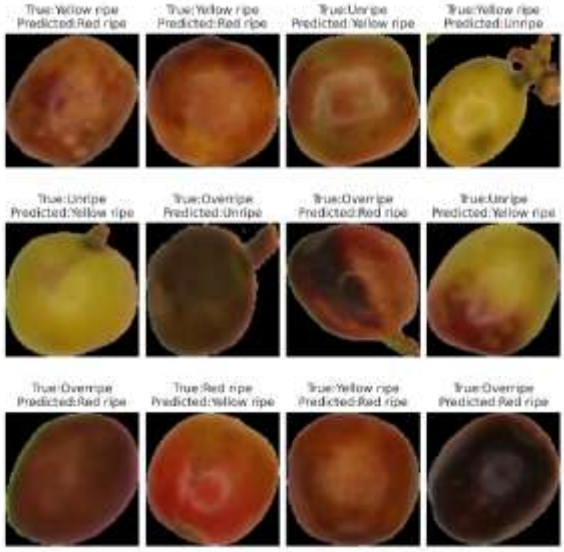
Figure 7B illustrates some incorrect classifications made by the model. It can be observed that some yellow ripe fruit that has a more brownish color was classified as red ripe fruit. This occurs due to the tendency of yellow ripe fruit to darken when it matures, which may have caused a bias in the model. Similarly, there is difficulty in classifying red ripe fruit in advanced stages of ripening, as it can be confused with overripe fruit.

Figure 7. Analysis of the classification model: (A) confusion matrix with the number of pieces of fruit detected for each class and (B) demonstration of misclassified fruit.

(A)



(B)



In the training of the classification model, the nearly ripe stage of the coffee fruit maturation cycle was not included as a class. This stage occurs when unripe fruit is becoming ripe (MORAIS et al., 2008). For genotypes of red ripe fruit, this stage can be confused with yellow ripe fruit, due to the similarity between them (Figure 8). Therefore, for this study, fruit in the nearly ripe stage may have been classified as in the yellow ripe stage.

The model was used to classify all segmented fruit that was obtained from the images captured by the phenotyping platform. As a result, 36,879 fruit images were obtained for different ripening stages, including 9,781 unripe, 11,653 red ripe, 5,072 yellow ripe, and 10,373 overripe pieces of fruit. The images vary in size, with widths ranging from 45 to 206 pixels and heights ranging from 44 to 202 pixels.

Figure 8. Fruit in the “yellow ripe” and “nearly red ripe” stage.

It should be noted that the model was trained on a dataset in which all images were acquired under specific lighting and acquisition conditions, so it may not be suitable for identifying data with other sampling distributions. However, the objective of this study is not to propose a general classification model, but rather a model for specific conditions.

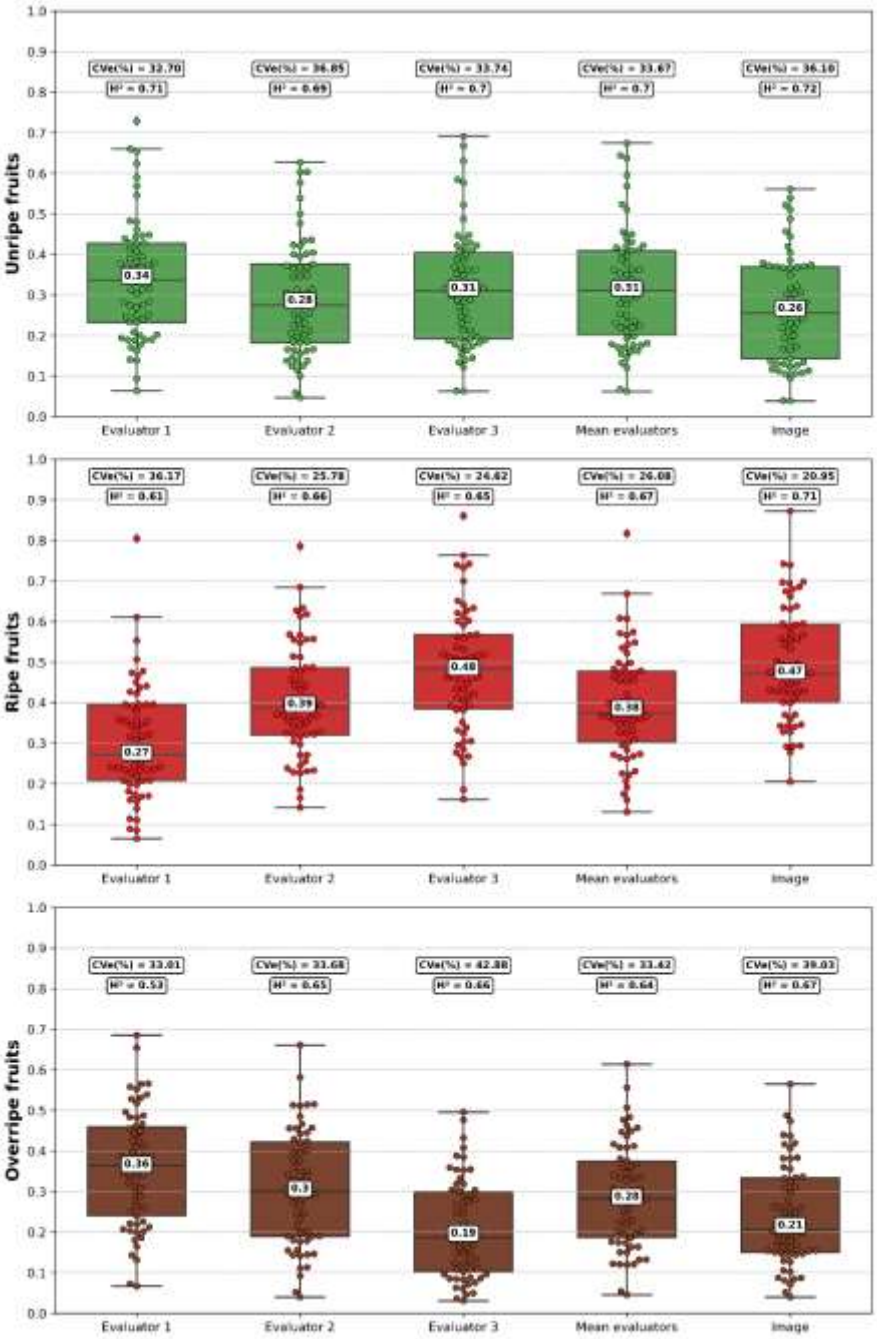
The classification of ripening stages from the model was compared with that from human evaluators, and significant differences in evaluations were found (p value < 0.01) for all ripening stages, as confirmed by ANOVA. Figure 9A shows the intervals for each ripening stage and evaluator, as well as the heritabilities and residual coefficients of variation.

The analysis through images resulted in higher heritabilities for all the ripening stages evaluated. The results for the residual coefficient of variation were similar for unripe fruit, with the lowest coefficient obtained by Evaluator 1 (32.70%). For ripe fruit, image analysis had the lowest coefficient (20.95%), while for overripe fruit, Evaluator 1 obtained the lowest coefficient (33.01%). The median evaluations for unripe fruit were similar, suggesting that evaluators tend to classify that fruit in a standardized way. However, that is not the case for the ripe and overripe stages, and it can be seen that Evaluator 1 tends to classify significantly more fruit as overripe than ripe, compared to Evaluators 2 and 3.

In Figure 9B, the differences in one of the evaluations are clearly seen. The evaluation of a sample by the three evaluators is shown in this case, highlighting the divergences that were suggested by ANOVA. These results indicate that there is a lack of consistency and agreement in the evaluations performed by different evaluators, which can affect the accuracy of the evaluation of ripening stages for the different genotypes.

Figure 9. Fruit classification for the ripening stages among different evaluators: (A) box plot showing the ripening stage classification performed by evaluators (1, 2, 3, and mean) and by the classification model (Image) and (B) classification of fruit from a sample performed by the three evaluators.

(A)



(B)

At research centers for evaluation of the different stages of maturation, it is common to use several evaluators to assess the samples, and each sample is evaluated only once by a single evaluator. Thus, there is not one sole evaluator responsible for evaluating the entire experiment, nor is more than one evaluation per sample performed; that results in a mix of evaluators. To simulate this scenario, 1000 simulations were performed with the data from the three evaluators, so that each evaluator evaluated at random 1/3 of the samples. Heritabilities and residual variation coefficients were recorded for each simulation, and then the 95% confidence interval of the estimate was calculated.

In the simulations for analysis of unripe fruit, an average CVe of 35% (32 - 38) and an average heritability of 0.42 (0.35 - 0.49) were obtained. For ripe fruit, the average CVe was 34% (28 - 40) and the average heritability was 0.28 (0.08 - 0.49). In overripe fruit, the average CVe was 42% (35 - 49) and the average heritability was 0.26 (0.07 - 0.46). Additionally, it was found that there was a change in the ranking of the top four genotypes with the highest value in the different maturity stages evaluated. These results show that the current way of evaluating the trait is not adequate, as there is no standardization among evaluations; and that can interfere in the selection of the genotypes evaluated.

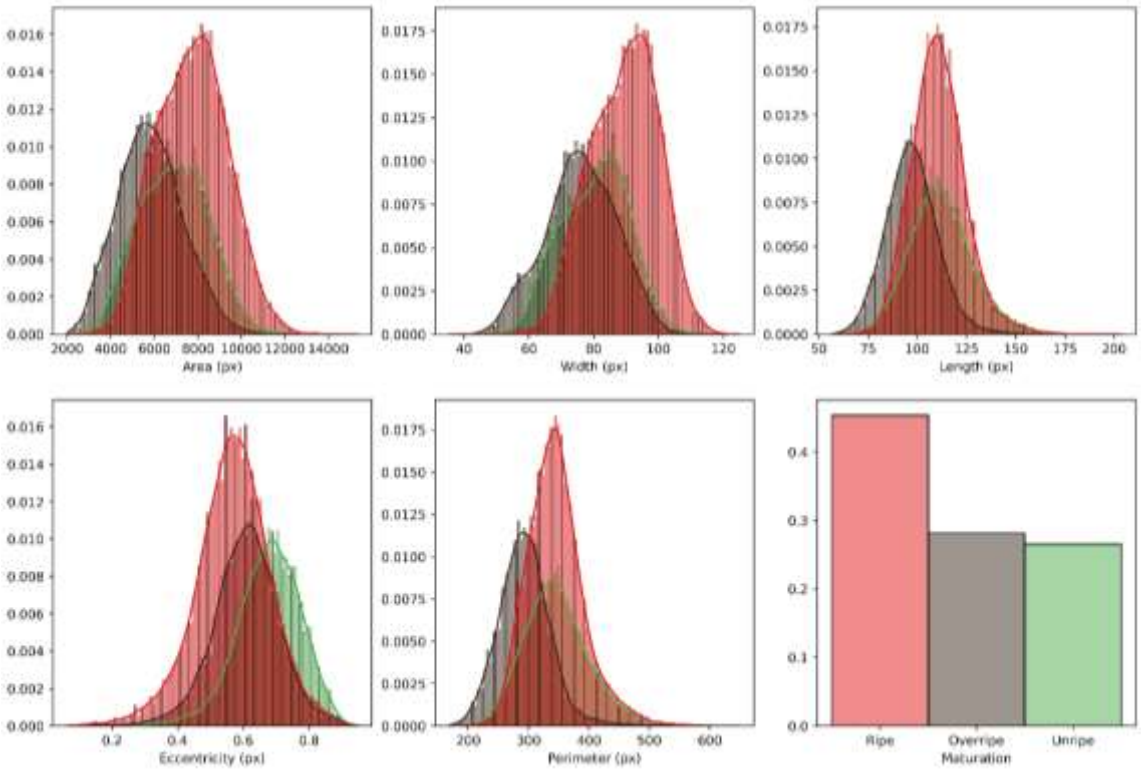
3.3 Morphological traits

It was found that the distributions of morphological traits for the different maturation stages in the samples are distinct from each other (Figure 10). For comparative analysis among cultivars, it is necessary that the stages of maturation have the same distribution to ensure valid comparison without the effect of the maturation cycle. Therefore, for all the morphological traits studied, the maturation stages were standardized separately.

The results indicate that ripe fruit has a significantly larger area and width compared to those of the other maturation stages. Generally, overripe fruit has smaller magnitudes for the

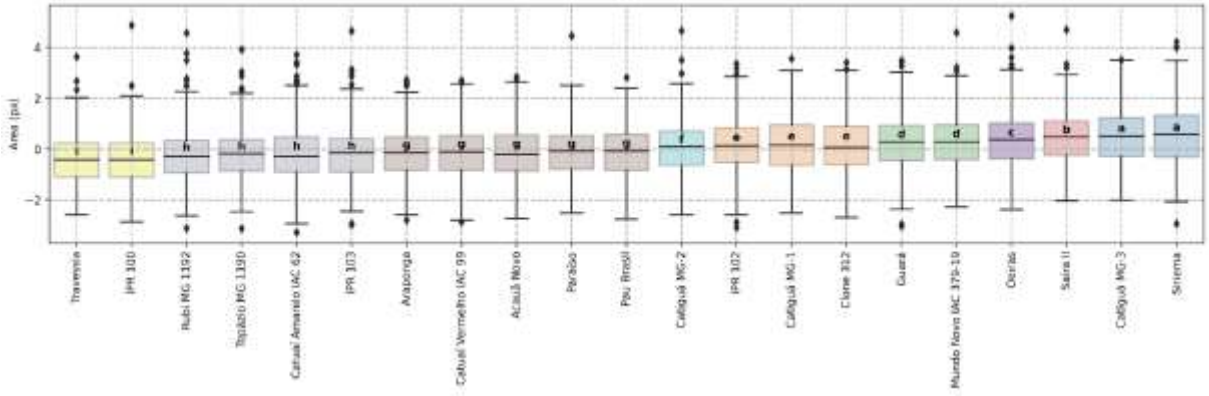
morphological traits evaluated, except for eccentricity. In contrast, unripe fruit had a larger magnitude only for the eccentricity variable.

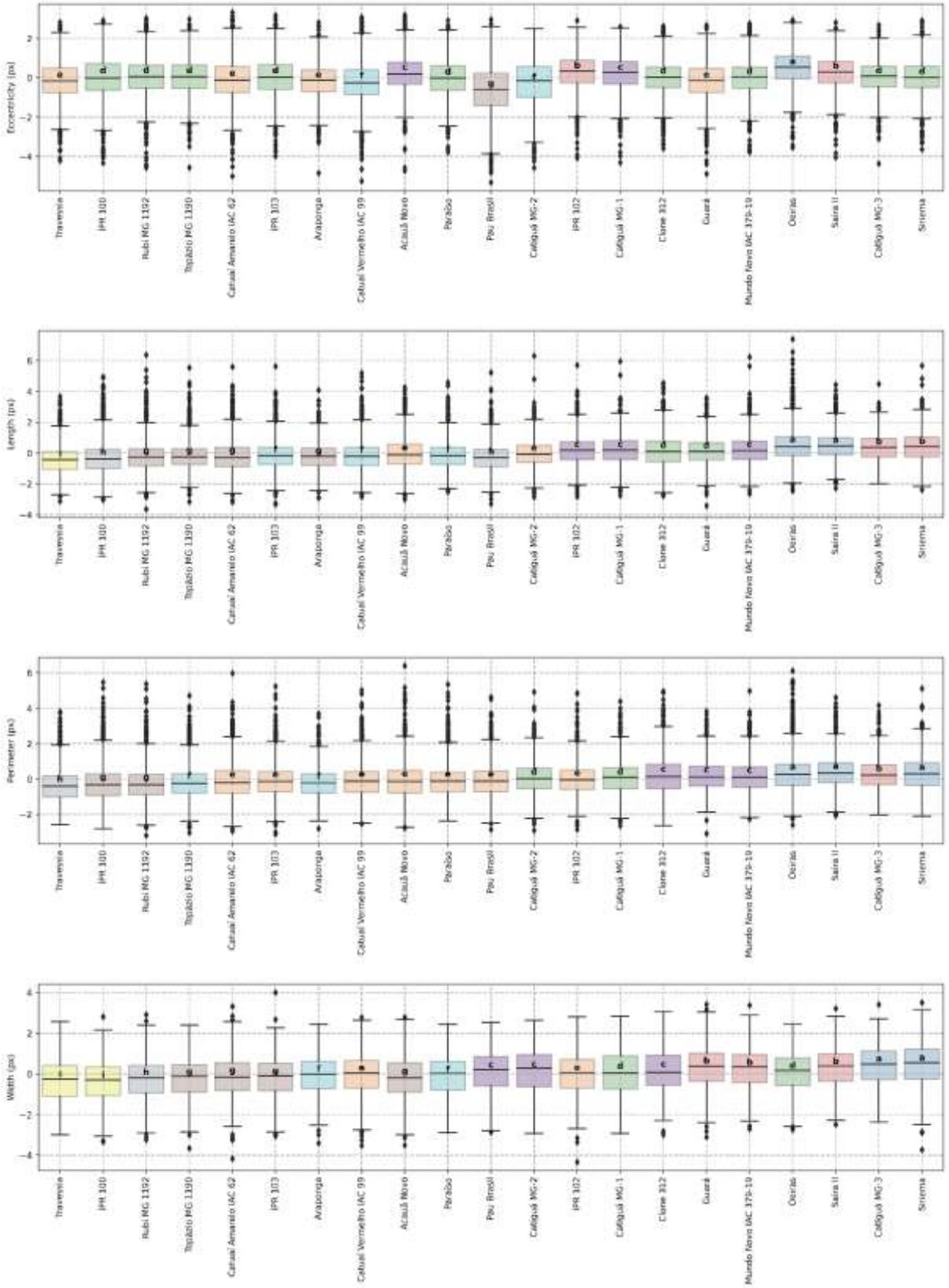
Figure 10. Histogram for the different morphological traits estimated.



Morphological traits were measured for each piece of fruit. For all traits, significant differences (p value < 0.01) were found among cultivars by ANOVA. Figure 11 shows the boxplot for all the variables measured, as well as the comparison made by the Scott-Knott grouping test, in order to express the statistical differences among cultivars.

Figure 11. Analysis of the variation among the fruit for the cultivars studied.





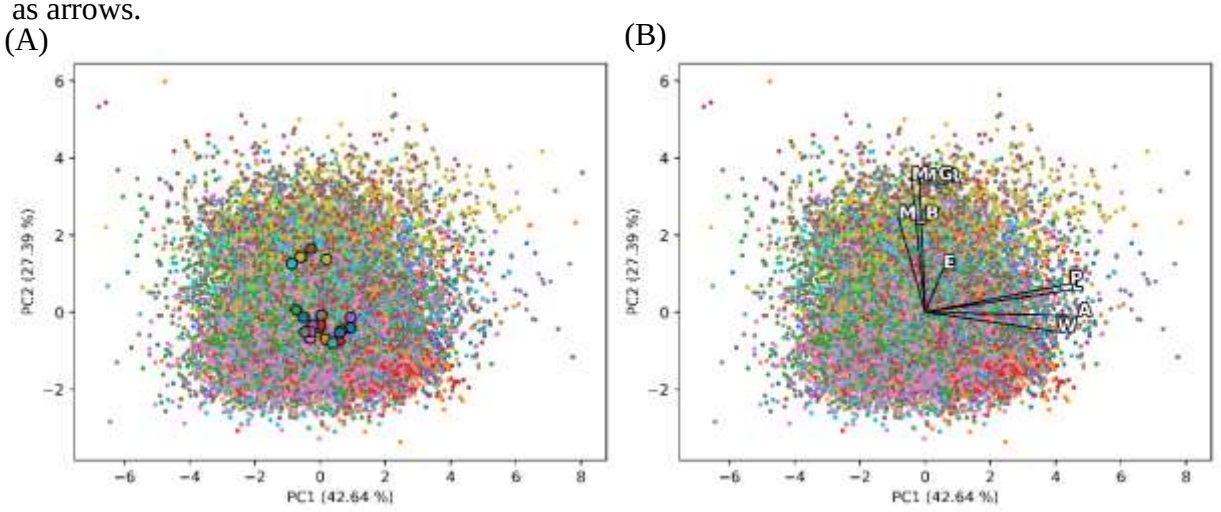
To obtain a better understanding of the fruit and also to check the possibility of identifying the cultivars studied through the morphological and color variables that were measured, multivariate analysis was performed. Figure 12A shows the results of principal

component analysis (PCA) using five morphological variables (length, width, area, perimeter, and eccentricity) and the median of each channel of the RGB color space for the ripe fruit.

The first two principal components (PC) explained 70% of all variation. It was also found that there are two groups of cultivars that can be distinguished by the variables studied: one group containing five cultivars that have ripe fruit with yellow color and the other group with cultivars that have ripe fruit with red color. In Figure 12A, we can observe these two groups; within one of the groups, we can observe 4 highlighted circles, while the fifth is overlapped. However, within the two groups formed, a well-defined separation of the cultivars was not obtained using the first two PCs.

The area variable is the one that most affected PC1, and the median of the green color space affected PC2. There were high magnitude correlations between fruit length and perimeter (0.93), area and diameter (0.92), median of the green color space and median of the red color space (0.88), and area and length (0.85).

Figure 12. Analysis of estimated morphological traits: **(A)** Principal component analysis (PCA) with morphological and color variables of coffee fruit. Each point represents a piece of fruit. The mean values of principal component 1 (PC1) and principal component 2 (PC2) for each cultivar are plotted as a large circle. **(B)** The eigenvectors of each morphological trait are drawn as arrows.



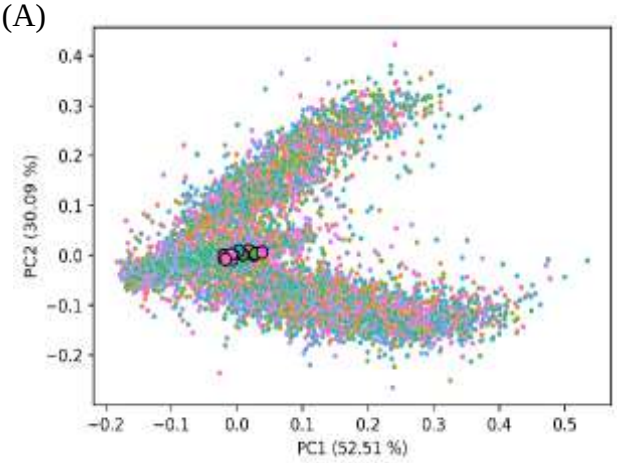
To better understand the fruit shape and also the distinction between cultivars, they were evaluated using the elliptic Fourier descriptors, followed by PCA, a procedure already used in other studies to determine fruit morphology (MAEDA; AKAGI; TAO, 2018; OSAKO et al., 2020). Although the PCA of the five morphological and color variables revealed two distinct groups of cultivars, the use of elliptic Fourier descriptors did not yield similar results, as

cultivars are condensed in relation to morphological and color variables for all ripening stages (Figure 13).

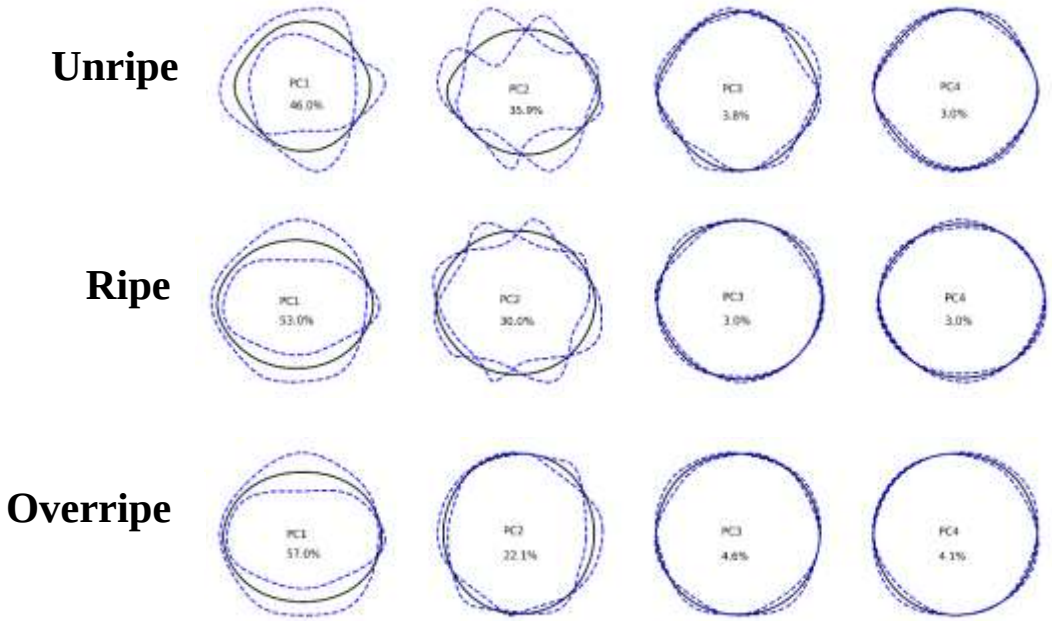
The variation in the shape of unripe, ripe, and overripe fruit for PC1, PC2, PC3, and PC4 is presented in Figure 13B. The contours were reconstructed considering the mean and two times the standard deviation. It is clear that PC1 is responsible for the differences in the oblong shape of the fruit, and PC2 seems to be involved with the irregularities at the edge of the fruit. However, the variation in unripe fruit is more pronounced than in ripe and overripe fruit.

Figure 13. Multivariate analysis using Fourier descriptors: **(A)** Principal component analysis (PCA) using variables estimated through elliptic Fourier descriptors. Each point represents a piece of ripe fruit. The mean values of principal component 1 (PC1) and principal component 2 (PC2) of each cultivar are plotted as a large circle. **(B)** Variation in the shape of unripe, ripe, and overripe fruit that can be explained by PC 1, 2, 3, and 4. The contours were reconstructed from the corresponding principal component and are equal to the mean (solid black line) or two times the standard deviation (dashed blue lines).

(A)

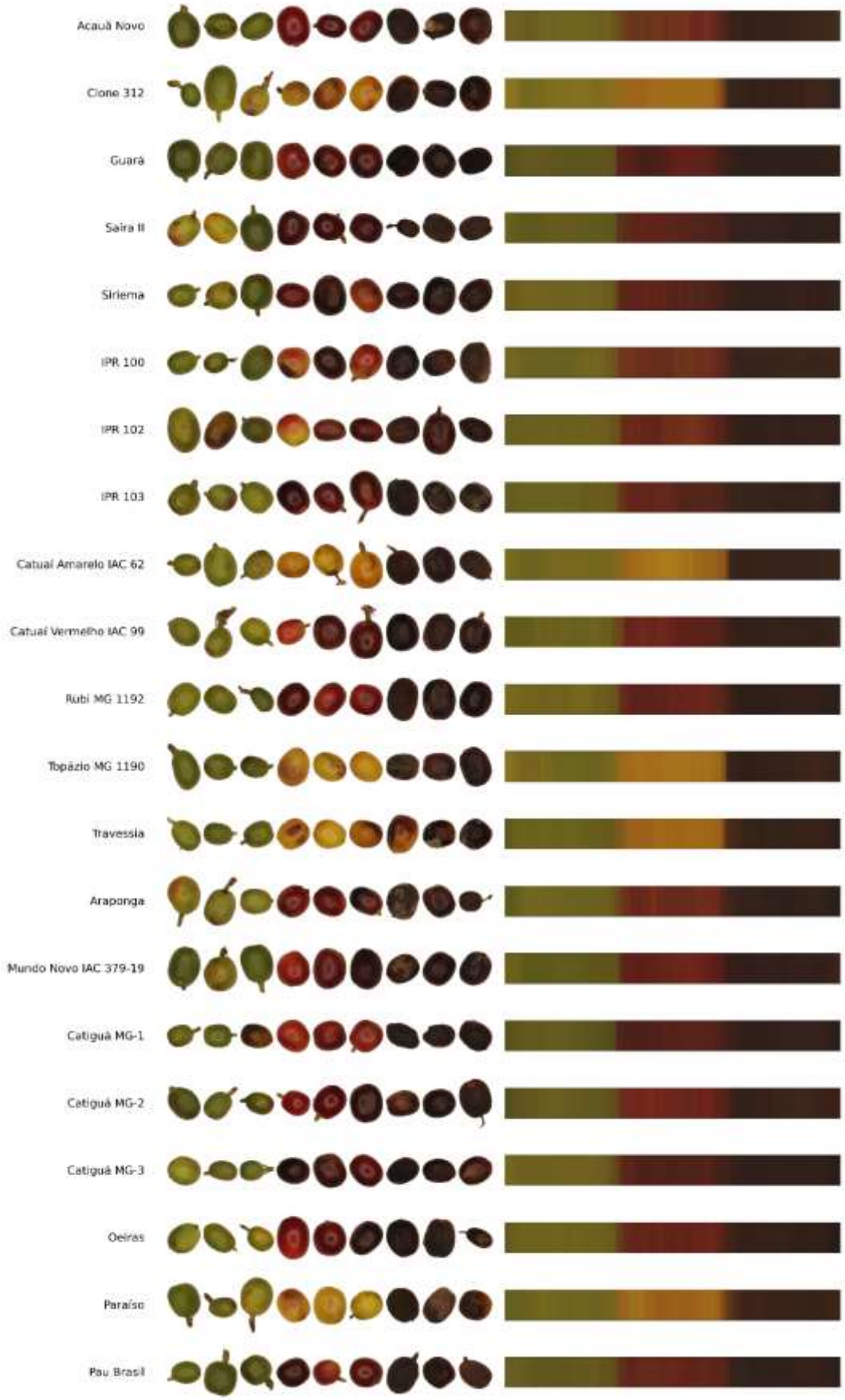


(B)



To obtain information about the process of color change in fruit maturation, Figure 14 shows three pieces of fruit at each maturation stage, chosen at random, as well as the color palette for the 21 cultivars studied. Through visual analysis, subtle differences can be seen among cultivars in fruit color as they mature. Five out of the 21 cultivars have yellow-colored fruit (Paraiso, Travessia, Topázio MG 1190, Catuaí Amarelo IAC 62, and Clone 312). Otherwise, all other cultivars have red-colored fruit.

Figure 14. Color palette and examples of fruit for the 21 cultivars analyzed.



3.4 Sieve classification

Three new variables were created based on the classification of bean size using sieves: large beans, composed of beans from sieves 18 and 17 (flat beans) and 13 and 12 (peaberry beans); medium beans, formed by sieves 15 and 16 (flat beans) and 10 and 11 (peaberry beans); and small beans, composed of beans from sieves 14 and 13 (flat beans) and sieve 9 (peaberry beans). The percentage of each new variable was then calculated. The main reason for the classification by sieves was to check for the existence of correlation between fruit size and bean size, to see if indirect selection through fruit size could be reflected in selection for bean size.

Significant differences (p value < 0.01) were found among cultivars, using ANOVA, for all three variables. The cultivar "Guará" presented the highest percentage of large beans (51.8%), and the cultivar "Catiguá MG-2" the lowest percentage (9.1%). The cultivar "Catiguá MG-2" obtained the highest percentage of medium-sized beans (73.1%), and "Guará" the lowest (38.7%). For small beans, the cultivar "IPR 100" had the highest percentage (21.1%), and "Clone 312" the lowest (6.7%) (Figure 15).

Figure 15. Classification of coffee beans by sieves for the cultivars studied.

Information about each individual piece of fruit was obtained through evaluation of morphological traits. However, the relationship between fruit size and corresponding bean size was not considered, making it impossible to establish a direct correlation between these two variables. To establish such a correlation, the sample distribution of bean size was examined and classified based on standard deviations. Thus, beans with values below -0.79 standard deviations were considered small, beans between -0.79 and 0.97 standard deviations were determined as medium, and beans above 0.97 standard deviations were considered large.

Based on this information, it was assumed that the sample distribution of fruit resembles the distribution for bean size. To mitigate the effect of ripening, morphological variables were standardized for each stage of ripeness. Using the standard deviation of each piece of fruit, they were categorized as a large, medium, and small bean according to the thresholds described above. The percentage of the three classes was then obtained for each treatment and replication.

Correlations were observed for each stage of ripeness, as well as for all stages together. Considering all stages, no high magnitude correlations were obtained between morphological traits and bean size variables. The highest correlation observed for large beans was with area, with a correlation coefficient of 0.5. For medium beans, the highest correlation was with perimeter (0.41); and for small beans, the highest was with diameter (0.66).

Analyzing correlations for each stage of ripeness showed that overripe fruit had higher magnitude correlations than unripe and ripe fruit. Therefore, when considering only overripe fruit, the fruit morphological traits and bean size variables of highest correlation for large and medium beans was with area, with correlation coefficients of 0.56 and 0.38, respectively. For small beans, the highest correlation found was with diameter, with a correlation coefficient of 0.70.

4 DISCUSSION

The phenotyping platform was constructed of low-cost materials accessible to all. It showed efficiency and ensured precise and reliable results in fruit evaluation. It creates high-quality uniform images, facilitating the analysis process. In addition, it was designed to allow images to be obtained of various traits for various crops that may be important and that have already been measured in other studies, such as biomass (GOLZARIAN et al., 2011), disease quantification (KHAN et al., 2020), seed morphology (CHERN et al., 2007; JAVANMARDI

et al., 2021; JOOSEN et al., 2012), and leaf morphology (NETO et al., 2006), increasing its possible uses.

The fruit analysis pipeline developed was efficient in segmentation, allowing precise definition of individual pieces of fruit. This made it possible to obtain accurate estimates of the traits being studied. The use of the watershed algorithm allowed very efficient segmentation of pieces of fruit that were in contact with each other. This allowed fruit to be arranged on the platform even in contact, increasing efficiency in the time required to analyze each sample as there is no need for human attention to separate each piece of fruit.

There were no previous studies of *C. arabica* that focused on in-depth analysis of its fruit. Within breeding programs, some traits are determined through visual evaluation of the fruit, such as the ripening cycle, ripening uniformity, and incidence of diseases (COSTA et al., 2013; DE LIMA; POZZA; DA SILVA SANTOS, 2012). The ripening cycle and ripening uniformity are determined by the ripening stages the fruit is in, and they are extremely important in breeding programs as they allow the selection of genotypes with specific objectives.

Usually, evaluators determine the ripening stages the fruit is in, and this is a completely subjective process, due to the complexity involved in color perception. Even experienced evaluators may disagree on classification. These divergences occur due to the undefined boundary between classes and the different personal experiences that each evaluator may have. In addition, in the process of evaluating ripening stages for an experiment, only one evaluator is often responsible for evaluating a sample. However, in order to optimize the phenotyping process, multiple evaluators are often used to evaluate samples. This method can generate random errors and interfere with the results obtained.

To increase the efficiency of visual evaluation, different evaluators can come to consensus on evaluation of a sample, reducing errors generated by individual evaluation. However, even with consensus among evaluators, visual evaluation still requires a lot of time, and there may still be divergences in evaluations as there is no description of the range of color for each ripening class.

The use of computer vision solves this problem, as analysis occurs in a non-subjective manner and with uniformity among evaluations. Among the techniques that can be used for determining the classes, the use of deep learning offers several advantages over other

approaches, making it possible to obtain higher accuracy and a more general result (O'MAHONY et al., 2020).

In this study, a classification model was adopted, as each piece of fruit was easily and efficiently analyzed, and its use allowed for satisfactory results in training the model. The classification model achieved excellent accuracy in the test images; however, the addition of other ripening classes, such as nearly ripe and nearly overripe fruit, may improve the efficiency of the model. Another point to consider is that some fruit showed signs of disease and may have interfered with the evaluations; the addition of new classes in which they can be inserted may also bring better efficiency to the model.

Analysis of the images allowed the creation of 36,879 images of coffee fruit in different ripening stages. These images may be important for future studies, such as the possibility of use in image synthesis. Some studies have already used synthetic images and obtained satisfactory results in validation on real images (MARGAPURI; NEILSEN, 2021; TODA et al., 2020; YANG et al., 2021).

One of the possibilities of this study was to identify cultivars based on the differences among their fruit. For that purpose, two approaches were used: principal component analysis with morphological and color traits, such as length, diameter, area, perimeter, and eccentricity, and median values of red, green, and blue color spaces and the use of elliptic Fourier descriptors. The approach using morphological and color traits showed better distinction among cultivars, resulting in two groups of cultivars: one containing cultivars with yellow ripe fruit and another with red ripe fruit. However, within the groups, there was no clear distinction among cultivars.

In the second approach, which also checked for major differences in coffee fruit shape, elliptic Fourier descriptors were used, but no discrimination among cultivars was achieved. This suggests that cultivars do not have distinct fruit shapes that can be used to detect different cultivars. In future studies, other approaches can be tested, such as Variational Autoencoders (VAE) (Kingma; Welling, 2013). This technique may be more efficient because, while the techniques used in this study extract a limited amount of information from the images, VAE can handle all the information contained in the image, and therefore detect more subtle differences among them.

Another question raised in the study was the possibility of indirect selection through the morphological fruit traits and whether they are reflected in selection for bean size. Some

researchers report that cultivars with large fruit produce large beans. However, the results of this study showed that the correlation between overripe fruit area and large beans was 0.56, indicating a significant and moderate correlation. However, indirect selection is not recommended for selecting large beans through overripe fruit area, as changes occurred in the ranking of the cultivars evaluated.

There are some problems in the way correlation was estimated, such as the variation among maturation stages, the assumption that fruit comes from the same probability distribution as beans, and the information condensation performed for fruit. Therefore, to have a clear idea of the possibility of indirect selection, it may be necessary to conduct a study with information on fruit size variation at different maturation stages, as well as to obtain and evaluate their respective beans.

This study demonstrated that the platform constructed has the ability to obtain high-quality and uniform images. Evaluators showed divergences in the evaluation of coffee fruit maturation stages, leading to classification divergences; therefore, computer vision proves to be an excellent alternative in mitigating these divergences. A dataset of labeled coffee fruit images in four maturation stages was created, which may be useful for future studies. In addition, differences in fruit morphology were shown for different maturation stages. The study made a detailed analysis of the fruit of 21 cultivars, providing valuable information regarding their morphological and color characteristics, which may serve as a reference and be useful for future studies.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv Prepr. arXiv1603.04467*.

Avendano, J., Ramos, P. J., and 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 Syst. Appl.* 88, 178–192. doi: https://doi.org/10.1016/j.eswa.2017.06.044.

Bazame, H. C., Molin, J. P., Althoff, D., and Martello, M. (2021). Detection, classification, and mapping of coffee fruits during harvest with computer vision. *Comput. Electron. Agric.* 183, 106066.

Bazame, H. C., Molin, J. P., Althoff, D., and Martello, M. (2022). Detection of coffee fruits on tree branches using computer vision. *Sci. Agric.* 80.

Bertrand, B., Alpizar, E., Lara, L., SantaCreo, R., Hidalgo, M., Quijano, J. M., et al. (2011). Performance of Coffea arabica F1 hybrids in agroforestry and full-sun cropping systems in comparison with American pure line cultivars. *Euphytica* 181, 147–158. doi:

10.1007/s10681-011-0372-7.

Bradski, G., and Kaehler, A. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc."

Bresilla, K., Perulli, G. D., Boini, A., Morandi, B., Corelli Grappadelli, L., and Manfrini, L. (2019). Single-shot convolution neural networks for real-time fruit detection within the tree. *Front. Plant Sci.* 10, 611.

Carvalho, C. H. S. de (2008). Cultivares de café: origem, características e recomendações. *Brasília Embrapa Café* 334.

Chern, C.-G., Fan, M.-J., Yu, S.-M., Hour, A.-L., Lu, P.-C., Lin, Y.-C., et al. (2007). A rice phenomics study—phenotype scoring and seed propagation of a T-DNA insertion-induced rice mutant population. *Plant Mol. Biol.* 65, 427–438. doi: 10.1007/s11103-007-9218-z.

Costa, J. C., Carvalho, C. H. S., Matiello, J. B., Almeida, S. R., Carvalho, S. P., and Baliza, D. P. (2013). Comportamento agronômico de progênies e cultivares de cafeeiro com resistência específica à ferrugem. *Coffee Sci. 1984-3909* 8, 183–191.

De Lima, L. M., Pozza, E. A., and Da Silva Santos, F. (2012). Relationship between incidence of brown eye spot of coffee cherries and the chemical composition of coffee beans. *J. Phytopathol.* 160, 209–211.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12.

Dyrmann, M., Mortensen, A. K., Midtiby, H. S., and Jørgensen, R. N. (2016). Pixel-wise classification of weeds and crops in images by using a fully convolutional neural network. in *Proceedings of the International Conference on Agricultural Engineering, Aarhus, Denmark*, 26–29.

Fiorani, F., and Schurr, U. (2013). Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol.* 64, 267–291.

Gentallan, R. P., Altoveros, N. C., Borromeo, T. H., Endonela, L. E., Hay, F. R., Lalusin, A. G., et al. (2019). An objective method of shape descriptor state establishment using elliptic Fourier analysis (EFA). *Plant Genet. Resour.* 17, 480–487.

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., and Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci.* 115, 4613–4618.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Golzarian, M. R., Frick, R. A., Rajendran, K., Berger, B., Roy, S., Tester, M., et al. (2011). Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant Methods* 7, 1–11.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. in *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

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*, 770–

778.

Javanmardi, S., Ashtiani, S.-H. M., Verbeek, F. J., and Martynenko, A. (2021). Computer-vision classification of corn seed varieties using deep convolutional neural network. *J. Stored Prod. Res.* 92, 101800.

Joosen, R. V. L., Arends, D., Willems, L. A. J., Ligterink, W., Jansen, R. C., and Hilhorst, H. W. M. (2012). Visualizing the genetic landscape of Arabidopsis seed performance. *Plant Physiol.* 158, 570–589.

Kamilaris, A., and Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Comput. Electron. Agric.* 147, 70–90.

Khan, M. A., Akram, T., Sharif, M., Javed, K., Raza, M., and Saba, T. (2020). An automated system for cucumber leaf diseased spot detection and classification using improved saliency method and deep features selection. *Multimed. Tools Appl.* 79, 18627–18656.

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv Prepr. arXiv1312.6114*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90.

Kuhl, F. P., and Giardina, C. R. (1982). Elliptic Fourier features of a closed contour. *Comput. Graph. image Process.* 18, 236–258.

Li, D., Cao, Y., Tang, X., Yan, S., and Cai, X. (2018). Leaf segmentation on dense plant point clouds with facet region growing. *Sensors* 18, 3625.

Li, Z., Guo, R., Li, M., Chen, Y., and Li, G. (2020). A review of computer vision technologies for plant phenotyping. *Comput. Electron. Agric.* 176, 105672.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Maeda, H., Akagi, T., and Tao, R. (2018). Quantitative characterization of fruit shape and its differentiation pattern in diverse persimmon (Diospyros kaki) cultivars. *Sci. Hortic. (Amsterdam).* 228, 41–48.

Margapuri, V., and Neilsen, M. (2021). Seed phenotyping on neural networks using domain randomization and transfer learning. in *2021 ASABE Annual International Virtual Meeting* (American Society of Agricultural and Biological Engineers), 1.

Mendoza, F., and Aguilera, J. M. (2004). Application of image analysis for classification of ripening bananasMENDOZA, F.; AGUILERA, J. M. Application of image analysis for classification of ripening bananas. Journal of food science, v. 69, n. 9, p. E471–E477, 2004. *J. Food Sci.* 69, E471–E477.

Meyer, F. (1994). Topographic distance and watershed lines. *Signal Processing* 38, 113–125. doi: https://doi.org/10.1016/0165-1684(94)90060-4.

Milioto, A., Lottes, P., and Stachniss, C. (2017). REAL-TIME BLOB-WISE SUGAR BEETS VS WEEDS CLASSIFICATION FOR MONITORING FIELDS USING CONVOLUTIONAL NEURAL NETWORKS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* 4.

Morais, H., Caramori, P. H., Koguishi, M. S., and Ribeiro, A. M. de A. (2008). Escala fenológica detalhada da fase reprodutiva de Cooffea arabica. *Bragantia* 67, 257–260.

Neto, J. C., Meyer, G. E., Jones, D. D., and Samal, A. K. (2006). Plant species identification using Elliptic Fourier leaf shape analysis. *Comput. Electron. Agric.* 50, 121–134. doi: https://doi.org/10.1016/j.compag.2005.09.004.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., et al. (2020). Deep learning vs. traditional computer vision. in *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1* (Springer), 128–144.

Osako, Y., Yamane, H., Lin, S.-Y., Chen, P.-A., and Tao, R. (2020). Cultivar discrimination of litchi fruit images using deep learning. *Sci. Hortic. (Amsterdam).* 269, 109360.

Parikh, D., and Zitnick, C. L. (2010). The role of features, algorithms and data in visual recognition. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE), 2328–2335.

Ramos, P. J., Avendano, J., and Prieto, F. A. (2018). Measurement of the ripening rate on coffee branches by using 3D images in outdoor environments. *Comput. Ind.* 99, 83–95.

Ramos, P. J., Prieto, F. A., Montoya, E. C., and Oliveros, C. E. (2017). Automatic fruit count on coffee branches using computer vision. *Comput. Electron. Agric.* 137, 9–22.

Sayıncı, B., Kara, M., Ercişli, S., Duyar, Ö., and Ertürk, Y. (2015). Elliptic Fourier analysis for shape distinction of Turkish hazelnut cultivars. *Erwerbs-Obstbau* 57, 1–11.

Sera, G. H., Sera, T., and Fazuoli, L. C. (2017a). IPR 102-Dwarf Arabica coffee cultivar with resistance to bacterial halo blight. *Crop Breed. Appl. Biotechnol.* 17, 403–407.

Sera, T., Sera, G. H., Fazuoli, L. C., Machado, A. C. Z., Ito, D. S., Shigueoka, L. H., et al. (2017b). IPR 100-Rustic dwarf Arabica coffee cultivar with resistance to nematodes Meloidogyne paranaensis and M. incognita. *Crop Breed. Appl. Biotechnol.* 17, 175–179.

Silva, V. A., Abrahão, J. C. de R., Reis, A. M., Santos, M. de O., Pereira, A. A., Botelho, C. E., et al. (2022). Strategy for Selection of Drought-Tolerant Arabica Coffee Genotypes in Brazil. *Agronomy* 12, 2167.

Sousa, T. V., Caixeta, E. T., Alkimim, E. R., Oliveira, A. C. B., Pereira, A. A., Sakiyama, N. S., et al. (2019). Early selection enabled by the implementation of genomic selection in Coffea arabica breeding. *Front. Plant Sci.* 9, 1934.

Toda, Y., Okura, F., Ito, J., Okada, S., Kinoshita, T., Tsuji, H., et al. (2020). Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. *Commun. Biol.* 3, 1–12.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., et al. (2014). scikit-image: image processing in Python. *PeerJ* 2, e453.

Williams, K., Munkvold, J., and Sorrells, M. (2013). Comparison of digital image analysis using elliptic Fourier descriptors and major dimensions to phenotype seed shape in hexaploid wheat (Triticum aestivum L.). *Euphytica* 190, 99–116.

Yang, S., Zheng, L., He, P., Wu, T., Sun, S., and Wang, M. (2021). High-throughput soybean seeds phenotyping with convolutional neural networks and transfer learning. *Plant*

Methods 17, 50.

YOSHIOKA, Y., IWATA, H., OHSAWA, R. Y. O., and NINOMIYA, S. (2004). Analysis of Petal Shape Variation of Primula sieboldii by Elliptic Fourier Descriptors and Principal Component Analysis. *Ann. Bot.* 94, 657–664. doi: 10.1093/aob/mch190.

ARTIGO 2 – DETECÇÃO E IDENTIFICAÇÃO DE FRUTOS DE CAFÉ UTILIZANDO IMAGENS SINTÉTICAS, YOLO E INFERÊNCIA POR FATIA.

Artigo redigido conforme as normas da revista Computers and Electronics in Agriculture

(VERSÃO PRELIMINAR)

RESUMO

Os produtores e os centros de pesquisas não possuem uma ferramenta eficiente na identificação e classificação dos frutos de café para os estágios de maturação, no qual sirva em diferentes ambientes e cenários. Neste estudo, é proposto um modelo de identificação e classificação dos estágios de maturação dos frutos de café, que pode ser utilizado em diversos cenários e ambientes. O modelo é baseado na arquitetura YOLOv5s e foi totalmente treinado em 10.000 imagens sintéticas geradas. Para validar o modelo foram utilizados 122 imagens reais. A inferência nas imagens reais, utilizou-se da técnica de hiper-inferência por fatias. O modelo treinado obteve uma acurácia média na identificação dos frutos de 92%. Para avaliar a classificação dos frutos foi estimado a correlação entre a estimação do modelo e a estimação real, a menor correlação observada foi na classificação dos frutos maduros de 75% e para os estágios de maturação secos e verdes foi de 88% e 98% respectivamente. Os resultados demostram potencialidade de adoção do modelo para avaliação da maturação dos frutos de café em diferentes cenários e ambientes.

Palavras-chaves: Visão computacional, Maturação, *Coffea arabica,* Fenotipagem, Melhoramento genético de plantas.

1 INTRODUÇÃO

A classificação dos frutos de café em relação aos estágios de maturação é um aspecto crucial em várias áreas da ciência e agricultura. A maioria dos agricultores e programas de melhoramento colhem os frutos todos de uma só vez, resultando em frutos com diferentes estágios de maturação, como verdes, maduros e secos. Isso ocorre porque o cafeeiro apresenta várias floradas, sendo uma principal e outras secundárias, que podem ocorrer em momentos distintos, dependendo das condições climáticas e da variabilidade genética da cultivar (Majerowicz e Söndahl, 2005). Entender o comportamento das plantas em relação ao ciclo fenológico, como a uniformidade de maturação e a duração do ciclo (precoce, médio e tardio), é um desafio para os centros de pesquisa. Além disso, a colheita realizada pelos agricultores, deve consistir principalmente de frutos maduros, uma vez que os frutos maduros produzem um produto final de maior qualidade e valor. A colheita de grandes quantidades de frutos verdes ou secos, pode levar a perdas qualitativas, pois os frutos nessas condições, sofrem alterações no tipo, sabor, aroma e qualidade da bebida (Reis e Cunha, 2010).

Uma prática comum na avaliação da maturação dos frutos é a adoção de uma avaliação subjetiva (Fazuoli et al., 1983; Petek et al., 2006; Sousa et al., 2019) ou por meio de uma amostragem dos frutos, sendo essa realizada por meio da quantificação do número de frutos nos diferentes estágios de maturação (Costa et al., 2013; Nogueira et al., 2005). Esses métodos, requerem uma grande quantidade de tempo e podem conter divergências e erros nas avaliações, pois a percepção das cores é totalmente subjetiva. Neste contexto, a visão computacional tem se destacado como uma técnica eficiente, especialmente na identificação de objetos em ambientes desafiadores, no qual é um componente fundamental para muitas tarefas de visão computacional e robótica. Todos os modelo de detecção de objetos líderes atuais, dependem de redes neurais convolucionais (Chen et al., 2019; Dai et al., 2016; He et al., 2017; Redmon e Farhadi, 2017; Ren et al., 2015). No entanto, no treinamento desses modelos, eles requerem grandes quantidades de dados de treinamento rotulados, o que geralmente é demorado e caro para ser criado.

A utilização de imagens sintéticas para o treinamento oferece uma solução alternativa a esses problemas, uma vez que, infinitas imagens podem ser geradas e as anotações serem obtidas automaticamente, sem a necessidade de trabalho humano. Além disso, vários cenários distintos podem ser simulados, o que geralmente é difícil de obter por meio de técnicas de aumento de imagem. Diversos autores utilizaram imagens sintéticas no treinamento de uma modelo e obtiveram bons resultados em imagens reais, como exemplo, Kuznichov et al. (2019) propuseram um método para segmentar e contar as folhas de *Arabidopsis*, abacate e banana, usando textura de folha sintética posicionada com diferentes tamanhos e ângulos para simular imagens obtidas em condições agrícolas reais. Toda et al. (2020) demonstraram que conjuntos de dados sintéticos, foram suficientes para treinar uma rede de segmentação de instâncias para sementes de cevada e testadas em imagens do mundo real. Yang et al. (2021) treinaram um modelo de segmentação de instâncias por meio de imagens sintéticas de sementes de soja e comprovaram sua eficiência em imagens reais. Portanto, conjuntos de imagens sintéticas têm um grande potencial na área da fenotipagem de plantas por meio do uso de visão computacional.

Um aspecto que pode vir a ser um problema na classificação e identificação dos frutos de café, é que cada fruto pode estar presente em uma pequena área da imagem, no qual a razão de altura e largura da caixa delimitadora de cada fruto com a altura e largura da imagem é menor que 0,1 ou a razão da área da caixa delimitadora do fruto com a área da imagem é menor que 0,03 (Chen et al., 2017). Esse aspecto leva os modelos empregados para detecção de objetos terem uma performance muito ruim em objetos pequenos comparado aos médios e grandes. A

principal razão para esse desempenho ruim é que objetos pequenos têm menor resolução e ocupam menos pixels do que objetos maiores. Além disso, existe a perda de informação de posição espacial ao realizar *down-sampling* e operações de *pooling* nas redes neurais convolucionais tornando mais desafiador a detecção e localização desses.

Diversos trabalhos tem focado em maneiras de lidar com esse problema. O método proposto em Kisantal et al. (2019) amostra imagens com objetos pequenos e as aumenta fazendo várias cópias desses objetos. Em Bosquet et al. (2018), é proposta uma rede convolucional para detecção de objetos pequenos que contém um mecanismo de atenção visual precoce que é proposto para escolher as regiões mais promissoras com objetos pequenos. Em Pang et al. (2019), é proposta uma nova rede (JCS-Net) para detecção de pedestres em pequena escala, que integra a tarefa de classificação e a tarefa de super-resolução em um *framework* unificado. O método de Chen et al. (2019) pode aprender recursos mais ricos de pequenos objetos a partir das áreas ampliadas, que são recortadas da imagem bruta. Em Van Etten (2019), é proposta uma técnica baseada em fatiamento e Akyon et al. (2022) propuseram uma generalização de inferência por cortes que pode ser utilizado a partir de qualquer modelo de detecção de objeto existente.

Existem algumas soluções publicadas que utilizam visão computacional para identificar e classificar os frutos de café de acordo com seus estágios de maturação. Alguns trabalhos se concentram na identificação e classificação dos frutos diretamente nas plantas de café (Bazame et al., 2022; Ramos et al., 2018, 2017). Outro, para classificar e identificar os frutos nos diferentes estágios de maturação durante a colheita por colhedoras (Bazame et al., 2021). Contudo apesar dos benefícios de avaliar os estágios de maturação na planta, esta prática não é eficaz para uma ampla avaliação da maturação, pois existe diferenças na posição dos frutos quanto aos estágios. Além disso, a classificação e identificação dos frutos durante a colheita por colhedoras não abrange todos os cenários possíveis para avaliação desse caráter. Por exemplo, no melhoramento de *Coffea* a colheita por produtores que não possuem colhedora, as plantas são colhidas manualmente ou por derriçadeiras manuais, fazendo necessário que o modelo abrange cenários diversos para classificação e identificação dos frutos.

Diante do exposto objetivou-se com esse trabalho o treinamento de um modelo de detecção e classificação de frutos de cafeeiro para os estágios de maturação verde, maduro e seco e que possa contemplar diferentes cenários na avaliação dos frutos. Para isso utilizou-se

da arquitetura YOLOv5s no qual o processo de treinamento se deu por meio de imagens sintéticas.

2 MATERIAL E MÉTODOS

2.1 Geração das imagens sintéticas

Imagens individuais de 36.879 frutos pertencendo a diferentes estágios de maturação, incluindo 9.781 frutos verdes, 11.653 frutos vermelhos, 5.072 frutos amarelos e 10.373 frutos secos foram utilizados. Essas imagens foram obtidas por meio do pipeline desenvolvido no capitulo 1. As imagens de fundo, foram obtidas de forma aleatória pelo https://picsum.photos/, totalizando 1000 imagens de tamanho 1000 pixels por 1000 pixels. Essas imagens foram utilizadas para a criação de um conjunto de dados de treinamento. As imagens sintéticas diferiram em tamanho variando de 400 pixel a 1200 pixels para altura e largura, sendo sintetizadas várias configurações

O procedimento de síntese das imagens foi realizado da seguinte forma. Primeiro, uma imagem de fundo é selecionada aleatoriamente e redimensionada de acordo com o tamanho pré-estabelecido, e então uma imagem de fruto é selecionada aleatoriamente e as coordenadas x e y, em que a imagem será colada, é determinada aleatoriamente, contudo, os valores das coordenadas são restringidos para que o fruto esteja dentro da imagem gerada. A quantidade de frutos colado em cada imagem variou de 10 a 120, e de acordo com o tamanho da imagem de fundo, a razão de sobreposição dos frutos também foi verificada, sendo a sobreposição máxima inicial de 0,1, para esse procedimento, caso a razão ultrapasse esse valor pré-determinado, a colagem é cancelada e outra coordenada é escolhida. Se depois de 200 iterações a colagem não tenha sido obtida, a razão de sobreposição é acrescida de 0,1.

Foram sintetizados um total de 11.000 imagens. Para cada imagem as coordenadas da posição de cada fruto foram obtidas. Exemplos dessas imagens se encontram na figura 1, nessas imagens estão contidas imagens criadas utilizando um procedimento de aumento por meio da biblioteca Albumentations (Buslaev et al., 2020) utilizando as funções RandomBrightnessContrast, Downscale, GaussNoise e RandomShadow, cada com uma probabilidade de ocorrência de 0,8.

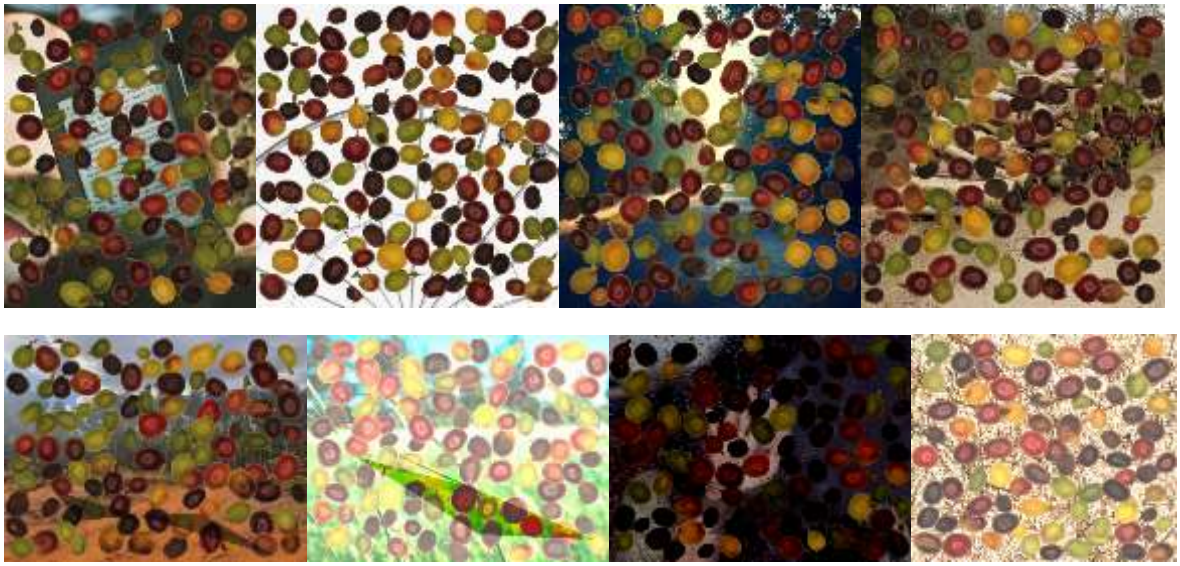


Figura 1. Exemplo das imagens sintéticas geradas.

2.2 Treinamento do modelo

You Only Look Once (YOLO) é um dos primeiros modelos de detecção de objetos que introduziu a ideia de combinar a predição por caixa e a classificação de objetos em uma única arquitetura (Redmon et al., 2016). A arquitetura foi introduzida como parte de um *framework* chamado *Darknet*. Ao longo do tempo, melhorias no YOLO foram implementadas e lançadas como pacotes de software distintos e independentes (Bochkovskiy et al., 2020; Redmon e Farhadi, 2018, 2017; Wang et al., 2022) e uma dessas é o YOLOv5 (Jocher et al., 2020), no qual consiste em quatro arquiteturas diferentes, YOLOv5s, YOLOv5m, YOLOv5l e YOLOv5x e estão disponíveis em https://github.com/ultralytics/yolov5. Cada uma das arquiteturas difere no número de camadas da rede neural, sendo que 5s, 5m, 5l e 5x contêm 283, 391, 499 e 607 camadas, respectivamente.

Para o presente estudo, utilizou-se do YOLOv5s e as imagens de entrada foram redimensionadas para 640x640 pixels. Para melhorar a precisão do modelo foi adotado a transferência de aprendizado. Portanto, para o treinamento, o modelo foi inicializado utilizando os parâmetros do modelo treinado no conjunto de dados *Microsoft Common Object in Context* (MS COCO) (80 classes, 1.5 milhões de exemplos, e 330 mil imagens) (Lin et al., 2014). O modelo foi treinado por 100 iterações e com um tamanho de lote de 12. Dez porcento das imagens foram separadas para teste, e o melhor modelo foi obtido pela precisão média dessas imagens.

2.3 Processo de inferência das imagens

Os algoritmos projetados para detecção de objetos têm baixo desempenho em imagens de alta resolução que contêm objetos pequenos e densos. Além deste problema, existem outros, como a perda de informações espacial causada pelo *down-sampling* e as operações de *pooling* nas redes neurais convolucionais, tornando desafiador a detecção e localização de objetos pequenos (Liang et al., 2022; Liu et al., 2021).

Portanto a detecção de pequenos objetos como pode ser o caso de frutos de café que são representados por uma pequena quantidade de pixels, requer técnicas especificas. A utilização da Hiper-Inferência Auxiliada por Fatias (SAHI) (Akyon et al., 2022) é uma solução para a detecção de objetos pequenos. Para lidar com o problema da detecção, os autores propõem um quadro genérico baseado em fatiamento. No qual divide as imagens de entrada em *patches* sobrepostos, resultando em áreas de pixels relativamente maiores para os objetos pequenos, em relação às imagens alimentadas na rede (Figura 2).

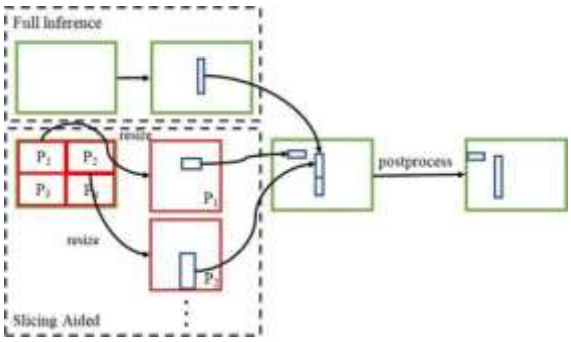


Figura 2. Estratégia da Hiper-Inferência Auxiliada por Fatias (figura retirada de Sun et al., 2023).

O *framework* SAHI pode ser encontrado em https://github.com/obss/sahi. Ao aplicar a estratégia SAHI para inferência, o tamanho da imagem fatiada é essencial para os resultados de detecção. Sob a condição de não alterar outros parâmetros, quando a imagem fatiada é menor, o modelo de inferência pode capturar mais recursos detalhados da imagem, o que também pode trazer mais detecções falsas. Em nosso estudo para a inferência as imagens foram fatiadas em 1000 por 1000 pixels, com uma sobreposição de 0,2 no qual dividiu a imagem em 20 partes. Esse valor foi determinado com base em alguns testes visuais nas imagens. Os resultados dos cálculos de inferência, são alimentados na supressão não máxima (NMS), removendo caixas de detecção de baixa confiança para o mesmo objeto e restaurando o tamanho original da imagem, neste estudo utilizou-se um NMS de 0,4.

2.4 Validação do modelo treinado

O treinamento do modelo foi realizado inteiramente nas imagens sintéticas geradas. Modelos treinados em imagens sintéticas têm sido cada vez mais utilizados na área de visão computacional, especialmente em tarefas de detecção e reconhecimento de objetos (Kuznichov et al., 2019; Toda et al., 2020; Yang et al., 2021).Esses modelos podem reduzir os custos e aumentar a eficiência do processo de treinamento. Contudo, um dos principais desafios é que as imagens sintéticas podem não capturar todas as variações e nuances das imagens reais, o que pode levar a uma limitação na capacidade dos modelos treinados generalizar para imagens reais (Taori et al., 2020).

Enquanto o uso de imagens sintéticas para treinar modelos de visão computacional oferece várias vantagens, é importante estar ciente dos desafios e limitações associados a este. Para garantir a generalização adequada dos modelos treinados em imagens sintéticas, é importante validar cuidadosamente o desempenho do modelo em situações reais antes de implantá-lo em uma aplicação real (Taori et al., 2020).

Foi utilizado um conjunto de dados reais para validar o modelo treinado, avaliando se ele é capaz de identificar os frutos contidos nas imagens e classificá-los corretamente para os estágios de maturação. O conjunto de dados consiste em imagens de frutos em vários estágios de maturação, no qual foi disposto sobre uma cartolina branca. As imagens foram capturadas por meio de uma câmera de celular e um total de 122 imagens foram analisadas. Para inferir o número de frutos em cada estágio de maturação foram utilizados os dados de classificação descrito no capítulo 1, no qual foi adquirido pela plataforma de fenotipagem. Alguns exemplos dessas imagens são apresentados na figura 3.

Figura 3. Exemplos das imagens reais.

3 RESULTADOS E DISCUSSÕES

Foram geradas 11.000 imagens sintéticas para o treinamento do modelo, no entanto, percebeu-se que não foi necessário utilizar todas elas. Com apenas 3000 imagens, o modelo alcançou a convergência em poucas iterações. Por isso, optou-se por utilizar apenas 3000 imagens. Essa escolha se deu pelo fato de que utilizar uma quantidade maior de imagens demandaria maior tempo de processamento, sem melhorar a performance do modelo.

Para avaliar a performance do modelo treinado na identificação dos frutos, realizou-se uma análise comparando o número total de frutos contabilizado pelo modelo treinado em relação a contagem realizada pela plataforma de fenotipagem demostrada no capítulo 1. As imagens contêm em média 270 frutos, com uma variação de 196 a 334 frutos. Constata-se que o modelo obteve acurácia média de 92% na identificação dos frutos, com uma variação entre 84% e 99%, conforme pode ser observado na figura 4. A correlação da contagem total dos frutos foi de 97%, esses resultados demonstram que o modelo foi capaz de identificar os frutos de café com uma boa precisão, para a situação estudada (Figura 5).

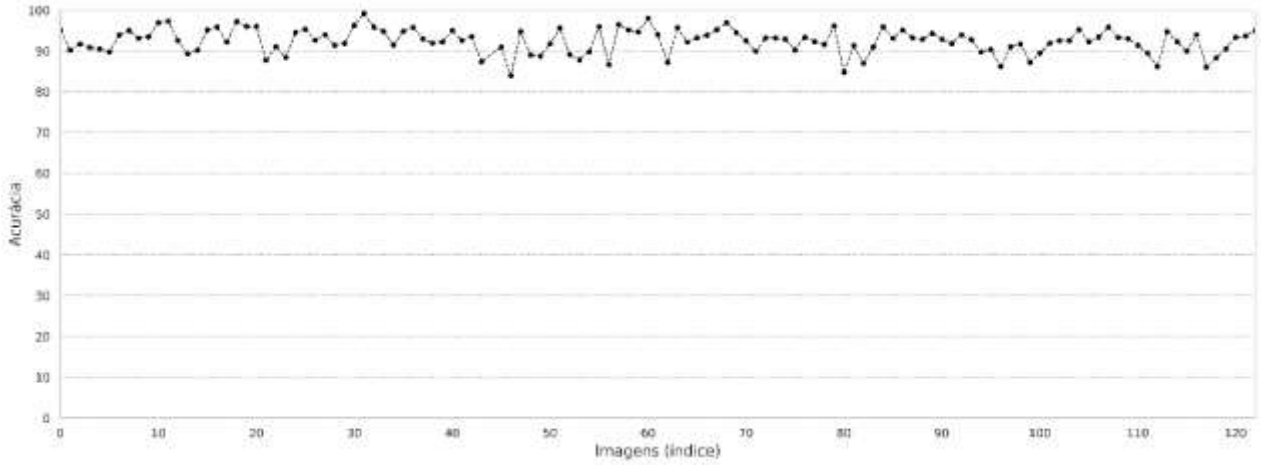


Figura 416. Acurácia do total de frutos identificados das 122 imagens.

No que se refere à classificação dos frutos, o modelo apresentou maior dificuldade na classificação dos frutos maduros, com uma correlação de 75%, enquanto para frutos secos a correlação foi de 88% e de frutos verdes de 98% (Figura 5). Assim, é importante destacar que ainda há espaço para melhorias no modelo, visando aumentar sua eficácia na classificação dos frutos. É possível que a dificuldade da classificação em ambientes diversos se deva às condições de iluminação constantes presentes nas imagens dos frutos utilizadas no treinamento. Para minimizar essa dificuldade, sugere-se a utilização de dados reais no treinamento do modelo, ou

a geração de imagens de frutos em diferentes condições de iluminação. Dessa forma, é possível obter um modelo mais generalista, capaz de capturar todas as nuances necessárias para uma classificação correta dos frutos em diferentes cenários.

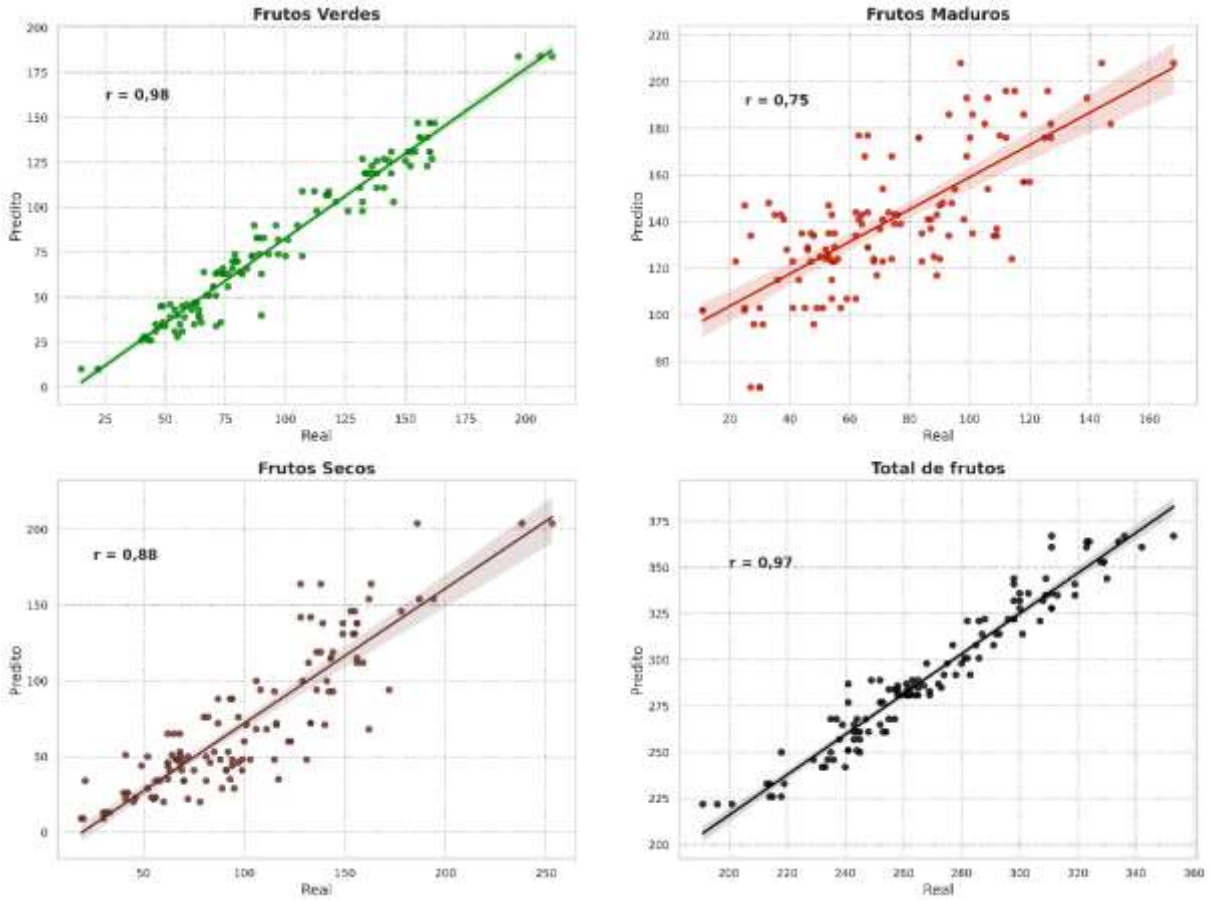


Figura 5. Correlação de Pearson entre as estimações do modelo e os valores reais para os diferentes estágios de maturação

O modelo foi treinado para classificar e identificar frutos de café nos estágios de maturação verde, maduro (amarelo e vermelho) e seco. A adição de outras classes pode melhorar a performance do mesmo, como demostrado na figura 4. Frutos verde-cana foram classificados como frutos maduros amarelos, devido as semelhanças existentes entre eles. Além disso, alguns frutos que claramente são maduros foram classificados como secos, evidenciando a necessidade da inserção de imagens de frutos em diferentes condições na síntese e treinamento em um novo conjunto de imagens.

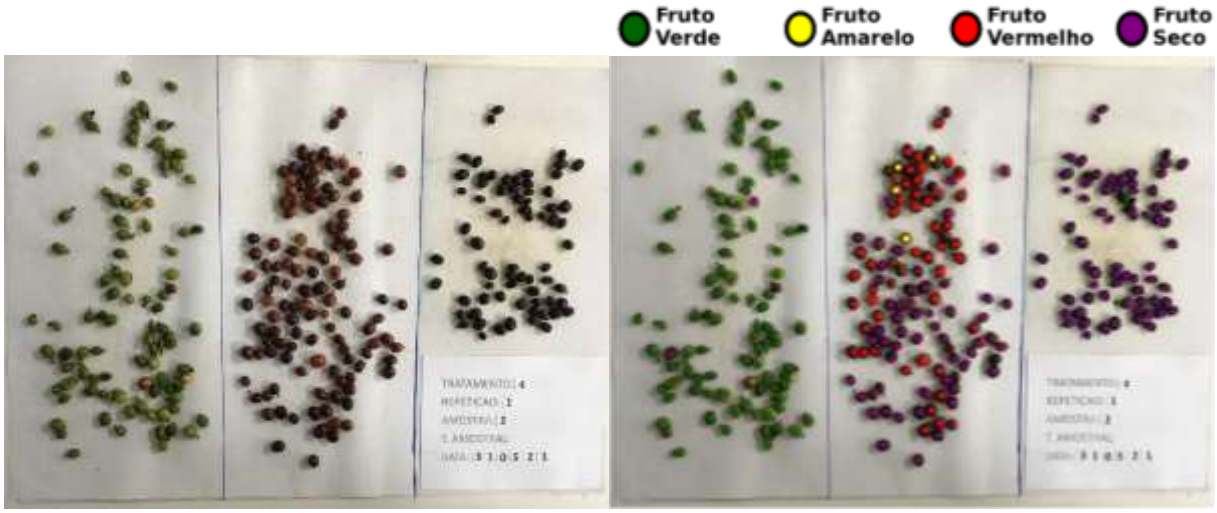


Figura 6. Classificação e identificação dos frutos de uma amostra.

Sete imagens são apresentadas para ilustrar a capacidade da avaliação dos frutos em diferentes condições (Figura 7). Temos com essas imagens diferentes contextos, incluindo frutos maduros e secos, frutos ainda dentro do saco de colheita, nas plantas de café, como também na plataforma de fenotipagem. Essas condições refletem a variedade de ambientes em que os frutos de café podem ser encontrados durante sua avaliação. O modelo treinado demonstra grande potencial de aplicação nesses diferentes ambientes.

Normalmente a fenotipagem dos frutos quanto aos estágios de maturação nos centros de pesquisa cafeeira, é realizada de maneira manual ou subjetiva (Costa et al., 2013; Nogueira et al., 2005; Petek et al., 2006; Sousa et al., 2019). Estas avaliações exigem uma grande quantidade de mão de obra e tempo para serem realizadas, e por muitas vezes possuem altos erros associados (NOGUEIRA *et al.*, 2005; SANTORO *et al.*, 2019). Isso faz com que a fenotipagem para o caractere muitas vezes não seja empregado, principalmente nos programas de melhoramento genético. O modelo treinado demonstra a potencialidade da fenotipagem dos frutos de maneira rápida, acessível e eficiente. Além disso, os resultados obtidos indicam que essa abordagem pode ser uma ferramenta valiosa para produtores em que necessitam da avaliação dos frutos para situações diversas.


Fruto Verde
Fruto Amarelo
Fruto Vermelho
Fruto Seco
9
04/10/2019
2
04/21/2016

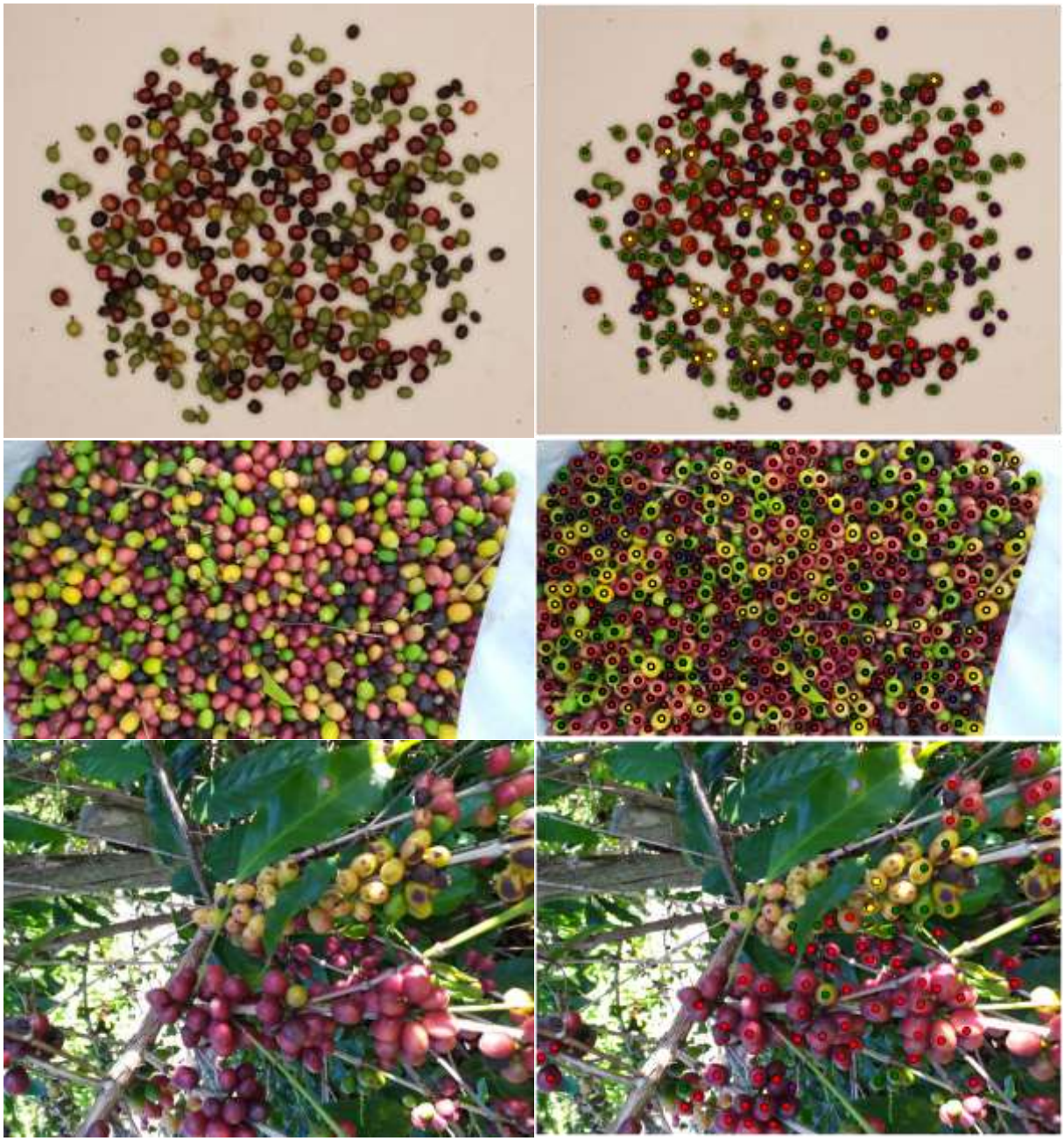


Figura 7. Imagens de frutos de café em vários ambientes e cenários e a identificação e classificação pelo modelo treinado.

A utilização da SAHI não é necessária para todos os conjuntos de dados, somente para aqueles em que os frutos ocupam uma pequena porção da imagem, no qual dificulta a identificação e classificação com a utilização do modelo treinado. Além disso a SAHI aumenta o tempo requerido para inferência de cada imagem dependendo do número de fatiamentos que é adotado, limitando com isso uma inferência em tempo real.

O tempo médio para inferência visual dos frutos para classificação e contabilização, para os diferentes estágios de maturação foi de 2:34 minutos para cada amostra. Esse tempo não inclui o processo de aquisição das amostras e todos as outras necessidades requeridas para avaliar as amostras visualmente. Para a inferência com modelo treinado juntamente com o *framework* SAHI o tempo foi de apenas 6 segundos, sendo que a tomada dos dados amostrais podem ser realizadas diretamente no campo. A nova abordagem é mais rápida e eficiente do que a avaliação visual, pois a avaliação dos frutos quanto aos estágios de maturação é realizada de maneira não subjetiva. Além disso, a avaliação pode ser realizada diretamente no campo, o que reduz o tempo e o custo de transporte das amostras.

Outros estudos já focaram na classificação e identificação dos frutos de café para diferentes estágios de maturação. Como exemplo, alguns autores utilizaram a visão computacional para identificar e classificar frutos diretamente nas plantas (Bazame et al., 2022; Ramos et al., 2018, 2017) outros em plataformas de colheitas por colhedeiras (Bazame et al., 2021). O modelo aqui treinado demostrou potencial de classificação e identificação dos frutos em diferentes ambientes e cenários. Além disso, ao estender o escopo da pesquisa desenvolvida neste trabalho para incluir imagens de frutos em uma ampla variedade de ambientes e condições reais, pode possibilitar o aprimoramento dessa abordagem de avaliação e ajudar a impulsionar a fenotipagem para os estágios de maturação.

4 CONCLUSÃO

O modelo treinado demonstrou eficiência para o conjunto de dados estudado, com potencialidade de aplicação em diferentes cenários e ambientes. Para uma adoção em larga escala necessita da validação em outros conjuntos de dados, a fim de verificar sua performance. Além disso, é importante desenvolver alternativas para melhorar a eficiência da classificação, principalmente dos frutos maduros.

REFERÊNCIAS

Akyon, F.C., Altinuc, S.O., Temizel, A., 2022. Slicing aided hyper inference and fine-tuning for small object detection, in: 2022 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 966–970.

Bazame, H.C., Molin, J.P., Althoff, D., Martello, M., 2022. Detection of coffee fruits on tree branches using computer vision. Sci. Agric. 80.

Bazame, H.C., Molin, J.P., Althoff, D., Martello, M., 2021. Detection, classification, and mapping of coffee fruits during harvest with computer vision. Comput. Electron. Agric. 183, 106066.

Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv Prepr. arXiv2004.10934.

Bosquet, B., Mucientes, M., Brea, V.M., 2018. STDnet: A ConvNet for Small Target Detection., in: BMVC. Northumbria, p. 253.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: fast and flexible image augmentations. Information 11, 125.

Chen, C., Liu, M.-Y., Tuzel, O., Xiao, J., 2017. R-CNN for small object detection, in: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13. Springer, pp. 214–230.

Chen, Z., Wu, K., Li, Y., Wang, M., Li, W., 2019. SSD-MSN: an improved multi-scale object detection network based on SSD. IEEE Access 7, 80622–80632.

Costa, J.C., Carvalho, C.H.S., Matiello, J.B., Almeida, S.R., Carvalho, S.P., Baliza, D.P., 2013. Comportamento agronômico de progênies e cultivares de cafeeiro com resistência específica à ferrugem. Coffee Sci. 1984-3909 8, 183–191.

Dai, J., Li, Y., He, K., Sun, J., 2016. R-fcn: Object detection via region-based fully convolutional networks. Adv. Neural Inf. Process. Syst. 29.

Fazuoli, L.C., Carvalho, A., Costa, W.M. da, Nery, C., Laun, C.R.P., Santiago, M., 1983. Avaliação de progênies e seleção no cafeeiro Icatu. Bragantia 42, 179–189.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

Jocher, G., Stoken, A., Jirka Borovec, NanoCode012, C., Changyu, L., Laughing, Tkianai, Adam Hogan, lorenzomammana, Y., AlexWang1900, Diaconu, L., Marc, Wanghaoyang0106, M., Doug, Fran-cisco Ingham, F., Guilhen, Hatovix, Poznansk, J., PetrDvoracek, Prashant, R., 2020. ultralytics/yolov5: v3.1 Bug Fixes and PerformanceImprovement. https://doi.org/https://doi.org/10.5281/zenodo.4154370

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., Cho, K., 2019. Augmentation for small object detection. arXiv Prepr. arXiv1902.07296.

Kuznichov, D., Zvirin, A., Honen, Y., Kimmel, R., 2019. Data augmentation for leaf segmentation and counting tasks in rosette plants, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. p. 0.

Liang, Y., Han, Y., Jiang, F., 2022. Deep Learning-based Small Object Detection: A Survey, in: Proceedings of the 8th International Conference on Computing and Artificial Intelligence. pp. 432–438.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014:

13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Liu, Y., Sun, P., Wergeles, N., Shang, Y., 2021. A survey and performance evaluation of deep learning methods for small object detection. Expert Syst. Appl. 172, 114602.

Majerowicz, N., Söndahl, M.R., 2005. Indução e diferenciação de gemas reprodutivas Coffea arabica L. Brazilian J. Plant Physiol. 17, 247–254.

Nogueira, Â.M., Carvalho, S.P. de, Bartholo, G.F., Mendes, A.N.G., 2005. Avaliação da maturação dos frutos de linhagens das cultivares Catuaí Amarelo e Catuaí Vermelho (Coffea arabica L.) plantadas individualmente e em combinações. Ciência e Agrotecnologia 29, 18–26.

Pang, Y., Cao, J., Wang, J., Han, J., 2019. JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. IEEE Trans. Inf. Forensics Secur. 14, 3322–3331.

Petek, M.R., Sera, T., Sera, G.H., Fonseca, I.C. de B., Ito, D.S., 2006. Seleção de progênies de Coffea arabica com resistência simultânea à mancha aureolada e à ferrugem alaranjada. Bragantia 65, 65–73.

Ramos, P.J., Avendano, J., Prieto, F.A., 2018. Measurement of the ripening rate on coffee branches by using 3D images in outdoor environments. Comput. Ind. 99, 83–95.

Ramos, P.J., Prieto, F.A., Montoya, E.C., Oliveros, C.E., 2017. Automatic fruit count on coffee branches using computer vision. Comput. Electron. Agric. 137, 9–22.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788. https://doi.org/10.1109/CVPR.2016.91

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv Prepr. arXiv1804.02767.

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7263–7271.

Reis, P., da Cunha, R., 2010. Café Arábica do plantio à colheita. Lavras.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection

with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Sousa, T.V., Caixeta, E.T., Alkimim, E.R., Oliveira, A.C.B., Pereira, A.A., Sakiyama, N.S., Zambolim, L., Resende, M.D.V., 2019. Early selection enabled by the implementation of genomic selection in Coffea arabica breeding. Front. Plant Sci. 9, 1934.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L., 2020. Measuring robustness to natural distribution shifts in image classification. Adv. Neural Inf. Process. Syst. 33, 18583–18599.

Toda, Y., Okura, F., Ito, J., Okada, S., Kinoshita, T., Tsuji, H., Saisho, D., 2020. Training instance segmentation neural network with synthetic datasets for crop seed phenotyping. Commun. Biol. 3, 1–12.

Van Etten, A., 2019. Satellite imagery multiscale rapid detection with windowed networks, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 735–743.

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 Prepr. arXiv2207.02696.

Yang, S., Zheng, L., He, P., Wu, T., Sun, S., Wang, M., 2021. High-throughput soybean seeds phenotyping with convolutional neural networks and transfer learning. Plant Methods 17, 50.

ARTIGO 3 – ESTIMAÇÃO DO TAMANHO AMOSTRAL E AVALIAÇÃO DE DADOS DE MATURAÇÃO EM C*offea Arabica.*

Artigo redigido conforme a NBR 6022 (ABNT, 2018) e formatado de acordo com o Manual da UFLA de apresentação de teses e dissertações.

RESUMO

O tamanho amostral influencia diretamente a precisão e a confiabilidade dos resultados obtidos em um estudo ou pesquisa. Ademais, a seleção criteriosa de genótipos com base em objetivos específicos é uma tarefa crucial nos programas de melhoramento genético de plantas, especialmente para os ciclos de maturação. Este estudo teve como objetivo verificar os erros associados a cada tamanho de amostra para estimar os estágios de maturação em experimentos de *Coffea arabica* e demonstrar uma alternativa de seleção e agrupamento de genótipos com base nos ciclos de maturação. Foram analisados dois experimentos, um com progênies segregantes e outro com 21 cultivares, no qual foram coletadas amostras no momento da colheita para determinar os estágios de maturação de cada parcela. Utilizando a abordagem bootstrap, foi possível verificar que um tamanho amostral de 500 ml de frutos é necessário para estimar a média da parcela com um erro de aproximadamente 5%. Além disso, a ferramenta *K-means* foi utilizada para analisar e agrupar os genótipos em três diferentes ciclos de maturação, permitindo seleções com base em objetivos específicos. Portanto, o estudo fornece informações valiosas para a coleta de dados e análises para maturação de frutos de *Coffea arabica*.

Palavras-chaves: K-means, Melhoramento de plantas, Bootstrap, Experimentação.

1 INTRODUÇÃO

No melhoramento genético do café arábica, é fundamental a avaliação de um grande número de características. Algumas das mais relevantes incluem: a produtividade, a resistência/tolerância a estresses bióticos e abióticos, o tamanho do grão, a uniformidade e o ciclo de maturação. O ciclo de maturação é particularmente importante, pois é essencial conduzir as seleções com base em objetivos específicos. O lançamento de cultivares que apresentem diferentes ciclos de maturação é fundamental para os produtores, pois permite escalonar a produção e maximizar a eficiência e lucratividade (CARVALHO, 2008).

Na fase reprodutiva, o cafeeiro pode apresentar várias floradas sendo uma principal, seguida de outras, cujo número depende das condições climáticas e da variabilidade genética da cultivar (MAJEROWICZ e SÖNDAHL, 2005). Em razão disso, durante a colheita do café, vários estágios de maturação dos frutos podem ser encontrados simultaneamente, como verdes, verde-cana, maduros (que pode ser vermelha ou amarela), passas e secos (PEZZOPANE *et al.*, 2003).

O conhecimento do comportamento dos genótipos em relação ao ciclo fenológico, como uniformidade de maturação e duração do ciclo (precoce, médio e tardio), e aos atributos agronômicos é essencial para subsidiar as pesquisas que são desenvolvidas visando o melhoramento genético. Uma prática comum na fenotipagem da maturação é a adoção de uma avaliação subjetiva (FAZUOLI *et al.*, 1983; PETEK *et al.*, 2006; SOUSA *et al.*, 2019) ou por meio de uma amostragem dos frutos, sendo essa realizada por meio da quantificação do número de frutos nos diferentes estágios de maturação (COSTA *et al.*, 2013; NOGUEIRA *et al.*, 2005). Porém, o tamanho da amostra para realizar a fenotipagem é determinado de forma empírica e seu tamanho ideal é desconhecido.

A amostragem é de extrema importância para o estudo de qualquer caractere, pois usualmente é impraticável o estudo de toda população. Assim, é necessário que estas amostras sejam representativas da população inteira, para que as conclusões sejam generalizadas e se possa ter uma estimativa dos parâmetros populacionais com maior precisão. Sabe-se que quanto maior é o tamanho amostral maior é a precisão e poder dos testes estatísticos. Contudo, aumentar o tamanho amostral por muitas vezes também aumenta os custos envolvidos no processo da tomada dos dados, portanto, tem de haver um balanço entre o tamanho amostral, a precisão e poder exigido nos testes estatísticos (ANDRADE, 2020).

Embora frequentemente se utiliza a porcentagem de frutos maduros como indicador ótimo da época de maturação, essa abordagem pode não ser suficiente para capturar a complexidade e a variabilidade dos estágios de maturação em programas de melhoramento genético. Isso porque, os programas de melhoramento trabalham com populações com grande variabilidade para os estágios de maturação, podendo conter genótipos precoces, médios e tardios na mesma população e que acabam sendo colhidos juntos. Portanto, é necessário utilizar abordagens mais precisas e holísticas para avaliar a maturação permitindo assim, selecionar genótipos com a época de maturação desejada.

A utilização de agrupamentos pode ser uma alternativa, uma vez que permite agrupar os genótipos para os diferentes ciclos de maturação. Isso pode ser realizado pelas técnicas de agrupamento estatístico, como o *K-means*, que permitem agrupar os genótipos de acordo com suas características de maturação. Dessa forma, os genótipos podem ser selecionados de acordo com o objetivo do pesquisador.

Mediante o exposto, o estudo teve por objetivo estabelecer o tamanho de amostra ideal na avaliação do caráter ciclo de maturação em *Coffea arabica* e verificar os erros associados em adotar cada tamanho, como também demonstrar que o método de agrupamento pode ser uma alternativa para auxiliar os pesquisadores na tomada de decisão acerca dos genótipos constituintes na população para avaliação desse caráter.

2 MATERIAL E MÉTODOS

2.1 Descrição dos experimentos

Dois experimentos distintos foram analisados no presente estudo, no primeiro foram avaliadas progênies da geração $F_{2:3}$ oriundas do primeiro ciclo de seleção recorrente do Programa de Melhoramento Genético do Cafeeiro (UFLA/EPAMIG). O experimento foi instalado no delineamento em blocos completos casualizados, contendo 45 progênies em três repetições, as parcelas foram implantadas com 10 plantas no espaçamento 3,5 x 0,7m. Desse experimento as 10 melhores progênies para produção de grãos foram selecionadas, para a realização do presente estudo.

Dessas progênies coletou-se uma amostra aleatória no momento da colheita contendo 1,8 litros de frutos de cada parcela, que foram separadas em seis sub-amostras, cada uma com 300 ml. Para avaliação da maturação foi realizada a contagem dos frutos por três avaliadores diferentes, os quais contabilizaram o número de frutos verdes (incluindo os verdes e verde-canas), maduros e secos (incluindo os passas e secos) para cada amostra. Para a estimação da porcentagem de maturação foi considerado a média dos três avaliadores.

Já o segundo experimento, constituiu de 21 cultivares instaladas em delineamento em blocos completos casualizados, com três repetições, as parcelas foram implantadas, oito plantas por parcela e espaçamento 3,6 x 0,7m. As cultivares utilizadas foram lançadas por diferentes empresas públicas (Tabela 1). Uma amostra aleatória de um litro de frutos foi coletada de cada parcela no momento da colheita, e essa amostra foi separada em 2 sub-amostras de 500 ml cada. Para a avaliação da maturação, foi adquirida uma imagem de cada amostra em uma plataforma de fenotipagem. Com essas imagens, a partir do procedimento desenvolvido no capitulo 1 foi possível calcular a porcentagem de frutos verdes (incluindo os verdes e verde-canas), maduros e secos (incluindo os passas e secos).

Tabela 1. Relação das cultivares lançadas por diferentes empresas públicas.

Epamig	Pro-Café	IAPAR	IAC + Epamig
Catiguá MG-1	Acauã Novo	IPR 100	Catuaí Vermelho IAC 62
Catiguá MG-2	Clone 312	IPR 102	Catuaí Amarelo IAC 99
Catiguá MG-3	Saíra II	IPR 103	Rubi MG 1190
Oeiras	Siriema		Topázio MG 1190
Paraíso	Guará		Travessia
Araponga			Mundo Novo IAC 379-19
Pau Brasil			

2.2 Simulação via bootstrap

Para determinar o tamanho ideal da amostra, foi utilizado o método de reamostragem *Bootstrap*. O *bootstrap* é uma técnica estatística para estimar a distribuição amostral de um estimador, amostrando com reposição a partir da amostra original, geralmente com o objetivo de obter estimativas robustas de erros padrão e intervalos de confiança de um parâmetro populacional. Esta técnica é especialmente útil quando a amostra original é pequena ou quando a distribuição subjacente é desconhecida. É uma ferramenta poderosa, flexível e intuitiva para inferência estatística (EFRON, 1992).

No presente estudo, foram realizadas 1.000 reamostragens com reposição para cada tamanho amostral, e em seguida, foi estimado o erro associado a estimativa da porcentagem de maturação a um intervalo de confiança de 95%. Cada parcela experimental foi considerada como uma amostra independente. O tamanho das amostras testadas variou de 20 a 900 frutos.

2.3 Análise estatística

Os dados foram analisados por meio da abordagem de modelos mistos considerando cada estágio de maturação como uma variável independente. Os componentes de variância foram estimados a partir da máxima verossimilhança restrita (REML) usando o algoritmo de expectativa-maximização (EM), e a predição dos valores genéticos via melhor preditor linear não-viesado (E-BLUP).

O modelo utilizado foi:

$$y = X\beta + Z\gamma + e$$

Em que: $y_{(nx1)}$ é o vetor das observações fenotípica; $\beta_{(bx1)}$ é o vetor do efeito fixo das repetições; $\gamma_{(gx1)}$ é vetor do efeito aleatório dos genótipos; $e_{(gx1)}$ é o vetor de erros; $X_{(nxb)}$ é a matriz de incidência dos blocos; $Z_{(nxg)}$ é a matriz de incidência dos efeitos de genótipos.

Após a obtenção dos E-BLUPs para cada genótipo assim como cada variável, utilizou-se do algoritmo de *K-means* para a classificação dos genótipos quanto ao status de maturação. *K-means* é um algoritmo de agrupamento não supervisionado de dados, com objetivo de separar um conjunto de dados em k grupos (k é um número especificado previamente) de maneira que os grupos sejam similares entre si (ABBAS, 2008). Para o presente estudo, k foi especificado como três, em que teoricamente os genótipos são agrupados em precoce, normal e tardio com base na informação de porcentagem de verdes, maduros e secos.

O algoritmo funciona inicializando k centroides aleatoriamente em que cada centroide representa o centro de um grupo. Em seguida, os pontos de dados são atribuídos ao grupo cujo centroide esteja mais próximo, com base em uma métrica estatística, sendo que, no presente estudo, utilizou-se da distância euclidiana. Os centroides então são recalculados com a média de todos os pontos do grupo, e o processo é repetido até que os grupos não alterem ou o número máximo de iterações seja atingido (ABBAS, 2008).

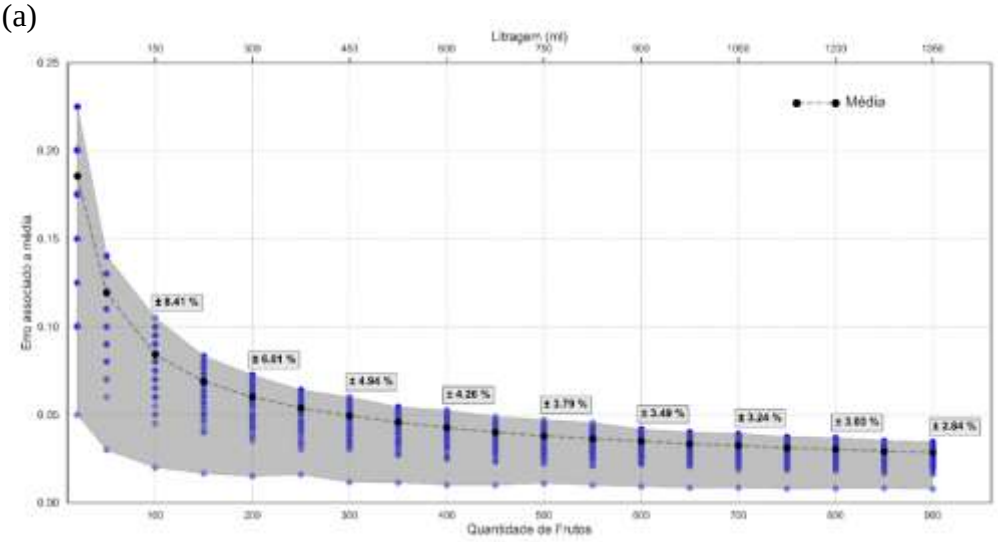
Utilizou-se do software Python, e a utilização da biblioteca *sklearn* para execução do algoritmo *K-means* (PEDREGOSA *et al.*, 2011), adicionalmente para criação das figuras utilizou-se do *matplotlib.pyplot* (HUNTER, 2007). E as análises estatísticas, foram realizadas no software R, com a utilização da biblioteca *sommer* (COVARRUBIAS-PAZARAN, 2018).

3 RESULTADOS

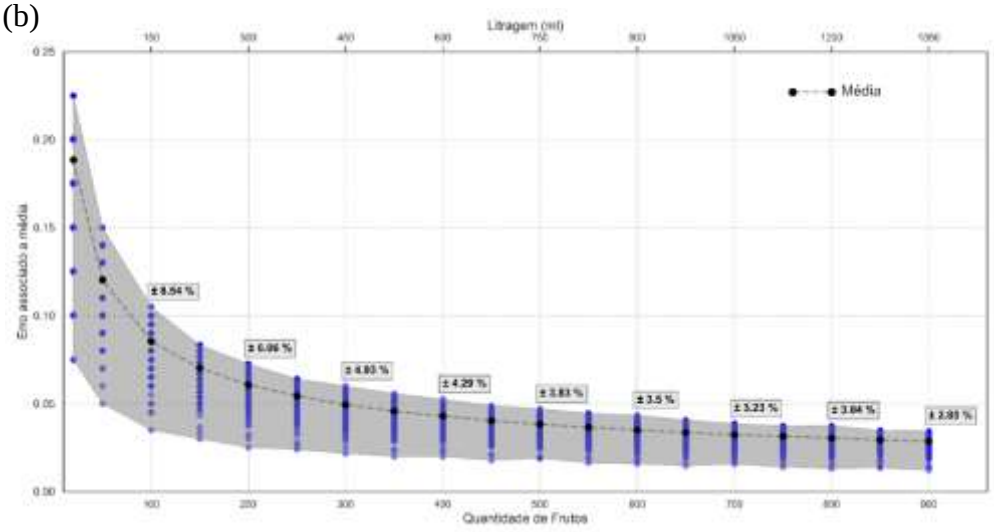
Em ambos os experimentos os resutados foram semelhantes. Observou-se uma redução considerável do erro associado até a avaliação com 300 frutos (±450 ml) e, após isso, a média do erro associado se tornou menor com o aumento do tamanho amostral. Ao adotar um tamanho amostral de 300 frutos pode-se estar cometendo, em média, um erro na estimativa da maturação de mais ou menos 5% (Figura 1). Como exemplo, para uma parcela que possui 50% de frutos no estágio maduro, pode-se afirmar que a média real estaria entre 45% e 55% em média.

Figura 1. Estimativa de erro associado a estimativa da média a 95% de probabilidade para os estágios de maturação (a) progênies e (b) cultivares. Cada ponto representa uma amostra. Os valores descritos nas caixas representam o erro médio associado médio ao tamanho amostral correspondente.

(a)



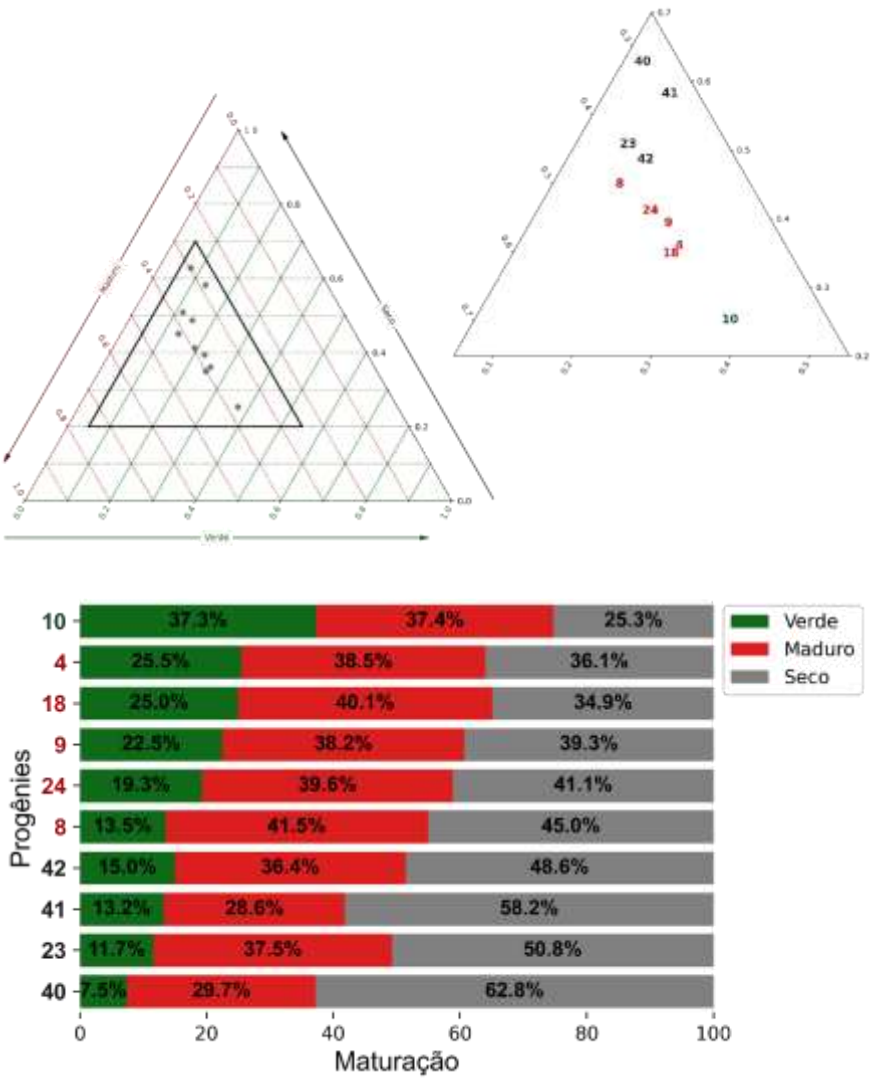
(b)



Fonte: Do autor (2023)

Para as progênies, a média para o estágio de maturação variou de 7,5% a 37,3% para os verdes, 28,6% a 41,5% para os maduros e de 18,9% a 40,1% para os secos (Figura 2). Com base na porcentagem de frutos em cada estágio pode-se classificar as progênies em grupos tardio, médio e precoce. A progênie 10 ficou no grupo tardio, uma vez que apresentou a maior porcentagem de frutos no estágio verde (40,2%). As progênies 4, 8, 9, 18 e 24 ficaram no grupo médio, sendo que a progênie 8 foi a que apresentou mais frutos no estágio maduro (41,5%) e a progênie 9 com menor porcentagem (38,2%). As progênies 23, 40, 41 e 42 ficaram no grupo precoce, em que a progênie 40 teve maior porcentagem de frutos no estágio seco (62,8%) e a 42 menor porcentagem (48,6%) (Figura 2).

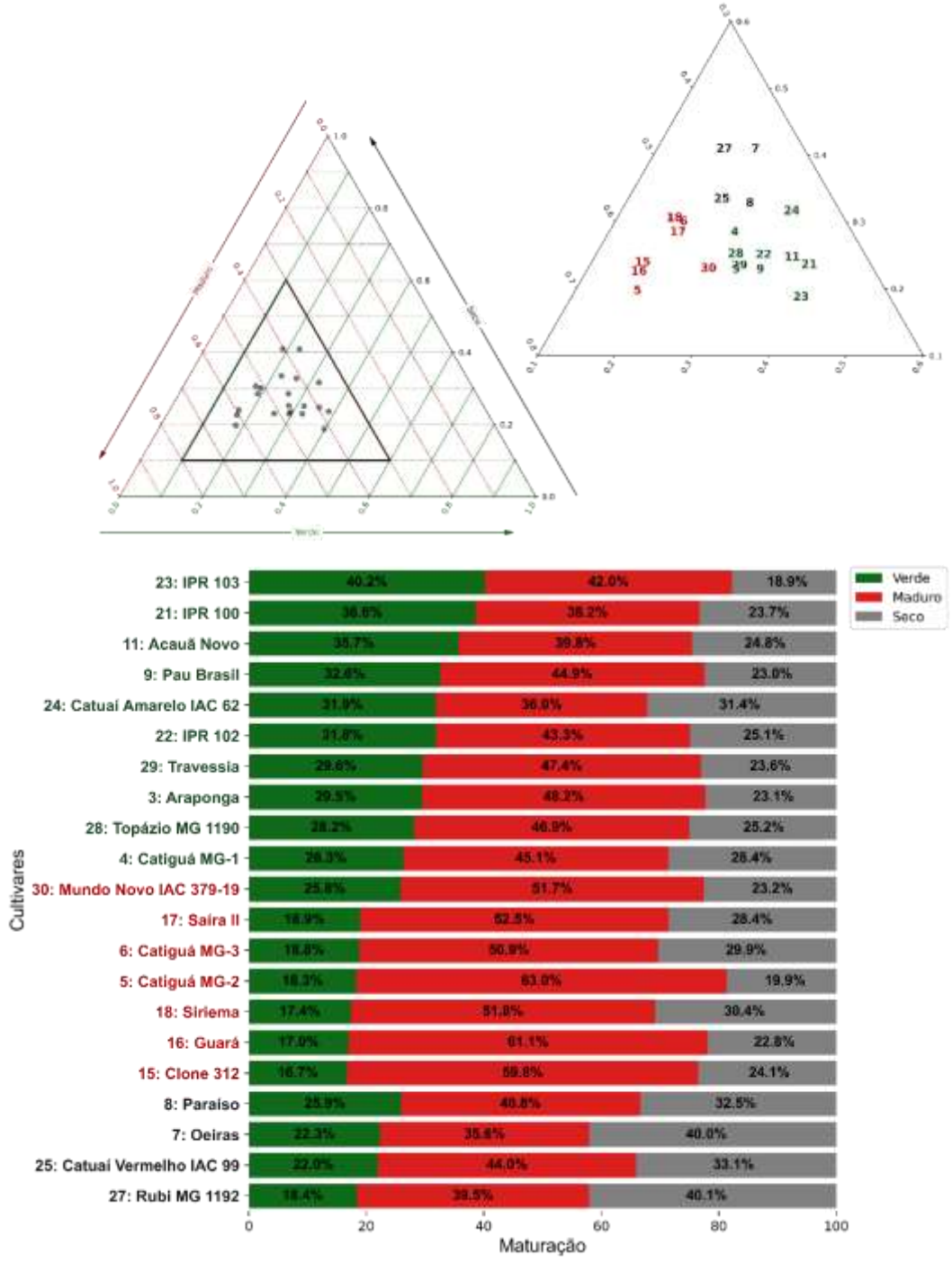
Figura 2. Análise das progênies estudadas quanto aos estágios de maturação, as cores representam os agrupamentos realizado pelo algoritmo de *K-means*.



Fonte: Do autor (2023)

Para as cultivares a média para o estágio de maturação variou de 16,7% a 40,2% para os verdes, 35,6% a 63% para os maduros e de 25,3% a 62,8% para os secos (Figura 3). Para as cultivares tem-se 10 cultivares que foram classificadas como tardia, sete como médias e quatro cultivares precoces. Nota-se que dentre as cultivares classificadas como tardia a cultivar IPR 103 foi a que mais teve frutos no estágio verde (40,2%) e a Catiguá MG-1 com menor porcentagem de frutos verdes (26,3%). Para as cultivares classificadas como época de maturação média a que apresentou maior porcentagem de frutos no estágio de maturação maduro foi a cultivar Catiguá MG-2 (63,0%) e a com menor foi a Catiguá MG-3 (50,9%). Já para as cultivares tidas como precoce as com maior porcentagem de frutos no estágio seco foi a Rubi MG 1192 (40,1%) e a com foi a Paraíso (32,5%) (Figura 3).

Figura 317. Análise das cultivares estudadas quanto aos estágios de maturação, as cores representam os agrupamentos realizado pelo algoritmo de *K-means*.



Fonte: Do autor (2023)

4 DISCUSSÃO

4.1 Tamanho Amostral

É comumente utilizado na fenotipagem para época de maturação a determinação subjetiva dos graus de maturação (PETEK *et al.*, 2006; SOUSA *et al.*, 2019). Contudo, esse procedimento não é adequado, pois, como já mencionado a avaliação se torna subjetiva e pequenas diferenças entre as progênies podem não serem detectadas. Com isso, a melhor maneira de avaliar o caráter é quantificar os frutos quanto ao grau de maturação em que esses estão, por meio de uma amostra representativa. Diversos trabalhos que estudaram a maturação dos frutos, adotaram diferentes tamanhos amostrais, que vão desde de 300 ml (COSTA *et al.*, 2013) a 1000 ml (NOGUEIRA *et al.*, 2005)

Determinar o tamanho de amostra ideal é crucial para avaliar um caráter de interesse. Diferentes abordagens podem ser usadas para essa determinação, no presente estudo utilizou-se da estimativa do erro associado a maturação por meio do método de *bootstrap*. Mas outros autores já utilizaram de outras abordagens como estimação de herdabilidade e coeficiente de variação experimental e suas mudanças com o tamanho amostral (COLOMBARI FILHO; GERALDI, 2014; DINIZ; PINTO; LAMBERT, 2006; KIM *et al.*, 2020). A abordagem utilizada neste estudo permitiu ter uma noção clara do erro associado a cada tamanho amostral, pois todos os outros parâmetros estimados dependem do valor fenotípico da parcela.

Observa-se que tanto para as cultivares quanto para as progênies o erro associado as estimativas de maturação foram bastante semelhantes entre si, para os diferentes tamanhos amostrais, esse fato é bastante notório e intrigante visto que as plantas das parcelas das progênies na geração $F_{2:3}$ segregam e que, portanto, pode-se ter diferente expressão de maturação entre as plantas da mesma parcela. Contudo, mesmo com a segregação dentro da parcela não se notou diferenças significativas quando se comparou com o experimento de cultivares, na qual não há segregação.

Quando a avaliação da maturação é realizada de forma manual, ou seja, contabilizando os frutos em cada estágio de maturação, o tamanho amostral pode ser um fator limitante, visto que, quanto maior o tamanho amostral mais oneroso a fenotipagem se torna. Além disso, é importante ter em mente que ao aumentar o tamanho amostral, aumenta-se também a chance de haver erros aleatórios associado a contagem dos avaliadores. Portanto, deve-se encontrar um equilíbrio entre o tamanho amostral e a precisão na avaliação quando essa é realizada de forma manual.

A utilização de técnicas de análise de imagem, como a visão computacional, para determinação da maturação, oferece uma abordagem precisa e eficiente para a avaliação dos frutos. Isso porque essas técnicas são capazes de capturar informações detalhadas sobre as características fenotípicas dos frutos de forma rápida e precisa, superando os desafios inerentes à avaliação manual. Além disso, as imagens geradas podem ser armazenadas e utilizadas no futuro ou até mesmo avaliar outros caracteres de interesse. A automação dos processos de avaliação reduz a margem de erro humano e aumenta a precisão e eficiência do processo, além de permitir aumento do tamanho amostral sem aumentar os custos da avaliação (BELAN; DE ARAÚJO; LIBRANTZ, 2012).

A partir da adoção de um tamanho amostral de 500 ml, foi possível observar que o erro associado à estimativa de maturação foi de aproximadamente 5%, o que indica um tamanho de amostra adequado para avaliação do caráter. Contudo, caso haja a possibilidade de analisar as amostras via imagem o tamanho amostral pode ser maior. Para obter uma estimativa de maturação com um erro associado de 1%, é necessário um número aproximado de 1300 frutos ou aproximadamente 2 litros. Isso proporcionaria maior confiança na estimativa da maturação dos genótipos.

4.2 Agrupamento

Nos programas de melhoramento, diversas progênies são avaliadas e podem apresentar diferenças significativas em relação às características avaliadas. A classificação dos genótipos quanto ao grau de maturação, é dependente da época em que os frutos são colhidos, e deve ser levado em consideração quando os genótipos são separados em grupos de maturação. O algoritmo de *K-means* pode ser uma alternativa pois é uma técnica utilizada para agrupar dados, e possibilita o agrupamento dos genótipos em classes de maturação independentemente da época de colheita, além do pesquisador conseguir enxergar quais genótipos mais se adequam a suas premissas, podendo esse selecionar os genótipos dentro dos agrupamentos que são formados, dependendo do seu objetivo.

Existem vários tipos de algoritmos de agrupamento, cada um com suas próprias características e aplicações. Alguns dos mais populares incluem o agrupamento hierárquico, *K-means* e o DBSCAN (Density-Based Spatial Clustering of Applications with Noise). O agrupamento hierárquico cria hierarquias de grupos de dados pela aglutinação ou divisão iterativa (MURTAGH; CONTRERAS, 2017). O DBSCAN é um algoritmo baseado em densidade que forma grupos com base na densidade dos pontos de dados no espaço (ESTER *et*

al., 1996). Cada um desses algoritmos tem suas vantagens e desvantagens, adotou-se *K-means* pela sua popularidade, simplicidade, escalabilidade e resultados interpretáveis.

Os agrupamentos de dados têm muitas aplicações no melhoramento de plantas. Alguns exemplos são, o uso do algoritimo de *K-means* para agrupar genótipos de um banco de germoplasma de soja com base em suas características fotossintéticas (SHAMIM *et al.*, 2022), utilização do algoritmo DBSCAN no auxílio a fenotipagem 3D em plantas de soja (WAN; WANG, 2020), e o uso do agrupamento hierárquico utilizado para agrupar indivíduos de trigo com base em suas características genéticas e fenotípicas (BONMAN *et al.*, 2015). Esses estudos mostram como os algoritmos de agrupamento podem ser usados para identificar grupos de genótipos ou indivíduos com características desejadas, o que é uma ferramenta valiosa para os programas de melhoramento de plantas.

A determinação do número de agrupamentos é sempre um desafio na abordagem de *K-means* e às vezes é atribuído de acordo com o número de classes presentes no conjunto de dados (ABBAS, 2008; JAIN, 2010). No entanto, esse não é o caso quando separa a época de maturação, em que o número de agrupamentos é determinado pelo número de classes de maturação que o pesquisador deseja.

No presente estudo, foram utilizados três agrupamentos para discriminar os genótipos em teoricamente precoce, médio e tardio. No entanto, é possível adotar um número maior de agrupamentos para obter uma discriminação ainda maior entre os genótipos. Por exemplo, caso seja do interesse do pesquisador categorias adicionais, como super precoce e super tardio, podem ser incorporadas. A abordagem por agrupamentos permite que isso seja feito.

Nos experimentos observa-se claramente uma manifestação fenotípica distinta para o caráter, assim a avaliação dos genótipos baseada somente na porcentagem de frutos vermelhos pode não ser a melhor opção, pois ela não leva em conta a variabilidade dos diferentes estágios de maturação dos genótipos. Portanto, a consideração dos diferentes estágios de maturação pode ser mais adequada. Isso permite uma avaliação mais precisa dos genótipos e uma seleção mais eficiente.

Para todos os genótipos dos dois experimentos, não foi obtida porcentagem maior que 63% de frutos no mesmo estágio de maturação, esses valores estão aquém dos preconizados como ótimos, que vão de 80 a 85% (BARTHOLO; GUIMARÃES, 1997). No entanto, não é objetivo tirar conclusões acerca dos genótipos estudados quanto ao grau de maturação, pois eles

foram avaliados somente durante uma colheita e, apesar da maturação dos frutos ser controlada geneticamente, ela é fortemente influenciada pelas condições edafoclimáticas e microclimáticas regionais (FAZUOLI *et al.*, 2002).

5 CONCLUSÃO

Amostras superiores a 500 ml de frutos, apresentou erro associado de aproximadamente 5% ou menos, valor esse, considerado aceitável para avaliar o caráter maturação. O uso da técnica de *K-means* para agrupar os dados nos diferentes ciclos de maturação pode ser uma excelente alternativa para os pesquisadores, permitindo uma análise e tomada de decisão precisa e eficiente.

REFERÊNCIAS

ABBAS, O. A. Comparisons between data clustering algorithms. International Arab Journal of Information Technology (IAJIT), v. 5, n. 3, 2008.

ANDRADE, C. Sample Size and Its Importance in Research. Indian journal of psychological medicine, v. 42, n. 1, p. 102–103, 6 jan. 2020. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/31997873>.

BARTHOLO, G. F.; GUIMARÃES, P. T. G. Cuidados na colheita e preparo do café. Informe Agropecuário, v. 18, n. 187, p. 33–42, 1997.

BELAN, P. A.; DE ARAÚJO, S. A.; LIBRANTZ, A. F. H. Técnicas de visão computacional aplicadas no processo de calibração de instrumentos de medição com display numérico digital sem interface de comunicação de dados. Exacta, v. 10, n. 1, p. 82–91, 2012.

BONMAN, J. M.; BABIKER, E. M.; CUESTA-MARCOS, A.; ESVELT-KLOS, K.; BROWN-GUEDIRA, G.; CHAO, S.; SEE, D.; CHEN, J.; AKHUNOV, E.; ZHANG, J. Genetic diversity among wheat accessions from the USDA National Small Grains Collection. Crop Science, v. 55, n. 3, p. 1243–1253, 2015.

CARVALHO, C. H. S. de. Cultivares de café: origem, características e recomendações. Brasília: Embrapa Café, v. 334, 2008.

COLOMBARI FILHO, J. M.; GERALDI, I. O. Sample size for the assessment of soybean inbred populations. Crop Breeding and Applied Biotechnology, v. 14, p. 71–75, 2014.

COSTA, J. C.; CARVALHO, C. H. S.; MATIELLO, J. B.; ALMEIDA, S. R.; CARVALHO, S. P.; BALIZA, D. P. Comportamento agronômico de progênies e cultivares de cafeeiro com resistência específica à ferrugem. Coffee Science-ISSN 1984-3909, v. 8, n. 2, p. 183–191, 2013.

COVARRUBIAS-PAZARAN, G. Software update: moving the R package sommer to multivariate mixed models for genome-assisted prediction. BioRxiv, p. 354639, 2018.

DINIZ, M. C. D. R.; PINTO, C. A. B. P.; LAMBERT, E. de S. Sample size for family evaluation in potato breeding programs. Ciência e Agrotecnologia, v. 30, p. 277–282, 2006.

EFRON, B. Bootstrap methods: another look at the jackknife. In: Breakthroughs in statistics. [s.l.] Springer, 1992. p. 569–593.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd, 34., 1996, [...]. 1996. v. 96, p. 226–231.

FAZUOLI, L. C.; CARVALHO, A.; COSTA, W. M. da; NERY, C.; LAUN, C. R. P.; SANTIAGO, M. Avaliação de progênies e seleção no cafeeiro Icatu. Bragantia, v. 42, p. 179–189, 1983.

FAZUOLI, L. C.; MEDINA FILHO, H. P.; GONÇALVES, W.; GUERREIRO FILHO, O.; SILVAROLLA, M. B. Melhoramento do cafeeiro: variedades tipo arábica obtidas no Instituto

Agronômico de Campinas. O estado da arte de tecnologias na produção de café, p. 253–287, 2002.

HUNTER, J. D. Matplotlib: A 2D graphics environment. Computing in science & engineering, v. 9, n. 03, p. 90–95, 2007.

JAIN, A. K. Data clustering: 50 years beyond K-means. Pattern recognition letters, v. 31, n. 8, p. 651–666, 2010.

KIM, D.-G.; LEE, S.-H.; CHO, B.-K.; BYEON, D.; LEE, J.; LEE, W.-H. Statistical analysis for determining optimal sample size for living modified organism (LMO) seed detection. Journal of Crop Science and Biotechnology, v. 23, n. 1, p. 1–7, 2020.

MAJEROWICZ, N.; SÖNDAHL, M. R. Indução e diferenciação de gemas reprodutivas Coffea arabica L. Brazilian Journal of Plant Physiology, v. 17, p. 247–254, 2005.

MURTAGH, F.; CONTRERAS, P. Algorithms for hierarchical clustering: an overview, II. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 7, n. 6, p. e1219, 2017.

NOGUEIRA, Â. M.; CARVALHO, S. P. de; BARTHOLO, G. F.; MENDES, A. N. G. Avaliação da maturação dos frutos de linhagens das cultivares Catuaí Amarelo e Catuaí Vermelho (Coffea arabica L.) plantadas individualmente e em combinações. Ciência e Agrotecnologia, v. 29, p. 18–26, 2005.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, v. 12, p. 2825–2830, 2011.

PETEK, M. R.; SERA, T.; SERA, G. H.; FONSECA, I. C. de B.; ITO, D. S. Seleção de progênies de Coffea arabica com resistência simultânea à mancha aureolada e à ferrugem alaranjada. Bragantia, v. 65, p. 65–73, 2006.

PEZZOPANE, J. R. M.; PEDRO JÚNIOR, M. J.; THOMAZIELLO, R. A.; CAMARGO, M. B. P. de. Escala para avaliação de estádios fenológicos do cafeeiro arábica. Bragantia, v. 62, p. 499–505, 2003.

RIBEIRO, A. C.; GUIMARÃES, P. T. G.; ALVAREZ, V. H. 5a aproximação. Viçosa: CFSEMG, p. 25–32, 1999.

SHAMIM, M. J.; KAGA, A.; TANAKA, Y.; YAMATANI, H.; SHIRAIWA, T. Analysis of Physiological Variations and Genetic Architecture for Photosynthetic Capacity of Japanese Soybean Germplasm. Frontiers in Plant Science, v. 13, 2022.

SOUSA, T. V.; CAIXETA, E. T.; ALKIMIM, E. R.; OLIVEIRA, A. C. B.; PEREIRA, A. A.; SAKIYAMA, N. S.; ZAMBOLIM, L.; RESENDE, M. D. V. Early selection enabled by the implementation of genomic selection in Coffea arabica breeding. Frontiers in Plant Science, v. 9, p. 1934, 2019.

WAN, S.; WANG, Y.-P. The comparison of density-based clustering approach among different machine learning models on paddy rice image classification of multispectral and hyperspectral image data. Agriculture, v. 10, n. 10, p. 465, 2020.