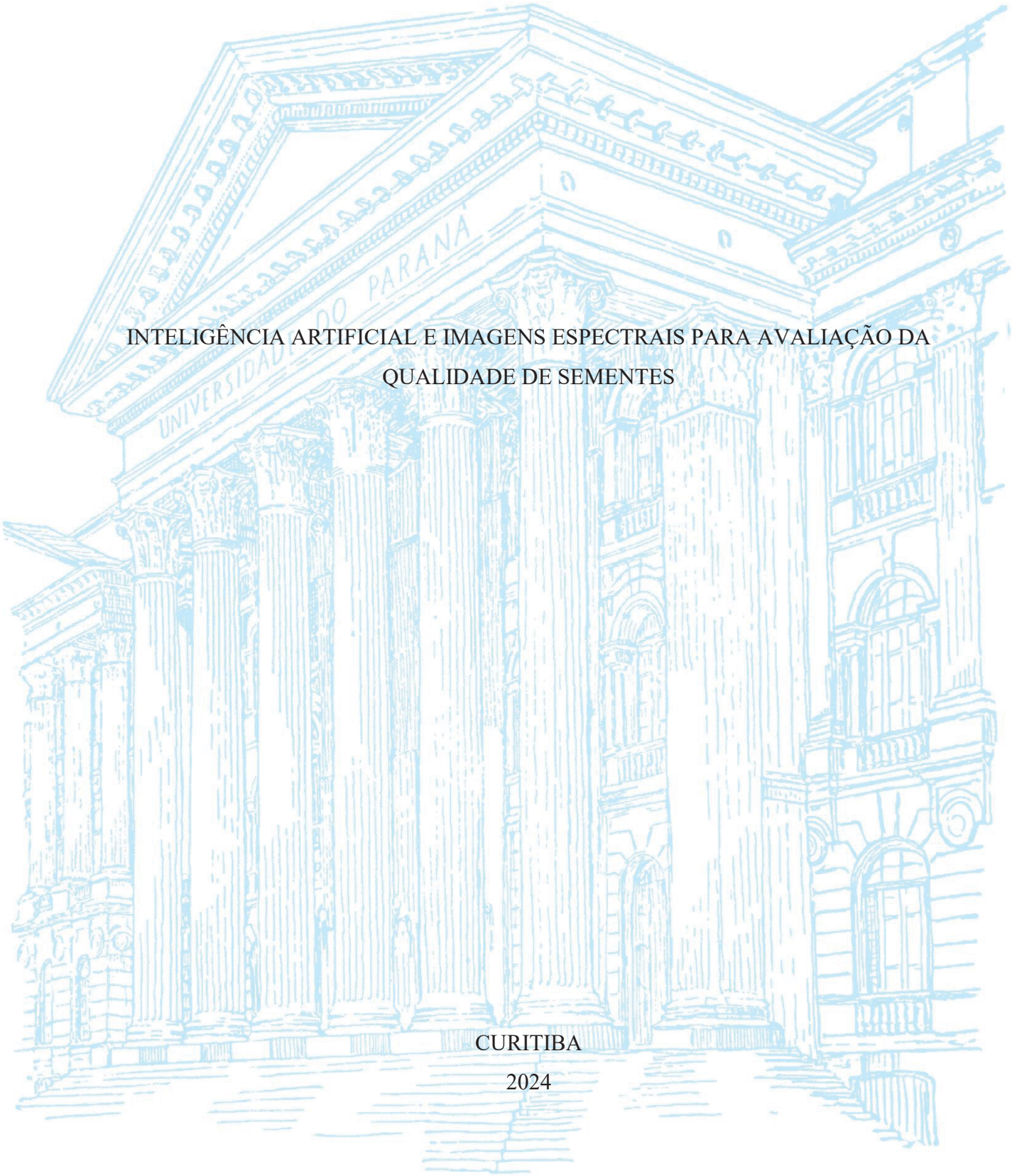
UNIVERSIDADE FEDERAL DO PARANÁ

THOMAS BRUNO MICHELON



INTELIGÊNCIA ARTIFICIAL E IMAGENS ESPECTRAIS PARA AVALIAÇÃO DA QUALIDADE DE SEMENTES

CURITIBA

2024

THOMAS BRUNO MICHELON

INTELIGÊNCIA ARTIFICIAL E IMAGENS ESPECTRAIS PARA AVALIAÇÃO DA QUALIDADE DE SEMENTES

Tese apresentada ao curso de Pós-Graduação em Agronomia, Área de concentração em Produção Vegetal, Área de Concentração em Produção Vegetal, Departamento de Fitossanidade e Fitotecnia, Setor de Ciências Agrárias, Universidade Federal do Paraná, como parte das exigências para obtenção do título de Doutor em Ciências.

Orientadora: Profª Drª Maristela Panobianco Vasconcellos

Coorientadora: Drª Elisa Serra Negra Vieira

Coorientadora: Drª Bárbara Blanco-Ulate

CURITIBA

2024

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS AGRÁRIAS

Michelon, Thomas Bruno
Inteligência artificial e imagens espectrais para avaliação da qualidade de sementes/ Thomas Bruno Michelon. – Curitiba, 2024.
1 recurso online: PDF.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Agrárias, Programa de Pós-Graduação em Agronomia (Produção Vegetal).
Orientadora: Profª Maristela Panobianco Vasconcellos
Coorientadora: Drª Elisa Serra Negra Vieira
Coorientadora: Drª Bárbara Blanco-Ulate

1. Análise espectral. 2. Inteligência artificial. 3. Sementes. 4. Sementes - Qualidade. I. Vasconcellos, Maristela Panobianco. II. Vieira, Elisa Serra Negra. III. Blanco-Ulate, Bárbara. IV. Universidade Federal do Paraná. Programa de Pós-Graduação em Agronomia (Produção Vegetal). V. Título.

Bibliotecária: Telma Terezinha Stresser de Assis CRB-9/944

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS AGRÁRIAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO AGRONOMIA
(PRODUÇÃO VEGETAL) - 40001016031P6

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação AGRONOMIA (PRODUÇÃO VEGETAL) da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **THOMAS BRUNO MICHELON** intitulada: **INTELIGÊNCIA ARTIFICIAL E IMAGENS ESPECTRAIS PARA AVALIAÇÃO DA QUALIDADE DE SEMENTES**, sob orientação da Profa. Dra. MARISTELA PANOBIANCO VASCONCELLOS, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 22 de Fevereiro de 2024.

Assinatura Eletrônica
26/02/2024 10:49:48.0
MARISTELA PANOBIANCO VASCONCELLOS
Presidente da Banca Examinadora

Assinatura Eletrônica
23/02/2024 12:11:45.0
ADRIANA MARTINELLI SENEME
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
23/02/2024 14:08:37.0
ELISA SERRA NEGRA VIEIRA
Avaliador Externo (EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA)

Assinatura Eletrônica
27/02/2024 13:34:32.0
FRANCISCO GUILHIEN GOMES JUNIOR
Avaliador Externo (ESCOLA SUPERIOR DE AGRICULTURA - LUIZ DE QUEIROZ -USP)

Essa tese é dedicada aos meus pais, Amilton Otávio Michelon e Maria Terezinha Della Riva Michelon, minha irmã Talita Gaby Michelon e à nova membra, minha sobrinha, Melina Michelon de Bortoli. Vocês são minha base.

AGRADECIMENTOS

Aos meus pais, Amilton Otávio Michelon e Maria Terezinha Della Riva Michelon, por sempre me incentivarem e me apoiarem nos meus estudos. Vocês têm meu agradecimento, admiração e amor.

À minha namorada Natália Menezello, por sempre estar comigo para me apoiar tanto nas aventuras quanto nos momentos difíceis.

À minha orientadora, Profª Drª Maristela Panobianco, pela sua orientação dedicada ao longo desses anos. Por ser uma ótima professora e estar sempre presente oferecendo apoio e compartilhando seu conhecimento. Obrigado por ter acreditado em mim e por me ajudar a crescer pessoal e profissionalmente.

À minha co-orientadora, Drª Elisa Serra Negra Vieira por sua generosidade em compartilhar seu conhecimento durante essa etapa e pela sabedoria de seus conselhos, fundamentais para o meu desenvolvimento.

À Profª Drª Bárbara Blanco-Ulate e ao MSc. Pedro Bello, por me receberem e orientarem na Universidade da Califórnia, Davis (UC Davis), permitindo assim a produção do capítulo três desta tese.

Ao Prof. Dr. Fushing Hisieh, por me orientar durante a etapa de análise de dados e pelas incontáveis reuniões que contribuíram imensamente para o capítulo três desta tese.

À Drª Andreza, pelos anos de amizade durante as etapas de mestrado e doutorado. Sua energia, entusiasmo e positividade foram fundamentais para tornar essa jornada mais agradável e produtiva.

Ao pesquisador associado da UC Davis, Adrian Sbodio, pela ajuda durante o período de pesquisa, a ótima amizade e pelos diversos cafés. Também, um agradecimento aos colegas de laboratório, cuja a amizade contribuiu imensamente durante o período no exterior: Isabel, Saskia, Elia, Yidou, Jaclyn.

Aos amigos Christian e Eva durante o período na Califórnia, por estarem sempre dispostos a nos receber e por propiciarem ótimos momentos juntos.

Ao Simone Piancatelli, pelos ótimos momentos juntos durante o período de internacionalização.

À Empresa Brasileira de Pesquisa Agropecuária - Embrapa, especificamente à Embrapa Florestas, por permitir o uso de suas instalações e laboratórios, que contribuíram muito para o decorrer da pesquisa realizada.

À empresa Videometer, especialmente aos proprietários Jens Michael e Nette Schultz, pelo acolhimento e pelo conhecimento técnico fornecido. Suas contribuições durante o período de internacionalização foram primordiais para a viabilização do segundo capítulo desta tese.

À Universidade da Califórnia, Davis (UC Davis), pelo uso das instalações e laboratórios.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela bolsa de estudos de doutorado e ao Conselho Nacional de Pesquisa e Desenvolvimento (CNPq), pela bolsa de doutorado-sanduíche, que possibilitou minha pesquisa na Universidade da Califórnia, Davis.

Aos demais que contribuiram com essa etapa.

RESUMO

A análise de imagem espectral é considerada a principal tecnologia emergente na avaliação da qualidade de sementes pois é capaz de capturar informações de diferentes naturezas, como espectrais, morfológicas e de autofluorescência. Assim, a tecnologia tem sido explorada com sucesso na indústria de sementes, visando complementar ou até substituir análises tradicionais, que são frequentemente demoradas, subjetivas e resultam na perda da semente. O objetivo principal dessa tese foi investigar a aplicação da análise de imagem espectral combinada com métodos de análise de dados multivariados na avaliação da qualidade de sementes. Os objetivos específicos de cada capítulo foram: (1) avaliar os principais procedimentos relacionados à análise de imagens espectrais e procedimentos quimiométricos aplicados na fenotipagem de sementes, bem como a sua aplicação prática; (2) identificar o potencial da análise de imagem espectral na distinção entre sementes híbridas de *Eucalyptus urograndis* (*Eucalyptus grandis* × *Eucalyptus urophylla*) e *Corymbia maculata* × *Corymbia torelliana* dos seus progenitores; e (3) avaliar a relação entre a respiração de sementes de soja e suas características biométricas do uso de imagens multiespectrais. Para avaliar o uso da técnica de imagens espectrais na fenotipagem de sementes, uma revisão sistemática baseada na metodologia PRISMA foi realizada. Um total de 1304 artigos foram inicialmente avaliados e 44 artigos foram selecionados conforme os critérios estipulados. Os resultados indicaram que a análise possui alta capacidade (93,33%) para classificar genótipos de sementes, incluindo cultivares, híbridos interespecíficos, progenitores e linhagens. Em relação à distinção de sementes híbridas florestais, foram realizados quatro experimentos com dois lotes separados e um combinado de sementes de *Eucalyptus urograndis* e um lote de *Corymbia maculata* × *Corymbia torelliana* e seus progenitores. Imagens multiespectrais foram capturadas e características espectrais e morfológicas foram extraídas. Algoritmos SVM, LDA e RF, foram utilizados para compor os modelos de classificação das sementes. O algoritmo LDA, combinado com características morfo-espectrais das sementes, foi o mais eficaz para ambos os gêneros com acurácia de 98,15% para as sementes *Corymbia* spp. e 92,75%, 85,38% e 86,00% para cada um dos lotes de *Eucalyptus* spp. e para eles misturados, respectivamente. A técnica se mostrou eficaz para a separação de sementes híbridas de *Corymbia* spp. e *Eucalyptus* spp. no contexto de programas de melhoramento florestal. Para o experimento com sementes de soja, 1806 sementes de seis lotes diferentes foram avaliadas. Imagens multiespectrais seguidas pela medição individual do consumo de oxigênio das sementes durante a germinação foram realizadas. Ao todo, 2775 pares de 75 medidas biométricas foram analisadas. Ambas as medidas de respiração e biometria foram categorizadas e associadas usando tabelas de contingência e análise de entropia. Os resultados revelaram diferenças nos padrões de respiração, especialmente em autofluorescência (365/600 nm, 430/700 nm, 450/700 nm, e 470/700 nm) e refletância (365 nm, 690 nm, e 405 nm). As características da semente de soja, em especial, sua informação espectral, estão fortemente correlacionadas com a respiração e qualidade da semente, e a análise de imagem espectral é uma ferramenta eficaz e não invasiva para sua avaliação.

Palavras-chave: *Glycine max,* análise multivariada de dados; análise de imagem multiespectral; aprendizado de máquina; *Eucalyptus* spp; *Corymbia* spp.

ABSTRACT

Spectral image analysis is considered the main emerging technology in seed quality assessment as it is capable of capturing information of different natures, such as spectral, morphological and autofluorescence information. The technology has therefore been successfully used in the seed industry to complement or even replace traditional analyses, which are often time-consuming, subjective and result in seed loss. The main objective of this thesis was to investigate the application of spectral image analysis combined with multivariate data analysis methods in seed quality assessment. The specific objectives of each chapter were: (1) to evaluate the main procedures related to spectral image analysis and chemometric procedures applied in seed phenotyping, as well as their practical application; (2) to identify the potential of spectral image analysis in distinguishing hybrid seeds of *Eucalyptus urograndis* (*Eucalyptus grandis* × *Eucalyptus urophylla*) and *Corymbia maculata* × *Corymbia torelliana* from their parents; and (3) to assess the relationship between soybean seeds respiration and its biometric features through multispectral imaging. To evaluate the use of spectral imaging techniques in seed phenotyping, a systematic review based on the PRISMA methodology was carried out. A total of 1304 articles were initially evaluated, and 44 articles were selected according to the stipulated criteria. The results indicated that the analysis has a high capacity (93.33%) for classifying seed genotypes, including cultivars, interspecific hybrids, parents and lines. Regarding distinguishing hybrid forest seeds, four experiments were carried out with two separate batches and one combined batch of *Eucalyptus urograndis* seeds and one batch of *Corymbia maculata* × *Corymbia torelliana* and their progenitors. Multispectral images were captured, and spectral and morphological characteristics were extracted. SVM, LDA and RF algorithms were used to compose the seed classification models. The LDA algorithm, combined with the morphological and spectral characteristics of the seeds, was the most effective for both genera, with an accuracy of 98.15% for the *Corymbia* spp. seeds and 92.75%, 85.38% and 86.00% for each of the Eucalyptus spp. lots and for them mixed, respectively. The technique proved to be effective for separating hybrid seeds of *Corymbia* spp. and *Eucalyptus* spp. in the context of forestry breeding programs. For the experiment with soybean seeds, 1806 seeds from six different lots were evaluated. Multispectral images followed by individual measurements of the seeds' oxygen consumption during germination were taken. In total, 2775 pairs of 75 biometric measurements were analyzed. Both respiration and biometric measurements were categorized and associated using contingency tables and entropy analysis. The results revealed differences in respiration patterns, especially in autofluorescence (365/600 nm, 430/700 nm, 450/700 nm, and 470/700 nm) and reflectance (365 nm, 690 nm, and 405 nm). Soybean seed characteristics, especially their spectral information, are strongly correlated with seed respiration and quality, and spectral image analysis is an effective and non-invasive tool for evaluating them.

Keywords: Forest seeds; *Glycine max*; Machine learning; Multispectral imaging analysis; Multivariate data analysis.

LISTA DE FIGURAS

Capítulo 1

Capítulo 2

Capítulo 3

LISTA DE TABELAS

Capítulo 1

Capítulo 2

Capítulo 3

SUMÁRIO

INTRODUÇÃO GERAL

Na agricultura contemporânea torna-se cada vez mais evidente a importância das sementes para o desenvolvimento econômico e a segurança alimentar. As sementes desempenham um papel fundamental no aumento constante da produção de alimentos, pois levam para o campo os ganhos obtidos pelo melhoramento genético realizado ao longo de anos de seleção. Consideradas como a forma natural mais eficaz de preservar a variabilidade genética, as sementes possibilitam o aprimoramento contínuo da espécie, garantindo a manutenção de níveis elevados na produção de alimentos e materiais essenciais. Essa característica é especialmente crucial diante dos desafios emergentes relacionados às mudanças climáticas (Bewley et al., 2013; Marcos-Filho, 2015).

Sob a perspectiva da exploração agrícola comercial, a qualidade da semente é um elemento decisivo. Sem uma semente de alta qualidade, a aplicação de técnicas modernas de produção e insumos não é suficiente, uma vez que a produtividade é limitada pela qualidade da semente. Uma semente de alta qualidade deve apresentar capacidade de germinação e vigor, ou seja, deve germinar de maneira rápida e uniforme, mesmo em condições adversas, além de ser isenta de impurezas, materiais genéticos indesejados e livre de patógenos. (Bewley et al., 2013; Caverzan et al., 2018; Marcos-Filho, 2015). O uso de sementes de alto vigor resulta frequentemente em plântulas mais resistentes, promovendo campos de melhor desempenho e, possivelmente, uma maior produção (Cheng et al., 2023).

No contexto dos programas de melhoramento vegetal, especialmente no setor florestal, torna-se crucial a identificação do material genético da semente a ser utilizado, dada a considerável demanda de tempo e custos envolvidos na produção. Para a produção de híbridos é necessário assegurar a identidade genética dos progenitores para a realização dos cruzamentos controlados e também confirmar a autenticidade dos genótipos resultantes Essa abordagem integrada, que engloba tanto a seleção criteriosa dos progenitores quanto a verificação pós-cruzamento, fortalece a integridade genética do programa de melhoramento, assegurando consistência e confiabilidade no desenvolvimento de sementes de alta qualidade (da Silva et al., 2022; Ramalho et al., 2022).

Métodos de avaliação do potencial fisiológico de um lote de sementes, como os testes de germinação e vigor, bem como a confirmação do material genético por meio de testes moleculares, são amplamente empregados. Entretanto, tais testes geralmente envolvem avaliações demoradas, trabalhosas e frequentemente resultam na perda da semente. Diante disso

há necessidade crescente de testes rápidos e não subjetivos determinar o atributo fisiológico e genético da qualidade de sementes (Elmasry et al., 2019; Xia et al., 2019).

Nesse contexto, a análise multiespectral de imagens surge como uma alternativa promissora, que combina a espetroscopia com a imagem digital. A particularidade desse método está na capacidade de diferentes objetos refletirem de maneira distinta quando expostos a determinados comprimentos de onda (por exemplo, ultravioleta ou infravermelho), influenciados por sua composição físico-química. Quando essa informação é integrada a uma imagem digital, a intensidade de luz refletida pode ser observada em cada pixel, proporcionando uma representação mais abrangente do objeto. A partir desse ponto, características como a informação espectral, juntamente com outras características extraídas da imagem, como textura e forma, podem ser empregadas para distinguir o objeto em questão de outros (Boelt et al., 2018; França-Silva et al., 2023).

A análise de imagem espectral é reconhecida como a principal tecnologia emergente na indústria de sementes, destacando-se por sua rapidez, flexibilidade e caráter não-destrutivo. Além disso, apresenta alta escalabilidade, uma vez que permite a automação do processo de coleta de informações e classificação das sementes de forma individualizada, especialmente quando integrada com a ciência de dados (Boelt et al., 2018; Elmasry et al., 2019; Xia et al., 2019; França-Silva et al., 2023).

Ao coletar informações de diversas naturezas, como espectrais, de autofluorescência e espaciais, a técnica tem sido rapidamente explorada em diversas áreas da indústria de sementes. Isso inclui sua aplicação na distinção de variedades, identificação de impurezas e separação de sementes híbridas de seus progenitores. A análise demonstrou sucesso ao distinguir 16 variedades de aveia, utilizando o algoritmo *Support Vector Machine* (SVM), alcançando uma precisão de 92.7% (Fu et al., 2023), assim como na identificação de sementes híbridas de quiabo em relação aos seus progenitores, também atingindo uma precisão de 95% (SVM) (Zhang et al., 2018).

Apesar de amplamente utilizada em estudos ligados à ciência de sementes, a aplicação da análise de imagens espectrais na avaliação do potencial fisiológico de sementes ainda está em desenvolvimento, devido à dificuldade de mensurar elementos ligados ao desempenho de uma semente.

Dessa forma, o objetivo principal do presente trabalho foi investigar a aplicação da análise de imagem espectral combinada com métodos de análise de dados multivariada na avaliação da qualidade de sementes. Os objetivos específicos de cada capítulo desta tese foram:

(1) avaliar os principais procedimentos relacionados à análise de imagens espectrais e procedimentos quimiométricos aplicados na fenotipagem de sementes, bem como a sua aplicação prática; (2) identificar o potencial da análise de imagem espectral na distinção entre sementes híbridas de *Eucalyptus urograndis* (*Eucalyptus grandis* × *Eucalyptus urophylla*), e *Corymbia maculata* × *Corymbia torelliana* dos seus progenitores; e (3) avaliar a relação entre o potencial fisiológico de sementes de soja e suas características biométricas individualmente por meio da respiração e do uso de imagens multiespectrais.

CAPÍTULO 1

Spectral Imaging and Chemometrics Applied at Phenotyping In Seed Science Studies: A Systematic Review[1]

1 Artigo publicado na revista *Seed Science Research*.

CAPÍTULO 1: SPECTRAL IMAGING AND CHEMOMETRICS APPLIED AT PHENOTYPING IN SEED SCIENCE STUDIES: A SYSTEMATIC REVIEW

Running title: Spectral imaging analysis in seeds

Thomas B. Michelon[1]*
https://orcid.org/0000-0002-7437-5062
Elisa Serra Negra Vieira[2]
https://orcid.org/0000-0002-0799-7654
Maristela Panobianco[1]
http://orcid.org/0000-0002-9990-2172

[1]Department of Plant Science – Federal University of Paraná – R. dos Funcionários, 1540, CEP 80035-050, Curitiba, PR, Brazil.
[2]Embrapa Forestry – Estrada da Ribeira, km 111 – 83411-000 – Colombo, PR – Brazil.

*Corresponding author <thomasnbrunomichelon@gmail.com>, phone +55 (41) 99166-5252.

Spectral imaging and chemometrics applied at phenotyping in seed science studies: a systematic review

Abstract

The evaluation of the genetic quality of a seed lot is crucial for the quality control process in its production and commercialization, as well as in the identification of superior genotypes and verification of the correct crossing in plant breeding programs. Current techniques, based on the identification of seed morphological characteristics, require skilled analysts, while biochemical methods are time-consuming and costly. The application of spectral imaging analysis, which combines digital imaging with spectroscopy, is gaining ground as a fast, accurate and non-destructive method. The success of this technique is closely linked to chemometric techniques, which use statistical and mathematical tools in data processing. The aim of the work was to evaluate the main procedures in terms of spectral imaging analysis and chemometric procedures applied in seed phenotyping and its practical application. A systematic review was conducted using PRISMA methodology, in which a total of 1304 articles were identified and screened to the inclusion of 44 articles pertaining to the scope. It was concluded that spectral imaging analysis has a high ability to classify seeds of different genotypes (93.33%) in a range of situations: between cultivars; hybrids and progenitors; and hybrids and lines, as well as in the separation of coated seeds. Accurate classification can be obtained by different strategies, such as the choice of the equipment type, the spectrum range and extra features, guided by the characteristics of the species. As well as in the choice of algorithms and dimensionality reduction procedures for the optimization of models when there is a large amount of data. Although the practical application of this technique in seed phenotyping still needs to be developed for use in laboratories with large volumes of analyses, lots, genotypes, and harvests. Research has been accelerated to overcome the practical challenges of this method, as seen in works using model update algorithms, online classification systems, and real-time classification maps. Thus, there are strong indications that the application of multispectral imaging analysis will reach the routine of seed analysis laboratories.

Keywords: deep learning, hyperspectral imaging, machine learning, multispectral imaging, seed classification**, s**pectroscopy

Nomenclature

BPNN - Back Propagation Neural Network

BPR - Biomimetic Pattern Recognition
BULDP - Biomimetic Uncorrelated Locality Discriminant Projection
CARS - Competitive Adaptive Reweighted Sampling
CCM - Correlation Coefficient Matrix
CDA - Canonical Discriminant Analysis
CNN - Convolutional neural network
DCNN - Deep convolutional neural network
DLJ4 - Deep Learning J4
EL - Ensemble learning
ELM - Extreme Learning Machine
FDA - Fisher's discriminant analysis
GDA - General Discriminant Analysis
JSWSA - Joint Skewness-based Wavelength Selection Algorithm
k-NN - k-nearest neighbor
LDA - linear discriminant analysis
LR - Logistic regression
LS-SVM - Least Squares support vector machine
MLDA - Multi-linear Discriminant Analysis
MSC - Multiplicative Scatter Correction
PCA - Principal Component Analysis
PLS-DA - Partial least square discriminant analysis
RBFNN - radial basis function neural network
RF - Random Florest
RSLD - random subspace linear discriminant
SIMCA - Soft Independent Modeling of Class Analogy
SNV - Standard Normal Variate
SPA - Successive projection algorithm
SVM - Support Vector Machine
SVM-DA - Support Vector Machine Discriminant Analysis
t-SNE - T-distributed stochastic neighborhood embedding

Introduction

Varietal sorting is an essential part of the quality control process of a seed lot, whether in germplasm bank management, production or commercialization, in order to identify its genetic quality and avoid species mixture (Elmasry *et al.,* 2019). For plant breeding programs, cultivar discrimination is also crucial to prove the correct crossing between plants, identify superior genotypes, and guarantee seed homogeneity according to their minimum descriptors for the purposes of registering new cultivars. For all these purposes, the process of separating

seeds by its morphological characteristics, such as color, texture, and shape, requires well-trained analysts and sometimes time-consuming and expensive biochemical and molecular techniques (Hansen *et al.,* 2016; Zhu *et al.,* 2020).

Thus, non-destructive, rapid, and non-subjective methods are of great interest for determining seed quality. (Xia *et al.*, 2019; Elmasry *et al.*, 2019). In this regard, multispectral imaging analysis is a promising alternative that combines spectroscopy with digital image. The technique is based on the reflectance of an object - the intensity that a given surface reflects a wavelength. An object can be illuminated by different wavelengths (e.g., visible light, near infrared), and when combined with a digital image, the reflectance of each pixel of this object's image can be measured to differentiate it from another (Boelt *et al.*, 2018; Xia *et al.*, 2019).

Since each pixel contains a dataset (reflectance from each wavelength) the result is a large amount of data proportional to the number of wavelengths used and the size of the image. As these data are considered chemical information, the role of chemometrics is to use statistical and mathematical tools to obtain the most important information from the dataset of each object (Amigo, 2020).

Spectral imaging analysis is considered one of the major emerging technologies in seed analysis and technology. Its versatility, non-destructive characteristics and rapid determination of quality attributes of a seed lot, combined with data science, make it possible to automate the entire seed sorting process. (Elmasry *et al.,* 2019; Xia *et al.,* 2019; Amigo, 2020; Zhou *et al.,* 2020).

The success in applying the technique lies in combining experimental issues with the process of extracting information from the seeds and the chemometric strategy used, which may include from the choice of classification algorithms to data dimensionality reduction processes. Therefore, the process of choosing each aspect involved in the analysis is not trivial and thus this systematic review aims to evaluate the main procedures in terms of spectral imaging analysis and chemometric procedures applied in seed phenotyping as well its practical application.

Materials and Methods

The study followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) methodology (Moher *et al.,* 2009; Page *et al.,* 2020), as it presents a clear and systematic research method with a focus on reproducibility.

Inclusion and exclusion criteria

The inclusion and exclusion criteria were based on literature type, access, period, language and scope (Table 1). The 15-year period was chosen to limit the search to new papers, given the recent expansion of spectral imaging technology in seed science. Regarding the scope, only papers on spectral imaging analysis (multispectral and hyperspectral) in seeds were considered; thus, papers using material not considered as seeds (i.e., grains) were not considered. Articles using spectral analysis only to quantify chemical components (e.g., oil, protein content) of seeds but did not classify them into different genotypes (e.g., cultivars or varieties) were not considered. Language was considered as a criterion to avoid bias in the translation of non-English language papers.

Table 1. Inclusion and exclusion criteria

Criterion	Eligibility	Exclusion
Literature type	Article	Reviews, conference paper and book chapter
Access	Ful-text available	
Period	Between 2006 and 2021	<2006
Language	English	Non-English
Scope	Uses spectral imaging (e.g., hyperspectral, multispectral imaging) applied to seed phenotyping	Did not use seeds; did not combine spectroscopy to image analysis; or just quantify certain component but did not differ cultivars, varieties, etc.

Search methodology

The keywords for the present work, as well as their synonyms, were obtained through prior review in studies related to the areas of seed science and technology and spectroscopy (Table 2). The databases used were the Web of Science Core Collection (WOS) and Scopus and were chosen according to previous research on the number of articles related to the scope present in each one. WOS is the database of Clarivate Analytics, has indexed more than 21,000 papers covering 256 disciplines, while the Scopus database belongs to Elsevier and is one of the most related to plant science with peer reviewed articles. In addition to covering a large quantity of articles related to the topic, these databases allow the inclusion of boolean operators

for the search strategy, as well as symbols that allow the inclusion of all possible terms with the same root.

Table 2. The search strategy used for the systematic review process.

Database	Search criteria
Web of Science	TS=((seed OR seeds) AND (multispectral OR hyperspectral OR spectral OR spectroscopy OR NIR OR "near infrared" OR nearinfrared OR "near-infrared" OR reflectance OR chemometrics) AND (variet* OR cultiv* OR phenot* OR breed* OR hybrid* OR transgenic*) AND (classification OR discrimination OR identification OR determination OR phenotyping))
Scopus	TITLE-ABS-KEY((seed OR seeds) AND (multispectral OR hyperspectral OR spectral OR spectroscopy OR NIR OR "near infrared" OR nearinfrared OR "near-infrared" OR reflectance OR chemometrics) AND (variet* OR cultiv* OR phenot* OR breed* OR hybrid* OR transgenic*) AND (classification OR discrimination OR identification OR determination OR phenotyping))

The search consisted of three steps: identification of potential articles, screening, and inclusion of articles (Fig. 1). A total of 1,304 articles were identified and duplicates were removed with the aid of the Mendeley Reference Manager management program (Dearden *et al.,* 2011). A total of 308 articles were eliminated based on their characteristics as per the exclusion criteria, while 508 articles were excluded as per the scope from the evaluation of the title and abstract. A total of 60 articles were evaluated in full and 44 were included in the review.

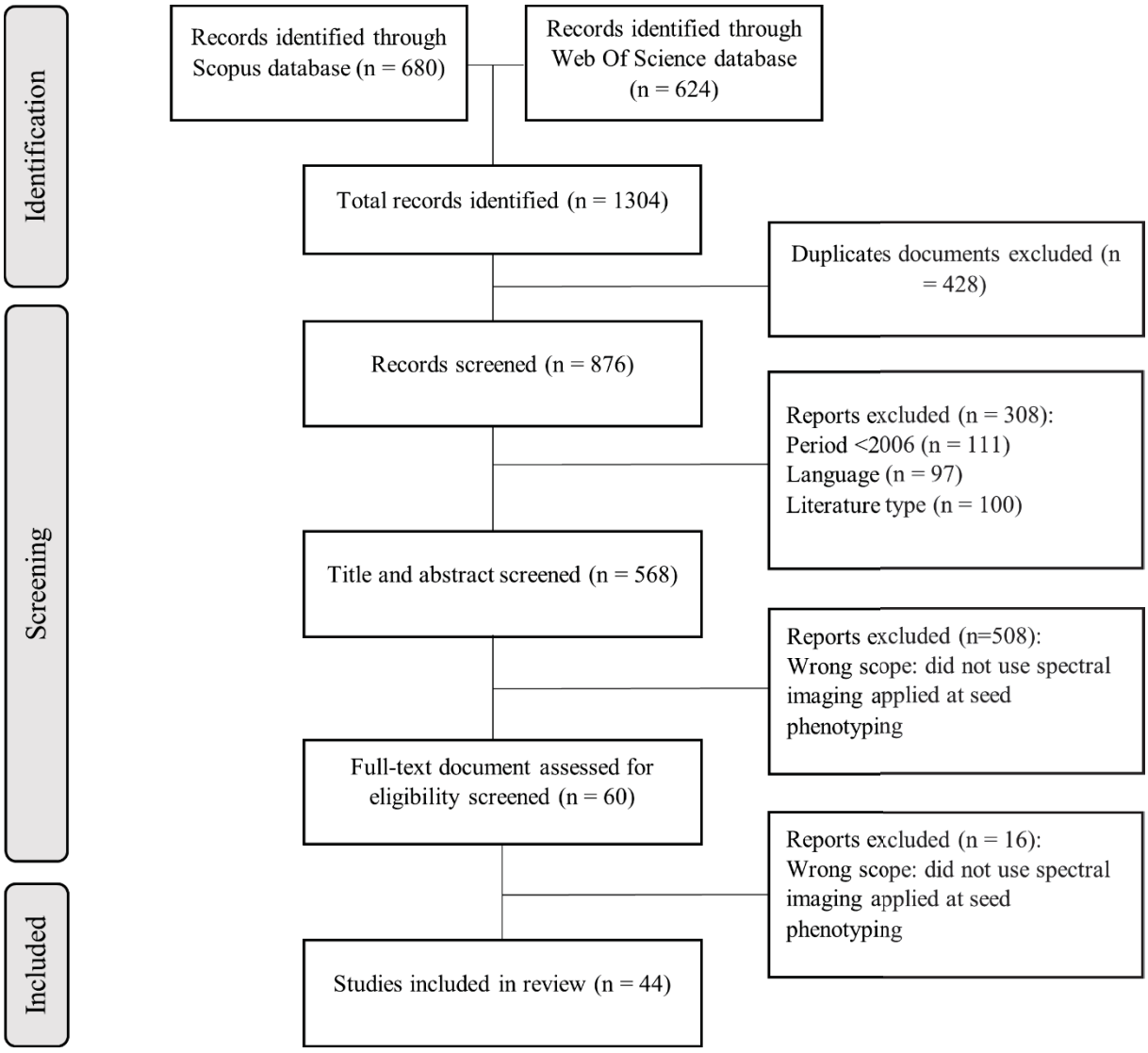


Figure 1. Selection of articles according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework.

Statistical analysis

From the articles evaluated, data were collected regarding the experiment, the best classification model obtained in each study, as well as other information deemed relevant (Table 3), to identify possible factors influencing the accuracy of seed classification through spectral imaging analysis. A multiple generalized linear regression model with gamma distribution and log-link function was used, due to the non-normality of the data, in conjunction with the stepwise feature selection algorithm (backward and forward) to select the final model. The algorithm adds and removes features and compares the models by means of Akaike's Selection Criterion (AIC), in order to obtain a final model with the feature (or combination of features)

best fitted (with the lowest AIC value) to predict the accuracy of spectral imaging analysis applied to seed phenotyping.

Table 3. Features that might affect the accuracy of seed distinction in spectral analysis applied to seed phenotyping.

Features	levels
Crop type	Agricultural crops, horticultural crops, fruit production, others
Application	Varietal discrimination, haploid, transgenic/non-transgenic, hybrid/progenitors
Spectrum	NIR, VIS-NIR
Sensor	Multispectral imaging, hyperspectral imaging
Number of wavelengths	19 - 700
Number of seed groups	2 - 90
Total seeds used	376 - 147096
Algorithm class	Machine learning, deep learning
Extra features (e.g., morphology, texture, color)	Present (1); Absent (0)
Wavelength selection and/or dimensionality reduction	Present (1); Absent (0)
Wavelength preprocessing	Present (1); Absent (0)

Results and applications

A total of 44 articles from the systematic review were included; since the authors reported more than one experiment in some papers, data from all the experiments performed was listed, including data from the best performing classification model (Table 4).

Table 4. Characteristics and applications of spectral imaging (HSI - hyperspectral imaging; MSI - multispectral imaging) applied in seed phenotyping.

Species	Method	n. wavelengths	Spectrum	Application	n. Groups	Total seeds	Training (Tr), Testing(Te), and Validation (V) proportion	Best classifier	Extra features	Wavelength (WL) selection/ Dimensionality reduction	Spectral preprocessing	Accuracy	References
Alfafa	MSI	19	365 - 970	Varietal discrimination	12	2400	70% Tr; 30% Te	SVM	Morphological; color	-	-	93.47%	Yang *et al.* (2020)
Cotton	HSI	256	1100 - 2500	Varietal discrimination	4	807	2:1:1	PLS-DA	-	-	Smoothing	98.00%	Soares *et al.* (2016)
Cotton	HSI	200	942 - 1646	Varietal discrimination	7	13160	3:1:1	CNN-SoftMax	-	-	Smoothing; normalization	88.84%	Zhu *et al.* (2019b)
Grape	HSI	240	914 - 1715	Varietal discrimination	4	56	60% Tr; 40% V	GDA	-	-	-	100.00%	Rodríguez-Pulido *et al.* (2013)
Grape	HSI	200	975 - 1646	Varietal discrimination	3	43357	2:1	SVM	-	10 WL (PCA)	Smoothing	88.70%	Zhao *et al.* (2018a)
Jatropha curcas	HSI	256	874 - 1734	Origin discrimination	4	240	2:1	LS-SVM	Morphological	10 WL (SPA)	-	93.75%	Gao *et al.* (2013)
Looffah	HSI	200	975 - 1645	Varietal discrimination	6	4128	2:1	DCNN	-	-	Smoothing	95.93%	Nie *et al.* (2019)
Maize	HSI	649	1110 - 2500	Varietal discrimination	4	80	1:1	SIMCA	-	PCA	Smoothing; first derivative; normalization	97.50%	Jia *et al.* (2015)
Maize	HSI	380	400 - 1000	Varietal discrimination	3	376	70% Tr; 30% Te	LS-SVM	-	-	Detrending	91.67%	Wang *et al.* (2016)

Maize	HSI	94	400 - 1000	Varietal discrimination	4	2000	2:1	LS-SVM	-	-	-	94.40%	Huang *et al.* (2016a)
Maize	HSI	94	400 - 1000	Varietal discrimination	4	2000	2:1	LS-SVM	-	-	-	98.30%	He *et al.* (2016)
Maize	HSI	233	400 - 1000	Varietal discrimination	17	1632	3:1	LS-SVM	Morphological	11WL (SPA); PCA	Normalization	92.65%	Huang *et al.* (2016b)
Maize	MSI	19	375 - 970	Haploid discrimination	2	240	1:1	CDA	Fluorescence excitation/emission	-	-	85.83%	De la Fuente *et al.* (2017)
Maize	HSI	94	400 - 1000	Varietal discrimination	4	3600	5:1; 5:2; 5:2; 5:4	LS-SVM	-	-	-	85.40%	Guo *et al.* (2017)
Maize	HSI	219	924 - 1657	Varietal discrimination	14	1120	3:1	LS-SVM	-	19 WL (JSWSA)	-	96.57%	Yang *et al.* (2017)
Maize	HSI	256	862.9 - 1704.2	Haploid discrimination	2	200	2:1	BPR	-	BULDP	Smoothing; first derivative; normalization	99.85%	Wang *et al.* (2018)
Maize	HSI	200	975 - 1646	Varietal discrimination	3	12900	2:1	RBFNN	-	15 WL (PCA)	Smoothing	91.00%	Zhao *et al.* (2018b)
Maize	HSI	233	400 - 1000	Varietal discrimination	17	1632	1:1	LS-SVM	Texture	10 WL (MLDA)	Normalization	99.13%	Xia *et al.* (2019)
Maize	HSI	200	975 - 1646	Varietal discrimination	3	5400	4:1:1	LR	-	-	-	96.67%	Zhang *et al.* (2020)
Maize	HSI	200	975 - 1646	Varietal discrimination	4	20400	2:1	RBFNN	-	-	Smoothing	98.09%	Bai *et al.* (2020)
Maize	HSI	420	450 - 979	Varietal discrimination	4	3200	6:2:2	DCNN	-	-	-	95.30%	Zhang *et al.* (2021)
Maize (silage maize)	HSI	200	975 - 1646	Varietal discrimination	4	20400	2:1	RBFNN	-	-	Smoothing	99.10%	Bai *et al.* (2020)
Maize (sweet maize)	HSI	700	480 - 1020	Varietal discrimination	9	810	4:1	SVM	-	23 WL (CARS)	Smoothing; first derivative	94.86%	Zhou *et al.* (2020b)

Maize and silage maize	HSI	200	975 - 1646	Varietal discrimination	8	40800	2:1	RBFNN	-	-	Smoothing	88.41%	Bai *et al.* (2020)
Maize and silage maize	HSI	200	975 - 1646	Varietal discrimination	2	40800	2:1	RBFNN	-	-	Smoothing	88.41%	Bai *et al.* (2020)
Maize Waxy maize	HSI	220	430 - 972	Varietal discrimination	8	800	4:1	FDA	-	t-SNE	Procrustes analysis (PA)	97.50%	Miao *et al.* (2018)
Oat	HSI	200	975 - 1646	Varietal discrimination	4	14846	3:1	DCNN	-	-	Smoothing	99.19%	Wu *et al.* (2019)
Okra	HSI	200	975 - 1645	Varietal discrimination	6	6136	2:1	DCNN	-	-	Smoothing	98.24%	Nie *et al.* (2019)
Pepper	MSI	19	365 - 970	Varietal discrimination	3	4416	9:1	SVM	-	-	-	97.70%	Li *et al.* (2020)
Rice	HSI	256	1039 - 1612	Varietal discrimination	4	225	2:1	RF	-	-	First derivative	100.00%	Kong *et al.* (2013)
Rice	MSI	19	365 - 970	Transgenic; non-transgenic	2	400		LS-SVM	Morphological	-	-	100.00%	Liu *et al.* (2014b)
Rice	MSI	19	365 - 970	Varietal discrimination	5	250	4:1	LS-SVM	Morphological; color	-	-	94.00%	Liu *et al.* (2016)
Rice	MSI	19	365 - 970	Varietal discrimination	20	600	-	k-NN + multiclass CDA	Morphological; color	-	-	93.00%	Hansen *et al.* (2016)
Rice	HSI	256	874.41 - 1733.91	Mutant discrimination	2	2640	2:1	ELM	-	-	Smoothing	91.75%	Feng el al. (2017)
Rice	HSI	256	975 - 1646	Varietal discrimination	4	20907	3:2	CNN	-	-	Smoothing	87.00%	Qiu *et al.* (2018)
Rice	HSI	256	385 - 1000	Varietal discrimination	90	8640	4:1	RF	Morphological	85 WL (LDA)	Normalization	79.64%	Fabiyi *et al.* (2020)

Soybean	MSI	19	365 - 970	Hybrid; progenitors	3	600		BPNN	Morphological	-	-	98.00%	Liu *et al.* (2014a)
Soybean	HSI	128	373 - 1043	Varietal discrimination	10	1200		EL	-	CARS	MSC	100.00%	Zhu *et al.* (2019a)
Soybean	HSI	200	975 - 1646	Varietal discrimination	3	5670		CNN	Pixel-wise	-	Smoothing; normalization	98.78%	Zhu *et al.* (2019c)
Soybean	HSI	128	400 - 1000	Varietal discrimination	10	1200	3:1:1	GS-SVM	-	-	First derivative	97.20%	Zhu *et al.* (2020)
Soybean	HSI	462	400 - 1000	Varietal discrimination	15	750	2:1	RSLD	-	155 WL (CCM)	-	99.20%	Wei *et al.* (2020)
Sunflower	MSI	-	450 - 1550	Varietal discrimination	6	12000	-	DLJ4	Morphological; texture	-	-	98.20%	Bantan *et al.* (2020)
Tomato	MSI	19	365 - 970	Varietal discrimination	11	2525	3:1	PLS-DA	Morphological; color	-	-	79.00%	Shrestha *et al.* (2015)
Tomato	MSI	19	365 - 970	Varietal discrimination	5	1236		SVM-DA	-	-	SNV; detrending	98.00%	Shrestha *et al.* (2016b)
Tomato	HSI	288	950 - 2500	Varietal discrimination	4	1366		PLS-DA	-	-	Smoothing; detrending	71.00%	Shrestha *et al.* (2016a)
Wheat	MSI	19	365 - 970	Varietal discrimination	7	1728	-	k-NN	Morphological; color, texture	-	-	69.53%	Vrešak *et al.* (2016)
Wheat	HSI	200	975 - 1660	Varietal discrimination	5	33494	9:1	SVM	-	-	-	87.81%	Bao *et al.* (2019)
Wheat	HSI	200	975 - 1645	Varietal discrimination	30	147096	2:1:1	CNN-ATT	-	-	Normalization	93.10%	Zhou *et al.* (2020a)

Accuracy, data splitting and validation methods

The average accuracy of the reviewed studies (considering all experiments listed in Table 4) was 93.33% (±7.07%). In some studies, the application of spectral image analysis resulted in 100% classification accuracy, as in Zhu et al. (2019a), on 10 soybean seed varieties, using the Ensemble Learning classification algorithm. A similar result was found for Liu et al. (2014b), whose study on transgenic and non-transgenic rice seeds, by means of the Least-Squares Support Vector Machine (LSSVM) algorithm, used both spectral information and biometric data regarding seed morphology. It was also the case of the study of Kong et al. (2013) on four rice seed varieties, using the Random Forest algorithm, and the study of Rodríguez-Pulido et al. (2013), which separated four grape varieties using General Discriminant Analysis (GDA).

Their high accuracy suggests a promising feature of spectral image analysis in distinguishing genotypes, but there are some concerns. The first is about the amount of classification groups: only 46% and 23% of the experiments had more than 5 and 10 categories, respectively. In works that used many categories, e.g., Fabiyi et al. (2020), with 90 cultivars, although the overall accuracy was relatively high (79.64%) using the Random Forest algorithm, for some cultivars accuracy was only 30% to 50%. The same result was found in the study of Zhou et al. (2020a), in which the overall accuracy was 93.10% using a deep learning algorithm (Convolutional Neural Network - CNN) for the classification of 30 cultivars, while there was variation of more than 20% in classification accuracy for certain cultivars.

It is not clear, in most of the reviewed papers, if spectral image analysis was applied owing to its agility and automation or if because of the ability to classify cultivars in situations in which classification by visual morphological characteristics was not possible. This point is important, because knowing whether the genotypes were chosen randomly or whether they were chosen arbitrarily from characteristics where separation would be possible even by eye, allows one to establish to what extent spectral image analysis is applicable in situations of cultivar diversity, as occurs in the seed industry.

Another point of concern is test and validation data: in most works, there was (a) absence of test data and (b) absence of test and/or validation lots (e.g., seeds from other harvests). Ideally, when enough data is available, seed samples should be divided into training, validation, and test data. Training data is used by the algorithm to estimate the model; validation data is not to be used in training, in order to gather unbiased information about the quality of the models developed. Validation data is used for predicting errors of each model for the

purpose of selecting the best model. Since validation data is used constantly (depending on the number of models to be tested), test data (i.e., data not yet used) is commonly used to obtain the true error of the final model (i.e., generalization error), and these data is used only once so as not to overestimate accuracy (Hastie et al., 2017).

In most works of the present review, despite the large number of seeds being used, test data were not used - only validation data (even when it was referred to as test data in the studies, owing to different definitions), which may lead to high accuracy. In studies with a small number of seeds, an alternative is to use cross-validation, in which samples go through *n* data splitting cycles (in training and validation), model building, and error computation, and final accuracy is determined from the average error of the *n* models obtained (Hastie et al., 2017). However, of the experiments that used approximately less than 100 seeds per classification category (referring to the first quartile of the variable number of seeds in Table 4), 37% did not perform cross-validation, which may cause overestimation of the resulting accuracy.

Another aspect regarding data division is that only 6% of the studies used validation and/or test lots (i.e., from other harvests and/or regions). In the works that did not use validation lots, high accuracy may have been due to a model overfitted to the lot; consequently, there may not be such accuracy in the classification of the same cultivars from other harvests and regions (He et al., 2016; Huang et al., 2016a). For example, Huang et al. (2016a), when classifying seeds of four wheat varieties and using - as test data - seeds from the same year as those used for training, found 100% classification accuracy using the LSSVM algorithm. However, when using seeds from other years, accuracy was only 75.4%. Similarly, Shrestha et al. (2016a) used tomato seeds of four cultivars from three harvest years, in experiments with seeds only from the same year and with the mixture of seeds from the other years, both in the test and training data. For the fitted and validated model with seeds from the same year, they found accuracy per cultivar from 73 to 100 %, whereas for the sample with mixed seeds from other harvests, accuracy ranged from 34 to 88 %.

There was great variation in the total number of seeds used per classified genotype, as observed in each experiment, even in those that used the same species. For example, in the works performed on wheat seeds, the number of seeds used per class ranged from 20 to 5,100 seeds. Few of the reviewed studies evaluated the influence of seed quantity on training samples. For instance, Qiu et al. (2018) tested training samples with different amounts for designing their classification models, ranging from 100 to 3000 seeds, for each of the four cultivars. They found when using more than 1,500 seeds, the increment in accuracy was not significant. Certainly,

the accuracy determined in experiments that used larger samples leads to more confidence, but the approach of studying the ideal number of seeds has more practical applicability, since the increase in the amount of samples generates extra processing costs, without necessarily leading to a significant increase in the accuracy of the models. Thus, stipulating the optimal number of seeds is important to achieve a balance between cost and performance of a model, which would, thus, facilitate the applicability of the analysis (Qiu et al., 2018).

One way to obtain more data without necessarily increasing the number of seeds in a sample is by using the spectral information of each seed pixel (i.e., pixel-wise spectrum), as opposed to averaging the seed spectrum (i.e., object-wise), as evaluated by Zhu et al. (2019c) in classifying three soybean cultivars. The authors used the pixel-wise spectrum of 60 seeds and reported equivalent performance of a sample with 810 seeds using object-wise spectrum. However, this technique requires a great deal of data processing, as there is a significant increase in the amount of information (i.e., equivalent to the number of pixels). Moreover, it also needs to be explored in different situations (e.g., species, cultivars, crops).

Crop type application

Out of the 44 articles evaluated, approximately 80% performed the analysis of agricultural crops species (e.g., soybean, maize, wheat), 11% of horticultural seeds, 5% of fruit production, and 5% of other classes (pasture and medicinal plants), while there was no work on forest seeds (Fig.2).

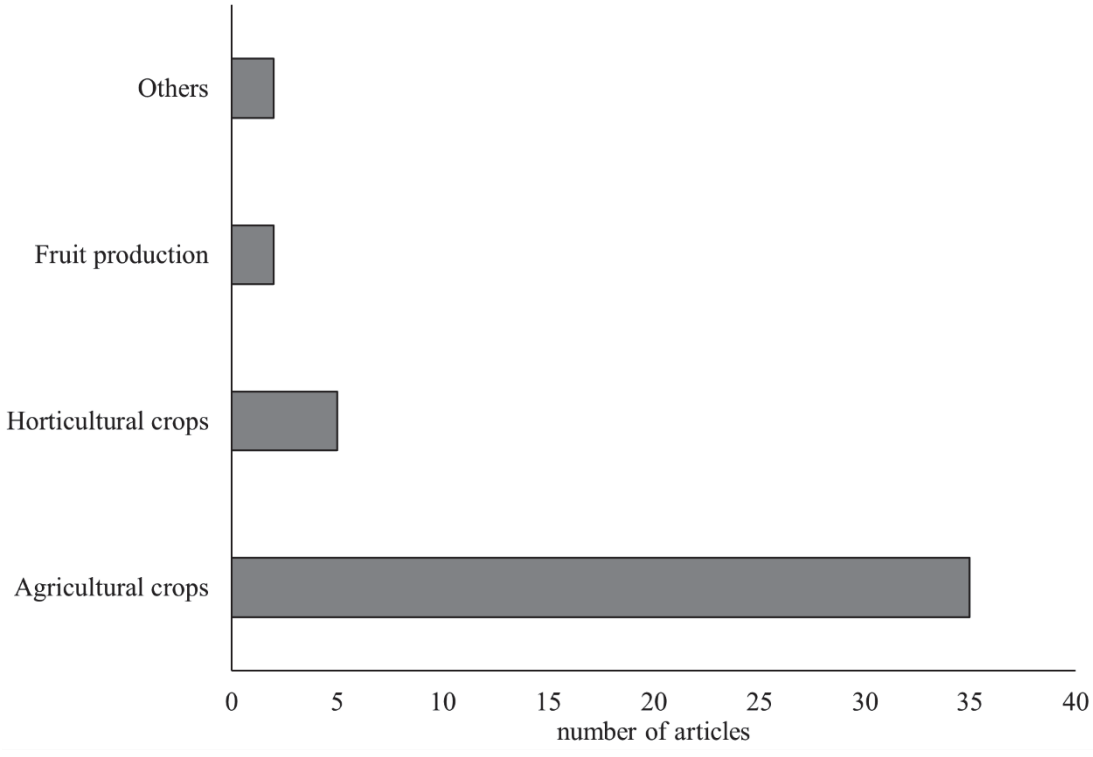


Figure 2. Species used in each article of the review.

As with the present study, Raman and Cho (2016), in a narrative review with 32 papers that applied seed variety identification using image analysis techniques, 31 focused on agricultural crops. Given the emerging feature of the spectral imaging analysis technique in seed phenotyping, the use of it in agricultural crop seeds over other seeds may be mainly linked to the economic appeal of these species, as well as to the greater number of plant breeding programs related to them.

Wavelength spectrum

Of the evaluated studies, 75% used hyperspectral equipment, and according to the density plot, the frequency distribution of the wavelengths applied in the studies varies according to type of equipment (Fig. 3). Commercial hyperspectral equipment operates in bands with greater amplitude in the near infrared (NIR) spectrum (750 - 2500 nm), whereas commercial multispectral imaging equipment concentrates on the visible light range up to the beginning of the NIR (350 - 950 nm).

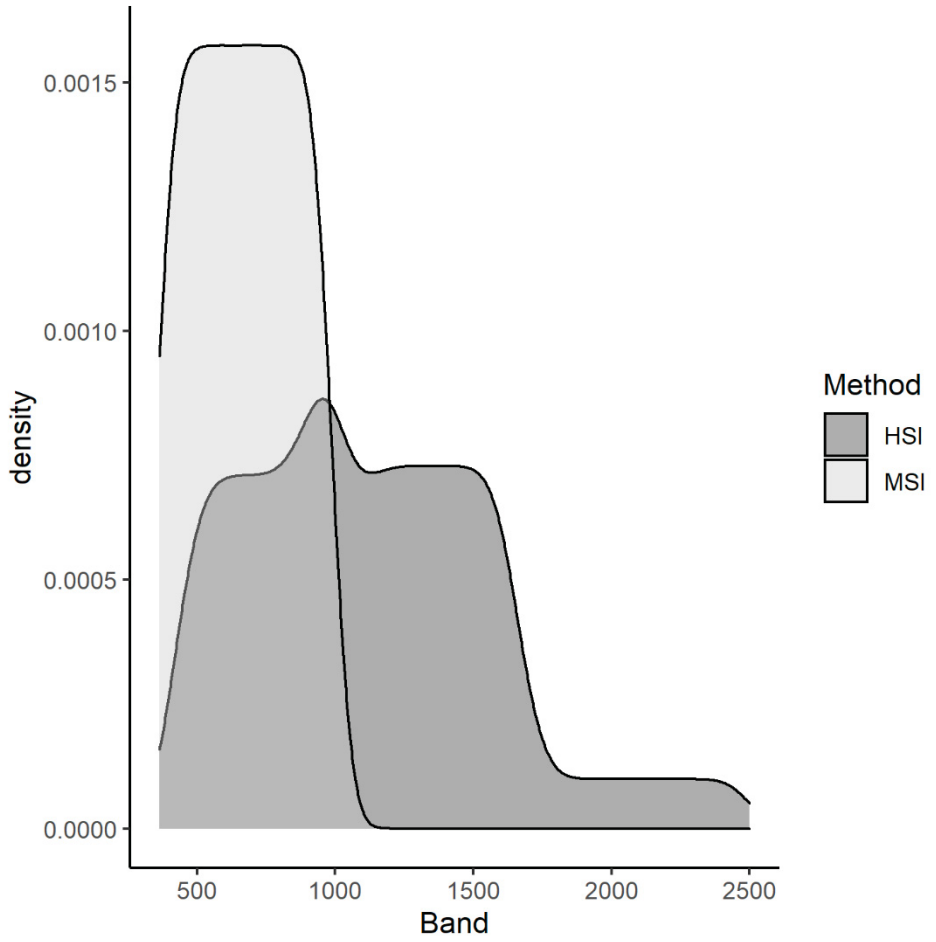


Figure 3. Density plot of the wavelengths used according to the hyperspectral (HSI) or multispectral (MSI) method.

The wavelength range used is closely linked with the components measured in the seeds, and the visible spectrum is related to superficial characteristics, such as pigmentation (e.g., flavonoids, carotenoids, chlorophyll) and oxidation. These characteristics are ideal for distinguishing seeds with marked physical characteristics, e.g., tegument color or texture. As regards the near infrared spectrum, this region is sensitive to the molecular overtone of hydrogen-containing groups, such as C-H, N-H, O-H chemical bonds, which represent seed starch, protein, and oil contents, and can penetrate deeper than visible light through the subsurface layer of seed coat (Rodríguez-Pulido *et al.*, 2013; Li *et al.,* 2014; Li *et al.,* 2020; Mortensen *et al.*, 2021). For works that used hyperspectral cameras, there was a peak near the 1000 nm range. In this range, the 1122, 1200, and 1314 nm bands (related to organic C-H compounds, such as starch) stand out, while the 1402 nm wavelength is associated with the O-

H region of carboxylic acids, as well as regions near the 1580 nm band (Osborne and Douglas, 1981; Lammertyn *et al.,* 1998; Serranti, *et al.,* 2013; Zhao *et al.,* 2014).

Shrestha et al. (2016a), using hyperspectral image analysis in the near infrared region for classification of four tomato seed varieties, found that the 1417, 1901, 2102 and 2238 nm bands, associated with protein and water content, and the 1222 and 1695 nm bands, associated with fatty acid content, represent an important spectral signature for this species. Rodríguez-Pulido et al. (2013), when separating seeds of three grape cultivars, with one coming from two different regions, using the Principal Component Analysis (PCA) score, found that the bands at 928, 940, 1148, 1620 and 1652 nm, referring to organic compounds with C-H chemical bonds, were primarily responsible for distinguishing the seeds. Zhao et al. (2018b), based on the score of the first six principal components of PCA, selected the bands in the 1100 and 1390 nm region and the bands at 1436, 1453, and 1554 nm (with the latter three corresponding to the first overtone of O-H stretching, to classify grape seeds of three cultivars. Zhang et al. (2021), using multispectral equipment and classifying four maize cultivars, found that the wavelength bands with the greatest contribution to the distinction of cultivars were 450 to 700 nm, related to the chlorophyll and β-carotene content of the endosperm, 730 and 785 nm, related to organic compounds with O-H and N-H bonding, and 850-950 nm, related to C-H hydrocarbons. Huang et al. (2016b) found 92.65% accuracy when they classified 17 corn cultivars, using 11 wavelengths selected by the Successive Projection Algorithm (SPA), located in the 500 to 750 nm region, which are sensitive to seed starch and oil contents. Similarly, Xia et al. (2019) classified 17 corn cultivars based on 10 optimal wavelengths, belonging to the regions of 410 to 470 nm, 524 to 790, and the wavelength of 988 nm, which represent seed texture, starch and oil content, and water content, respectively.

Thus, since the near infrared region is sensitive to organic compounds in seeds in deeper layers than visible light, this region seems to be a good strategy to differentiate seeds with similar surface characteristics (i.e., where the visible spectrum region acts more intensely) (Willian and Norris, 2001, Rodríguez-Pulido et al., 2013). For example, Wang et al. (2018) separated haploid from diploid maize seeds, whose visual similarity makes it difficult to separate them by traditional or machine vision methods. Using hyperspectral image analysis in the near infrared region (860 - 1700 nm), they were able to identify differences in oil content and other organic components, differentiating the seeds with 99.85% accuracy.

As for the equipment, the basic difference between multispectral and hyperspectral devices is in the number of wavelengths that each one can measure. Multispectral equipment

measures up to 20 wavelengths, while hyperspectral cameras can reach higher values, as reported in the work of Zhou *et al.* (2020b) with sweet corn and 700 wavelengths measured. The use of hyperspectral equipment results in a larger amount of data and, consequently, more time for processing and development of the classifier models. Therefore, all the studies that performed some form of wavelength selection or dimensionality reduction used hyperspectral equipment. Dimensionality reduction aims to mitigate the problem of correlation between predictor variables as well as model overfitting (Wu *et al.,* 2019; Friend, 2020). For example, Gao *et al.* (2013) used SPA to reduce from 256 to 10 wavelengths and obtained 93.75% accuracy.

Classifiers

For seed classification based on the selected wavelength and other features, the evaluated papers used machine learning and deep learning class algorithms on 30 and 17 occasions, respectively; in 2017, the percentage of papers that used machine learning was 95% (Fig. 4).

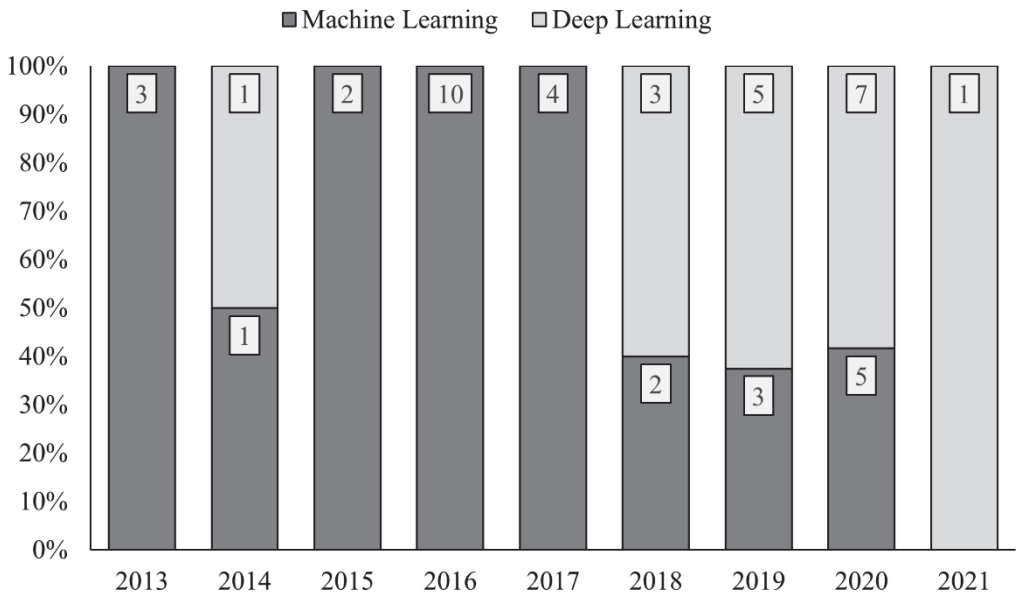


Figure 4. Proportion and absolute quantity (number in square) of the classification algorithm classes used in each paper over the years.

In 2018 and later, the number of papers using deep learning not only increased but was proportionally higher than the number of papers using machine learning. Deep learning is an unsupervised classification method (the class of seeds is not previously provided to the

algorithm) and brings the advantage of identifying abstract patterns in a large amount of data that supervised methods would not be able to find (i.e., deep features) (Gheisari *et al.,* 2017; Wu *et al.*, 2019). However, to achieve successful classification, deep learning algorithms preferentially need a larger volume of data, and this is represented in the average number of seeds used in the evaluated papers that applied machine learning: 3,897, compared to 22,765 in deep learning.

The larger the amount of data, the greater the demand for technology and processing time, which may be linked to the low frequency of use of deep learning in previous studies. In contrast to processing time, this class of algorithms seems to be more advantageous in seed classification as highlighted by Zhu *et al.* (2020), who found that all tested deep learning algorithms had higher accuracy than machine learning algorithms. Similarly, Qiu *et al.* (2018), comparing deep learning algorithm CNN with the machine learning algorithms support vector machine (SVM) and K-nearest neighbor (K-NN), found that, as the training samples increased, the CNN model outperformed the others. Nie *et al.* (2019), when classifying hybrid okra and loofah seeds using the deep learning model deep convolutional neural network (DCNN) and comparing it to Partial least square discriminant analysis (PLS-DA) and SVM, found that the number of varieties increased from two to six. The authors reported that with increased complexity (number of varieties), the accuracy of the DCNN model remains more stable than that of the others.

Thus, the main advantage of using deep learning algorithms lies in their ability to integrate the steps of feature learning, feature extraction, dimensionality reduction, and classification into just one system, which brings greater convenience in the use of data with more complexity (i.e., a larger number of features), as occurs when hyperspectral images are used (Wu et al., 2019).

Chemometric features

The adjusted model, which was identified by the stepwise algorithm with the lowest AIC (-07.9015), used the following features: number of seeds, seed classification groups, and use of methods for wavelength selection and/or data dimensionality reduction. Only the first feature was significant (Table 5). According to the estimated and exponentialized coefficient (to reverse the logarithmic scale) of the number of groups (0.998), as the number of seed classification groups increases, the final model accuracy tends to decrease by approximately 0.2% with each new group.

Table 5. Coefficients estimated from the adjusted model identified by the stepwise algorithm with the features that may influence the accuracy of the spectral imaging analysis of the evaluated studies.

Coefficients	Estimate (log scale)	Std. Error	t	p-value
Intercept	-0.0643	0.0138	-4.651	3.02E-05
n. Groups	-0.0020	0.0009	-2.263	0.0286
WL Selection1*	0.0378	0.0256	1.476	0.1471

***** Papers that used wavelength selection or dimensionality reduction procedure to obtain the most accurate model.

Classification accuracy tends to naturally decrease as the number of possible groups increases, but the non-significant influence of the other factors is due to the fact that the technique can result in high classification models using different strategies in the process (e.g., method, features selection, preprocessing), i.e., an isolated factor is not enough to determine the accuracy of an analysis. It must be clear that the final model only indicates a possible relationship between the variables, since other factors not listed may be relevant to determine classification accuracy (e.g., species, seed quality); moreover, further research is needed to make a robust analysis.

Discussion

Overall strengths

In the 44 evaluated studies, it was clear that the information collected through spectral image analysis, both reflectance and biometric measures of morphology and texture, are sufficient to classify seeds of different genotypes. Although one needs to further explore the ability to generalize the use of the analysis between seeds from other regions and/or harvests, as well as make different combinations of genotypes in future work, the fact is that well-fitted classification models have high accuracy in several situations: between cultivars, hybrids and progenitors, and hybrids and lines; transgenic and non-transgenic seeds.

Spectral image analysis allows the separation of genotypes even in coated seeds, mainly by using the near infrared, which penetrates beyond the surface of the seed. Coated seeds are common in the industry, as coating provides protection against fungi and microorganisms, and aids germination by supplying nutrients and amino acids, among other benefits. In the case of these treated seeds, even when dyes are applied for identification, classification using spectral image analysis has still proved possible (Jia et al., 2015; Zhang et

al., 2020). However, for small and/or non-uniform seeds, which are coated with thicker layers (which occurs by the encrustation or pelleting process), the analysis may not be applicable.

Reflectance offers sufficient information for separation of the seeds of different genotypes, and spectral information can be collected quickly, through an image or set of images according to the number of bands measured; seeds remain intact and there is no need for prior treatment. Therefore, multispectral image analysis has a huge advantage over conventional tests, because its limitations refer to treatment and processing of data rather than data collection. Traditional methods, such as molecular markers, are indeed highly reliable, robust methods; thus, they can hardly be replaced. However, in routine work in seed analysis laboratories, when identifying hybrid seeds in breeding programs or in the identifying cultivar mixtures in purity testing, traditional methods are not necessary if there is an alternative way that is reliable, fast and agile enough to meet the industry's demands (Bao et al., 2019; Zhang et al., 2020; Shrestha et al., 2015).

Challenges and limitations

a) Phenotypic variation

The main challenge of the spectral image analysis technique applied to seed phenotyping surely lies in the extrapolation of the fitted model to seeds coming from other harvests/regions (Zhu et al., 2020). The development of a prediction model is somewhat difficult and laborious and requires professionals specialized in data analysis. In the process, seeds need to be used to train the proposed model, test different forms of data processing, and validate the model with new seeds. This is time consuming, and not easily adapted to the work routine in seed producing companies. Thus, the model created to classify produced cultivars, must be able to classify seeds over the years and also those grown in different regions (He et al., 2016).

When it comes to biological data, there is great variation among seeds from different years and regions, since morphophysiological characteristics are highly affected by climate, soil, parent plant characteristics, among other factors. Importantly, the biometric data obtained by spectral image analysis are sensitive to morphophysiological characteristics (e.g., pigmentation and organic compounds); therefore, variations in the characteristics of cultivars obtained in other harvests may be enough to misclassify them (Shrestha et al., 2016).

The intensity of phenotypic influence varies with species and the characteristics used for classification of cultivars. For some species, there are cultivars with outstanding characteristics that facilitate differentiation, as is the case with the tegument color of bean seeds

or peanut seeds. However, in many species, with subtle differences among cultivars, the influence on phenotypic variation can be a problem. For example, in some maize cultivars, the balance between sugar and amylopectin, which can be used as a spectral marker, can be affected by variations in water content in the seed from different regions, which can influence classification (Wang et al., 2016).

An alternative would be to use seeds from cultivars from other years and/or regions to train the classification model, but this poses some difficulties. The first is to obtain a sufficient number of seeds from other years and/or regions, since a large volume of samples is needed to overcome the effect of location. In addition, the use of seeds from a seed bank or archive samples, which were stored in different periods, could influence the classification of newly harvested seeds (Shrestha et al., 2016).

Some studies suggest the use of model updating, which seems to be a promising alternative. In this method, the original model, previously prepared using seeds from the same harvest, is updated with seeds from the following harvests, in order to have a more accurate model, without the need to perform the whole process of adjusting a new model. Some studies show 10 to 35% increase in overall accuracy when classifying cultivars coming from other years, when compared to a non-updated model (Guo et al., 2016; He et al., 2016; Huang et al., 2016a). Such a practice would partly solve the model portability problem; however, most model updating methods need to be updated with previously classified seeds, which requires time for sampling and classification. An alternative to updating the model is through semi-supervised classification algorithms that use the pre-label approach, i.e., they use the original model to classify a new seed sample (from another year), based on retraining the model with seeds classified with a high degree of confidence. However, these methods still need to be evaluated for different species and situations (Guo et al., 2016).

b) Seed shell influence

The external structures of the seeds can cause great influence on the analysis, either by using the visible spectrum, which is sensitive to their surface characteristics, or by using the near infrared spectrum, which is able to penetrate to the subsurface layer and is sensitive to the organic compounds present in these structures, for example, the bands of 1180 and 1470 nm, which are sensitive to the presence of lignin and fiber, commonly present in the seed coat of several species. Thus, among species whose external structures are the same, as occurs in palea and lemma in hybrid and self-pollinating cereals, the influence of these structures can be a limiting factor (Blackwell et al., 1977; Gao et al., 2013; Feng et al., 2017; Caporaso et al., 2021).

Thus, when it comes to the differentiation of genotypes that present external structures without enough distinguishing characteristics, the use of spectral imaging is not the best choice, since it would merely describe the composition of these structures. For this type of seed, processing would be necessary, but the commercial use of spectral image analysis requires processing, which leads to extra costs and is time consuming. In addition, the removal of the husk in cereal seeds limits the use of them, since the husk has an important protective function against fungi and insects (Abebe et al., 2004; Mortensen et al., 2021).

c) Seed orientation

The area exposed to the analysis may influence classification accuracy, given the sensitivity of the spectra used by the analysis to seed surface and subsurface compounds. The influence of orientation was reported in work using models fitted with measurements obtained from corn caryopses with the embryo facing up and the face facing down. The endosperm and the embryo have different compositions; therefore, choosing between the face of the caryopsis that has both structures (i.e., the face in which the embryo is facing up) or the face with only the endosperm, can influence one's ability to distinguish different genotypes. The influence of orientation can vary across genotypes, and in the case of differentiation between cultivars, there was an average variation of 10% (Miao et al., 2018; Tang et al., 2020). Sorting seeds with a certain orientation is a laborious job; therefore, when seed orientation plays a role but does not impair genotype differentiation, loss of accuracy can be accepted.

However, when the difference between genotypes is found in the embryo, as occurs between haploid and diploid maize caryopses, and separation is performed in breeding programs of the species for different purposes, seed orientation is essential. Thus, seed orientation can be a limiting factor when it comes to haploid seed identification if there is no processing before the analysis is performed; however, on a large scale, processing the seeds may be impractical (De la Fuente et al., 2017; Wang et al., 2018).

d) Specificity

Unlike other techniques, such as molecular markers, in which we can state with great certainty that a seed belongs to a particular genotype from a segment of its genetic code, in spectral image analysis we cannot use reflectance as an absolute marker of species/cultivar, because it is influenced by seed composition. In other words, to identify an unknown seed, it must always come from a sample in which all the possible classifications are known, which were previously defined considering lot-specific issues (e.g., harvest, year, storage). This is not a major limitation in a seed producing company, where the cultivars produced are known, but

in situations when one must identify possible seeds and/or adulterants from an unknown sample, the use of spectral image analysis is impractical.

Adulterant genotypes can be identified in a seed lot when this genotype is commonly used in the trade of the adulterated species. In this case, spectral image analysis tends to have good applicability, as the marked difference among species allows a model fitted to a particular variety or crop of the adulterated species to be distinguished with some ease from the adulterant species, as reported by Faqeerzada et al. (2020) in separating seeds of two varieties of almonds from adulterant apricot seeds.

Perspectives

a) Open data base and key wavelength

There has been increased interest in sharing the data collected through spectral image analysis - be it reflectance or biometric data regarding morphological characteristics of the seeds of the species used in the experiments (e.g., diameter, texture) - through online repositories. Data sharing can leverage the use of the analysis by directly allowing researchers to (a) test different chemometric techniques (e.g., preprocessing, classifier algorithms) on real data, without the need to perform a new experiment, and (b) more accurately identify key bands in certain species and cultivars when comparing different experiments.

The identification of key bands would help in the transition from hyperspectral equipment to multispectral equipment with more accurate and relevant bands in seed phenotyping. Hyperspectral equipment has the ability to measure many bands. However, many of these bands contain redundant or unnecessary information for the classification of most species; in addition, hyperspectral equipment is very expensive and more difficult to handle, since the reflectance of the various wavelengths are usually obtained by the point by point or line by line system, in which the object moves and reflectance is obtained for every pixel or line of pixels at a time, making the process more time consuming (Jaillais et al., 2015; Zhou et al., 2020a).

Thus, the migration to multispectral equipment seems to be the most obvious trend, since it requires fewer wavelengths that are applied to all pixels of the image at once, and it is more agile and suitable for application in the seed industry, especially in sectors that work with large numbers of cultivars and lots. A fast identification system is essential, especially in real time, and multispectral equipment is ideal for this purpose, (Elmarsy et al., 2019).

However, in order to efficiently develop multispectral equipment with key wavelengths, a deeper understanding is needed of the interaction of the different wavelengths

with the organic compounds of the different evaluated genotypes. To this end, an open database would facilitate such understanding (Elmarsy et al., 2019). Some technologies greatly benefit from an open database, e.g., to share data from Raman spectrometry and X-ray diffraction, which can be combined to identify different materials (Mendili et al., 2019). Naturally, when it comes to seeds, external factors have a great influence on the analysis (e.g., environment, parent plant) and consequently on their ability to be distinguished. However, with a large amount of data, one can identify relevant patterns between genotypes and at least direct the development of equipment, even if specific to certain species, to obtain a system capable of providing sufficient information for decision making in accepting or rejecting a seed lot, which would save a great deal of time and money (Elmarsy et al., 2019; Xia et al., 2019).

b) Field of application

One of the areas where spectral image analysis presents great potential is in breeding programs, especially in the production of hybrid seeds. Differentiating hybrid seeds from seeds generated by unwanted pollination, either from their parents or from self-pollination, is indispensable. This means differentiating between a few classes from samples with high genetic purity and coming from areas with production control and, thus with less variability among seeds, which is ideal for applying spectral analysis (Nie et al., 2019).

Forest species, for example, have a great lack of quality control methods. For species with great economic importance, such as the species of the genus Eucalyptus spp, the use of seeds is especially important in breeding programs for production of hybrids. The correct hybridization must be confirmed, given the difficulties of controlling pollination, either in indoor orchards or in the field. Thus, spectral image analysis has a great potential to meet this need and bring great advances to forest improvement programs (Ribeiro-Oliveira and Ranal, 2014).

Another relevant point that makes spectral image analysis an important tool in breeding programs is that images show individual morphological features of seeds, since the analysis allows the collection of both reflectance and spatial biometric data. The use of morphological features is especially important to check the homogeneity of seed morphological descriptors, because morphology is an attribute relatively unaffected by environmental issues and could be used to evaluate the genetic quality of a lot, which decreases with every generation (Mortensen et al., 2021).

c) Online and real time sorting systems

Probably the most promising aspect of spectral image analysis is the possibility of integration with an online system that allows real-time estimation of the quality of a seed lot. According to the International Seed Testing Association (ISTA, 2020), a certain amount of mixing of other cultivars is allowed in a seed lot. This is evaluated through purity analysis; however, there is great difficulty in determining the presence of other cultivars mixed in a lot in certain species, since the analysis depends on the analysts experience and their ability to identify cultivars by eye (Elmarsy et al., 2019).

Although each company presents a specific situation (i.e., different combinations of genotypes, number of genotypes, presence of different years and/or regions) and it is not clear to what extent spectral image analysis is able to handle these different situations, the fact is that in the studies identified in the present review, the analysis was effective. This means that, at least in certain situations, the analysis could be integrated into a system to estimate the genetic quality of a seed lot, since the alternative way (i.e., through purity analysis) is extremely laborious and, in many situations, impractical (Elmarsy et al., 2019).

Some researchers, such as Faqeerzada et al. (2020), reported the feasibility of an online system for real-time classification of seeds moved by a conveyor belt, in which the classification model previously adjusted using hyperspectral images in the infrared region was transferred to an online system. However, some problems still need to be overcome; for example, the speed of the conveyor belt, the variation in light, the overlapping of the seeds on the belt, among other points described by the authors.

Several studies have shown that the models developed from spectral information of seeds are robust enough for large-scale application of the analysis in real-time seed phenotyping, through the design of classification maps. In this way, the seeds are determined in real time as to the probability of belonging to a certain class, based on the color scale stipulated for each class. This approach enables decision making by the analyst and would act as a powerful tool to differentiate cultivars that would hardly be identified with the naked eye (Wang et al., 2016; Zhao et al., 2018a; Zhao et al., 2018b; Zhang et al., 2021).

Conclusions

The present review evaluated 44 papers that applied spectral image analysis in seed phenotyping; they were selected among 1304 papers identified in the main journal databases. The review sought to identify the main characteristics of the experiments described in the published papers, as well as to guide researchers in the choice of strategies for experimental

design and data analysis, since there are many ways to obtain a highly accurate classification model. Thus, after analyzing the papers, the following points summarize the main findings:

- All the evaluated studies presented satisfactory final accuracy; however, few used test data, as well as test and/or validation lots, which may have contributed to the high accuracy reported.
- As the application of multispectral analysis is relatively new in seed phenotyping, the works are still focused on agricultural species with greater economic appeal.
- Most studies have focused on the use of hyperspectral equipment, which works mainly in the near infrared region, and is sensitive to seed organic compounds and able to penetrate the subsurface layer. The use of the near infrared region seems to be a good strategy to identify differences between genotypes with similar surface structures, where visible light acts with greater intensity.
- The use of deep learning algorithms has been a trend in recent years, mostly because of its ability to work with more complex data, e.g., data collected by using hyperspectral cameras.
- Reflectance and biometric data on seed morphology provides sufficient information to separate different genotypes in several situations: among cultivars; hybrids and progenitors; and hybrids and lines, as well as in the separation of coated seeds.
- The main challenge of the analysis is certainly the phenotypic variation of the seeds, which implies the difficulty of using the adjusted model in the classification of cultivars from other harvest, years and/or regions. The main limitations refer to the sensitivity of reflectance to seed compounds, which are highly influenced by environmental issues; the influence of seed coat on the classification of genotypes with similar external characteristics; and the influence of seed orientation when the information needed for classification is on a certain face of the seed (e.g., face with the embryo).

Thus, the present review allowed a critical analysis of the use of spectral imaging in seed phenotyping, as well as a thorough evaluation of the limitations of this method. The practical application of this technique needs to be developed for use in laboratories with large volumes of analyses, lots, genotypes, and harvests. However, research has been accelerated to overcome the practical challenges of this method, as seen in work using model update algorithms, online classification systems, real-time classification maps; also, spectral information of genotypes is being shared through online repositories. Thus, there are strong

indications that the application of multispectral image analysis will become a part of the routine of seed analysis laboratories.

Acknowledgements

The authors thank the Coordination for the Improvement of Higher Education Personnel (CAPES) for granting the scholarships.

Conflicts

None declared.

References

Abebe T, Skadsen RW and Kaeppler HF (2004) Cloning and Identification of Highly Expressed Genes in Barley Lemma and Palea. *Crop Science* **44**, p. 942-950. doi.org/10.2135/cropsci2004.9420

Ali A, Mashwani WK, Tahir MH, Belhaouari SB, Alrabaiah H, Naeem S, Nasir JA, Jamal F and Chesneau C (2021) Statistical features analysis and discrimination of maize seeds utilizing machine vision approach. *Journal of Intelligent and Fuzzy Systems* **40**(1), 703–714.

Amigo J (2020) Data Handling in Science and Technology: Hyperspectral Imaging. Elsevier **32**, 630 p.

Bai X, Zhang C, Xiao Q, He Y and Bao Y (2020) Application of near-infrared hyperspectral imaging to identify a variety of silage maize seeds and common maize seeds. *RSC Advances* **10**(20), 11707–11715.

Bantan RAR, Ali A, Naeem S, Jamal F, Elgarhy M and Chesneau C (2020) Discrimination of sunflower seeds using multispectral and texture dataset in combination with region selection and supervised classification methods. *Chaos* **30**(11). doi:10.1063/5.0024017

Bao Y, Mi C, Wu N, Liu F and He Y (2019) Rapid classification of wheat grain varieties using hyperspectral imaging and chemometrics. *Applied Sciences (Switzerland)*, **9**(19). doi:10.3390/app9194119

Blackwell J, Cael JJ and Koenig JL (1977) Infrared and Raman-spectroscopy of cellulose. *American Chemical Society* **13**, p. 206-218.

Boelt B, Shrestha S, Salimi Z, Jørgensen, J, Nocolaisen M, Cartensen JM (2018) Multispectral imaging – a new tool in seed quality assessment? *Seed Science Research* **28**(3), 222–228. https://doi.org/10.1017/ S0960258518000235

Caporaso N, Whitworth MB and Fisk ID (2021) Total lipid prediction in single intact cocoa beans by hyperspectral chemical imaging. *Food Chemistry* **344**, 128663. doi.org/10.1016/j.foodchem.2020.128663

Carreiro Soares SF, Medeiros EP, Pasquini C, De Lelis Morello C, Harrop Galvão RK and Ugulino Araújo MC (2016) Classification of individual cotton seeds with respect to variety using near-infrared hyperspectral imaging. *Analytical Methods* **8**(48), 8498–8505.

De La Fuente GN, Carstensen JM, Edberg MA and Lü bberstedt T (2017) Discrimination of haploid and diploid maize kernels via multispectral imaging. *Plant Breeding* **136**(1), 50–60.

Dearden P, Kowalski B, Lowe J, Roland R, Surridge M, Thomas S, Jones S (2011) Mendeley Reference Manager. Mendeley Support Team London UK.

Elmasry G, Mandour N, Al-Rajaie S, Belin E, Rousseau D (2019) Recent Applications of Multispectral Imaging in Seed Phenotyping and Quality Monitoring — An Overview. *Sensors* **19**(5), 1–32. doi:10.3390/s19051090

Elmasry G, Mandour N, Al-Rejaie, S, Belin E and Rousseau D (2019) Recent applications of multispectral imaging in seed phenotyping and quality monitoring — An overview. *Sensors (Switzerland)* **19**(5) p. 1090.

Fabiyi SD, Vu H, Tachtatzis C, Murray P, Harle D, Dao TK, Andonovic I, Ren J and Marshall S (2020) Varietal Classification of Rice Seeds Using RGB and Hyperspectral Images. *IEEE ACCESS*, **8**, 22493–22505.

Faqeerzada MA, Perez M, Lohumi S, Lee H, Kim G, Wakholi C, Joshi R and Cho BK (2020) Online application of a hyperspectral imaging system for the sorting of adulterated almonds. *Applied Sciences (Switzerland)*, **10**(18), 1–16.

Feng X, Peng C, Chen Y, Liu X, Feng X and He Y (2017) Discrimination of CRISPR/Cas9-induced mutants of rice seeds using near-infrared hyperspectral imaging. *Scientific Reports* **7**. doi:10.1038/s41598-017-16254-z

Gao J, Li X, Zhu F and He Y (2013) Application of hyperspectral imaging technology to discriminate different geographical origins of Jatropha curcas L. seeds. *Computers And Electronics in Agriculture* **99**, 186–193.

Gheisari M, Wang G and Bhuiyan MDZA (2017) A Survey on Deep Learning in Big Data. p. 173 In IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), August 2017, Guangzhou China.

Guo D, Zhu Q, Huang M, Guo Y and Qin J (2017) Model updating for the classification of different varieties of maize seeds from different years by hyperspectral imaging coupled with a pre-labeling method. *Computers and Electronics in Agriculture* **142**, 1–8.

Hansen MA E, Hay FR and Carstensen JM (2016) A virtual seed file: The use of multispectral image analysis in the management of genebank seed accessions. *Plant Genetic Resources: Characterisation and Utilisation* **14**(3), 238–241.

Hastie T, Tibshirani R and Friedman J (2017) The Elements of Statistical Learning: Data, Inference, and Prediction (12 ed). Springer, 737 p.

He C, Zhu Q, Huang M and Mendoza F (2016) Model updating of hyperspectral imaging data for variety discrimination of maize seeds harvested in different years by clustering algorithm. *Transactions of the ASABE*, **59**(6), 1529–1537.

Huang M, He C, Zhu Q and Qin J (2016b) Maize seed variety classification using the integration of spectral and image features combined with feature transformation based on hyperspectral imaging. *Applied Sciences (Switzerland)* **6**(6). doi:10.3390/app6060183

Huang M, Tang J, Yang B and Zhu Q (2016a) Classification of maize seeds of different years based on hyperspectral imaging and model updating. *Computers and Electronics in Agriculture* **122**, 139–145.

ISTA (2020) The International Seed Testing Association. International Rules for Seed Testing. International Rules for Seed Testing, 300p.

Jaillais B, Roumet P, Pinson-Gadais L and Bertrand D (2015) Detection of Fusarium head blight contamination in wheat kernels by multivariate imaging. *Food Control* **54**, p. 250–258.

Jia S, An D, Liu Z, Gu J, Li S, Zhang X, Zhu D, Guo T and Yan Y (2015) Variety identification method of coated maize seeds based on near-infrared spectroscopy and chemometrics. *Journal of Cereal Science* **63**, 21–26.

Kong W, Zhang C, Liu F, Nie P and He Y (2013) Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. *Sensors (Basel Switzerland)* **13**(7), 8916–8927.

Lammertyn J, Nicolai B, Ooms K, De Smedt V, De Baerdemaeker J (1998) Non-destructive measurement of acidity soluble solids and firmness of Jonagold apples using NIR spectroscopy. *Trans. ASAE* **41**, 1089–1094

Li H, Jiang D, Cao J and Zhang D (2020) Near-infrared spectroscopy coupled chemometric algorithms for rapid origin identification and lipid content detection of *Pinus koraiensis* seeds. *Sensors (Switzerland)* **20**(17), 1–17.

Li L, Zhang Q and Huang D (2014) A review of imaging techniques for plant phenotyping. *Sensors (Switzerland)* **14**, 20078–20111

Li X, Fan X, Lili Zhao Huang S, He Y and Suo X (2020) Discrimination of pepper seed varieties by multispectral imaging combined with machine learning. *Applied Engineering in Agriculture* **36**(5), 743–749.

Liu C, Liu W, Lu X, Chen W, Chen F, Yang J and Zheng L (2014a) Non-destructive discrimination of conventional and glyphosate-resistant soybean seeds and their hybrid descendants using multispectral imaging and chemometric methods. *Journal of Agricultural Science* **154**(1), 1–12.

Liu C, Liu W, Lu X, Chen W, Yang J and Zheng L (2014b) Nondestructive determination of transgenic Bacillus thuringiensis rice seeds (*Oryza sativa* L.) using multispectral imaging and chemometric methods. *Food Chemistry* **153**, 87–93.

Liu W, Liu C, Hu X, Yang J and Zheng L (2016) Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics. *Food Chemistry* **210**, 415–421.

Liu W, Liu C, Ma F, Lu X, Yang J and Zheng L (2016) Online Variety Discrimination of Rice Seeds Using Multispectral Imaging and Chemometric Methods. *Journal of Applied Spectroscopy* **82**(6), 993–999.

Manattayil JK, Ravichandran NK, Wijesinghe RE, Shirazi MF, Lee SY, Kim P, Jung HY, Jeon M and Kim J (2018) Non-destructive classification of diversely stained capsicum annuum seed specimens of different cultivars using near-infrared imaging based optical intensity detection. *Sensors (Switzerland)* **18**(8). doi:10.3390/s18082500

Mbanjo EG N, Jones H, Caguiat XG I, Carandang S, Ignacio JC, Ferrer MC, Boyd LA and Kretzschmar T (2019) Exploring the genetic diversity within traditional Philippine pigmented Rice. *Rice* **12**(1). doi:10.1186/s12284-019-0281-2

Mendili YE, Vaitkus A, Merkys A, Grazulis S, Chateigner, D, Mathevet, F, Gascoin S, Petit S, Bardeau JF, Zanatta M, Secchi M, Mariotto G, Kumar A, Cassetta M, Lutterotti L, Borovin E, Orberger B, Simon P, Hehlen B and Le Guen M (2019) Raman Open Database: first interconnected Raman–X-ray diffraction open-access resource for material identification. *Journal of applied crystallographic* **52**, 618–625. doi.org/10.1107/S1600576719004229

Miao A, Zhuang J, Tang Y, He Y, Chu X and Luo S (2018) Hyperspectral image-based variety classification of waxy Maize seeds by the t-SNE model and procrustes analysis. *Sensors (Switzerland)* **18**(12), 11–14.

Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PloS Medicine* **6**, 1-6.

Mortensen AK, Gislum R, Jørgensen JR and B Boelt (2021) The Use of Multispectral Imaging and Single Seed and Bulk Near-Infrared Spectroscopy to Characterize Seed Covering Structures: Methods and Applications in Seed Testing and Research. *Agriculture* **11(**301), p.1–18. doi.org/10.3390/agriculture11040301

Nie P, Zhang J, Feng X, Yu C and He Y (2019) Classification of hybrid seeds using near-infrared hyperspectral imaging technology combined with deep learning. *Sensors and Actuators B: Chemical* **296**, 126630.

Osborne BG, Douglas S (1981) Measurement of the degree of starch damage in flour by near infrared reflectance analysis. *J. Sci. Food Agr.* **32**, 328–332.

Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald SMcGuiness LA, Stewart

LA, Thomas J, Tricco AC, Welch VA, Whiting P, McKenzie JE (2021) PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BJM* **372**, 1-36.

Qiu Z, Chen J, Zhao Y, Zhu S, He Y and Zhang C (2018) Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. *Applied Sciences (Switzerland)* **8**(2), 1–12.

Rahman A, Cho BK (2016) Assessment of seed quality using nondestructive measurement techniques: a review. *Seed Science Research* **26**, 285–305.

Ribeiro-Oliveira JP and Ranal MA (2014) Brazilian forest seeds: a precarious beginning, a heady present and the future, will it be promising? *Ciência Florestal* **24**(3), p. 771-784.

Rodríguez-Pulido FJ, Barbin DF, Sun DW, Gordillo B, González-Miret ML and Heredia FJ (2013) Grape seed characterization by NIR hyperspectral imaging. *Postharvest Biology and Technology* **76**, 74–82.

Serranti S, Cesare D, Marini F, Bonifazi G (2013) Classification of oat and groat kernels using NIR hyperspectral imaging. *Talanta* **103**, 276–284.

Shrestha S, Deleuran LC and Gislum R (2016b) Classification of different tomato seed cultivars by multispectral visible-near infrared spectroscopy and chemometrics. *Journal of Spectral Imaging* **5**. doi:10.1255/jsi.2016.a1

Shrestha S, Deleuran LC, Olesen MH and Gislum R (2015) Use of Multispectral Imaging in Varietal Identification of Tomato. *Sensors* **15**(2), 4496–4512.

Shrestha S, Knapič M, Žibrat U, Deleuran LC and Gislum R (2016a) Single seed near-infrared hyperspectral imaging in determining tomato (Solanum lycopersicum L.) seed quality in association with multivariate data analysis. *Sensors and Actuators B: Chemical* **237**, 1027–1034.

Tang Y, Cheng Z, Miao A, Zhuang J, Hou C, He Y, Chu X and Luo S (2020) Evaluation of cultivar identification performance using feature expressions and classification algorithms on optical images of sweet corn seeds. *Agronomy* **10**(9). doi:10.3390/agronomy10091268

Vrešak M, Olesen MH, Gislum R, Bavec F and Jørgensen JR (2016) The use of image-spectroscopy technology as a diagnostic method for seed health testing and variety identification. *PLoS ONE* **11**(3), 1–10.

Wang L, Sun DW, Pu H and Zhu Z (2016) Application of Hyperspectral Imaging to Discriminate the Variety of Maize Seeds. *Food Analytical Methods* **9**(1), 225–234.

Wang Y, Lv Y, Liu H, Wei Y, Zhang J, An D and Wu J (2018) Identification of maize haploid kernels based on hyperspectral imaging technology. *Computers and Electronics in Agriculture* **153**, 188–195.

Wei Y, Li X, Pan X and Li L (2020) Nondestructive classification of soybean seed varieties by hyperspectral imaging and ensemble machine learning algorithms. *Sensors (Switzerland)* **20**(23), 1–12.

Williams P and Norris K (2001) Near-Infrared Technology: In the Agricultural and Food Industries (2nd ed). American Association of Cereal Chemists, 296 p.

Wu N, Zhang Y, Na R, Mi C, Zhu S, He Y and Zhang C (2019) Variety identification of oat seeds using hyperspectral imaging: Investigating the representation ability of deep convolutional neural network. *RSC Advances* **9**(22), 12635–12644.

Xia C, Yang S, Huang M, Zhu Q, Guo Y and Qin J (2019) Maize seed classification using hyperspectral image coupled with multi-linear discriminant analysis. *Infrared Physics and Technology* **103**, 103077.

Xia Y, Xu Y, Li J, Zhang C, Fan S (2019) Recent advances in emerging techniques for non-destructive detection of seed viability: A review. *Artificial Intelligence In Agriculture* **1**.

Yang L, Zhang Z and Hu X (2020) Cultivar discrimination of single alfalfa (Medicago sativa l.) seed via multispectral imaging combined with multivariate analysis. *Sensors (Switzerland)* **20**(22), 1–14.

Yang S, Zhu QB, Huang M and Qin JW (2017) Hyperspectral Image-Based Variety Discrimination of Maize Seeds by Using a Multi-Model Strategy Coupled with Unsupervised Joint Skewness-Based Wavelength Selection Algorithm. *Food Analytical Methods* **10**(2) 424–433.

Zapotoczny P, Żuk-Gołaszewska K and Ropelewska E (2016) Discrimination based on changes in the physical properties of fenugreek (*Trigonella foenum-graecum* L.) seeds subjected to various cultivation conditions. *European Food Research and Technology* **242**(3) 405–414.

Zhang C, Zhao Y, Yan T, Bai X, Xiao Q, Gao P, Li M, Huang W, Bao Y, He Y and Liu F (2020) Application of near-infrared hyperspectral imaging for variety identification of coated maize kernels with deep learning. *Infrared Physics and Technology* **111**. doi:10.1016/j.infrared.2020.103550

Zhang J, Dai L and Cheng F (2021) Corn seed variety classification based on hyperspectral reflectance imaging and deep convolutional neural network. *Journal of Food Measurement and Characterization* **15**(1), 484–494.

Zhao H, Guo B, Wei Y, Zhang B (2014) Effects of grown origin genotype harvest year and their interactions of wheat kernels on near infrared spectral fingerprints for geographical traceability. *Food Chem* **152** 316–322.

Zhao Y, Zhang C, Zhu S, Gao P, Feng L and He Y (2018a) Non-destructive and rapid variety discrimination and visualization of single grape seed using near-infrared hyperspectral imaging technique and multivariate analysis. *Molecules* **23**(6). doi:10.3390/molecules23061352

Zhao Y, Zhu S, Zhang C, Feng X, Feng L and He Y (2018b) Application of hyperspectral imaging and chemometrics for variety classification of maize seeds. *RSC Advances* **8**(3), 1337–1345.

Zhou L, Zhang C, Taha MF, Wei X, He Y, Qiu Z and Liu Y (2020a) Wheat Kernel Variety Identification Based on a Large Near-Infrared Spectral Dataset and a Novel Deep Learning-Based Feature Selection Method. *Frontiers in Plant Science* **11**. doi:10.3389/fpls.2020.575810

Zhou Q, Huang W, Fan S, Zhao F, Liang D and Tian X (2020b) Non-destructive discrimination of the variety of sweet maize seeds based on hyperspectral image coupled with wavelength selection algorithm. *Infrared Physics and Technology* **109**.

Zhu S, Chao M, Zhang J, Xu X, Song P, Zhang J and Huang Z (2019a) Identification of soybean seed varieties based on hyperspectral imaging technology. *Sensors (Switzerland)*, **19**(23). doi:10.3390/s19235225

Zhu S, Zhang J, Chao M, Xu X, Song P, Zhang J and Huang Z (2020) A rapid and highly efficient method for the identification of soybean seed varieties: Hyperspectral images combined with transfer learning. *Molecules* **25**(1). doi:10.3390/molecules25010152

Zhu S, Zhou L, Gao P, Bao Y, He Y and Feng L (2019b) Near-infrared hyperspectral imaging combined with deep learning to identify cotton seed varieties. *Molecules* **24**(18). doi:10.3390/molecules24183268

Zhu S, Zhou L, Zhang C, Bao Y, Wu B, Chu H, Yu Y, He Y and Feng L (2019c) Identification of soybean varieties using hyperspectral imaging coupled with convolutional neural network. *Sensors (Switzerland)* **19**(19). doi:10.3390/s19194065

CAPÍTULO 2

Multispectral imaging for distinguishing hybrid forest seeds of Corymbia spp. and Eucalyptus spp. from their progenitors[1]

[1]Capítulo em formato de artigo conforme submetido na revista *Computers and Electronics in Agriculture*.

CAPÍTULO 2: MULTISPECTRAL IMAGING FOR DISTINGUISHING HYBRID FOREST SEEDS OF CORYMBIA SPP. AND EUCALYPTUS SPP. FROM THEIR PROGENITORS

Multispectral imaging for distinguishing hybrid forest seeds of *Corymbia* spp. and *Eucalyptus* spp. from their progenitors

Thomas Bruno Michelon*, Jens Michael Carstensen[a,b], Elisa Serra Negra Vieira[c], Maristela Panobianco[d]

[a] Videometer A/S, Lyngsø Allé 3, DK- 2970 Hørsholm, Denmark; [b] DTU Compute, Technical University of Denmark, DK-2800, Kongens Lyngby, Denmark

[c] Embrapa Forestry – Estrada da Ribeira, km 111, CEP 83411-000, Colombo, PR, Brazil.

[d] Department of Plant Science, Federal University of Paraná, R. dos Funcionários, 1540, CEP 80035-050, Curitiba, PR, Brazil

* Department of Plant Science, Federal University of Paraná, R. dos Funcionários, 1540, CEP 80035-050, Curitiba, PR, Brazil. <thomasbrunomichelon@gmail.com>

Abstract

In the forest industry, interspecific hybridization, such as *Eucalyptus urograndis* (*Eucalyptus grandis* × *Eucalyptus urophylla*) and *Corymbia maculata* × *Corymbia torelliana*, has led to the development of high-performing F1 generations. The successful breeding of these hybrids relies on verifying progenitor origins and confirming post-crossing, but conventional genotype identification methods are resource-intensive and result in seed destruction. As an alternative, multispectral imaging analysis has emerged as an efficient and non-destructive tool for seed phenotyping. This approach has demonstrated success in various crop seeds. However, identifying seed species in the context of forest seeds presents unique challenges due to their natural phenotypic variability and the striking resemblance between different species. This study evaluates the efficacy of spectral imaging analysis in distinguishing hybrid seeds of *E. urograndis* and *C. maculata* × *C. torelliana* from their progenitors. Four experiments were conducted: one for *Corymbia* spp. seeds, one for each *Eucalyptus* spp. batch separately, and one for pooled batches. Multispectral images were acquired at 19 wavelengths within the spectral range of 365 to 970 nm. Classification models based on Linear Discriminant Analysis

(LDA), Random Forest (RF), and Support Vector Machine (SVM) was created using reflectance and reflectance features, combined with color, shape, and texture features, as well as nCDA transformed features. The LDA algorithm, combining all features, provided the highest accuracy, reaching 98.15% for *Corymbia* spp., and 92.75%, 85.38, and 86.00 for *Eucalyptus* batch one, two, and pooled batches, respectively. The study demonstrated the effectiveness of multispectral imaging in distinguishing hybrid seeds of *Eucalyptus* and *Corymbia* species. The seeds' spectral signature played a key role in this differentiation. This technology holds great potential for non-invasively classifying forest seeds in breeding programs.

Keywords: Breeding; Machine learning; Machine vision; Phenotyping; Spectral imaging.

INTRODUCTION

Interspecific hybridization has played a significant role in the advancement of the forest sector industry. Crossbreeding within genera consistently results in F1 generations that manifest heterotic effects, surpassing the performance of their progenitors in critical aspects such as growth rate, disease resistance, and overall wood quality, making them the preferred choice for plantation forestry (Ibarra et al., 2023; Ramalho et al., 2022). Notably, the cross between *Eucalyptus grandis* and *Eucalyptus urophylla* demonstrates heightened resistance to prevalent diseases, making it a widely embraced choice in both Brazilian and South African contexts. Furthermore, *Corymbia torelliana* hybrids have found extensive use in tropical and subtropical areas, primarily due to their significantly higher growth rate (da Silva et al., 2022; Ramalho et al., 2022; Van Den Berg et al., 2015).

To ensure the production of hybrid seeds within the context of a breeding program, it is important to guarantee the origin of the selected progenitors, confirming their alignment with the desired genotypes. Additionally, post-crossing verification is essential to confirm the authenticity of the resulting genotypes and prevent the use of seeds from inadvertent crosses. This verification is crucial whether the crosses occur within controlled indoor environments or open field crossing orchards. In the latter case, the inherent challenges related to fertilization control, makes the risk of unintended crossbreeding with other species higher (Dickinson et al., 2013; Ramalho et al., 2022). Moreover, the seed production process in tree cultivation is marked by high costs and extended growth timelines. Consequently, the ability to distinguish seeds plays a pivotal role in the success of breeding programs within the forest sector.

Traditionally, genotype identification hinges on resource-intensive methods involving molecular markers and biochemical assays, which not only incur high costs and demand significant labor but also result in the destruction of seeds (Boelt et al., 2018; Shrestha and Hardeberg, 2015). Therefore, an urgent demand exists for the development of non-invasive and efficient techniques to assess seed genotypes (Elmasry et al., 2019; Michelon et al., 2023; Xia et al., 2019).

In the field of seed analysis and technology, spectral imaging analysis emerges as a cutting-edge tool. This innovative technique seamlessly combines spectroscopy with digital imaging, empowering rapid quantification of unique phenotypic traits within individual seeds. This, in turn, expedites the efficient differentiation of seeds from one another. The swiftness and non-destructive nature of this method, coupled with its capacity for automating seed identification processes, have led to its widespread adoption and utilization (Elmasry et al., 2019; Michelon et al., 2023; Xia et al., 2019; Zhou et al., 2020).

Multispectral image analysis has proven to be an exceptional tool for distinguishing between various seed genotypes, as demonstrated by prior research (Elmasry et al., 2019). Notably, this technique achieved a remarkable 98% accuracy in distinguishing between conventional and glyphosate-resistant soybean seeds and their hybrid offspring (Liu et al., 2014). Furthermore, in a study covering a wide range of cultivars, multispectral analysis effectively separated 12 distinct varieties of alfalfa, with an accuracy rate of 93.47% (Yang et al., 2020). It is notable that a systematic review comprising 11 studies that employed multispectral analysis for seed phenotyping identified a surprising average accuracy of 91.34% (Michelon et al., 2023).

Although this technology has been explored with success in seed phenotyping over different crops, there appears to be a gap in the literature concerning its utilization in the phenotyping of forest seeds. Identifying seed genotypes presents an added layer of complexity in the realm of forest seeds, primarily due to the absence of domestication which amplifies the natural variability within species. This challenge is especially pronounced when dealing with the *Eucalyptus* spp. and *Corymbia* spp. seeds, which are inherently difficult to discern due to their diminutive size and striking similarity to seeds of other species (Ibarra et al., 2023). As such, this study seeks to assess the potential in distinguishing hybrid seeds of *Eucalyptus urograndis* (*Eucalyptus grandis* × *Eucalyptus urophylla*), as well *Corymbia maculata* × *Corymbia torelliana* from their progenitors through the spectral imaging analysis technique.

MATERIAL AND METHOD

Seed Samples

In the conducted experiment, two samples per seed genotype of *Eucalyptus grandis*, *Eucalyptus urophylla*, and the hybrid *Eucalyptus urograndis* (*Eucalyptus grandis* × *Eucalyptus urophylla*) as well as one sample of *Corymbia maculata*, *Corymbia torelliana*, and the hybrid *Corymbia maculata* × *Corymbia torelliana* were evaluated (Fig. 1). Each seed sample weighed approximately 50g and was obtained from the reduction of batches corresponding to each genotype.

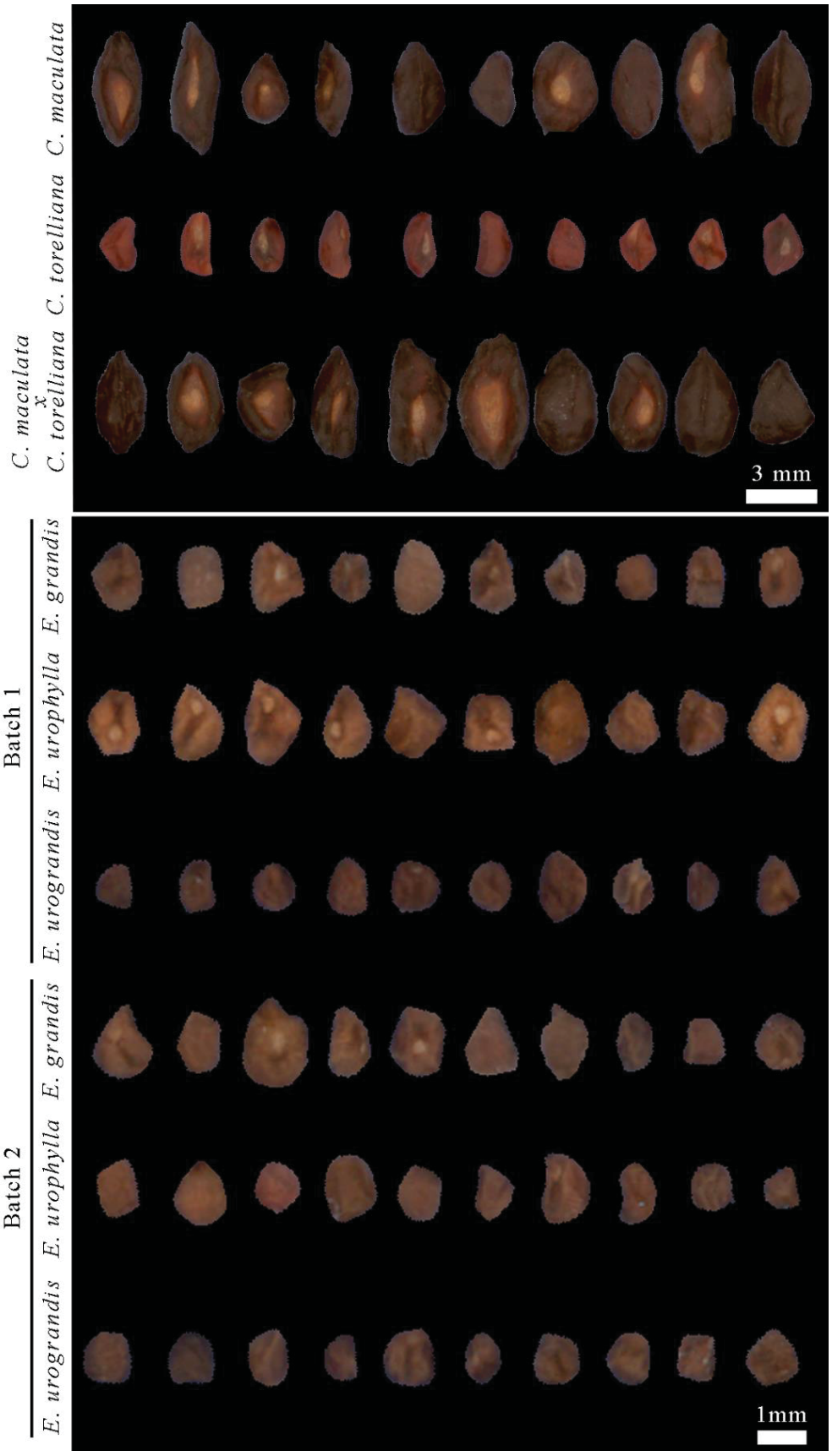


Figure 1. *Corymbia maculata, Corymbia torelliana, Corymbia maculata x Corymbia torelliana, Eucalyptus grandis, Eucalyptus urophylla, Eucalyptus urograndis* batch one and two seeds.

The seeds were obtained from forestry breeding programs, with parent trees located in the southern region of Brazil, and were produced in 2020. It is important to note that, in the case of *Corymbia* spp. seeds, they were obtained from five different parent trees. The *Eucalyptus* seeds were categorized into two batches, identified as batch one and batch two, each containing a sample of the hybrid and its progenitors. This classification took into account the respective breeding programs, resulting in samples with similar production characteristics, geographical origin, and harvest season, as illustrated in Fig. 2. All seeds were packaged in plastic containers and stored with temperature and humidity controlled until the start of the experimental activities.

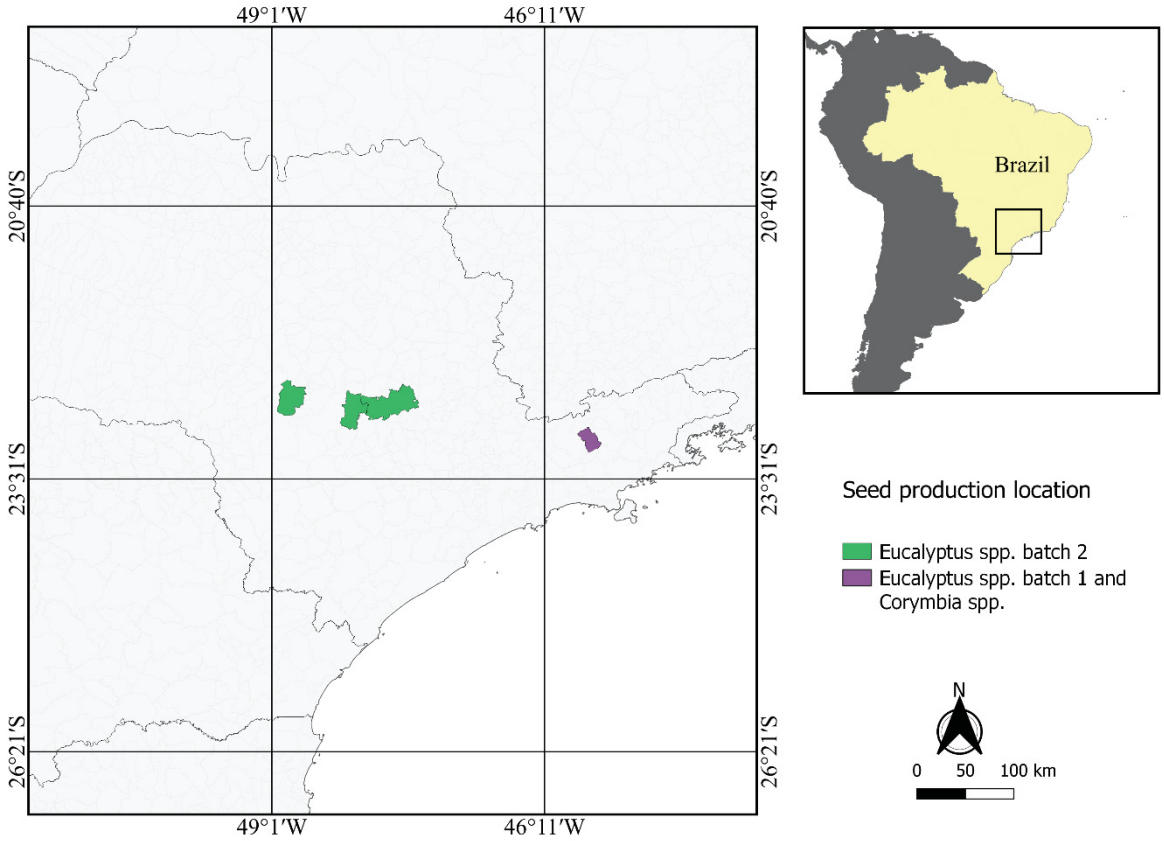


Figure 2. *Corymbia maculata*, *Corymbia torelliana*, *Corymbia maculata x Corymbia torelliana*, *Eucalyptus grandis*, *Eucalyptus urophylla*, *Eucalyptus urograndis* batch one and two seed production sites.

From each homogenized sample, seeds were separated from impurities (e.g., empty seeds and chaff) with a aim of tweezers, a magnifying glass, and sieves. Subsequently, only healthy seeds without visible damage were chosen to form the dataset, as indicated in Table 1.

Table 1. Overview of genotypes, batches and quantity of seeds used in the experiment.

Batch	Genotype	Number of seeds	Batch	Genotype	Number of seeds	Genotype	Number of seeds
1	*E. grandis*	245	2	*E. grandis*	311	*C. maculata*	495
	E. urophylla	245		*E. urophylla*	309	*C. toreliana*	500
	E. urograndis	245		*E. urograndis*	310	*C. maculata x C. torelliana*	519
Total		735	Total		930	Total	1514

Each dataset was split into a training set (75%) and a validation set (25%) to assess the models' performance.

Multispectral imaging system

The VideometerLab4 system (Videometer, Hørsholm, Denmark) was used for capturing the multispectral image of the samples. The system consisted of a coated matte sphere with LEDs along the rim and a monochromatic camera with high spatial resolution (40 µm/pixel and 2192 × 2192 pixels). The system underwent radiometric, geometric, and light setup calibration before capturing the images. The samples were illuminated by the LEDs at 19 wavelengths, which included ultraviolet (365, 405 nm), visible (430, 450, 470, 490, 515, 540, 570, 590, 630, 645, 660, and 690 nm), and near-infrared (780, 850, 880, 890, and 970 nm), resulting in 19 monochrome pictures per sample.

Multispectral imaging analysis

The VideometerLab software (version 3.24.11) was employed for image segmentation and feature extraction. A predetermined mask was utilized to isolate regions of interest (ROIs), which corresponded to seeds, while simultaneously eliminating the background (the blue plate and petri dish). Spectral features, including reflectance mean and standard deviation, were extracted from these ROIs. In addition, a set of morphological features was selected based on visual inspection of the seeds' characteristics. These morphological features encompassed area, autocorrelation energy, beta shape a and b, CIE color space components, compactness (circle and ellipse), eccentricity, hue intensity and saturation mean, length, max edge distance, moment y ratio, non-convex area, pointedness, width-to-length ratio, region color band local standard deviation, region horizontal length mean, skew (y and x), vertical orientation, vertical skewness,

and width. Additionally, Normalized Canonical Discriminant Analyses (nCDA) based on reflectance mean were used to create new features that increase the differences between the classes. The nCDA is a supervised model used to minimize the distance between observations within classes and maximize the distance between observations between classes (Cruz-Castillo et al., 1994). As a result, three extra features were created: hybrid vs. progenitor one, hybrid vs. progenitor two, and progenitor one vs. progenitor two. Features with near-zero variance were excluded, leaving a total of 89 features per seed, including 38 spectral, 48 morphological, and three nCDA-transformed features. All features were exported to an Excel file for data analysis.

Multivariate analysis methods

Seed samples were used to design four experiments: one for *Corymbia* spp. seeds, one for each *Eucalyptus* spp. batch separately, and one for pooled batches. Three algorithms, Linear Discriminant Analysis (LDA), Random Forest, and Support Vector Machine (SVM), were tested using two models each. The first model used only reflectance data (mean and standard deviation), resulting in 38 features, while the second model included all features (reflectance, morphology, and nCDA features), resulting in 89 features. In total, six models were created and evaluated for accuracy using the 10-fold cross-validation method. Sensitivity and specificity values were calculated for each class. The analyses were performed using R version 3.5.2 and RStudio software version 2022.02.3, along with caret package version 6.0-92.

RESULTS

The average spectral profiles of each *Eucalyptus* spp. and *Corymbia* spp., along with the Principal Component Analysis (PCA) plot derived from the set of 89 measured features per seed, are illustrated in Fig. 3. In Fig. 3A, the mean reflectance values for the *Corymbia* spp. species reveal a notable divergence among the species. Notably, *Corymbia torelliana* exhibits an overall higher reflectance than the other species, particularly in the wavelength range of 515 – 970 nm. The *Corymbia* hybrid seeds share more spectral similarities with *Corymbia maculata*, particularly within the 370 – 690 nm region. Looking at the additional features through PCA plots (Fig. 3B), it becomes evident that the hybrid closely aligns with *C. maculata*, as indicated by the overlapped region, while C. *torelliana* exhibits distinct differences from both species, forming a separate cluster. In terms of *Eucalyptus* spp., both batches displayed consistent mean reflectance patterns within their respective species, as shown in Fig. 3C and Fig. 3E. In both batches, the hybrid seeds exhibited lower mean reflectance compared to their parental species,

resulting in a distinct gap between them, while the parental species showed a higher degree of similarity across the spectra range. The PCA plots (Fig. 3D and Fig. 3F) reveal a substantial overlap within species, with batch two exhibiting a more pronounced overlap. Both reflectance and PCA indicate high similarities within the genotypes from *Eucalyptus* spp. seeds.

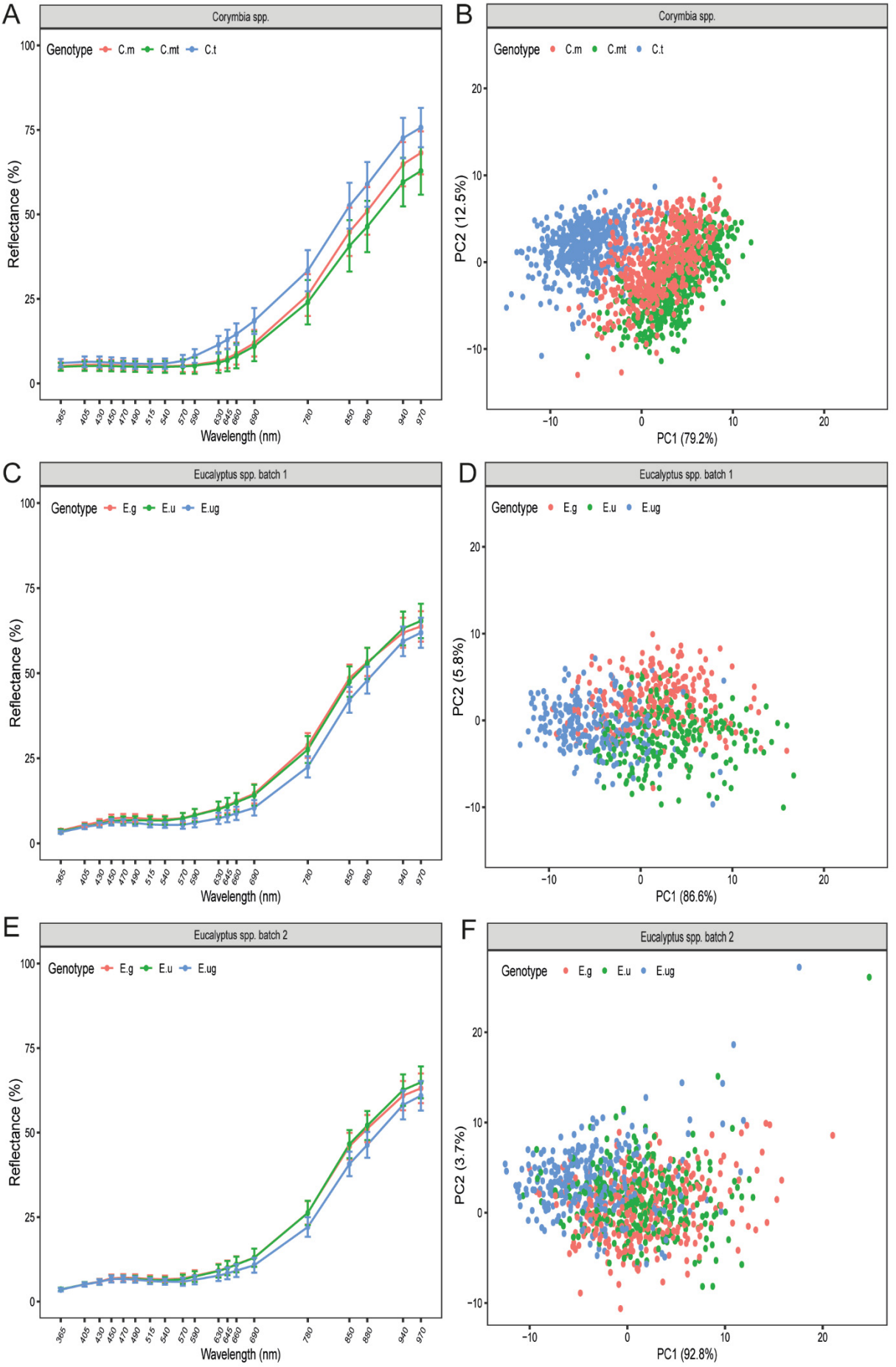

A
Corymbia spp.
Genotype C.m C.mt C.t
Reflectance (%)
Wavelength (nm)
B
Corymbia spp.
Genotype C.m C.mt C.t
PC2 (12.5%)
PC1 (79.2%)
C
Eucalyptus spp. batch 1
Genotype E.g E.u E.ug
Reflectance (%)
Wavelength (nm)
D
Eucalyptus spp. batch 1
Genotype E.g E.u E.ug
PC2 (5.8%)
PC1 (86.6%)
E
Eucalyptus spp. batch 2
Genotype E.g E.u E.ug
Reflectance (%)
Wavelength (nm)
F
Eucalyptus spp. batch 2
Genotype E.g E.u E.ug
PC2 (3.7%)
PC1 (92.8%)

Figure 3. The average reflectance spectrum (A, C, E) and PCA using all features (B, D, F) from *C. maculata* (C.m), *C. maculata x torelliana* (C.mt), *C. torelliana* (C.t), *E. urophyllia* (E.u), *E. urograndis* (E.ug) and *E. grandis* (E.g) seeds batch 1 and 2.

Fig. 4 illustrates the distinctions between two genotypes within the reflectance spectrum of individual pixels, achieved through the application of normalized canonical discriminant analysis (nCDA) transformation. Comparing the hybrid genotype to its progenitor *C. maculata* (Fig. 4A), an array of pixels displaying colors spanning both extremes of the scale is noticeable within the seeds. This suggests greater similarity between their reflectance. Conversely, when comparing hybrid seeds from *Corymbia* spp. to *C. torelliana* (Fig. 4B), predominance of dark blue pixels is observed on the hybrid picture, reflecting a contrast to the predominantly red/yellow pixels of its progenitor. This discrepancy points to a notable divergence in spectral reflectance between these two genotypes. These results corroborate with Fig. 1A where the hybrid genotype's reflectance spectrum is more related to *C. maculata* than *C. torelliana*.

The transformed reflectance images of the *Eucalyptus* seeds from lots 1 and 2 are depicted in Fig. 4C-F. Overall, a higher heterogeneity is observed among each contrast, indicated by pixels ranging from blue to red within seeds. Furthermore, a noteworthy aspect is the presence of seeds with uniform coloration but belonging to opposite classes on the rating scale. Those characteristics are more evident in lot 2 and suggest a stronger resemblance between hybrid genotypes and their progenitors.

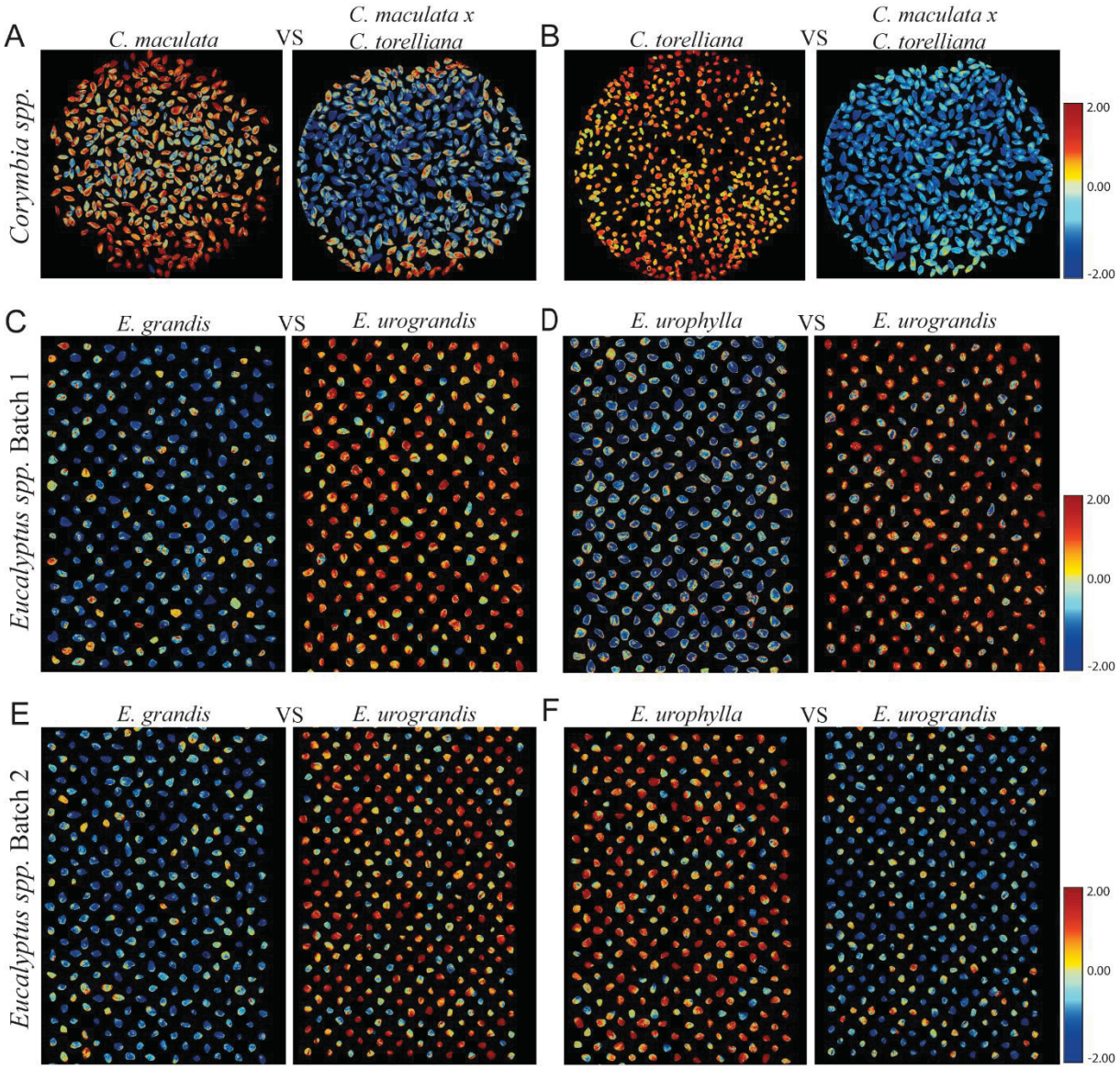


Figure 4. nCDA transformed multispectral images from hybrid seeds of *Corymbia* spp. and *Eucalyptus* spp. batch one and batch two versus its progenitors.

The performance of genotype classification algorithms for *Eucalyptus* spp., using both spectral and morphological features, is summarized in Tab. 2. The results indicate that the LDA algorithm outperformed the SVM and RF algorithms for both individual lots and their combined data. Lot 1 achieved a notable overall accuracy of 92.7%, with *E. urograndis* demonstrating the highest rate of correctly classified seeds, reaching 94.69% sensitivity. In contrast, Lot 2 exhibited lower accuracy compared to Lot 1 and the combined lots, at 85.38%, with *E. urophylla* showing a higher incidence of misclassification, as evident in the confusion matrix. Ultimately, the combined lots attained an overall accuracy of 86.01%, with hybrid seeds

exhibiting the lowest error rates according to the confusion matrix and displaying higher indices of sensitivity and specificity at 89.91% and 94.05%, respectively.

Table 2. Confusion matrix and overall accuracy based on Random Forest (RF), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) with spectral and morphological features of *Eucalyptus* spp. batch 1, batch 2 and pooled batches.

		Batch 1			Batch 2			Pooled		
		E. grandis	*E. urophylla*	*E. urograndis*	*E. grandis*	*E. urophylla*	*E. urograndis*	*E. grandis*	*E. urophylla*	*E. urograndis*
RF	*E. grandis*	197	27	19	269	50	32	462	83	47
	E. urophylla	19	210	19	22	217	27	53	420	48
	E. urograndis	29	8	207	20	42	251	41	51	460
	Sensitivity (%)	80.41	85.71	84.49	86.50	70.23	80.97	83.09	75.81	82.88
	Specificity (%)	90.61	92.24	92.45	86.75	92.11	90.00	88.28	90.91	91.71
	Overwall accuracy (%)		83.54			79.25			80.60	
SVM	*E. grandis*	210	13	12	266	29	22	463	58	36
	E. urophylla	15	226	11	26	252	24	56	451	26
	E. urograndis	20	6	222	19	28	264	37	45	493
	Sensitivity (%)	85.71	92.24	90.61	85.53	81.55	85.16	83.27	81.41	88.83
	Specificity (%)	94.90	94.69	94.69	91.76	91.95	92.42	91.52	92.62	92.61
	Overall accuracy (%)		89.52			84.09			84.50	
LDA	*E. grandis*	221	11	5	276	32	29	478	59	34
	E. urophylla	9	229	8	25	253	16	52	455	22
	E. urograndis	15	5	232	10	24	265	26	40	499
	Sensitivity (%)	90.20	93.47	94.69	88.75	81.88	85.48	85.97	82.13	89.91
	Specificity (%)	96.73	96.53	95.92	90.15	93.40	94.52	91.61	93.34	94.05
	Overall accuracy (%)		**92.79**			**85.38**			**86.01**	

Tab. 3 represents the performance of classification algorithms for *Eucalyptus* spp. seeds using only spectral variables in the model. The LDA algorithm showed higher accuracy in both isolated batches, as well as when combined, with respective accuracies of 91.70%, 82.15%, and 82.94%, followed by SVM and the RF algorithm. When comparing the accuracy of the LDA algorithm with the results of the same algorithm using a combination of spectral and morphological features (Tab. 2), it is observed that the classification of genotypes was minimally affected by using only spectral features. This fact becomes more evident in batch 1, where the accuracy decreased by only 1.02%, while maintaining the same sensitivity and specificity for hybrid seeds.

Table 3. Confusion matrix and overall accuracy based on Random Forest (RF), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) with spectral features of *Eucalyptus* spp. batch 1, batch 2 and pooled batches.

		Batch 1			Batch 2			Pooled		
		E. grandis	*E. urophylla*	*E. urograndis*	*E. grandis*	*E. urophylla*	*E. urograndis*	*E. grandis*	*E. urophylla*	*E. urograndis*
RF	*E. grandis*	172	30	27	231	56	47	400	96	68
	E. urophylla	37	208	16	52	211	29	91	390	56
	E. urograndis	36	7	202	28	42	234	65	68	431
	Sensitivity (%)	70.20	84.90	82.45	74.28	68.28	75.48	71.94	70.40	77.66
	Specificity (%)	88.37	89.18	91.22	83.36	86.96	88.71	85.21	86.77	88.02
	Overall accuracy (%)		79.18			72.69			73.33	
SVM	*E. grandis*	213	19	8	258	36	28	442	68	37
	E. urophylla	14	222	7	34	226	29	72	420	45
	E. urograndis	18	4	230	19	47	253	42	66	473
	Sensitivity (%)	86.94	90.61	93.88	82.96	73.14	81.61	79.50	75.81	85.23
	Specificity (%)	94.49	95.71	95.51	89.66	89.86	89.35	90.53	89.47	90.27
	Overall accuracy (%)		90.48			79.25			80.18	
	E. grandis	213	14	5	268	46	30	454	59	34

LDA	*E. urophylla*	14	229	8	29	241	25	67	444	38
	E. urograndis	18	2	232	14	22	255	35	51	483
	Sensitivity (%)	86.94	93.47	94.69	86.17	77.99	82.26	81.65	80.14	87.03
	Specificity (%)	96.12	95.51	95.92	87.72	91.30	94.19	91.61	90.55	92.25
	Overwall accuracy (%)	**91.70**			**82.15**			**82.94**		

The performance of models utilizing both morphological and reflectance features for the classification of *Corymbia* spp. genotypes is presented in Tab. 4. Overall, all the models exhibited exceptional performance, achieving an accuracy rate exceeding 90% across all algorithms. Notably, the RF algorithm demonstrated the lowest accuracy, with the SVM and LDA algorithms following suit. Remarkably, LDA achieved an accuracy of 98.71% in classification and excelled in classifying hybrid *Corymbia* seeds, with only 10 misclassifications out of 519 seeds. *C. torelliana* seeds were consistently and accurately classified across all algorithms, with a sensitivity rate reaching 99%.

Table 4. Confusion matrix and overall accuracy based on Random Forest (RF), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) with spectral and morphological features of *Corymbia* spp.

		C. maculata	*C. maculata x C. torelliana*	*C. torelliana*
RF	*C. maculata*	448	52	5
	C. maculata x C. torelliana	44	465	0
	C. torelliana	3	2	495
	Sensitivity (%)	90.51	89.60	99.00
	Specificity (%)	94.41	95.58	99.51
	Overwall accuracy (%)		93.00	
SVM	*C. maculata*	471	27	4
	C. maculata x C. torelliana	22	492	0
	C. torelliana	2	0	496
	Sensitivity (%)	95.15	94.80	99.20
	Specificity (%)	96.96	97.79	99.80
	Overall accuracy (%)		96.37	
LDA	*C. maculata*	482	10	3
	C. maculata x C. torelliana	13	509	2

C. torelliana	0	0	495
Sensitivity (%)	97.37	98.07	99.00
Specificity (%)	98.72	98.49	100.00
Overall accuracy (%)		**98.15**	

The performance of models using just the reflectance features for the classification of *Corymbia* spp. genotypes is presented in Tab. 5. The model employing the RF algorithm exhibited lower performance compared to SVM and LDA, primarily due to misclassifications between hybrid seeds and those of *C. maculata*. In contrast, SVM and LDA yielded similar results, achieving overall accuracies of 96.63% and 97.49%, respectively. Furthermore, when comparing these models to those incorporating all features (as shown in Tab. 4), RF demonstrates a substantial reduction in accuracy, while LDA effectively maintains its performance with only a marginal 0.66% reduction. Notably, SVM outperforms its previous model in this context. These results underscore the significance of seed spectrum reflectance as a primary component in the screening of *Corymbia* spp. genotypes.

Table 5. Confusion matrix and overall accuracy based on Random Forest (RF), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) with spectral features of *Corymbia* spp.

		C. maculata	*C. maculata x C. torelliana*	*C. torelliana*
RF	*C. maculata*	412	103	29
	C. maculata x C. torelliana	75	414	2
	C. torelliana	8	2	469
	Sensitivity (%)	83.23	79.77	93.80
	Specificity (%)	87.05	92.26	99.01
	Overall accuracy (%)		85.54	
SVM	*C. maculata*	471	18	7
	C. maculata x C. torelliana	20	501	2
	C. torelliana	4	0	491
	Sensitivity (%)	95.15	96.53	98.20
	Specificity (%)	97.55	97.79	99.61
	Overall accuracy (%)		96.63	
LDA	*C. maculata*	477	10	8
	C. maculata x C. torelliana	18	509	2
	C. torelliana	0	0	490
	Sensitivity (%)	96.36	98.07	98.00

Specificity (%)	98.23		97.99	100.00
Overall accuracy (%)		**97.49**		

The bidimensional plot of LDA functions LD1 and LD2, based on spectral and morphological features of *Corymbia* spp. and *Eucalyptus* spp. seeds (pooled batches), as well as the individual feature contributions are present in Fig. 5. Overall, the LDA functions explained nearly 100% of the total variation and were successful in separating the *Corymbia* genotypes, particularly *C. torelliana*, with no overlapping observed within the 95% confidence interval (Fig. 5A). Notably, the spectral signature transformed by nCDA emerged as the feature with the highest contribution for all genotypes (Fig. 5C), followed by shape characteristics such as length and area, texture attributes, including moment ratio, as well as color and reflectance features from the red spectrum (645, 630, and 660 nm). Additionally, the wavelengths within the near-infrared spectrum, specifically those between 940 and 970 nm, made a substantial contribution to the model's performance regarding the hybrid genotype and *C. torelliana*.

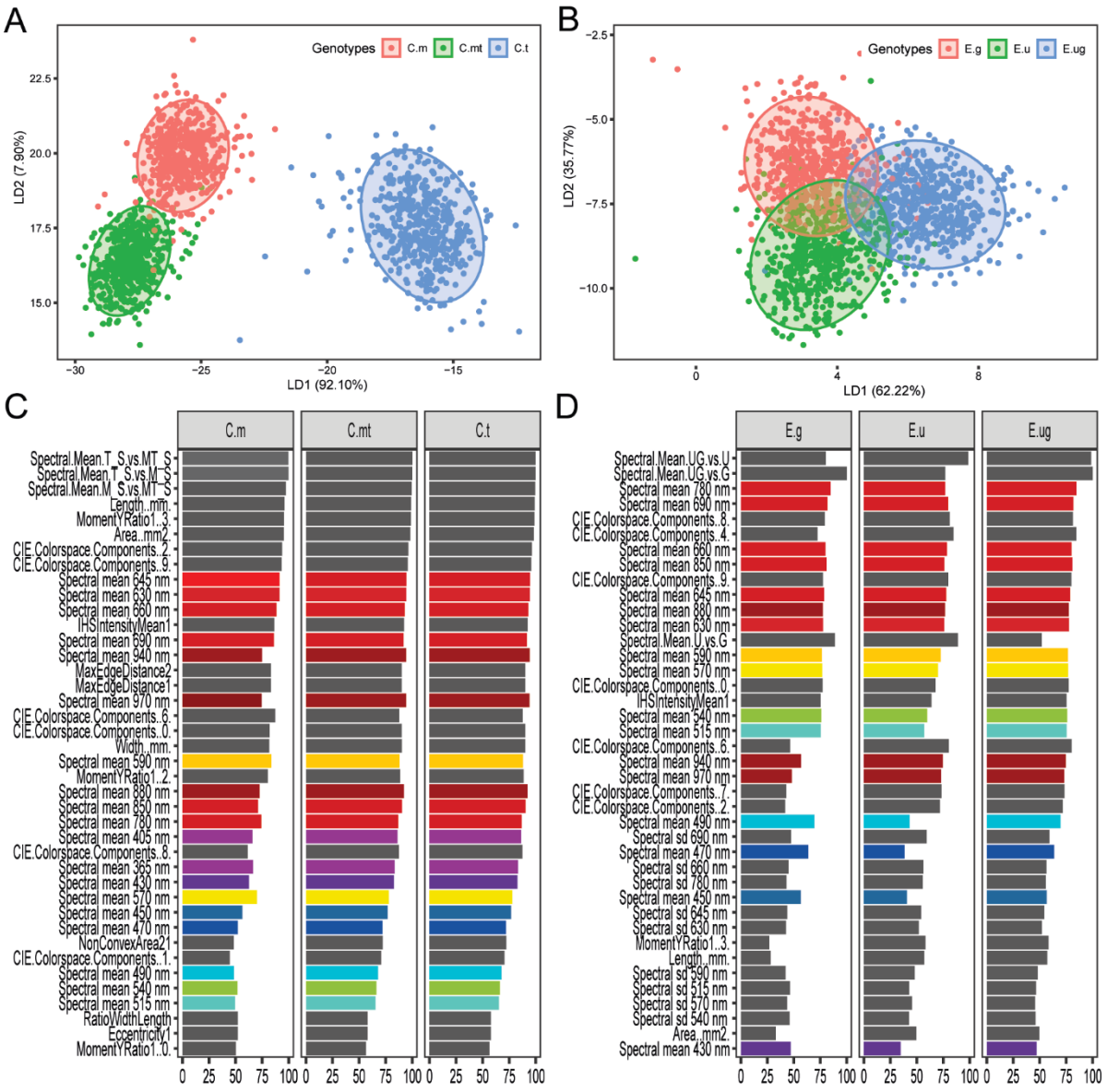


Figure 5. Linear Discriminant Analysis (LDA) score plot based on spectral and morphological features from *Corymbia* spp. (A) and *Eucalyptus* spp. (B) seeds. The ellipse shows the 95% confidence interval. The individual contribution from the 40 more important variables from the LDA model for *Corymbia* spp. (C) and *Eucalyptus* spp. (D). C.m, C.mt, C.t stands for *Corymbia maculata*, *Corymbia maculata* × *Corymbia torelliana*, *Corymbia torelliana*, where E.g., E.u, and E.ug stand for *Eucalyptus grandis*, *Eucalyptus urophylla*, and *Eucalyptus urograndis*, respectively.

For the *Eucalyptus* genotypes pooled batches, the LDA functions were able to explain 98% of the total variation (Fig. 5B). The plotting of both functions revealed overlapping cluster regions, indicating a notable degree of similarity among these classes. Notably, the nCDA-

transformed spectral signature emerged as the most significant feature for distinguishing between the hybrids and their progenitors (Fig. 5D), followed by wavelengths' reflectance from the near-infrared range (780 and 850 nm), the red spectrum (660, 690 nm) and the seed color components.

DISCUSSION

In this study, we explored the application of multispectral imaging to distinguish hybrid forest seeds from their progenitors, specifically focusing on Corymbia and Eucalyptus species. The developed models exhibited exceptional performance, with accuracy consistently exceeding 90%. The utilization of single-seed multispectral information consistently sustained the high accuracy in genotype classification. Additionally, the nCDA supervised algorithm played a crucial role in discriminating between genotypes, by enhancing classification through the reduction of spectral redundancy between adjacent bands (Wei et al., 2020). It is worth noting the significance of specific wavelength ranges for each species, notably 630, 645, 660, 690, 940, and 970 nm for *Corymbia* spp. and 630, 645, 660, 690, 780, 850, and 880 nm for *Eucalyptus* spp. The reflectance values at these wavelengths exhibited high sensitivity to distinguishing key features in the seed coat of hybrid genotypes from their progenitors. The region from 630 to 690 nm corresponds to the red spectrum and is associated with seed pigments, e.g., tannins and anthocyanin. In contrast, the near-infrared region (780 – 970 nm) is partially absorbed by C-H, N-H, and O-H bonds, while the remainder reflects and enable the measurement of water and organic compounds, including proteins, carbohydrates, and lipids (Esteve Agelet and Hurburgh, 2014; Mortensen et al., 2021; Sendin et al., 2018).

The *Eucalyptus* spp. exhibited lower accuracy compared to the Corymbia seeds, and a significant variation in accuracy was observed between different seed batches. This observation aligns with the noise evident in the Eucalyptus nCDA images (Fig. 4 C-F). The noise in the nCDA images of Eucalyptus seeds can be attributed to several factors. Firstly, Eucalyptus is an allogamous plant with substantial genetic diversity, which can lead to the presence of seeds resulting from undesirable crosses and/or natural phenotypic variation. This variation may be due to the varying degree of resemblance to one of the parents for the hybrid genotype, in addition to the absence of domestication of the species. Oliveira et al. (2023) highlighted the presence of genetic admixture in *Eucalyptus* spp., which could result in an unexpected ancestral genomic composition of interspecific hybrids. This genetic admixture, especially among phylogenetically closer species that easily hybridize in exotic environments, is expected and

could contribute to some of the observed noise. Furthermore, Van der Berg (2015) estimated the genetic parameters based on a large population of *E. urograndis* seedlings and reported that the genetic variance mostly arises from non-additive variance, particularly dominance variance. High dominance variance often results in a greater range of phenotypic variation within a population, where individuals with the same genotype for a trait may exhibit a wide range of trait values due to the influence of dominant alleles.

Despite the observed noise, the genotype signal seems to be significantly stronger. The mean reflectance patterns were similar between both batches of Eucalyptus spp. (Fig. 3 C and Fig. 3 E). Furthermore, the LDA model achieved an impressive 92.79% and 85.38% classification accuracy for *Eucalyptus* spp. batches one and two, respectively. These results indicate that the seeds within each batch exhibit clear similarities. The samples share similar production regions and harvest periods within their respective batches, strongly suggesting that genotype is the primary distinguishing factor. Identifying hybrid seeds can be challenging due to the potential overlap in material content inherited from the parent plants, which may cause difficulty in distinguishing them from their progenitors (Zhang et al., 2018). However, previous studies have revealed distinct spectral signatures within genotypes, leading to highly accurate classifications of parents and their offspring. This phenomenon has been observed in various crops and vegetables such as soybean, okra and tomatoes (Liu et al., 2016; Shrestha et al., 2015; Zhang et al., 2018).

The combination of spatial and spectral features contributed to the model's accuracy and robustness, as exemplified by the noteworthy performance of the 98.15% LDA model applied to *Corymbia* spp (Fig. 5C). Parameters such as length, area, and texture contributed to the model's efficacy, a result that aligns with expectations, particularly considering the smaller size of *C. torelliana* seeds. Additionally, the model performed consistently well regardless of the seeds' orientation, showing that even differences observed between *Corymbia* seeds facing up or down did not present an obstacle. This feature could greatly enhance the scalability of the analysis. Furthermore, the common presence of impurities and empty seeds, a typical challenge encountered with forest seeds, may be a challenge to the initial phase of image segmentation. For *Corymbia* spp., this challenge is readily surmountable, as the discernible distinction between the seeds and impurities facilitates separation using a sieve. Conversely, when dealing with *Eucalyptus* seeds, the structural similarities necessitate more intricate screening procedures. However, it is worth noting that the presence of chaff and empty seeds, common

challenges in the context of *Eucalyptus* genus, and a pre-screening method can be justified particularly in the context of breeding programs.

In summary, the combination of multispectral imaging with multivariate analysis methods emerges as a powerful tool for seed phenotyping in the context of hybrid forest seeds. The effectiveness and non-invasive characteristics of this technology play an innovative role in breeding programs, providing a new and straightforward step to verify the alignment of the genotype to be used. The analysis demonstrated a high classification accuracy for both genera, with *Corymbia* spp. standing out, where the seed spectrum proved to be the most significant factor. For future studies, we recommend the use of hybrid seeds obtained from controlled pollination orchards or incorporating a single-seed genotyping test to confirm the seed genotype in the training set. This step will enhance the model's reliability, increase its accuracy, and make it an even more robust tool for distinguishing hybrid seeds.

CONCLUSION

The study has demonstrated the effectiveness of multispectral imaging in distinguishing hybrid seeds of *Eucalyptus urograndis* and *Corymbia maculata* × *Corymbia torelliana* from their progenitors. The spectral signature of the seeds genotypes significantly contributes to the high performance of the models. These results highlight the potential of multispectral imaging as a powerful and non-invasive tool for classifying forest seed genotypes in the context of breeding programs.

AKNOWLEDGEMENTS

To the Coordination for the Improvement of Higher Education Personnel (CAPES – Finance Code 001) for providing a scholarship to the first author, the company Suzano for donating seeds used in the experiment, and the company Videometer for providing the facilities and equipment for the experiment.

REFERENCES

Boelt, B., Shrestha, S., Salimi, Z., Jorgensen, J.R., Nicolaisen, M., Carstensen, J.M., 2018. Multispectral imaging - A new tool in seed quality assessment? Seed Sci. Res. 28, 222–228. https://doi.org/10.1017/S0960258518000235

da Silva, P.H.M., Lee, D.J., Amancio, M.R., Araujo, M.J., 2022. Initiation of breeding programs for three species of Corymbia: Introduction and provenances study. Crop Breed. Appl.

Biotechnol. 22, 1–9. https://doi.org/10.1590/1984-70332022v22n1a01

Cruz-Castillo, J.G.; Ganeshanandam, S.; MacKay, B.R.; Lawes, G.S.; Lawoko, C.R.O.; Woolley, D.J. Applications of canonical discriminant analysis in horticultural research. HortScience 1994, 29, 1115–1119.

Dickinson, G.R., Wallace, H.M., Lee, D.J., 2013. Reciprocal and advanced generation hybrids between Corymbia citriodora and C. torelliana: Forestry breeding and the risk of gene flow. Ann. For. Sci. 70, 1–10. https://doi.org/10.1007/s13595-012-0231-2

Elmasry, G., Mandour, N., Al-Rejaie, S., Belin, E., Rousseau, D., 2019. Recent applications of multispectral imaging in seed phenotyping and quality monitoring—An overview. Sensors (Switzerland). https://doi.org/10.3390/s19051090

Esteve Agelet, L., Hurburgh, C.R., 2014. Limitations and current applications of Near Infrared Spectroscopy for single seed analysis. Talanta 121, 288–299. https://doi.org/10.1016/j.talanta.2013.12.038

Ibarra, L., Hodge, G., Acosta, J.J., 2023. Quantitative Genetics of a Hybrid Population of Eucalyptus nitens × Eucalyptus globulus: Estimation of Genetic Parameters and Implications for Breeding Strategies. Forests 14. https://doi.org/10.3390/f14020381

Liu, C., Liu, W., Lu, X., Chen, W., Chen, F., Yang, J., Zheng, L., 2016. Non-destructive discrimination of conventional and glyphosate-resistant soybean seeds and their hybrid descendants using multispectral imaging and chemometric methods. J. Agric. Sci. 154, 1–12. https://doi.org/10.1017/S0021859614001142

Liu, C., Liu, W., Lu, X., Chen, W., Chen, F., Yang, J., Zheng, L., 2014. Non-destructive discrimination of conventional and glyphosate-resistant soybean seeds and their hybrid descendants using multispectral imaging and chemometric methods. J. Agric. Sci. 154, 1–12. https://doi.org/10.1017/S0021859614001142

Michelon, T.B., Serra Negra Vieira, E., Panobianco, M., 2023. Spectral imaging and chemometrics applied at phenotyping in seed science studies: a systematic review. Seed Sci. Res. 33, 9–22. https://doi.org/10.1017/s0960258523000028

Mortensen, A.K., Gislum, R., Jørgensen, J.R., Boelt, B., 2021. The use of multispectral imaging and single seed and bulk near-infrared spectroscopy to characterize seed covering structures: Methods and applications in seed testing and research. Agric. 11. https://doi.org/10.3390/agriculture11040301

Ramalho, M.A.P., Santos, H.G., Souza, T. da S., 2022. Eucalyptus breeding programs: a proposal for the use of inbred progenies. Cerne 28, 1–9.

https://doi.org/10.1590/01047760202228013049

Sendin, K., Manley, M., Williams, P.J., 2018. Classification of white maize defects with multispectral imaging. Food Chem. 243, 311–318. https://doi.org/10.1016/j.foodchem.2017.09.133

Shrestha, R., Hardeberg, J.H., 2015. An experimental study of fast multispectral imaging using LED illumination and an RGB camera. Final Progr. Proc. - IS T/SID Color Imaging Conf. 2015-Janua, 36–40.

Shrestha, S., Deleuran, L.C., Olesen, M.H.H.H.H., Gislum, R.R., 2015. Use of multispectral imaging in varietal identification of tomato. SENSORS 15, 4496–4512. https://doi.org/10.3390/s150204496

Van Den Berg, G.J., Verryn, S.D., Chirwa, P.W., Van Deventer, F., 2015. Genetic Parameters of Interspecific Hybrids of Eucalyptus grandis and E. urophylla Seedlings and Cuttings. Silvae Genet. 64, 291–308. https://doi.org/10.1515/sg-2015-0027

Wei, Y., Li, X., Pan, X., Li, L., 2020. Nondestructive classification of soybean seed varieties by hyperspectral imaging and ensemble machine learning algorithms. Sensors (Switzerland) 20, 1–12. https://doi.org/10.3390/s20236980

Xia, Y., Xu, Y., Li, J., Zhang, C., Fan, S., 2019. Recent advances in emerging techniques for non-destructive detection of seed viability : A review Arti fi cial Intelligence in Agriculture Recent advances in emerging techniques for non-destructive detection of seed viability : A review. Artif. Intell. Agric. 1. https://doi.org/10.1016/j.aiia.2019.05.001

Yang, L., Zhang, Z., Hu, X., 2020. Cultivar discrimination of single alfalfa (Medicago sativa l.) seed via multispectral imaging combined with multivariate analysis. Sensors (Switzerland) 20, 1–14. https://doi.org/10.3390/s20226575

Zhang, J., Feng, X., Liu, X., He, Y., 2018. Identification of hybrid okra seeds based on near-infrared hyperspectral imaging technology. Appl. Sci. 8. https://doi.org/10.3390/app8101793

Zhou, Q., Huang, W., Fan, S., Zhao, F., Liang, D., Tian, X., 2020. Non-destructive discrimination of the variety of sweet maize seeds based on hyperspectral image coupled with wavelength selection algorithm. Infrared Phys. Technol. 109, 103418. https://doi.org/10.1016/j.infrared.2020.103418

CAPÍTULO 3

Soybean single-seed respiration evaluation through spectral imaging [1]

[1]Formato de artigo conforme as normas da revista *Seed Science and Research*.

CAPÍTULO 3: SOYBEAN SINGLE-SEED RESPIRATION EVALUATION THROUGH SPECTRAL IMAGING

Title: Soybean single-seed respiration evaluation through spectral imaging

Running title: Seed respiration through spectral imaging

Thomas B. Michelon[1]*
https://orcid.org/0000-0002-7437-5062
Fushing Hsieh[2]
https://orcid.org/0000-0002-9292-6980
Pedro Bello[3]
Bárbara Blanco-Ulate[3]
https://orcid.org/0000-0002-8819-9207
Maristela Panobianco[1]
http://orcid.org/0000-0002-9990-2172

[1]Department of Plant Science – Federal University of Paraná – R. dos Funcionários, 1540, CEP 80035-050, Curitiba, PR, Brazil.
[2] Department of Statistics, University of California Davis, 95616
[3] Department of Plant Science, University of California Davis, 95616

*Corresponding author <thomasnbrunomichelon@gmail.com>, phone +55 (41) 99166-5252.

Soybean single-seed respiration evaluation through spectral imaging

ABSTRACT

Soybean (*Glycine max* (L.) Merril) is an important global crop for oil and protein production, requiring high-vigor seeds for optimal growth in various conditions. Traditional seed quality assessments are time-consuming and subjective. Thus, spectral imaging, which combines spectroscopy and digital imaging, emerges as a promising tool for seed analysis. While widely used in seed technology, its application to physiological quality assessment is limited. The study aimed to assess. The study aimed to assess the relationship between soybean seeds respiration and its biometric features through multispectral imaging. Multispectral images were captured from 1808 seeds, and 75 features were extracted. The seeds, followed by oxygen consumption during germination, were computed individually. The respiration data and biometric measurements were categorized. The biometric measurements were paired and clustered, resulting in more than 8,000 unique traits, where the respiration curves were classified into 12 groups and three clusters were selected. The association between respiration patterns and biometric features was conducted using contingency tables and entropy analysis. The results demonstrated differences among respiration patterns specially in autofluorescence excitation-emission at 365/600 nm, 430/700 nm, 450/700 nm, and 470/700 nm, as well as differences in reflectance at 365 nm, 690 nm, and 405 nm. Additionally, two-way interaction of features highlighted different patterns on seed biometric features composition leading to the same physiological quality. In conclusion, soybean seed appearance and spectral data correlate strongly with respiration and seed quality. Multispectral imaging is non-invasive and efficient in identifying traits linked to respiration patterns and visualizing their relationship, enhanced by autofluorescence and reflectance.

Keywords: *Glycine max* (L.) Merril, machine vision, multivariate analysis, oxygen consumption, seed respiration spectroscopy, vigor

INTRODUCTION

Soybean (*Glycine max* (L.) Merril) stands out as one of the most crucial crops globally, contributing significantly to oil and protein production and serving as a vital source of food for both humans and livestock. The success of soybean cultivation hinges significantly on the use of high-vigor seeds which enable the rapid and uniform establishment of plant stands, even in

diverse environmental conditions, including under biotic and abiotic stress. High-vigor seeds often lead to more resistant seedlings, promoting better field performance and potentially higher yields (Caverzan et al., 2018; Cheng et al., 2023; ISTA, 2020). Conventional methods for assessing a seed lot's physiological quality, such as germination and vigor tests, are widely applied in the seed industry. However, these methods rely heavily on time-consuming, labor-intensive, and subjective assessments. Additionally, they lack automation, are destructive, and may require specialized training. Hence, there is a pressing need for rapid and non-subjective tests to efficiently determine the physiological quality of seeds (Elmasry et al., 2019; Xia et al., 2019).

Spectral imaging analysis has emerged as an innovative tool in the context of seed science. This technique combines spectroscopy with digital imaging, enabling the measurement of spatial characteristics, spectral features, and auto-fluorescence of individual seeds. Notably, the method is both rapid and non-destructive, facilitating automation and scalability in analysis processes (Boelt et al., 2018). By measuring various traits of a single seed, this technique is widely employed for seed classification in areas such as distinguishing varieties, identifying hybrid seeds from their progenitors and detecting adulterants in seed samples (Faqeerzada et al., 2020; Fu et al., 2023; Zhang et al., 2018).

Although spectral image analysis is being widely used in seed technology, studies relating this analysis to physiological quality are incipient, as it is difficult to measure elements that are related to germination and vigor (Caverzan et al., 2018; Elmasry et al., 2019; França-Silva et al., 2023; Xia et al., 2019). Studies employing spectral imaging analysis in seed physiological quality have primarily focused on distinguishing pre-established viability and/or vigor classes through accelerated aging and tetrazolium tests, or by correlating them with biochemical compounds extracted from a seed population (Barboza da Silva et al., 2021; Jin et al., 2022; Kandpal et al., 2016; Olesen et al., 2015; Qi et al., 2023; Shrestha et al., 2017). While beneficial for distinguishing seed lots, measurements on seed populations have limited ability to link various characteristics with specific aspects of seed quality. Without assessing individual seeds, determining which seeds contribute to observed changes in characteristics becomes challenging (Bradford, 2018; Bradford et al., 2013).

In this context single-seed respiration plays a key role in seed vigor evaluation. The relationship between seed quality and respiration is already known through direct and indirect measures, such as tetrazolium tests. The capacity of a dry seed in repairing its respiratory systems post-imbibition indicates seed quality, with high-vigor seeds typically exhibiting

elevated respiration rate patterns. While poor seed quality is often linked to either a delay in the onset of respiration or the incapacity of vital seed tissues to initiate respiratory activity (Bello and Bradford, 2016; Bradford et al., 2013).

Connecting the characteristics of an individual seed, as revealed by its spectral image, with its respiration pattern could help in understanding the factors influencing this pattern. This linkage could enable predictions of the respiration patterns of seeds based on their characteristics, offering insights into their vigor. The study aimed to assess the relationship between soybean seeds respiration and its biometric features through multispectral imaging.

MATERIAL AND METHODS

Seed sample

Six samples totaling 1808 soybean seeds of the cultivar 55i57 RSF ipro, harvested in the 2021/2022 season and produced in the region of Ponta Grossa, southern Brazil (25°05'52.1"S 50°09'25.7"W), were used (Fig. 1). The seeds obtained in a seed laboratory originated from the reduction of six distinct seed lots, resulting in six individual samples of approximately 500g each. Each sample was numbered from 1 to 6 and stored in plastic bags at 10 °C.

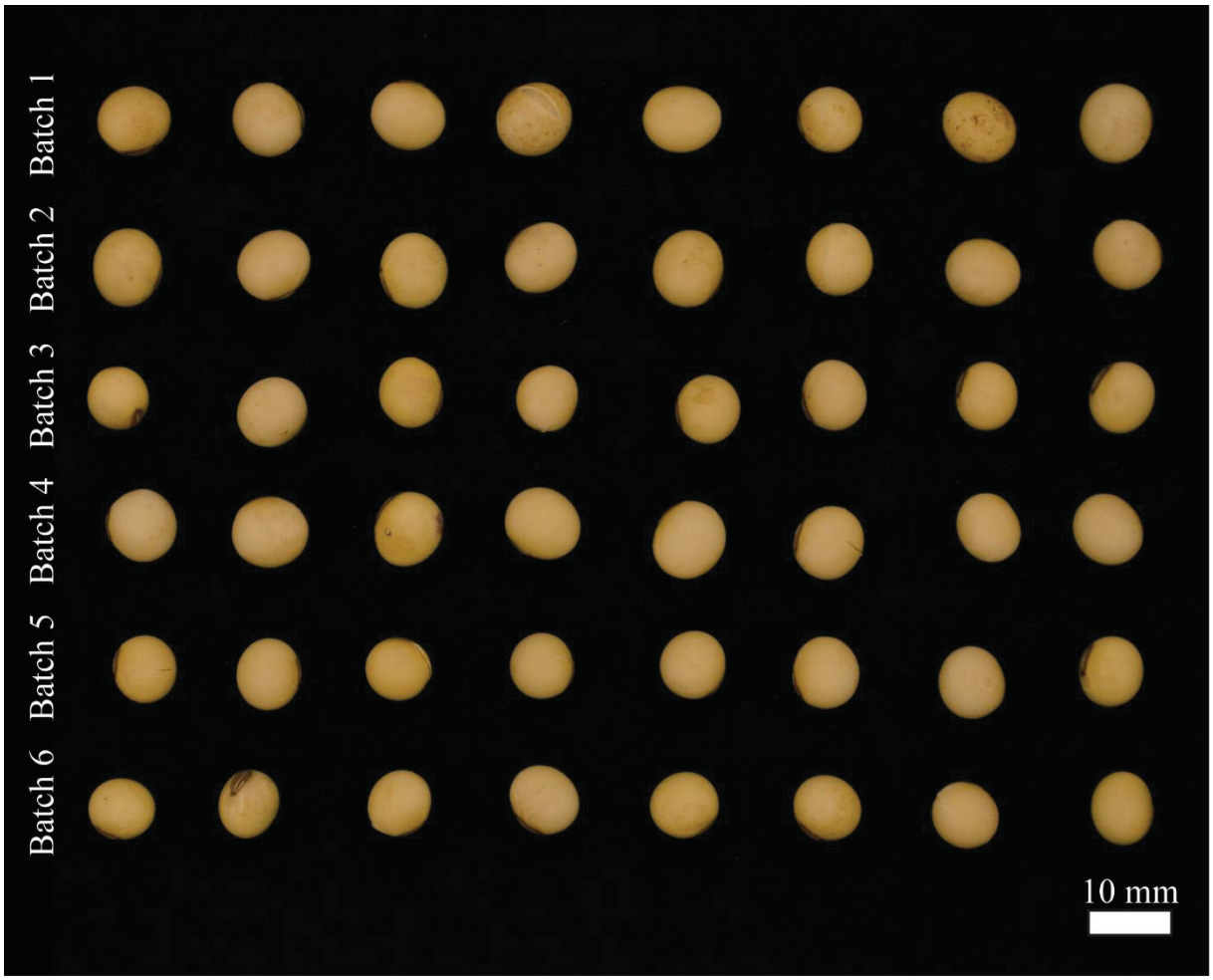


Figure 1. Six batches of dryseeds of soybean used on the experiment.

The experiment was conducted in six rounds, determined by the measurement capabilities of the equipments. Each round involved approximately 300 seeds and comprised seeds from all six batches.

Multispectral system

The VideometerLab4 system (Videometer, Hørsholm, Denmark) was responsible for capturing the multispectral images. This system comprised a coated matte sphere with LEDs arranged along its perimeter and a monochromatic camera positioned on top, offering high spatial resolution (40 µm per pixel and a resolution of 2192 × 2192 pixels). Before capturing the images, the system underwent calibration to ensure radiometric accuracy, geometric alignment, and proper lighting setup. The samples were exposed to 19 distinct wavelengths of LED illumination, covering ultraviolet (365, 405 nm), visible (430, 450, 470, 490, 515, 540, 570, 590, 630, 645, 660, and 690 nm), as well as near-infrared (780, 850, 880, 890, and 970

nm). Additionally, four long-pass filters with cut-off wavelengths at 400, 500, 600, and 700 nm were employed to measure fluorescence emitted from the seed surface. The filters were combined with different excitation wavelengths, providing 30 excitation-emission combinations (e.g., 365/400 nm). As a result, 19 images were collected for each illuminated wavelength, in addition to 31 images based on distinct excitation-emission combinations, resulting in a total of 50 monochromatic images for each sample.

Multispectral Imaging

The VideometerLab software (version 3.24.11) was utilized for image segmentation, seed labelling and feature extraction. A predefined mask was applied to isolate regions of interest (ROIs), specifically seeds, and to eliminate the background (consisting of the blue plate and petri dish). Autofluorescent-spectral features were extracted from these ROIs, in addition to morphological features, such as area, autocorrelation energy components, CIE color space components, width-to-length ratio, and width (Tab. 1). Subsequently, all the biometric features were exported to an Excel file for further data analysis.

Table 1. Biometric features extracted from dry soybean seeds description.

Feature	n. features	Description
Autofluorescence (AF)	31	Average autofluorescence signals from the combination of each wavelength excitation, with a cutoff emission at 400 nm, 500 nm, 600 nm, and 700 nm.
360-970 (RF)	19	Average reflectance of specific wavelengths (in nanometers) for individual seeds.
CIELab/CIELCh (CIE)	10	Refers to a color space defined by the International Commission on Illumination (CIE): L for lightness from black (0) to white (100), A from green (-) to red (+), and B from blue (-) to yellow (+). It calculates the mean and standard deviation of each L, a, B, C, and h component of CIE.
Auto correlation energy (ACE)	10	Seed texture component. Returns auto-correlation energy for vertical and horizontal directions in the seed blob. Result[0]=vCorr/hCorr, [1]=hCorr, [2]=vCorr, [3]=(hCorr+vCorr)/2, [4]=hCov., [5]=vCov, [6]=(hCov+vCov)/2, [7]=hor.cv, [8]=ver.cv, [9]=(hor.cv+ver.cv)/2.
Area	1	Individual projected area of the seed (mm2).
Length	1	Individual projected length of the seed (mm).
Ratio width lenght (RatioWL)	1	Width-to-length ratio calculated per seed.

Width	1	Individual projected width of the seed (mm).
Compactness circle (CC)	1	Ratio of the seed area to the area of a circle with the same length (isolength quotient).
Total	75	

Seed respiration assessment

After capturing multispectral images using VideometerLab, individual seeds were transferred to 5 mL screw-cap vials, each containing 800 µL of agar (0.4% w/v) and 0.2% Plant Preservative Mixture (PPMTM) to prevent fungal growth. The seeds were placed in vials within plates, maintaining the same position as they were initially positioned in VideometerLab. The vials were sealed with caps featuring a fluorescent polymer dot on their inner side. This polymer contains a dye that changes its fluorescence properties in response to oxygen concentration. The Seed Respiration Analyzer (Fytagoras B.V., Leiden, The Netherlands) was employed to measure the oxygen consumption (respiration) rates of the individual seeds during the processes of imbibition and germination. As the seeds respire, oxygen in the sealed vial is depleted, causing a detectable change in the fluorescence intensity of the dye. This change is monitored by a light source focused on the dot and a sensor that measures the fluorescence intensity. A robotic arm systematically guides the light source and sensor over each vial, enabling the measurement of oxygen concentration within (Bradford et al., 2013). Measurements were recorded at 30-minute intervals over a duration of 100 hours to construct time courses of oxygen consumption activity. The sample temperature was controlled at 20 ± 0.5°C using Peltier heating/cooling units and fans to maintain a stable environment. The recorded data was extracted into Excel files and subsequently subjected to data analysis.

Clustering analysis

In this study, both oxygen consumption curves and biometric features were converted into categorical variables through hierarchical clustering. The clustering of the oxygen consumption curves was performed using the 1^{st}-order difference in O_2 levels between two consecutive time points for each individual soybean, aiming to obtain less correlated data. The 1^{st}-order difference in O_2 respiration rate calculated for each i-th individual soybean was defined by $\Delta Xi(t) = Xi(t) - Xi(t-1)$, where Xi represents the O_2 level of the i-th soybean, t ($\leq$ T) denotes a specific time point, and T is the total number of recorded time points. The O_2 level 1^{st}-order difference was computed at intervals of every 3 hours, starting from 6 hours after imbibition and continuing until 100 hours. The biometric feature clustering process involved analyzing

both one-way and two-way interactions of the features. The one-way analysis focused on each of the 75 individual features independently. On the other hand, the two-way interaction analysis considered the clustering of combinations formed by pairing every two features, resulting in a total of $\binom{75}{2} = 2775$ combinations.

The hierarchical clustering analysis was based on the Euclidean distance metric and Ward.D2 linkage method, sourced from the R package *stats* (version 4.2.0). This approach was applied to group oxygen consumption curves, as well as the one-way and two-way interactions of biometric features, into *n* distinct classes (Fig. 2A and Fig. 2B). The determination of the optimal number of categories involved a visual inspection of their distribution and the dendrogram generated from the clustering analysis.

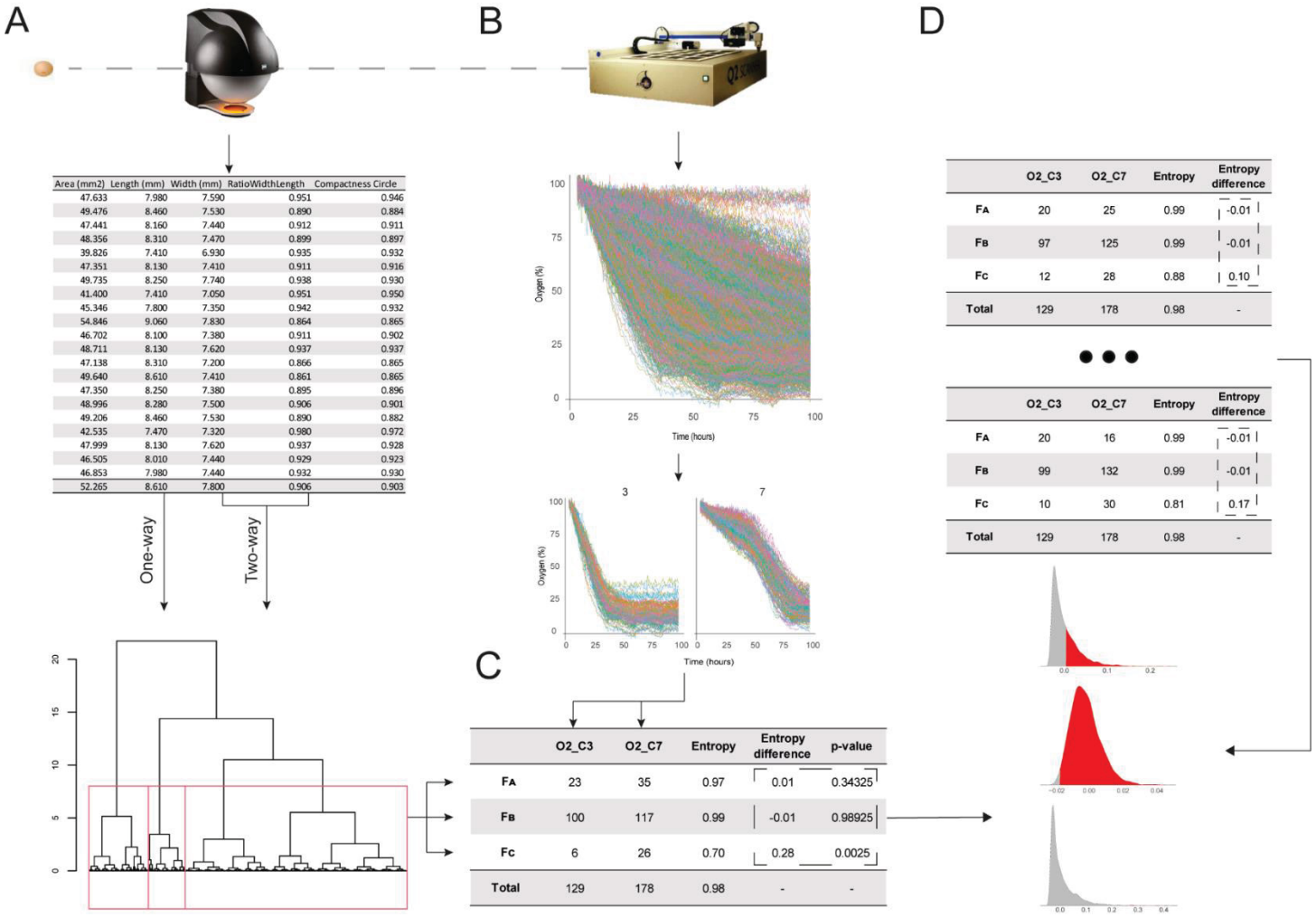


Figure 2. Seed biometric features (A) and oxygen consumption (B) data acquisition and clustering process. Contingency table and entropy evaluation (C), and contingency table simulation and entropy difference density plot (D).

The output from the clustering process was used to create contingency tables. The tables were designed with the one-way or two-way interaction of biometric features classes on the row-axis, while two oxygen consumption classes contrast were positioned on the column-axis (Fig. 2D). To quantify the association between the biometric features and the oxygen consumption classes, Shannon's entropy of the contingency tables was computed both before and after the inclusion of each biometric features classes (Shannon, 1948). The conditional entropy of Γ given C=c is denoted by $H_{\Gamma|c} = -\sum(\frac{n_{r|c}}{N_c} * log2\left(\frac{n_{r|c}}{N_c}\right))$. Here, $n_{r|c}$ is the number of seeds with a biometric characteristic class *c* and within each respiration pattern of Γ= r with $r \in \{r_1, r_2\}$, where N_c is the total number of seeds with that biometric characteristic class. Entropy is a metric for uncertainty, capturing the degree of disorder and uncertainty in the association between variables. A decrease in entropy following the inclusion of a feature suggests a potential association with the respiration pattern.

To assess the significance of entropy reduction within each biometric feature class, a simulation was conducted (Fig. 2C). This involved generating 4000 random contingency tables for each biometric feature, using a multinomial distribution to sample values based on the original table parameters. The multinomial distribution is denoted as $MN(n_r, [p_{c1}, \dots, p_{ck}])$, where n_r represents the total number of seeds on each respiration pattern r, p is the probability of a seed being in each biometric class *c*, and *k* is the total number of biometric feature classes. Entropy reduction was computed for each randomly generated table and compared to the original values. Significance was determined by evaluating whether the observed entropy reduction exceeded 5% of the random entropy reduction values, corresponding to a significance level of 0.05.

Visualization of O_2 consumption and biometric measurements relationship

Data mechanics visualization was implemented to explore the relationship between seeds based on the presence or absence of biometric characteristics previously identified when contrasting two respiration patterns. This was achieved by creating a binary matrix (1 – presence or 0 – absence of the biometric feature class) of size $N_r \times C$ were N_r were the seeds from the respiration patterns r_1 and r_2 and *C* denotes the significant biometric feature classes identified.

The data mechanics procedure consisted of computing two Euclidean distance matrices: one quantifying the similarity between pairwise combinations of seeds, and the second quantifying the similarity between pairwise combinations of biometric characteristics. These

distance matrices were then used to generate two independent hierarchical clustering trees using the Ward D2 linkage method. By cutting both trees at a fixed number of clusters, the original distance matrices were updated. This involved using the clustering structure between seeds to create a weighted distance matrix between categories and vice versa (Fushing and Chen, 2014; McVey et al., 2021). This procedure allowed for the integration of information on seed similarities with biometric category similarities, thereby enhancing the visualization of the seed pattern.

Inference process

A voting-system inference process was developed to assess the predictability of an unknown respiration pattern based on the presence or absence of significant biometric features within a given seed. The leave-one-out cross-validation method was employed, utilizing binary matrices constructed. For each seed, the Euclidean distance was computed, identifying its 20 nearest neighbors. Subsequently, for each of these nearest neighbors, their respective nearest neighbors were determined, and those that included the unknown seed were tallied as votes for its respiration pattern. The determination of the unknown seed's respiration pattern was made based on the class with the highest number of votes. The analyses were performed using R version 3.5.2 and RStudio software version 2022.02.3.

RESULTS

The normalized distribution of all 1808 seeds' biometric features and their correlogram are presented in Fig. 3. Notably, the distribution of biometric features highlighted the heterogeneity within the seed population (Fig. 1A). For instance, the distribution of size attributes (area, width, and length) showed high deviation, indicating the presence of seeds from a broad size range. The spectral features (RF-365 to RF-970) exhibited similar distribution patterns, although they had distinct median values. In contrast, morphological, autofluorescence, and CIE components displayed diverse distribution patterns and median values. Furthermore, spectral features, autofluorescence, and autocorrelation components (ACE) features exhibited high correlation within each respective class (Fig. 1B).

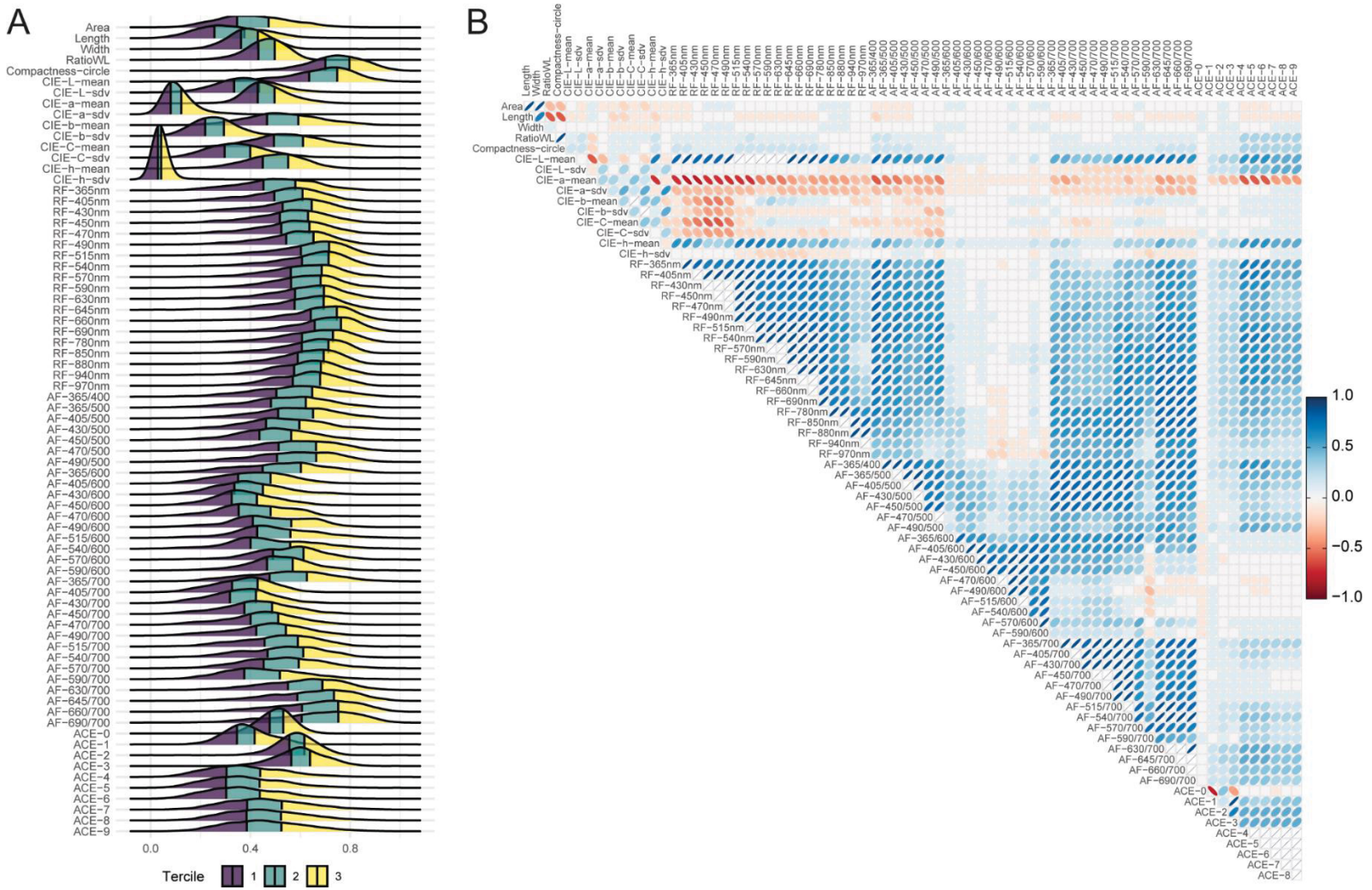


Figure 3. Soybean morphological, spectral (RF), autofluorescence (AF) and texture (ACE) normalized (min-max normalization) features distribution (A) and correlation (B).

The time-course oxygen consumption curves were classified into 12 categories based on a visual examination of the hierarchical cluster tree (Fig. 4A). This clustering revealed distinct patterns in the oxygen consumption of the seeds (Fig. 4B). Clusters one to five shared the same branch in the dendrogram and exhibited a common characteristic of rapid oxygen consumption initially, followed by stabilization. Clusters six and seven, located on the second branch of the dendrogram, shared a pattern characterized by a slight initial linear rate of respiration phase, with slower oxygen consumption at the beginning, followed by a steeper slope and a plateau. Finally, seeds from clusters eight to 12, positioned on the third dendrogram branch, displayed varying patterns among themselves but collectively demonstrated a slow and consistent oxygen consumption over time. The seeds extracted from cluster trees seven and ten exhibit a notable contrast in their respiration patterns, revealing both homogeneous patterns within each cluster and distinctive characteristics in terms of the transition from a fast (cluster three) to intermediate (cluster seven) to a slow (cluster ten) respiration rate (Fig. 4B).

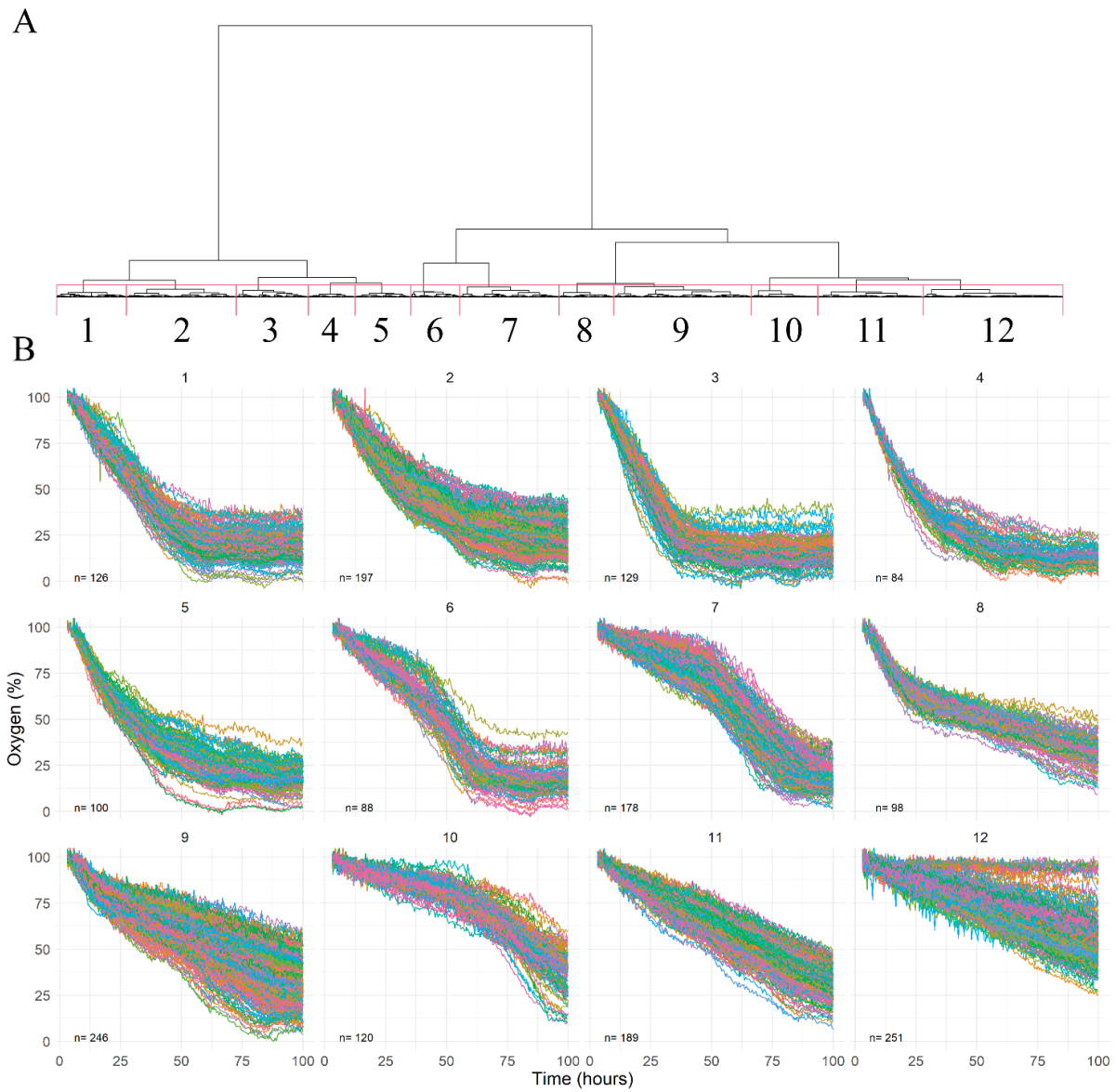


Figure 4. Time course of oxygen consumption activity for individual soybean seeds, including a dendrogram (A) and curves segmented into 12 clusters (B). Each curve's color corresponds to a distinct seed.

The evaluation of individual biometric features revealed distinctions among seeds with different respiration patterns (Fig. 5). Notably, 40 characteristics significantly (p-value<0.05) distinguished between seeds with a fast respiration pattern from those with a slow respiration pattern. Additionally, 23 characteristics differed between seeds with a slow and intermediate respiration pattern, while 51 characteristics distinguished between fast and intermediate respiration patterns. It is noteworthy that seeds with a certain level of attributes related to size, such as width, area, and lengths, as well as reflectance at 365, 405, 660, and 690 nm, along with

autofluorescence features, presented higher odds of displaying a particular respiration pattern. For instance, seeds with a smaller width are five times more likely to exhibit an intermediate respiration pattern compared to a faster respiration pattern, as well as among the seeds with smaller areas and lengths (Fig. 5C). This result corroborates with Fig. 5A, where seeds with larger areas or width are nearly twice as likely to demonstrate a fast respiration pattern, while smaller seeds are twice as likely to exhibit a slow respiration pattern. Additionally, seeds that presented higher values of autofluorescence 470/700 excitation-emission are less likely to present a fast respiration pattern.

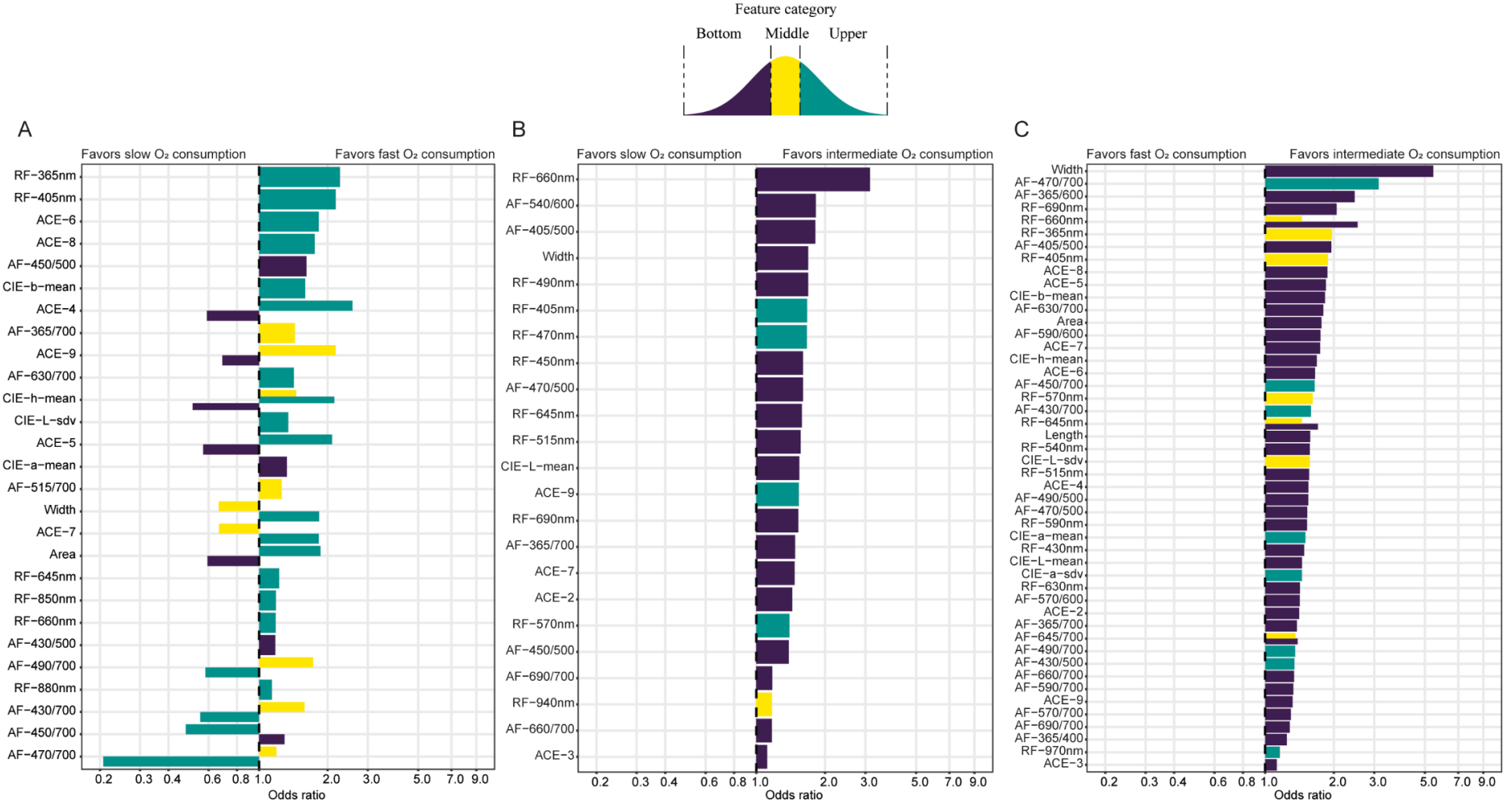


Figure 5. Significant biometric feature classes odds ratio between seeds with fast and slow oxygen consumption (A), slow and intermediate oxygen consumption (B) and fast and intermediate oxygen consumption (C). The x-axis is presented on logarithmic scale. The odds ratio represents the

odds of seeds with a particular feature between two respiration patterns, divided by the odds of the total number of seeds in each respective respiration pattern.

Differences in the two-way interaction of biometric features between seeds displaying fast and intermediate respiration patterns are illustrated in Fig. 6. Notably, a total of 3296 characteristics significantly (p-value < 0.05) distinguished between the respiration patterns, and these are arranged along the heatmap's column-axis. These distinctions are visually presented on the heatmap, where the presence or absence of the significant biometric features among seeds from the two respiration patterns is color-coded as red or blue, respectively. Seeds are grouped on the row-axis according to the similarities in their characteristics. Notably, seeds with the same respiration pattern tend to share a common branch on the dendrogram, while rounds and batches remain randomly distributed. This observation underscores the similarity among seeds from the same respiration pattern in terms of the presence or absence of selected characteristics, indicating a robust relationship between respiration and biometric features. Moreover, examining the block-patterns on the presence of characteristics revealed by the row and column-axis dendrograms suggests that different combinations of characteristics contribute to either higher or intermediate respiration.

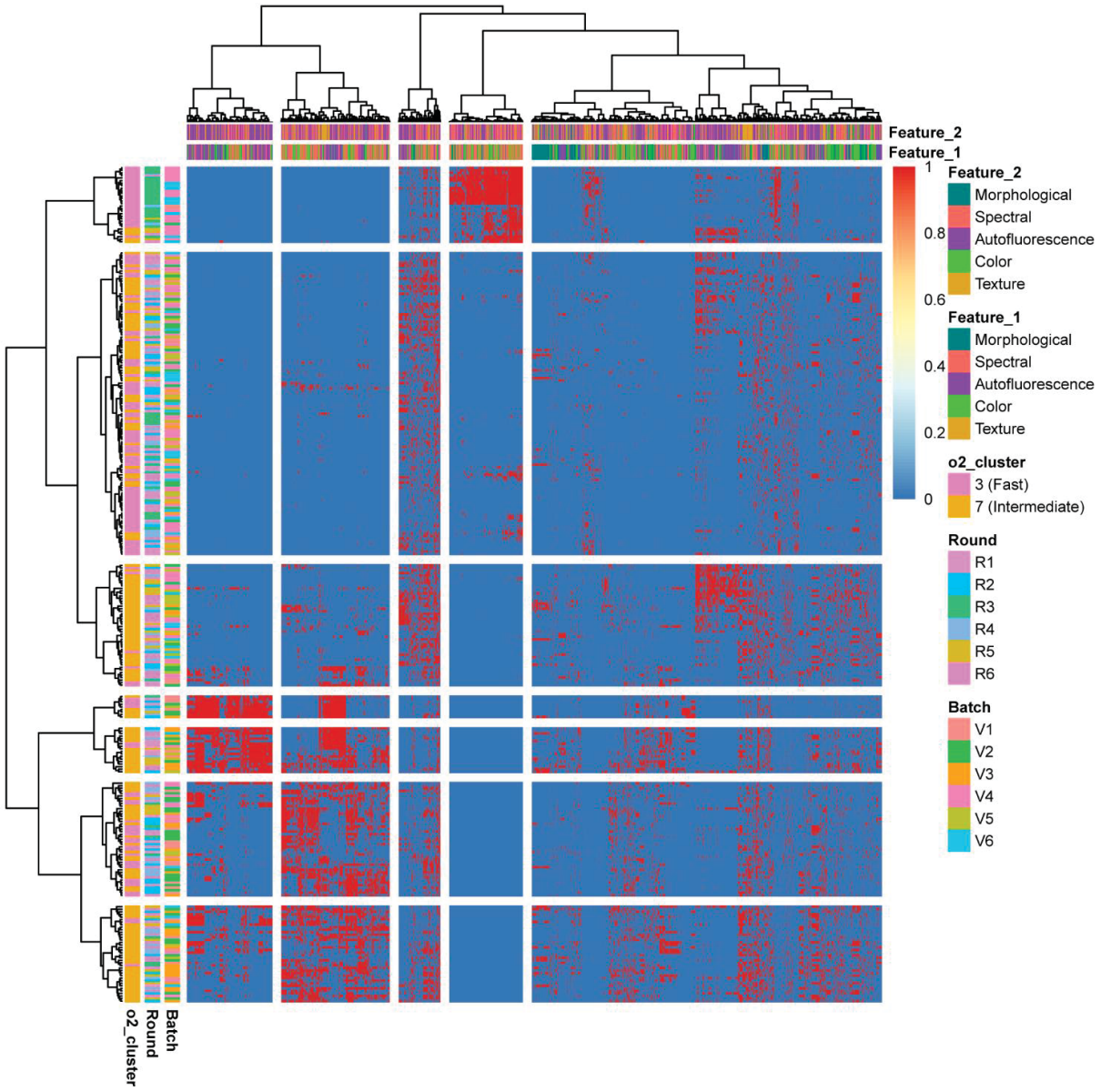


Figure 6. Data Mechanics visualization of the significant (p-value < 0.05) differences in the two-way interaction of biometric features between soybean seeds displaying fast (O_2 cluster 3) and intermediate (O_2 cluster 7) respiration patterns. The row-axis represents each soybean, color-coded according to its respiration pattern (O_2 cluster), as well as its round and batch. The column-axis represents each significant biometric characteristic, color-coded based on its corresponding two-way interaction category. The presence or absence of the characteristic in the seed is color-coded as red or blue, respectively.

A total of 2765 two-way interactions of biometric features differed significantly between seeds with fast and slow respiration patterns (Fig. 7). The seeds from the same

respiration pattern tend to share similar branches on the dendrogram, as observed in Fig. 6, indicating that the selected biometric feature characteristics are also associated to the seed respiration pattern. Notably, a high presence of characteristics linked to autofluorescence in seeds with a slower respiration pattern is observed in the block-pattern formation in the last row, third column. This result corroborates with Fig. 3A, where autofluorescence features showed higher odds toward slower oxygen consumption.

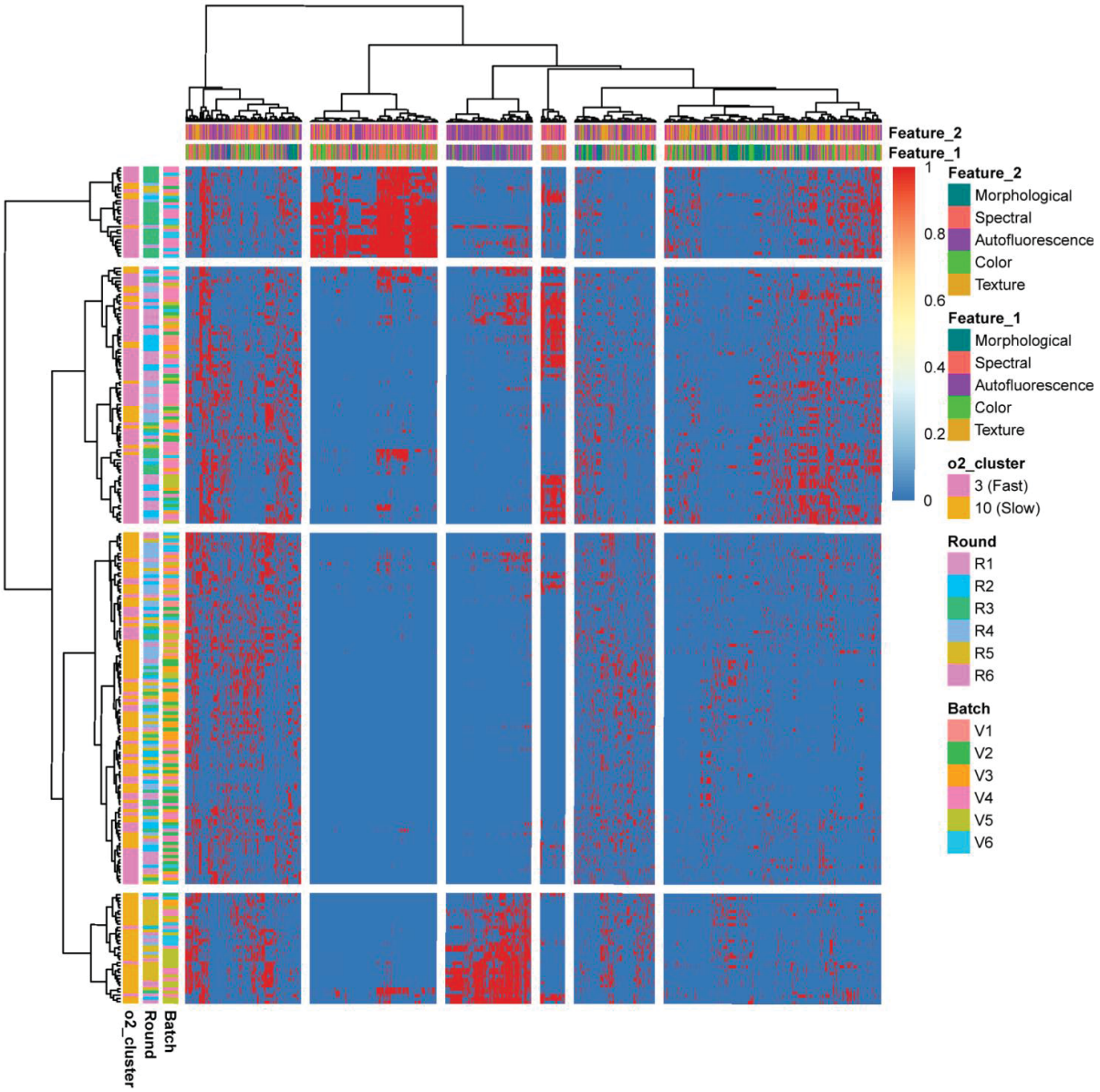


Figure 7. Data Mechanics visualization of the significant (p-value < 0.05) differences in the two-way interaction of biometric features between soybean seeds displaying fast (O_2 cluster 3) and slow (O_2 cluster 10) respiration patterns. The row-axis represents each soybean, color-

coded according to its respiration pattern (O_2 cluster), as well as its round and batch. The column-axis represents each significant biometric characteristic, color-coded based on its corresponding two-way interaction category.

Seeds with slow and intermediate respiration patterns significantly distinguished themselves in 1650 two-way biometric feature combinations. The fewer characteristics identified in those two respiration patterns, in comparison with the contrasts fast and intermediate, and fast and slow respiration patterns, suggest a higher similarity between them. This similarity is also observed in Fig. 4, where the time-course oxygen consumption curves from clusters 7 and 10 (intermediate and slow respiration patterns) share more similarities than with cluster 3 (fast). Notably, seeds from the same respiration pattern sharing branches on the dendrogram are observed, and a higher number of block patterns are formed. This indicates that seeds with the same respiration pattern share similar characteristics; however, not one selection of characteristics can be attributed to differentiate each respiration pattern.

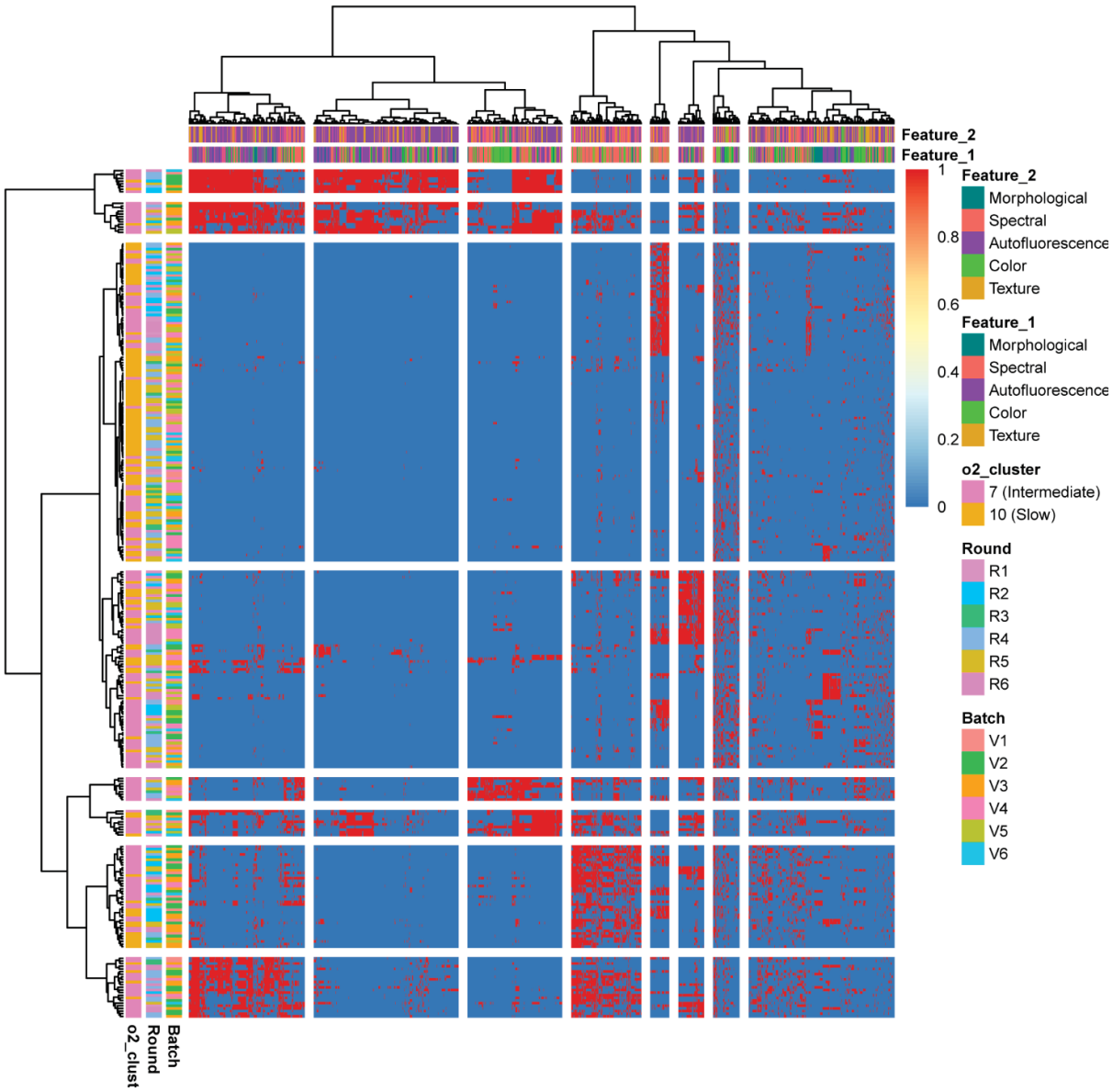


Figure 8. Data Mechanics visualization of the significant (p-value < 0.05) differences in the two-way interaction of biometric features between soybean seeds displaying intermediate (O_2 cluster 7) and slow (O_2 cluster 10) respiration patterns. The row-axis represents each soybean, color-coded according to its respiration pattern (O_2 cluster), as well as its round and batch. The column-axis represents each significant biometric characteristic, color-coded based on its corresponding two-way interaction category.

The performance of seed respiration pattern classification, based on significant two-way biometric features for each respiration contrast, is presented in Tab. 2. The mean accuracy for classifying seeds between two respiration patterns, according to selected biometric features,

achieved 76.17%. Overall, the tree contrast showed similar accuracy, and the result aligns with the observations in Figures 6 through 8, where seeds with the same respiration pattern tend to share similar branches on the dendrogram.

Table 2. Confusion matrix and overall accuracy based on the inference procedure using significant biometric characteristics between seed respiration pattern contrast fast, intermediate and slow (O_2 cluster 3, 7 and 10, respectively).

	O_2 Cluster 3 (fast)	O_2 Cluster 7 (intermediate)
O_2 Cluster 3 (fast)	89	40
O_2 Cluster 7 (intermediate)	34	143
Sensitivity (%)	72.30%	
Specificity (%)	78.10%	
Overall accuracy (%)	75.80%	
	O_2 Cluster 3 (fast)	O_2 Cluster 10 (slow)
O_2 Cluster 3 (fast)	95	33
O_2 Cluster 10 (slow)	31	88
Sensitivity (%)	75.40%	
Specificity (%)	72.70%	
Overall accuracy (%)	74.10%	
	O_2 Cluster 7 (intermediate)	O_2 Cluster 10 (slow)
O_2 Cluster 7 (intermediate)	150	27
O_2 Cluster 10 (slow)	36	82
Sensitivity (%)	80.60%	
Specificity (%)	75.20%	
Overall accuracy (%)	78.60%	

DISCUSSION

The study extensively explored the relationship between soybean seed respiration and its biometric features at the single-seed level. The research covered 75 biometric features, including morphological, textural, spectral, and autofluorescence aspects, considering both one-way and two-way interactions, leading to a set of 2,775 feature combinations. Each feature and

feature pair were grouped, resulting in more than 8,000 unique seed biometric characteristics that were evaluated. The study focused on three distinctive respiration patterns, and the differences in the seed population characteristics from each pair of respiration patterns were efficiently identified. The selected two-way interaction features were able to distinguish between two respiration patterns with an overall accuracy of more than 75%, highlighting the potential for utilizing these shared traits as discriminative markers, facilitating a more efficient classification of seeds based on their respiratory behavior.

The soybean seed population used in the study effectively highlighted a range of respiration patterns, as illustrated in Fig. 4. The differences observed in the respiration patterns are directly related to physiological status of the soybean seeds at various developmental stages. Typically, the time-course of oxygen consumption in seeds displays a sigmoid pattern. The initial stage involves a linear respiration rate, followed by a second stage characterized by a steep slope until oxygen is depleted. The seeds' oxygen consumption during the imbibition process and prior to radicle emergence is frequently associated with seed vigor. (Bello and Bradford, 2021; Corbineau, 2012; Tu et al., 2023; Xin et al., 2013). A brief initial linear rate of respiration indicates seeds with a robust capacity to repair their respiratory system. On the other hand, a steeper slope is associated with embryo axis development, radicle emergence, and seedling growth, consequently accelerating oxygen consumption (Bello and Bradford, 2016; Bradford et al., 2013). Both situations are commonly linked to high-vigor seeds. For instance, studies with sweet-corn, pepper, wheat, watermelon, onion and Brassicas demonstrates that high vigor seeds tend to consume more oxygen during the germination process than low vigor seeds (Bradford et al., 2013; He et al., 2019; Tu et al., 2023). Interestingly, a linear rate of respiration was frequently observed among the seed population until the end of the measured time (Fig. 4B – clusters 9, 11, and 12). A linear respiration rate is often associated with absence of embryo axis growth and radicle emergence (i.e., germination *strictu sensu*) (Bello and Bradford, 2016).

Clear distinctions emerged in the composition of characteristics between soybean seed populations exhibiting rapid oxygen consumption, indicating high-vigor seeds, and those displaying slower oxygen consumption. Notably, variations were observed in the size, autofluorescence, and seeds' reflectance. Autofluorescence features are an established markers of seed maturity and quality, such as chlorophyll, lignin, carotenoids and phenols fluorescence (Barboza da Silva et al., 2021; Donaldson, 2020; França-Silva et al., 2023; Jalink et al., 1998). Autofluorescence excitation-emission combinations of 365/600 exhibited a positive correlation

with slower oxygen consumption. The excitation wavelength at UV-light (365 nm) has been primarily associated to chlorophyll fluorescence (Donaldson, 2020; Li et al., 2014). Additionally, a strong correlation (p=0.91) with hydrogen peroxide (H_2O_2) levels in soybeans has also been reported (Barboza da Silva et al., 2021). Previous studies have linked lower autofluorescence intensity at a 365 nm excitation with lower vigor seeds. For instance, Barboza da Silva et al. (2021) and Batista et al. (2022) noted reduced autofluorescence at 365 nm excitation in aged soybean seeds, which also presented a better discrimination among seed aging classes compared to early germination tests, corroborating with the present study. Similarly, Li et al. (2019) observed that aged and non-viable soybean seeds exhibited lower fluorescence intensity at the 365 nm excitation wavelength. An interesting finding was the similarities observed at excitation wavelengths of 430 nm, 450 nm, and 470 nm, where higher fluorescence signal was associated with lower oxygen consumption rates. This association was evident both between seeds with fast and intermediate respiration patterns and between seeds with fast and slower respiration patterns (Fig. 5A and 5C). The wavelength spectrum of 430-470 nm has been reported to be associated with lignin, ferulic acid and flavonoids fluorescence and to exhibit a strong correlation (Pearson correlation > 0.95) with lignin content in soybean seed coats (Barboza da Silva et al., 2021; Batista et al., 2022; Donaldson, 2020). This is an interesting finding, since the role of those compounds is not completely elucidate on seed vigor (Batista et al., 2022).

Concerning seed reflectance, wavelengths of 365 nm, 405 nm, and 690 nm exhibited a positive correlation with seed oxygen consumption. These wavelengths are associated with the absorbance peaks of chlorophyll *a* (Donaldson, 2020; França-Silva et al., 2023). Therefore, higher reflectance at these wavelengths indicates lower light absorption due to lower chlorophyll *a* content, likely attributable to the presence of more mature and vigorous seeds. Also, Barboza da Silva et al. (2021) reported that non-aged seeds have less chlorophyll a content than aged ones, suggesting that the aging process may also contribute to the signals observed at the aforementioned wavelengths Additionally, as expected, size-related features such as area, width, and length showed a direct correlation with faster oxygen consumption. Larger seeds typically have a greater mass and, consequently, more respiratory tissues, leading to increased oxygen consumption within the vial (Xin et al., 2013).

Although the biometric characteristics investigated in this study serve as valuable indicators for estimating seed performance, the quality of soybean seeds is susceptible to

various factors, including its maturation stage, mechanical damage, seed aging, greenish seeds, and the presence of pathogens (de Medeiros et al., 2020). Consequently, individual markers often prove insufficient for a comprehensive evaluation of seed quality (Corbineau, 2012). This limitation is clearly observed in the biometric features two-way interactions map, where different compositions of seed characteristics lead to the same respiration rate (Figures 6 to 8). Additionally, different composition patterns of features were observed among seeds from the same batch, possibly due to the presence of subpopulations among them. It is known that seed baches present mixtures of multiple subpopulations, influenced by factors such as different genotypes, seed locations in the fruit or on the mother plant or commercial seed lots blending. Thus, the integrated assessment of different features in a single-seed approach is highly needed to understand whether a characteristic or combination of characteristics can lead to a certain seed performance (Bello and Bradford, 2016; Corbineau, 2012; de Medeiros et al., 2020).

In summary, the described methodology proves efficient in mapping differences in seed characteristics within populations and tracing them back to individual seeds. It enables the visualization of seed clusters with similar physiological quality but distinct characteristics, which is valuable for seed vigor studies and subpopulation identification. Notably, despite focusing on the two-way interaction of biometric features, the methodology achieved an average 75% separation between classes. We suggest conducting studies to evaluate the interation of three or more factors, which could significantly enhance seed class separation, even among more similar respiration patterns.

CONCLUSION

In conclusion, soybean seed appearance, along with its spectral information, is strongly correlated with seed respiration and, consequently, with their physiological quality. Multispectral imaging is a convenient and non-invasive method that can be utilized to identify seed traits related to individual seed respiration patterns and to visually explore the relationship between these characteristics. Autofluorescence and seed reflectance significantly contribute to the differentiation of seed physiological qualities.

AKNOWLEDGEMENT

To the Coordination for the Improvement of Higher Education Personnel (CAPES – Finance Code 001), for providing a scholarship to the first author and the seed lab Apasem (Ponta Grossa/PR) donating the seeds to the experiment.

CONFLICTS

None declared.

REFERENCES

Barboza da Silva, C., Oliveira, N.M., de Carvalho, M.E.A., de Medeiros, A.D., de Lima Nogueira, M., dos Reis, A.R., 2021. Autofluorescence-spectral imaging as an innovative method for rapid, non-destructive and reliable assessing of soybean seed quality. Sci. Rep. 11, 1–12. https://doi.org/10.1038/s41598-021-97223-5

Batista, T.B., Mastrangelo, C.B., de Medeiros, A.D., Petronilio, A.C.P., Fonseca de Oliveira, G.R., dos Santos, I.L., Crusciol, C.A.C., Amaral da Silva, E.A., 2022. A Reliable Method to Recognize Soybean Seed Maturation Stages Based on Autofluorescence-Spectral Imaging Combined With Machine Learning Algorithms. Front. Plant Sci. 13, 1–14. https://doi.org/10.3389/fpls.2022.914287

Bello, P., Bradford, K.J., 2021. Relationships of brassica seed physical characteristics with germination performance and plant blindness. Agric. 11, 1–21. https://doi.org/10.3390/agriculture11030220

Bello, P., Bradford, K.J., 2016. Single-seed oxygen consumption measurements and population-based threshold models link respiration and germination rates under diverse conditions. Seed Sci. Res. 26, 199–221. https://doi.org/10.1017/S0960258516000179

Boelt, B., Shrestha, S., Salimi, Z., Jorgensen, J.R., Nicolaisen, M., Carstensen, J.M., 2018. Multispectral imaging - A new tool in seed quality assessment? Seed Sci. Res. 28, 222–228. https://doi.org/10.1017/S0960258518000235

Bradford, K.J., 2018. Interpreting biological variation: Seeds, populations and sensitivity thresholds. Seed Sci. Res. 28, 158–167. https://doi.org/10.1017/S0960258518000156

Bradford, K.J., Bello, P., Fu, J.C., Barros, M., 2013. Single-seed respiration: A new method to

assess seed quality. Seed Sci. Technol. 41, 420–438. https://doi.org/10.15258/sst.2013.41.3.09

Caverzan, A., Giacomin, R., Müller, M., Biazus, C., Lângaro, N.C., Chavarria, G., 2018. How does seed vigor affect soybean yield components? Agron. J. 110, 1318–1327. https://doi.org/10.2134/agronj2017.11.0670

Cheng, H., Ye, M., Wu, T., Ma, H., 2023. Evaluation and Heritability Analysis of the Seed Vigor of Soybean Strains Tested in the Huanghuaihai Regional Test of China. Plants 12. https://doi.org/10.3390/plants12061347

Corbineau, F., 2012. Markers of seed quality: From present to future. Seed Sci. Res. https://doi.org/10.1017/S0960258511000419

de Medeiros, A.D., Capobiango, N.P., da Silva, J.M., da Silva, L.J., da Silva, C.B., dos Santos Dias, D.C.F., 2020. Interactive machine learning for soybean seed and seedling quality classification. Sci. Rep. 10, 1–10. https://doi.org/10.1038/s41598-020-68273-y

Donaldson, L., 2020. Autofluorescence in plants. Molecules. https://doi.org/10.3390/molecules25102393

Elmasry, G., Mandour, N., Al-Rejaie, S., Belin, E., Rousseau, D., 2019. Recent applications of multispectral imaging in seed phenotyping and quality monitoring—An overview. Sensors (Switzerland). https://doi.org/10.3390/s19051090

Faqeerzada, M.A., Perez, M., Lohumi, S., Lee, H., Kim, G., Wakholi, C., Joshi, R., Cho, B.K., 2020. Online application of a hyperspectral imaging system for the sorting of adulterated almonds. Appl. Sci. 10, 1–16. https://doi.org/10.3390/APP10186569

França-Silva, F., Gomes-Junior, F.G., Rego, C.H.Q., Marassi, A.G., Tannús, A., 2023. Advances in imaging technologies for soybean seed analysis. J. Seed Sci. https://doi.org/10.1590/2317-1545v45274098

Fu, X., Bai, M., Xu, Y., Wang, T., Hui, Z., Hu, X., 2023. Cultivars identification of oat (Avena sativa L.) seed via multispectral imaging analysis. Front. Plant Sci. 14. https://doi.org/10.3389/fpls.2023.1113535

Fushing, H., Chen, C., 2014. Data Mechanics and Coupling Geometry on Binary Bipartite

Networks. PLoS One 9, e106154. https://doi.org/10.1371/journal.pone.0106154

He, Y., Ye, Z., Ying, Q., Ma, Y., Zang, Y., Wang, H., Yu, Y., Zhu, Z., 2019. Glyoxylate cycle and reactive oxygen species metabolism are involved in the improvement of seed vigor in watermelon by exogenous GA3. Sci. Hortic. (Amsterdam). 247, 184–194. https://doi.org/10.1016/j.scienta.2018.12.016

ISTA, 2020. International Rules for Seed Testing. International Seed Testing Association, Bassersdorf, Switzerland.

Jalink, H., Van Der Schoor, R., Frandas, A., Van Pijlen, J.G., Bino, R.J., 1998. Chlorophyll fluorescence of Brassica oleracea seeds as a non-destructive marker for seed maturity and seed performance. Seed Sci. Res. 8, 437–443. https://doi.org/10.1017/s0960258500004402

Jin, B., Qi, H., Jia, L., Tang, Q., Gao, L., Li, Z., Zhao, G., 2022. Determination of viability and vigor of naturally-aged rice seeds using hyperspectral imaging with machine learning. Infrared Phys. Technol. 122, 104097. https://doi.org/10.1016/j.infrared.2022.104097

Kandpal, L.M., Lohumi, S., Kim, M.S., Kang, J.S., Cho, B.K., 2016. Near-infrared hyperspectral imaging system coupled with multivariate methods to predict viability and vigor in muskmelon seeds. Sensors Actuators, B Chem. 229, 534–544. https://doi.org/10.1016/j.snb.2016.02.015

Li, L., Zhang, Q., Huang, D., 2014. A review of imaging techniques for plant phenotyping. Sensors (Switzerland). https://doi.org/10.3390/s141120078

Li, Y., Sun, J., Wu, X., Chen, Q., Lu, B., Dai, C., 2019. Detection of viability of soybean seed based on fluorescence hyperspectra and CARS-SVM-AdaBoost model. J. FOOD Process. Preserv. 43. https://doi.org/10.1111/jfpp.14238

McVey, C., Hsieh, F., Manriquez, D., Pinedo, P., Horback, K., 2021. Livestock Informatics Toolkit: A Case Study in Visually Characterizing Complex Behavioral Patterns across Multiple Sensor Platforms, Using Novel Unsupervised Machine Learning and Information Theoretic Approaches. Sensors 22, 1. https://doi.org/10.3390/s22010001

Olesen, M.H., Nikneshan, P., Shrestha, S., Tadayyon, A., Deleuran, L.C., Boelt, B., Gislum, R., 2015. Viability prediction of ricinus cummunis L. Seeds using multispectral imaging.

Sensors (Switzerland) 15, 4592–4604. https://doi.org/10.3390/s150204592

Qi, H., Huang, Z., Sun, Z., Tang, Q., Zhao, G., Zhu, X., Zhang, C., 2023. Rice seed vigor detection based on near-infrared hyperspectral imaging and deep transfer learning. Front. Plant Sci. 14, 1–13. https://doi.org/10.3389/fpls.2023.1283921

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shrestha, S., Deleuran, L.C., Gislum, R., 2017. Separation of viable and non-viable tomato (Solanum lycopersicum L.) seeds using single seed near-infrared spectroscopy. Comput. Electron. Agric. 142, 348–355. https://doi.org/10.1016/j.compag.2017.09.004

Tu, K., Yin, Y., Yang, L., Wang, J., Sun, Q., 2023. Discrimination of individual seed viability by using the oxygen consumption technique and headspace-gas chromatography-ion mobility spectrometry. J. Integr. Agric. 22, 727–737. https://doi.org/10.1016/j.jia.2022.08.058

Xia, Y., Xu, Y., Li, J., Zhang, C., Fan, S., 2019. Recent advances in emerging techniques for non-destructive detection of seed viability : A review Arti fi cial Intelligence in Agriculture Recent advances in emerging techniques for non-destructive detection of seed viability : A review. Artif. Intell. Agric. 1. https://doi.org/10.1016/j.aiia.2019.05.001

Xin, X., Wan, Y., Wang, W., Yin, G., McLamore, E.S., Lu, X., 2013. A real-time, non-invasive, micro-optrode technique for detecting seed viability by using oxygen influx. Sci. Rep. 3. https://doi.org/10.1038/srep03057

Zhang, J., Feng, X., Liu, X., He, Y., 2018. Identification of hybrid okra seeds based on near-infrared hyperspectral imaging technology. Appl. Sci. 8. https://doi.org/10.3390/app8101793

CONSIDERAÇÕES FINAIS

Ao longo deste trabalho, o uso da análise de imagem espectral, em conjunto com técnicas de análise de dados multivariados, foi extensivamente explorado para avaliação de diferentes componentes da qualidade de sementes. Essa tecnologia permite a extração de uma grande quantidade de informações espaciais e espectrais de forma individual das sementes. Em especial, informações espectrais trazem informações importantes de pigmentação, reserva e estruturais das sementes. Dessa forma, demonstrou-se altamente capaz de distinguir genótipos de sementes, abrangendo diferentes espécies, sementes híbridas e variedades, como observado nos diversos trabalhos levantados.

Foi demonstrado com sucesso que a análise de imagem espectral pode ser empregada na classificação de sementes híbridas de eucalipto, que apresentam desafios adicionais devido à grande quantidade de material inerte, ao seu pequeno tamanho e a semelhança fenotípica com seus progenitores. Isso sugere que a aplicação dessa técnica por empresas de melhoramento florestal deve ser considerada. Os procedimentos de avaliação dos progenitores a serem cruzados e a confirmação do cruzamento, comuns no processo de melhoramento, podem ser realizados por meio dessa análise, com a vantagem significativa de evitar a perda da amostra analisada. Esse benefício é especialmente relevante em programas de melhoramento florestal, nos quais os custos e o tempo necessários são elevados.

Além da capacidade de distinguir entre diferentes genótipos, este estudo também destacou o potencial da técnica na avaliação da qualidade fisiológica, estabelecendo relações entre características biométricas-chave e o processo de respiração de sementes de soja. Tais características fornecem informações valiosas para o controle de qualidade, permitindo a avaliação do estado de maturação das sementes e o nível de deterioração, o que é crucial para o monitoramento durante a pré-colheita, análise da qualidade e armazenamento, .

A análise de imagem espectral oferece vantagens significativas em comparação com os métodos tradicionais. A integridade da semente é preservada, a avaliação é rápida e o método pode ser facilmente escalável. Além disso, sua aplicação traz possibilidades inovadoras, onde cada característica-chave identificada pode ser utilizada para o desenvolvimento de novos equipamentos especializados mais acessíveis e simplificados.

A capacidade de extrair grandes volumes de dados individuais das sementes e transferi-los automaticamente para o meio virtual também abre grandes perspectivas na área da

tecnologia da informação. Informações extraídas das sementes, como características biométricas, geolocalização e dados climáticos, podem ser incluídas em bancos de dados colaborativos e utilizados por empresas produtoras de sementes para estimar a qualidade futura dos lotes

Embora a análise de imagem espectral se mostre uma técnica promissora na análise de sementes, é importante ressaltar que não substitui o trabalho humano, mas o complementa. Apesar da possibilidade de extrair e selecionar informações automaticamente para compor modelos de classificação, os algoritmos não supervisionados dificilmente superarão o conhecimento e a experiência de um analista. Portanto, a tecnologia é mais eficiente quando guiada por conhecimento biológico prévio, otimizando sua aplicação em laboratórios com analistas experientes.

REFERÊNCIAS GERAIS

BARBOZA DA SILVA, C.; OLIVEIRA, N. M.; DE CARVALHO, M. E. A.; et al. Autofluorescence-spectral imaging as an innovative method for rapid, non-destructive and reliable assessing of soybean seed quality. **Scientific Reports**, v. 11, n. 1, p. 1–12, 2021. Nature Publishing Group UK. Disponível em: <https://doi.org/10.1038/s41598-021-97223-5>. .

BATISTA, T. B.; MASTRANGELO, C. B.; DE MEDEIROS, A. D.; et al. A Reliable Method to Recognize Soybean Seed Maturation Stages Based on Autofluorescence-Spectral Imaging Combined With Machine Learning Algorithms. **Frontiers in Plant Science**, v. 13, n. June, p. 1–14, 2022.

BELLO, P.; BRADFORD, K. J. Single-seed oxygen consumption measurements and population-based threshold models link respiration and germination rates under diverse conditions. **Seed Science Research**, v. 26, n. 3, p. 199–221, 2016.

BELLO, P.; BRADFORD, K. J. Relationships of brassica seed physical characteristics with germination performance and plant blindness. **Agriculture (Switzerland)**, v. 11, n. 3, p. 1–21, 2021. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102923122&doi=10.3390%2Fagriculture11030220&partnerID=40&md5=d4a3b834d6aad0545e6855d2fb84f4f0>. .

VAN DEN BERG, G. J.; VERRYN, S. D.; CHIRWA, P. W.; VAN DEVENTER, F. Genetic Parameters of Interspecific Hybrids of Eucalyptus grandis and E. urophylla Seedlings and Cuttings. **Silvae Genetica**, v. 64, n. 1–6, p. 291–308, 2015.

BEWLEY, J. D.; BRADFORD, K. J.; HILHORST, H. W. M.; NONOGAKI, H. **Seeds: Physiology of Development, Germination and Dormancy, 3rd Edition**. 3rd ed. New York, NY: Springer New York, 2013.

BOELT, BIRTE; SHRESTHA, S.; SALIMI, Z.; et al. Multispectral imaging - A new tool in seed quality assessment? **Seed Science Research**, v. 28, n. 3, p. 222–228, 2018.

BOELT, B; SHRESTHA, S.; SALIMI, Z.; et al. Multispectral imaging - A new tool in seed quality assessment? **Seed Science Research**, v. 28, n. 3, p. 222–228, 2018. Disponível

em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85049315069&doi=10.1017%2FS0960258518000235&partnerID=40&md5=9f2956fc27c419b0003dbd7fe1ce8848>. .

BRADFORD, K. J. Interpreting biological variation: Seeds, populations and sensitivity thresholds. **Seed Science Research**, v. 28, n. 3, p. 158–167, 2018.

BRADFORD, K. J.; BELLO, P.; FU, J. C.; BARROS, M. Single-seed respiration: A new method to assess seed quality. **Seed Science and Technology**, v. 41, n. 3, p. 420–438, 2013.

CAVERZAN, A.; GIACOMIN, R.; MÜLLER, M.; et al. How does seed vigor affect soybean yield components? **Agronomy Journal**, v. 110, n. 4, p. 1318–1327, 2018.

CHENG, H.; YE, M.; WU, T.; MA, H. Evaluation and Heritability Analysis of the Seed Vigor of Soybean Strains Tested in the Huanghuaihai Regional Test of China. **Plants**, v. 12, n. 6, 2023.

CORBINEAU, F. Markers of seed quality: From present to future. **Seed Science Research**, 2012.

DICKINSON, G. R.; WALLACE, H. M.; LEE, D. J. Reciprocal and advanced generation hybrids between Corymbia citriodora and C. torelliana: Forestry breeding and the risk of gene flow. **Annals of Forest Science**, v. 70, n. 1, p. 1–10, 2013.

DONALDSON, L. Autofluorescence in plants. **Molecules**, 2020.

ELMASRY, G.; MANDOUR, N.; AL-REJAIE, S.; BELIN, E.; ROUSSEAU, D. Recent applications of multispectral imaging in seed phenotyping and quality monitoring—An overview. **Sensors (Switzerland)**, 2019.

ESTEVE AGELET, L.; HURBURGH, C. R. Limitations and current applications of Near Infrared Spectroscopy for single seed analysis. **Talanta**, v. 121, p. 288–299, 2014. Elsevier. Disponível em: <http://dx.doi.org/10.1016/j.talanta.2013.12.038>. .

FAQEERZADA, M. A.; PEREZ, M.; LOHUMI, S.; et al. Online application of a hyperspectral imaging system for the sorting of adulterated almonds. **Applied Sciences (Switzerland)**, v. 10, n. 18, p. 1–16, 2020.

FRANÇA-SILVA, F.; GOMES-JUNIOR, F. G.; REGO, C. H. Q.; MARASSI, A. G.; TANNÚS, A. Advances in imaging technologies for soybean seed analysis. **Journal of Seed Science**, 2023. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2317-15372023000100302&tlng=en>. .

FU, X.; BAI, M.; XU, Y.; et al. Cultivars identification of oat (Avena sativa L.) seed via multispectral imaging analysis. **Frontiers in Plant Science**, v. 14, n. February, 2023.

FUSHING, H.; CHEN, C. Data Mechanics and Coupling Geometry on Binary Bipartite Networks. (E. Hernandez-Lemus, Org.)**PLoS ONE**, v. 9, n. 8, p. e106154, 2014. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0106154>. .

HE, Y.; YE, Z.; YING, Q.; et al. Glyoxylate cycle and reactive oxygen species metabolism are involved in the improvement of seed vigor in watermelon by exogenous GA3. **Scientia Horticulturae**, v. 247, n. April 2018, p. 184–194, 2019. Elsevier. Disponível em: <https://doi.org/10.1016/j.scienta.2018.12.016>. .

IBARRA, L.; HODGE, G.; ACOSTA, J. J. Quantitative Genetics of a Hybrid Population of Eucalyptus nitens × Eucalyptus globulus: Estimation of Genetic Parameters and Implications for Breeding Strategies. **Forests**, v. 14, n. 2, 2023.

ISTA. **International Rules for Seed Testing**. Bassersdorf, Switzerland: International Seed Testing Association, 2020.

JALINK, H.; VAN DER SCHOOR, R.; FRANDAS, A.; VAN PIJLEN, J. G.; BINO, R. J. Chlorophyll fluorescence of Brassica oleracea seeds as a non-destructive marker for seed maturity and seed performance. **Seed Science Research**, v. 8, n. 4, p. 437–443, 1998.

JIN, B.; QI, H.; JIA, L.; et al. Determination of viability and vigor of naturally-aged rice seeds using hyperspectral imaging with machine learning. **Infrared Physics and**

Technology, v. 122, n. February, p. 104097, 2022. Elsevier B.V. Disponível em: <https://doi.org/10.1016/j.infrared.2022.104097>. .

KANDPAL, L. M.; LOHUMI, S.; KIM, M. S.; KANG, J. S.; CHO, B. K. Near-infrared hyperspectral imaging system coupled with multivariate methods to predict viability and vigor in muskmelon seeds. **Sensors and Actuators, B: Chemical**, v. 229, p. 534–544, 2016. Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.snb.2016.02.015>. .

LI, L.; ZHANG, Q.; HUANG, D. A review of imaging techniques for plant phenotyping. **Sensors (Switzerland)**, 2014.

LI, Y.; SUN, J.; WU, X.; et al. Detection of viability of soybean seed based on fluorescence hyperspectra and CARS-SVM-AdaBoost model. **Journal Of Food Processing And Preservation**, v. 43, n. 12, 2019.

LIU, C.; LIU, W.; LU, X.; et al. Non-destructive discrimination of conventional and glyphosate-resistant soybean seeds and their hybrid descendants using multispectral imaging and chemometric methods. **Journal of Agricultural Science**, v. 154, n. 1, p. 1–12, 2014.

LIU, C.; LIU, W.; LU, X.; et al. Non-destructive discrimination of conventional and glyphosate-resistant soybean seeds and their hybrid descendants using multispectral imaging and chemometric methods. **Journal Of Agricultural Science**, v. 154, n. 1, p. 1–12, 2016.

MARCOS-FILHO, J. **Fisiologia de Sementes de Plantas Cultivadas**. 2 ed. ed. Londrina: ABRATES, 2015.

MCVEY, C.; HSIEH, F.; MANRIQUEZ, D.; PINEDO, P.; HORBACK, K. Livestock Informatics Toolkit: A Case Study in Visually Characterizing Complex Behavioral Patterns across Multiple Sensor Platforms, Using Novel Unsupervised Machine Learning and Information Theoretic Approaches. **Sensors**, v. 22, n. 1, p. 1, 2021. Disponível em: <https://www.mdpi.com/1424-8220/22/1/1>. .

DE MEDEIROS, A. D.; CAPOBIANGO, N. P.; DA SILVA, J. M.; et al. Interactive machine learning for soybean seed and seedling quality classification. **Scientific Reports**, v. 10, n. 1, p. 1–10, 2020. Nature Publishing Group UK. Disponível em: <https://doi.org/10.1038/s41598-020-68273-y>. .

MICHELON, T. B.; SERRA NEGRA VIEIRA, E.; PANOBIANCO, M. Spectral imaging and chemometrics applied at phenotyping in seed science studies: a systematic review. **Seed Science Research**, v. 33, n. 1, p. 9–22, 2023.

MORTENSEN, A. K.; GISLUM, R.; JØRGENSEN, J. R.; BOELT, B. The use of multispectral imaging and single seed and bulk near-infrared spectroscopy to characterize seed covering structures: Methods and applications in seed testing and research. **Agriculture (Switzerland)**, v. 11, n. 4, 2021.

OLESEN, M. H.; NIKNESHAN, P.; SHRESTHA, S.; et al. Viability prediction of ricinus cummunis L. Seeds using multispectral imaging. **Sensors (Switzerland)**, v. 15, n. 2, p. 4592–4604, 2015.

QI, H.; HUANG, Z.; SUN, Z.; et al. Rice seed vigor detection based on near-infrared hyperspectral imaging and deep transfer learning. **Frontiers in Plant Science**, v. 14, n. October, p. 1–13, 2023.

RAMALHO, M. A. P.; SANTOS, H. G.; SOUZA, T. DA S. Eucalyptus breeding programs: a proposal for the use of inbred progenies. **Cerne**, v. 28, n. 1, p. 1–9, 2022.

SENDIN, K.; MANLEY, M.; WILLIAMS, P. J. Classification of white maize defects with multispectral imaging. **Food Chemistry**, v. 243, n. June 2017, p. 311–318, 2018. Elsevier. Disponível em: <http://dx.doi.org/10.1016/j.foodchem.2017.09.133>. .

SHANNON, C. E. A Mathematical Theory of Communication. **Bell System Technical Journal**, v. 27, n. 3, p. 379–423, 1948. Disponível em: <https://ieeexplore.ieee.org/document/6773024>. .

SHRESTHA, R.; HARDEBERG, J. H. An experimental study of fast multispectral imaging using LED illumination and an RGB camera. **Final Program and Proceedings - IS and T/SID Color Imaging Conference**, v. 2015-Janua, n. 1 of 1, p. 36–40, 2015.

SHRESTHA, S.; DELEURAN, L. C.; GISLUM, R. Separation of viable and non-viable tomato (Solanum lycopersicum L.) seeds using single seed near-infrared spectroscopy. **Computers and Electronics in Agriculture**, v. 142, p. 348–355, 2017. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-

85029577373&doi=10.1016%2Fj.compag.2017.09.004&partnerID=40&md5=8291fd36a13835af3a17584f5f139de8>. .

SHRESTHA, S.; DELEURAN, L. C.; OLESEN, M. H. H. H. H.; GISLUM, R. R. Use of multispectral imaging in varietal identification of tomato. **SENSORS**, v. 15, n. 2, p. 4496–4512, 2015.

DA SILVA, P. H. M.; LEE, D. J.; AMANCIO, M. R.; ARAUJO, M. J. Initiation of breeding programs for three species of Corymbia: Introduction and provenances study. **Crop Breeding and Applied Biotechnology**, v. 22, n. 1, p. 1–9, 2022.

TU, K.; YIN, Y.; YANG, L.; WANG, J.; SUN, Q. Discrimination of individual seed viability by using the oxygen consumption technique and headspace-gas chromatography-ion mobility spectrometry. **Journal of Integrative Agriculture**, v. 22, n. 3, p. 727–737, 2023. CAAS. Publishing services by Elsevier B.V. Disponível em: <http://dx.doi.org/10.1016/j.jia.2022.08.058>. .

WEI, Y.; LI, X.; PAN, X.; LI, L. Nondestructive classification of soybean seed varieties by hyperspectral imaging and ensemble machine learning algorithms. **Sensors (Switzerland)**, v. 20, n. 23, p. 1–12, 2020.

XIA, Y.; XU, Y.; LI, J.; ZHANG, C.; FAN, S. Recent advances in emerging techniques for non-destructive detection of seed viability : A review Arti fi cial Intelligence in Agriculture Recent advances in emerging techniques for non-destructive detection of seed viability : A review. **Artificial Intelligence in Agriculture**, v. 1, n. May, 2019. Elsevier B.V. Disponível em: <https://doi.org/10.1016/j.aiia.2019.05.001>. .

XIN, X.; WAN, Y.; WANG, W.; et al. A real-time, non-invasive, micro-optrode technique for detecting seed viability by using oxygen influx. **Scientific Reports**, v. 3, n. 12, 2013.

YANG, L.; ZHANG, Z.; HU, X. Cultivar discrimination of single alfalfa (Medicago sativa l.) seed via multispectral imaging combined with multivariate analysis. **Sensors (Switzerland)**, v. 20, n. 22, p. 1–14, 2020.

ZHANG, J.; FENG, X.; LIU, X.; HE, Y. Identification of hybrid okra seeds based on near-infrared hyperspectral imaging technology. **Applied Sciences (Switzerland)**, v. 8, n. 10, 2018.

ZHOU, Q.; HUANG, W.; FAN, S.; et al. Non-destructive discrimination of the variety of sweet maize seeds based on hyperspectral image coupled with wavelength selection algorithm. **Infrared Physics and Technology**, v. 109, n. April, p. 103418, 2020. Elsevier. Disponível em: <https://doi.org/10.1016/j.infrared.2020.103418>. .