UNIVERSIDADE FEDERAL DE VIÇOSA

Machine learning and digital phenotyping for soybean characterization and classification

Wagner Faria Barbosa
Doctor Scientiae

VIÇOSA - MINAS GERAIS
2024

WAGNER FARIA BARBOSA

Machine learning and digital phenotyping for soybean characterization and classification

Thesis submitted to the Applied Statistics and Biometry Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Cosme Damiao Cruz

Co-adviser: Ivan Ricardo Carvalho

VIÇOSA - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

Barbosa, Wagner Faria, 1985-

B238t 2024

Machine learning and digital phenotyping for soybean characterization and classification / Wagner Faria Barbosa. – Viçosa, MG, 2024.

1 tese eletrônica (111 f.): il. (algumas color.).

Orientador: Cosme Damião Cruz.

Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2024.

Inclui bibliografia.

DOI: https://doi.org/10.47328/ufvbbt.2025.554

Modo de acesso: World Wide Web.

1. Soja - Variedades - Métodos estatísticos. 2. Aprendizado do computador. I. Cruz, Cosme Damião, 1958-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 633.3470727

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

WAGNER FARIA BARBOSA

Machine learning and digital phenotyping for soybean characterization and classification

Thesis submitted to the Applied Statistics and Biometry Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: December 19, 2024.

Assent:

Wagner Faria Barbosa
Author

Cosme Damiao Cruz
Adviser

To God and to my parents.

ACKNOWLEDGMENTS

I give thanks to God, who created me, shaped me, and renewed me through His Son, who gave His life so that I might have life. I am grateful to God, who sustains me through His Holy Spirit, strengthens me, and brings comfort in moments of sorrow and trial. I praise the Almighty God, who fills my heart with joy through the salvation secured and preserved in Christ. To the Triune God – Father, Son, and Holy Spirit – the eternal and sovereign King, both transcendent and near, be all honor, glory, strength, and power forever.

I am deeply grateful to my mother, Ana Virgínia Faria Barbosa, my father, Venilton Santos Barbosa (*in memoriam*), and my brother, João Victor Faria Barbosa, for their love, care, dedication, and encouragement in my studies.

I thank the new friends I made during the course, who shared with me not only the challenges but also the joys of graduate school.

I am grateful to my fellowship group friends for their friendship, prayers, and for walking alongside me, teaching me to walk with Jesus.

A special thanks to my friend Lorena Lisbetd, who dedicated her time to bless me with her prayers and shared in my joy.

I thank the Universidade Federal de Viçosa for the opportunity to pursue a PhD in Applied Statistics and Biometrics.

My gratitude extends to the professors of the PPESTBIO for their dedication to teaching.

I also thank the program secretary, Júnior José Pires, for his readiness and diligence in assisting with various bureaucratic tasks.

I am thankful to the members of my defense committee – Prof. Ivan, Prof. Prof. Ithalo, Prof. Peternelli, Prof. Matheus, and Prof. Éder – for their valuable corrections and suggestions.

I express my profound gratitude to my advisor, Prof. Cosme Damião Cruz, whose dedication and enthusiasm for research provided me with exceptional guidance, support, advice, and confidence.

This work has been sponsored by the following Brazilian research agencies: Coordination for the Improvement of Higher Education Personnel (CAPES; Financing code 001), Minas Gerais State Foundation for Research Aid (FAPEMIG) and National Council of Scientific and Technological Development (CNPq).

“Ele é a imagem do Deus invisível, o primogênito sobre toda a criação, pois por meio dele foram criadas todas as coisas nos céus e na terra, as visíveis e as invisíveis, sejam tronos, sejam soberanias, quer poderes, quer autoridades; todas as coisas foram criadas por ele e para ele. Ele é antes de todas as coisas, e por meio dele tudo subsiste. Ele é a cabeça do corpo, que é a igreja; é o princípio e o primogênito dentre os mortos, para que em tudo tenha a supremacia. Pois foi do agrado de Deus que nele habitasse toda a plenitude e por meio dele reconciliasse consigo todas as coisas, tanto as que estão na terra quanto as que estão nos céus, estabelecendo a paz por meio do seu sangue derramado na cruz.”

Colossenses 1:15-20 (NVI)

ABSTRACT

BARBOSA, Wagner Faria, D.Sc., Universidade Federal de Viçosa, December, 2024. **Machine learning and digital phenotyping for soybean characterization and classification**. Adviser: Cosme Damiao Cruz. Co-adviser: Ivan Ricardo Carvalho.

Digital phenotyping has revolutionized the study of plant phenotypic traits, particularly in crops such as soybean (*Glycine max* (L.) Merr.). Combining digital phenotypic descriptors with machine learning enabled significant methodological advances, increasing the reliability and precision of classifications. Thus, this study aimed to develop a protocol for extracting various phenotypic traits from soybean leaflets, including shape attributes, elliptical Fourier descriptors (EFDs), Haralick texture features, and vegetation indices (VIs). The study also assessed the potential of these traits using several statistical methods and applied machine learning to discriminate the ancestry of the genotypes. The protocol proved effective in facilitating the acquisition of different sets of features for modeling and classification tasks, providing researchers with a robust and versatile framework that can be easily adapted to similar applications. The results indicated that VIs were the most frequent attributes, while EFDs exhibited the least redundancy, highlighting each dataset's potential to identify genotypic patterns. However, using Random Forest (RF) as a classification method demonstrated high efficiency in handling data, maximizing accuracy and specificity in the developed models. Specifically, texture and shape-related attributes were crucial for successful discrimination. This study concludes that the integrated approach of digital phenotyping and machine learning represents a powerful tool for plant breeding, providing practical solutions to cultivar identification and characterization challenges.

Keywords: image-based phenotyping; Fourier descriptors; vegetative indices; Haralick textures; genotypic discrimination

RESUMO

BARBOSA, Wagner Faria, D.Sc., Universidade Federal de Viçosa, dezembro de 2024. **Aprendizado de máquina e fenotipagem digital para caracterização e classificação da soja**. Orientador: Cosme Damiao Cruz. Coorientador: Ivan Ricardo Carvalho.

A fenotipagem digital tem revolucionado o estudo de características fenotípicas de plantas, particularmente em culturas como a soja (*Glycine max* (L.) Merr.). A combinação de descritores fenotípicos digitais com aprendizado de máquina tem possibilitado avanços metodológicos significativos, aumentando a confiabilidade e a precisão das classificações. Este estudo teve como objetivo criar um protocolo para extração de diversas características fenotípicas de folíolos de soja, incluindo atributos de forma, descritores elípticos de Fourier (DEFs), características de textura de Haralick e índices vegetativos (IVs). O estudo também avaliou a potencialidade dessas características através de métodos estatísticos e utilizou aprendizado de máquina para discriminar a ascendência dos genótipos. O protocolo mostrou facilitar a aquisição dos diferentes conjuntos de características para tarefas de modelagem e classificação, oferecendo aos pesquisadores uma estrutura robusta e versátil que pode ser facilmente adaptada a aplicações semelhantes. Os resultados indicaram que os IVs foram os atributos mais frequentes, enquanto os DEFs apresentaram menor redundância, destacando o potencial de cada conjunto de dados para identificar padrões genotípicos. O uso do método de classificação Random Forest (RF) demonstrou alta eficiência no tratamento dos dados, maximizando a acurácia e a especificidade nos modelos desenvolvidos. Especificamente, os atributos relacionados à textura e forma foram cruciais para a discriminação bem-sucedida. Este estudo conclui que a abordagem integrada de fenotipagem digital e aprendizado de máquina representa uma ferramenta poderosa para o melhoramento de plantas, oferecendo soluções práticas para os desafios de identificação e caracterização de cultivares.

Palavras-chave: fenotipagem baseada em imagens; descritores de Fourier; índices vegetativos; texturas de Haralick; discriminação genotípica

SUMMARY

INTRODUCTION

The evolution of digital technologies has brought significant innovations to agriculture, with notable impacts on the systematic study of phenotypes, known as phenomics. This approach enables large-scale characterization of observable traits in organisms, linking them to genetic and environmental factors (HOULE; GOVINDARAJU; OMHOLT, 2010). Within this context, digital image analysis stands out as a promising tool to overcome the limitations of traditional methods, such as manual measurements, which are often time-consuming, costly, and prone to human error (COBB et al., 2013; WALTER; LIEBISCH; HUND, 2015).

Image-based phenotyping enables the precise and rapid extraction of morphological and structural plant traits using methods that combine digital capture, automated processing, and advanced computational analyses. It allows consistently handling large datasets while reducing subjectivity inherent to manual methods. Additionally, image phenotyping identifies complex traits not easily discernible through conventional analyses, such as geometric patterns, colorimetric variations, and textural features in plant tissues (AMARAL et al., 2024; HARALICK; SHANMUGAM; DINSTEIN, 1973; OLIVOTO, 2022).

Geometric patterns can be analyzed using conventional measures such as area, perimeter, distances, and Fourier Elliptical Descriptors (EFDs). These mathematical equations describe the closed two-dimensional contour of objects in an image, proving useful in studies involving genotype selection and classification (MOLLMAN; ÇIFTÇI; EROL, 2023; VISCOSI; FORTINI, 2011). Similarly, texture demonstrates significant potential for distinguishing complex phenotypic traits (MONTES; PAUL; MELCHINGER, 2007), quantified using Haralick's equations, which characterize tonal variation patterns in images (HARALICK; SHANMUGAM; DINSTEIN, 1973). Lastly, colorimetric patterns in plant tissues can be identified using Vegetative Indices (VIs). These indices, derived from spectral combinations of image bands, maximize sensitivity to specific material traits (FANG; LIANG, 2014), making them valuable as discriminative features in biological material studies.

This array of image-derived attributes is instrumental in recognizing patterns within germplasm banks, enabling the identification of cultivars that combine higher productivity with desirable traits, such as drought tolerance and pest resistance (REN; WERADUWAGE; SHARKEY, 2019; ROWLAND et al., 2020). However, robust genotype discrimination requires the application of suitable methodologies to establish similarity patterns. Among emerging approaches, machine learning methods, such as Random Forest (RF), have efficiently handled large datasets and interrelated characteristics. RF enables modeling non-linear relationships and identifying critical features for genotypic distinction, fostering more robust

analyses (BREIMAN, 1996; LIAW; WIENER, 2015). Nonetheless, RF can be negatively affected by imbalanced datasets, where dominant classes tend to be favored (TANG; HENDERSON; GARDNER, 2021; WANG et al., 2021).

Soybean is one of the crops that has benefited significantly from image analysis. This crop is crucial in the global economy and human and animal nutrition (ANDERSON et al., 2019). Its relevance extends to various domains, serving as a source of proteins and nutrients for human consumption, an essential input for animal feed, and a raw material for biofuel production.

Therefore, this study aims to:

1. Develop a robust method for extracting phenotypic attributes from soybean leaflets, encompassing VIs, texture descriptors, and EFDs.
2. Evaluate the discriminative capacity of different attribute sets in classifying genotypes and identifying distinct patterns among cultivars within a soybean germplasm bank.
3. Propose improvements to Random Forest's biometric strategy to address imbalanced datasets.

The presented results aim to contribute to advances in genetic improvement, integrating digital technologies and machine learning into practical and innovative solutions for cultivar identification, selection, and characterization challenges.

REFERENCES

AMARAL, L. R. et al. Remote sensing imagery to predict soybean yield: a case study of vegetation indices contribution. **Precision Agriculture**, v. 25, n. 5, p. 2375–2393, 2024.

ANDERSON, E. J. et al. Soybean [*Glycine max* (L.) Merr.] breeding: History, improvement, production and future opportunities. In: AL-KHAYRI, J. M.; JAIN, S. M.; JOHNSON, D. V (Org.). **Advances in Plant Breeding Strategies: Legumes: Volume 7**. Cham: Springer International Publishing, 2019. p. 431–516. Disponível em: <https://doi.org/10.1007/978-3-030-23400-3_12>.

BREIMAN, L. Bagging predictors. Machine Learning 24 (2): 123–140. **Google Scholar Google Scholar Digital Library Digital Library**, 1996.

COBB, J. N. et al. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. **Theoretical and Applied Genetics**, v. 126, p. 867–887, 2013.

FANG, H.; LIANG, S. Leaf Area Index Models☆. **Reference Module in Earth Systems and Environmental Sciences**. [S.l.]: Elsevier, 2014. Disponível em: <https://www.sciencedirect.com/science/article/pii/B978012409548909076X>.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural Features for Image Classification. **IEEE Transactions on Systems, Man, and Cybernetics**, v. SMC-3, n. 6, p. 610–621, 1973.

HOULE, D.; GOVINDARAJU, D. R.; OMHOLT, S. Phenomics: the next challenge. **Nature reviews genetics**, v. 11, n. 12, p. 855–866, 2010.

LIAW, A.; WIENER, M. randomForest: Breiman and Cutler's random forests for classification and regression. **R package version**, v. 4, p. 14, 2015.

MOLLMAN, R.; ÇIFTÇI, A.; EROL, O. Variable leaf shape on short and long shoots: an elliptic Fourier analysis of *Prunus microcarpa* CA Mey. **Brazilian Journal of Botany**, v. 46, n. 1, p. 113–125, 2023.

MONTES, J. M.; PAUL, C.; MELCHINGER, A. E. Determination of chemical composition and nutritional attributes of silage corn hybrids by near-infrared spectroscopy on chopper: evaluation of traits, sample presentation systems and calibration transferability. **Plant Breeding**, v. 126, n. 5, p. 521–526, 2007.

OLIVOTO, T. Lights, camera, pliman! An R package for plant image analysis. **Methods in Ecology and Evolution**, v. 13, n. 4, p. 789–798, 1 abr. 2022. Disponível em: <https://doi.org/10.1111/2041-210X.13803>.

REN, T.; WERADUWAGE, S. M.; SHARKEY, T. D. Prospects for enhancing leaf photosynthetic capacity by manipulating mesophyll cell morphology. **Journal of Experimental Botany**, v. 70, n. 4, p. 1153–1165, 2019.

ROWLAND, S. D. et al. Leaf shape is a predictor of fruit quality and cultivar performance in tomato. **new phytologist**, v. 226, n. 3, p. 851–865, 2020.

TANG, J.; HENDERSON, A.; GARDNER, P. Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced data sets. **Analyst**, v. 146, n. 19, p. 5880–5891, 2021.

VISCOSI, V.; FORTINI, P. Leaf shape variation and differentiation in three sympatric white oak species revealed by elliptic Fourier analysis. **Nordic Journal of Botany**, v. 29, n. 5, p. 632–640, 1 out. 2011. Disponível em: <https://doi.org/10.1111/j.1756-1051.2011.01098.x>.

WALTER, A.; LIEBISCH, F.; HUND, A. Plant phenotyping: from bean weighing to image analysis. **Plant Methods**, v. 11, n. 1, p. 14, 2015. Disponível em: <https://doi.org/10.1186/s13007-015-0056-8>.

WANG, L. et al. Review of classification methods on unbalanced data sets. **Ieee Access**, v. 9, p. 64606–64628, 2021.

CHAPTER 1

THOROUGH METHOD FOR COMPREHENSIVE MORPHOLOGICAL CHARACTERIZATION OF SOYBEAN LEAVES

ABSTRACT

Computational analysis and interpretation of images are pivotal in enhancing various agricultural procedures. The morphological characterization of plants is widely recognized as essential for initiatives aimed at improving crops. Parameters such as leaf area and shape are unequivocally linked to the efficiency of photosynthetic energy capture and, consequently, to overall crop productivity. Additionally, traits like texture, vegetative indexes (VIs), morphological descriptors, and Fourier elliptical descriptors possess significant discriminatory power for differentiating genotypes. Despite this, methodologies for extracting comprehensive variable sets from leaf images still need to be more adequately explored and are often subjective. Thus, this method offers: 1) an option to store images in the cloud, allowing them to be deleted from the local directory once they have been processed; 2) to meticulously detail the extraction of diverse characteristic sets from RGB images of soybean leaflets, and; 3) to facilitate the acquisition of these variable sets for modeling and classification tasks, providing researchers with a thorough and adaptable protocol for similar applications.

Keywords: Fourier elliptical descriptor, Haralick's texture features, vegetative indexes, phenotype, RGB image.

INTRODUCTION

Agriculture has experienced profound advancements through technological integration, yet these innovations bring forth novel challenges. Digital images and computational analysis have become a cornerstone in managing various agricultural tasks (TILLETT, 1991).

The composition of a digital image presents a myriad of mathematical relationships among the values assigned to various bands of each pixel, offering the potential to unveil or accentuate features within the depicted scene. These mathematical relationships, often termed indexes or vegetative indexes (VIs) are meticulously crafted to enhance sensitivity to vegetation attributes while mitigating confounding factors such as soil reflectance, directional effects, or atmospheric interference (FANG; LIANG, 2014).

Hence, digital imagery has become valuable in contemporary agricultural experimentation, particularly plant breeding programs (PRADEBON et al., 2024; ZHAO et al., 2019). The utilization of images offers heightened measurement reliability, enabling precise, accurate, and swift phenotyping of targeted components while yielding a vast array of diverse data. Consequently, phenotyping through imagery has progressively emerged as an alternative to conventional evaluations, which typically entail more significant labor, time, and susceptibility to experimental errors (CORTES et al., 2017).

Attributes like leaf area and shape are directly related to a plant's capacity to capture solar energy, which affects its productivity (ROWLAND et al., 2020). However, beyond conventional measurements like area, perimeter, and length obtained through image object segmentation, Elliptical Fourier descriptors (EFDs) offer an additional dimension to plant morphological analysis (NETO et al., 2006). EFDs are mathematical equations that delineate two-dimensional closed contours by decomposing them into a series of harmonically related ellipses (KUHL; GIARDINA, 1982). Consequently, they prove invaluable in describing various aspects of organisms, individual parts, or entire structures (MOLLMAN; ÇIFTÇI; EROL, 2023; VISCOSI; FORTINI, 2011).

Texture represents another vital aspect in plant morphological characterization and differentiation (BEGHIN et al., 2010). Despite being an inherent property of all surfaces and easily discernible by human observers, texture necessitates analytical measurement in digital images to be effectively harnessed (HARALICK; SHANMUGAM; DINSTEIN, 1973). Within the realm of plant analysis, Haralick's texture features have seen some application (EKIZ; ARICA, 2022; RAJU; BALACHANDER; NEEHARIKA, 2022), though their utilization in this domain appears to remain relatively modest.

Despite the range of potential variable sets for plant phenotyping derived from images, their extraction needs more standardization in the literature, often resulting in replication efforts grounded in subjective and superficial methodologies. Recognizing this gap, we established a robust method for extracting diverse variable sets from soybean leaflets (*Glycine max* (L.) Merr.), readily adaptable by researchers tackling similar challenges. Soybean was chosen as the model organism because it is one of the world's largest commodities, boasting a global production of approximately 397 million tons in 2023/2024 (UNITED STATES DEPARTMENT OF AGRICULTURE, 2024).

The method encompassed the extraction of variable sets comprising VIs, shape measures, EFDs, and Haralick's texture characteristics. Extraction procedures were conducted using the R software [v. 4.4.0], renowned for its user-friendly interface and open-access nature, prized for its data manipulation, analysis, and graphics (POSIT TEAM, 2024). Additionally, the pliman package, tailored specifically for plant image analysis, was employed (OLIVOTO, 2022).

To ensure a proper understanding of this variable extraction method, the R project described in this work, along with some images of soybean leaflets, is available in the following online repository: https://github.com/barbosawf/ImageAnalysis.

METHOD DETAILS

General Instructions for Method Execution

It is essential to consider necessary procedures for obtaining information from a set of images of objects of interest, which in this case were soybean leaflets. Blocks were formed to facilitate reading and identify execution sections in the scripts. They can be identified by a specific title between the hashtag character (#) and a sequence of dashes (e.g., # Packages for usage ---------). In R Studio, these script blocks can be created using the shortcut "Control + Shift + R". These blocks can be opened or closed to adjust the scope of script visualization.

It is also recommended that you carefully read the comments identified solely by the hashtag character (#) above the scripts' command lines, as they assist in understanding the choice of arguments for functions and indicate essential procedures or changes in some steps.

Furthermore, some general recommendations are indispensable for better understanding the procedures and are added as notes at the end of the different steps of this method. Please read them carefully before executing any procedure.

Acquisition of Leaflet Images

Leaves from the middle third of three plants of each soybean line were harvested during stage R5 when the plant has attained its maximum height, number of nodes, and leaf area. The leaf sampling occurred between 8 a.m. and 4 p.m., and the collected leaves were transported to the laboratory for RGB image capture. Photos of the leaves were taken under 22ºC and diffused artificial lighting. Three leaflets from each plant (totaling nine) were arranged in three horizontal sections on red cardboard alongside an 18 cm^2 yellow reference (**Figure 1**). Leaflet images were captured using a digital camera positioned 50 cm above the cardboard on appropriate equipment for image capture (**Figure 2**). The images were acquired at a resolution of 1280 × 1024 pixels and 96 dpi (dots per inch), then named and stored on Google Drive.

Notes: Ensure that the objects to be photographed, in this case, leaflets, are well-framed within the camera's view. Avoid overlapping of objects and maintain a relative distance between them. Depending on the characteristics to be extracted from the object (e.g., area), prevent their edges from folding to ensure subsequent measurements are accurate. Gently dry any wet objects (in this case, it was necessary due to rain in some days prior to leaf collection). Establish conditions for diffuse lighting to minimize shadowing. The background color can be any, if it contrasts with the objects under study (in this case, the leaflets) and the reference. Change the background (while keeping the same color) whenever excess spots are observed, which may occur due to pigment detachment from objects (e.g., leaflets) during friction.

Figure 1. Soybean (*Glycine max* (L.) Merr) leaflets at R5 stage. Three leaflets from three plants of a single line are horizontally arranged on red cardboard along with a yellow reference measuring 18 cm^2.

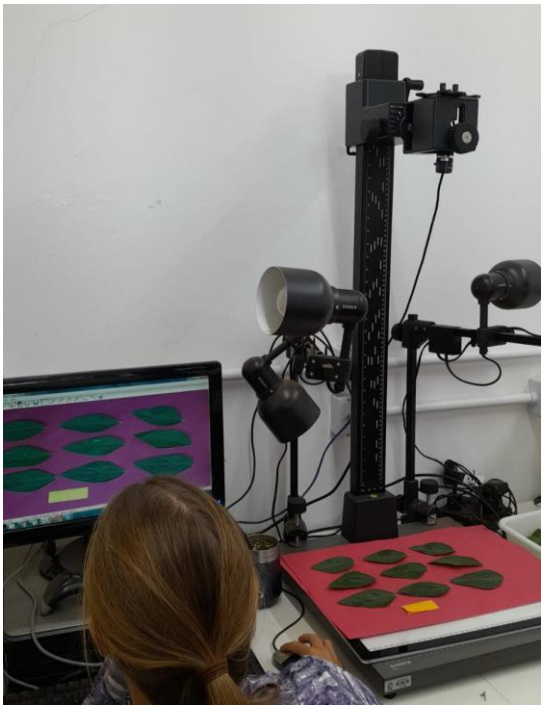

Figure 2. Setup assembled for capturing images of soybean (*Glycine max* (L.) Merr.) leaflets at R5 stage.

Software

The R software [v. 4.4.0] was used to establish a method for extracting different sets of variables. It is accessible on the website https://cran.r-project.org/. To facilitate the development of this work, the RStudio [v. 2024.4.1.748] (POSIT TEAM, 2024) integrated development environment (IDE) was utilized. It is available at https://posit.co/download/rstudio-desktop/. RStudio seamlessly integrates with R upon the installation of both software. These tools were employed to acquire image characteristics and manipulate data.

Package Installation

The extraction and manipulation of data in this work were assisted using the following packages: pliman [v. 2.1.0] (OLIVOTO, 2022), tidyverse [v. 2.0.0] (WICKHAM et al., 2019), usethis [v. 2.2.3] (WICKHAM et al., 2024), googledrive [v. 2.1.1] (D'AGOSTINO MCGOWAN; BRYAN, 2023), and writexl [v. 1.5.0] (OOMS, 2024). To install the packages, execute the following commands in the RStudio Console:

```
# Install the packages required for this method.
install.packages("remotes")
packages <- c("pliman", "tidyverse", "usethis", "googledrive", "writexl")
```

```
versions <- c("2.1.0", "2.4.5", "2.0.0", "2.2.3", "2.1.1", "1.5.0")

for (i in seq_len(length(packages))) {
  if (!requireNamespace(packages[i])){
    remotes:: install_version(packages[i], versions[i])
  }
}
```

Notes: It is recommended that the package installation is executed only once. Reinstallations or updates of packages may cause execution errors in the scripts presented in this work. Therefore, if any of the packages are already installed, uninstall them and then reinstall them using the provided script to ensure they are in the required version specified by this method.

Creating a Project

It is recommended that a project using RStudio be created to facilitate the organization of the work environment and the definition of directories within the project. Creating a project also helps change directories on the computer, avoiding the need to specify the project's main directory when opening it in RStudio. Two ways to create a project in RStudio are suggested:

1. In RStudio, follow these steps: File ➜ New Project... ➜ New Directory ➜ New Project. In "Directory name:" define the project name (suggested: ImageAnalysis). Select the option Create a git repository only if you are familiar with local code versioning (git installation is required in this case: https://git-scm.com/downloads).

2. This procedure can be performed using the following command, which can be entered in the RStudio Console:

```
# Load the "usethis" package
library(usethis)

# Creating a project. The final directory name will serve as both
# the project name (with the extension "Rproj") and the containing folder.
# Suggested project name: ImageAnalysis.
create_project(path = "C:/Users/Administrador/Documents/ImageAnalysis")
```

Notes: Avoid using spaces, accents, and non-alphabetic characters in project names. To reopen the project after closing it, double-click the project file (with the extension "Rproj") within the project folder or use the path "File ➜ Open Project..." in RStudio.

Downloading the Images

This method was developed to facilitate the use of images stored in the cloud, specifically on Google Drive. After downloading and processing the images, they can be deleted from the local directory, freeing up space on the user's storage device (e.g., Solid State Drive). However, authorization to Google Drive for the Tidyverse API Packages will be required (**Figure 3**) to enable the R software to access and download images stored in the cloud. This access includes entering a Google email, password, and verification code, signing into Tidyverse API Packages, and granting Tidyverse API Packages permission to see, edit, create, and delete files on Google Drive. Once authorization has been granted, subsequent accesses will be readily allowed using the "drive_auth()" function, except when attempted on different computers or after R updates.

In addition, the provided script can be run whenever new image files are added to the Google Drive folder. The code will compare the contents of the Google Drive folder with the local folder within the project folder and download only the images not already in the local folder.

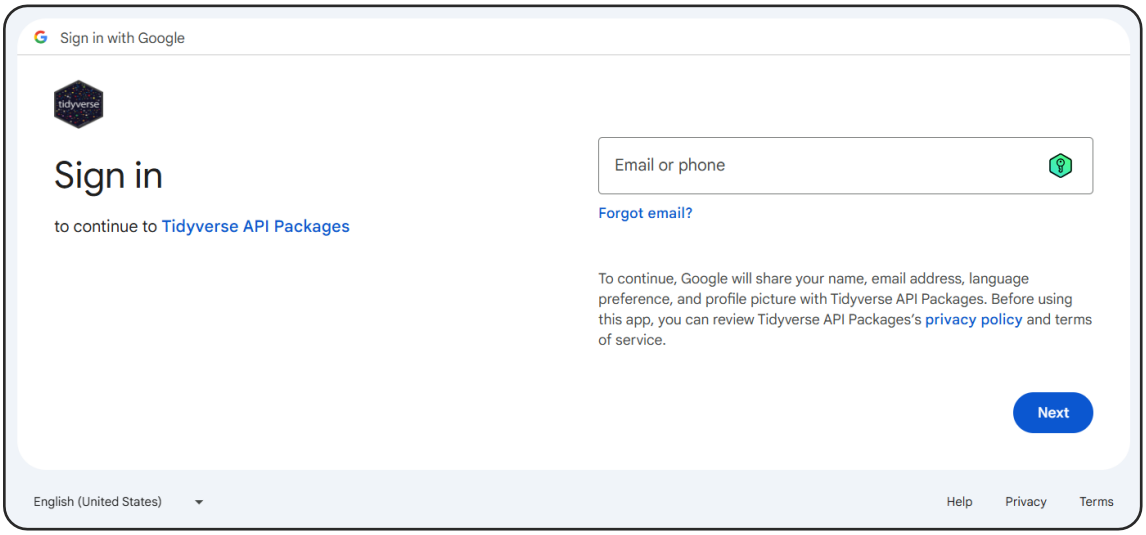

Figure 3. The first step is authorizing access to Google Drive for the Tidyverse API Packages.

```
# Creating the file for the image download script -------------------------

# Create a file where the script below can be entered and then saved
# within the project directory.
# Type the code line below into the Console after opening RStudio.
usethis::use_r("ImageDownloads")


# Packages for usage ------------------------------------------------------

library("googledrive")
library("tidyverse")
```

```
# Obtaining authorization for Google Drive access ------------------------

# Please provide a Google access email. Email information is only
# required for the initial access.
drive_auth(email = "my_gmail_account@gmail.com")

# After the first access, if you need to access Google Drive repeatedly,
# just run the authorization function without specifying the email.
# The function automatically prompts users to choose the previous
# access option.
drive_auth()


# Creating a folder to store the images ----------------------------------

# Create a folder named "ImagesProj" within the project directory to host
# the images downloaded from Google Drive.
if (!dir.exists("ImagesProj")) {
  dir.create("ImagesProj")
}

# Listing the names of images on the computer and Google Drive ----------

# Lists and stores the names of images within the "ImagesProj" folder.
HD_files <- list.files("ImagesProj", pattern = ".jpg")

# Lists and stores the names of images within the "ImagesGoogle" folder.
# Note: "ImagesGoogle" is a suggested folder to store images
# within Google Drive.
Google_files <- drive_ls("ImagesGoogle")
GD_files <- Google_files[[1]]


# Distinguishing images by storage location ------------------------------

# Obtains the differentiation between image files stored in the
# "ImagesGoogle" folder on Google Drive and those in the
# "ImagesProj" folder within the project.
dif <- setdiff(GD_files, HD_files)

dif_in_Google_files <- Google_files |>
  dplyr::filter(name %in% dif)


# Downloading images from Google Drive to the computer -------------------

# Download the image files from Google Drive to the "ImagesProj" folder.

map2(dif_in_Google_files$id,
     dif_in_Google_files$name,
     ~drive_download(as_id(.x), path = file.path("ImagesProj", .y)))


# Verifying if all images on Google Drive have been downloaded ------------

# Verifies whether all image files from Google Drive have been downloaded
# to the "ImagesProj" folder set within the project directory. If TRUE,
# executing the setequal() function will return TRUE.
HD_files_2 <- list.files("ImagesProj", pattern = ".jpg")
setequal(GD_files, HD_files_2)
```

Notes: Adopt the same folder naming convention proposed in this method for storing images, both on Google Drive (i.e., "ImagesGoogle") and the local drive (i.e., "ImagesProj"). This

practice ensures seamless use of the code without requiring any modifications. In addition, an important consideration at this stage is the potential conflict of image names stored on Google Drive. The provided script addresses this by resolving duplicate image names through a unique key assigned to each image upon upload to Google Drive. However, it is preferable to avoid these conflicts altogether by ensuring that image names are distinct, even if they are stored in different folders on Google Drive. Therefore, if there are other images on Google Drive besides those used for variable extraction, ensure that the images for your R project are named uniquely. In addition, keep in mind that deleting all images from the local disk should only be done after the project is complete and all images have been processed.

Testing Object Segmentation

The ultimate goal of the segmentation test is to identify an index that effectively segments both the objects of interest (leaflets) and the reference alongside another index that solely isolates the reference (**Figure 4**). In the former scenario, the objects of interest and the reference should be distinctly highlighted, with the image's background appearing white (**Figure 4**). Conversely, the reference will be depicted in black against a white background (**Figure 4**). It is crucial to emphasize that the final image (**Figure 4**) should be derived from the initial segmented image (**Figure 4**).

Figure 4. Segmentation test for both the objects of interest (leaflets) and the reference (yellow rectangle) (A) and the reference only (B).

This step is pivotal for selecting VIs that effectively segment the objects and references within the images. If the images were acquired under controlled lighting conditions and retained consistent background colors and references, subsequent steps involving extracting variable sets from the studied objects are expected to require no adjustments.

Before starting the segmentation test, researchers with prior knowledge of VIs that might be useful can add these indexes to the “indexes.csv” file in the pliman package. To access the directory containing the indexes.csv file, run the command utils::browseURL(system.file(package="pliman")) in the RStudio Console. Adding new VIs to the indexes.csv file will include them in the set of variables that will be extracted subsequently.

The pliman package includes 55 indexes, yet only 43 belong to one of the three color systems (RGB, HSB, and CIELab) capable of segmenting three-band images. In this study, 11 new VIs from the RGB color system were appended to the final lines of the “indexes.csv” file, bringing the total count of useful VIs for segmenting three-band images to 54 (see Supplementary CSV File available at https://github.com/barbosawf/ImageAnalysis).

```
# Creating the file for the segmentation test script ----------------------

# Enter the line below in the RStudio Console to create the script file.
usethis::use_r("SegmentationTests")


# Packages for use --------------------------------------------------------

library("pliman")
library("tidyverse")


# Choose an image from those stored in the "ImagesProj" folder ------------

# List the names of the first ten images
# Modify the extension pattern of your images (e.g., png or tiff) if
# it differs from what was used in this procedure
list.files("ImagesProj", pattern = "jpg")[1:10]


# Finding the best VIs for segmentation -----------------------------------

# Importing the image for segmentation tests. You should choose an image
# from the "ImagesProj" folder and replace the name "10.jpg"
# in the function image_import()
img <- image_import("10.jpg", path = "ImagesProj", plot = T)


# Segmentation with all VIs and false for inversion (invert = F).
# Look for VIs that segment the objects under study and the reference.
# The background will be white. The objects under study and the reference
# will be highlighted.
image_segment(
  # image chosen for segmentation
  img,
  # index = "all" performs segmentation using all available VIs
  index = "all",
  # Processes the images asynchronously in parallel
  parallel = T,
  # Display the image segmented with all available VIs
  plot = T,
  # invert = F does not invert the segmentation
  invert = F
)
```

```
# Segmentation with all VIs and true for inversion (invert = T).
# Look for VIs that segment the objects under study and the reference.
# If the inversion parameter is enabled (invert = TRUE), the segmentation
# in the image will occur in the region complementary to the desired object.
# For the selected indexes, the objects under study and the reference would
# appear white, with the background highlighted. However, when inversion
# is activated, the opposite occurs.
image_segment(
  # image chosen for segmentation
  img,
  # index = "all" performs segmentation using all available VIs
  index = "all",
  # Processes the images asynchronously in parallel
  parallel = T,
  # Display the image segmented with all available VIs
  plot = T,
  # invert = T inverts the segmentation
  invert = F
)


# Expanded view of segmentations with the chosen VIs ----------------------

# VIs for segmenting the objects under study and the reference when
# inversion is false (invert = F).
# Replace these indexes to fit the segmentation of your image.
VIs_obj_ref_inv_false <- c("SCI", "HUE", "R-G", "RGRI", "MyIndex")


# Display the segmentation of the study objects (leaflets) and the reference
# comprehensively when inversion is set to false (invert = F).
map(VIs_obj_ref_inv_false,
    ~ image_segment(
      img,
      index = .x,
      plot = T,
      filter = 4,
      invert = F
    ))


# VIs for segmenting the objects under study and the reference when
# inversion is true (invert = T).
# Replace these indexes to fit the segmentation of your image.
VIs_obj_ref_inv_true <-
  c("NG",
    "GLI",
    "NGRDI",
    "VARI",
    "GLAI",
    "G-R",
    "MNGRDI",
    "ExG",
    "VEG")


# Display the segmentation of the study objects (leaflets) and the reference
# comprehensively when inversion is set to true (invert = T).
map(VIs_obj_ref_inv_true,
    ~ image_segment(
      img,
      index = .x,
      plot = T,
      filter = 4,
      invert = T
    ))
```

```
# Finding the optimal segmentation for the reference only -----------------

# First, segment the study objects and the reference.
# Use an index that supports this (e.g., SCI).
# Specify whether inversion is true or false for the chosen index.
seg_obj_ref <-
  image_segment(
    img,
    index = "SCI",
    plot = T,
    filter = 4,
    invert = F
  )


# Indexes to try to segment only the reference.
# Only "RGB", "HSB", and "CIELab" band indexes segment three-band images
paste0(system.file(package = "pliman"), "/indexes.csv") |>
  read_csv2(show_col_types = FALSE) |>
  filter_at(vars(Band), ~ . %in% c("RGB", "HSB", "CIELab")) |>
  select("Index") |>
  pull() ->
  indexes_for_use

names(indexes_for_use) <- indexes_for_use


# Saves VIs that can be used to segment three-band images
# This file will also be used in the VarExtraction.R and
# VarExtractionFailedObj.R scripts
saveRDS(indexes_for_use, "indexes_for_use.rds")


# Determine which VIs can segment only the reference using the
# image_binary() function with inversion set to false (invert = F).
# The reference should appear black while the rest of the image is white.
seg_only_ref_inv_false <-
  map(indexes_for_use, \(x)
      try(image_binary(
        seg_obj_ref,
        index = x,
        plot = T,
        filter = 4,
        parallel = T,
        invert = F
      ))
  )

# Determine which VIs can segment only the reference using the
# image_binary() function with inversion set to true (invert = T).
# The reference should appear black while the rest of the image is white.
seg_only_ref_inv_true <-
  map(indexes_for_use, \(x)
      try(image_binary(
        seg_obj_ref,
        index = x,
        plot = T,
        filter = 4,
        parallel = T,
        invert = T
      ))
  )


# Reference pixels about the VIs -----------------------------------------

# Here, you can find the number of reference pixels corresponding to
# each VI chosen for segmentation.
```

```
# Reference segmentation using the "R-B" index when invert = FALSE
length(which(seg_only_ref_inv_false$`R-B`$`R-B` != 1))

# Reference segmentation using the "HUE2" index when invert = TRUE
length(which(seg_only_ref_inv_true$HUE2$HUE2 != 1))

# Reference segmentation using the "B-R" index when invert = TRUE
length(which(seg_only_ref_inv_true$`B-R`$`B-R` != 1))

# Reference segmentation using the "b*" index when invert = TRUE
length(which(seg_only_ref_inv_true$`b*`$`b*` != 1))

# Reference segmentation using the "b*-a" index when invert = TRUE
length(which(seg_only_ref_inv_true$`b*-a`$`b*-a` != 1))
```

Extraction of Variable Sets

Various variable sets can be extracted after identifying optimal VIs for segmenting the objects under study (e.g., leaflets) and the reference. The algorithm detailed in this study was engineered to afford safeguarding measures and preclude redundant processing of images within a designated directory. This design obviates the necessity of manually removing processed images from the folder. The implementation ensures that newly added images are consistently appended to the end of a list, enabling the algorithm to discern and circumvent those that have been previously processed.

Furthermore, an algorithmic measure for computational processing protection has been integrated. Given the capability of the “analyse_objects()” function within the pliman package to process all images within a designated directory, the computational demands, particularly regarding RAM allocation, may become excessive when confronted with a large volume of images. A protective mechanism was established to reduce this risk by organizing images into batches of 150 within a list object in R. These batches are subsequently accessed through a loop created by the for() function. Depending on the computational capabilities at the researcher’s disposal, the batch size accessed during each iteration can be adjusted accordingly.

Ultimately, after diverse variable sets were extracted from the elements depicted in the images, they could be archived in Microsoft Excel spreadsheet format, denoted by the extension ".xlsx" in the "Sheets" folder.

```
# Creating the script file for variable extraction ------------------------

# Enter the line below in the RStudio Console to create the script file.
usethis::use_r("VarExtraction")


# Packages for usage ------------------------------------------------------

library("pliman")
library("tidyverse")
library("writexl")
```

```
# Files for analysis ----------------------------------------------------

# Specifies the maximum number of images to be analyzed simultaneously.
# The value of n_imgs can be modified based on the user"s
# computational capabilities.
n_imgs <- 150

# Saves the initial data processing date.
# Additionally, it lists and saves the initial set of images for processing.
if (!file.exists("initial_data.rds")) {
  # Saves the initial date of algorithm usage.
  saveRDS(Sys.time(), file = "initial_data.rds")

  # Create a vector containing the names of all images within the
  # ImagesProj directory.
  img_files <- list.files("ImagesProj")

  # Create a list that limits the number of images analyzed at a time.
  # The user can adjust the number of images (n_imgs) as needed.
  split_file_list <-
    split(img_files, ((seq_along(img_files) - 1) %/% n_imgs) + 1)

  # Save the list for future use.
  saveRDS(split_file_list, file = "split_file_list.rds")

}


# Determine the difference between the current list of images and
# the previous list. Then, append the new images to the end of the
# previous list
if (Sys.time() > readRDS("initial_data.rds")) {
  # Read the previously saved list of images.
  split_file_list <- readRDS("split_file_list.rds")

  # Retrieve a vector of the listed images
  last_imgs <- as.vector(unlist(split_file_list))

  # Retrieve the length of the previous list of images.
  l <- length(split_file_list)

  # Get the distinction between the current images in the ImagesProj
  # folder and the previous image list.
  img_diff <- setdiff(list.files("ImagesProj"), last_imgs)

  # Create a list that limits the number of images analyzed at a time.
  # The user can adjust the number of images (n_imgs) as needed.
  split_file_diff <-
    split(img_diff, ((seq_along(img_diff) - 1) %/% n_imgs) + 1 + l)

  # Concatenate the previous list with the new one obtained by
  # calculating the list difference
  split_file_list <- c(split_file_list, split_file_diff)

  # Save the list for future use.
  saveRDS(split_file_list, file = "split_file_list.rds")

}


# Acquisition of variable sets ------------------------------------------

# The project directory required internally for the routine established
# by the for() function.
project_path <- rstudioapi::getActiveProject()
```

```
# Create the "Features" directory to store the objects saved within
# the for() loop routine.
if (!dir.exists("Features")) {
  dir.create("Features")
}

# Available vegetative indexes
# Only "RGB", "HSB", and "CIELab" band indexes segment three-band images
# Reads VIs that can be used to segment three-band images
indexes_for_use <- readRDS("indexes_for_use.rds")

# Padding width of the numerical sequence to name the
# result files in order within the for() routine.
pad_width <- nchar(max(as.numeric(names(split_file_list)))) + 1

# Saves the padding width to be used also
# in VarExtractionFailedObj.R script
saveRDS(pad_width, "pad_width.rds")

# Routine for obtaining sets of variables.
# Adjust the "split_file_list" object as necessary.
for (i in seq_len(length(split_file_list))) {

  # Item from the running list
  list_item <- names(split_file_list[i])

  # This condition controls the creation of objects from the
  # analyse_objects() function. When i ≥ 1, all items in the
  # "split_file_list" will be processed. If i > 1 (as long as it is ≤ the
  # size of the list), only a subset of the items will be processed.
  # Hence, if your processing unexpectedly halts (e.g., due to a power
  # outage), you can review the processed items on the list and resume
  # from where you left off, ensuring continuity in processing the
  # remaining items.
  if (i >= 1) {

    # Creates a folder for temporarily copying n images (as defined in the
    # n_imgs object above) to be processed at once.
    dir_ <- "ImagesProj/tmp"
    dir.create(dir_)

    # Copies the n images to the "tmp" folder.
    # Modify the "split_file_list" object as needed.
    p <- paste0(project_path, "/ImagesProj/", split_file_list[[i]])
    file.copy(p, dir_)


    # Function for extracting variable sets.
    analyze_objects(
      # When pattern = "", all images currently in the directory dir_
      # will be analyzed.
      pattern = "",
      # Directory from which the images will be analyzed.
      dir_original = dir_,
      # Directory where the processed images will be saved.
      dir_processed = "Proc_ImagesProj",
      # Determines if the processed image has a reference.
      reference = TRUE,
      # Reference area.
      reference_area = 18,
      # Index for removing only the background from the images.
      # In this work, it was determined with invert = FALSE.
      back_fore_index = "SCI",
      # Index is used to remove the background and the studied objects.
      # In this work, it was determined with invert = TRUE.
      fore_ref_index = "HUE2",
      # When objects in the image are not closely spaced, it"s preferable
      # to set watershed = FALSE. However, if they are very close or
```

```
      # slightly overlapping, it"s advisable to switch it to TRUE.
      watershed = FALSE,
      # Adds identification to the objects under study.
      marker = "id",
      # Size of the marker in the objects under study.
      marker_size = 4,
      # Filter used in image processing.
      filter = 4,
      # Filling in holes within the objects being studied. Specifying
      # fill_hull = TRUE will account for internal holes, possibly
      # created by insects in the total area of the leaflets.
      fill_hull = TRUE,
      # Executes image processing in parallel, utilizing multiple
      # processor cores simultaneously.
      parallel = TRUE,
      # Do not display the processed image (the default setting of the
      # function is TRUE).
      plot = FALSE,
      # Saves the processed images. This enables verification of the
      # image processing quality later on.
      save_image = TRUE,
      # Vector to specify the inversion of indexes. In this work, FALSE
      # was set for the "SCI" index and TRUE for the "HUE2" index.
      invert = c(FALSE, TRUE),
      # Color of the background of the processed and saved image.
      col_background = "white",
      # Outline the thickness of the objects under study in the processed
      # and saved images.
      contour_size = 2,
      # Vegetative indexes to be used as one of the variable sets.
      # This parameter of the function is defined as a vector of VIs.
      object_index = indexes_for_use,
      # Specify if texture characteristics will be computed.
      haralick = TRUE,
      # Specifies whether Fourier coefficients will be extracted.
      efourier = TRUE,
      # Defines the number of harmonics for the Fourier coefficients.
      # Objects with deeper contours need a greater number of harmonics
      nharm = 10
    ) -> features

    # Saves the "features" object in the Features folder
    saveRDS(features, file = paste0(
      "Features/features_",
      str_pad(list_item, width = pad_width, pad = "0"),
      ".rds"
    ))

    # Removes the last object created by the analyse_objects() function.
    rm("features")

    # Removes the "tmp" directory.
    unlink(dir_, recursive = TRUE)

    # Cleans up the system memory during the routine created
    # by the for() function.
    gc()

  }
}


# Combining the files of all processed images ------------------------------

str_c("Features/features_",
      str_pad(
        names(split_file_list),
        width = pad_width,
```

```
        pad = "0"),
      ".rds") |>
  map(readRDS) ->
  all_features

# Display the objects within the "all_features" list
all_features |>
  map(names)


# Images that failed to detect the number of objects under study ----------

# Lists the images where there was a failure to detect the number of
# objects under study. In this work, the failure occurred when a number
# other than nine leaflets were identified.

# Creates a list that can substitute the "split_file_list" in the routine
# created using the "for()" function.

failed_imgs_list <- list()

all_features |>
  map("count") |>
  bind_rows() |>
  filter(Objects != 9) ->
  failed_imgs

failed_imgs


# Saves a table named "failed_imgs", containing the identification of
# images with errors in object counting.
saveRDS(failed_imgs, "failed_imgs.rds")


# Counts the number of images that had failures.
failed_imgs$Image |> length()


# Stores the names of the images that failed for re-inclusion in the
# routine created by the "for()" function.
if (!is_empty(failed_imgs$Image)) {
  failed_imgs$Image |>
    sort() |>
    paste0(".jpg") ->
    failed_imgs_list[["0"]]

  print(failed_imgs_list)

}


# Saves the list of images with failures.
if (!is_empty(failed_imgs$Image)) {
  saveRDS(failed_imgs_list, "failed_imgs_list.rds")
}


# Shape and texture features of images without failures -------------------

all_features |>
  map_dfr("results") |>
  filter(!img %in% c(failed_imgs$Image)) |>
  as_tibble() |>
  mutate(img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
         id = as_factor(id)) |>
  arrange(img) ->
  results_imgs
```

```
results_imgs


# Descriptive statistics of the area of images without failures ----------

all_features |>
  map_dfr("statistics") |>
  pivot_wider(values_from = value, names_from = stat) |>
  filter(!id %in% c(failed_imgs$Image)) |>
  as_tibble() |>
  mutate(id = fct_relevel(id, \(x) str_sort(x, numeric = T))) |>
  arrange(id) ->
  statistics_imgs

statistics_imgs


# Mean of the VIs for each object being studied within each image ---------

all_features |>
  map_dfr("object_index") |>
  filter(!img %in% c(failed_imgs$Image)) |>
  as_tibble() |>
  mutate(img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
         id = as_factor(id)) |>
  arrange(img) ->
  object_index_imgs

object_index_imgs


# Fourier coefficients for each object within each image ------------------

all_features |>
  map_dfr("efourier") |>
  filter(!img %in% c(failed_imgs$Image)) |>
  as_tibble() |>
  mutate(img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
         id = as_factor(id)) |>
  arrange(img) ->
  efourier_imgs

efourier_imgs


# Normalized Fourier coefficients for each object within each image -------

all_features |>
  map_dfr("efourier_norm") |>
  filter(!img %in% c(failed_imgs$Image)) |>
  as_tibble() |>
  mutate(img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
         id = as_factor(id)) |>
  arrange(img) ->
  efourier_norm_imgs

efourier_norm_imgs


# Error between original and reconstructed contours by DEF ----------------

all_features |>
  map_dfr("efourier_error") |>
  filter(!img %in% c(failed_imgs$Image)) |>
  as_tibble() |>
  mutate(img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
         id = as_factor(id)) |>
```

```
  arrange(img) ->
  efourier_error_imgs

efourier_error_imgs


# Minimum harmonics ------------------------------------------------------

all_features |>
  map_dfr("efourier_minharm") |>
  filter(!img %in% c(failed_imgs$Image)) |>
  as_tibble() |>
  mutate(img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
         id = as_factor(id)) |>
  arrange(img) ->
  efourier_minharm_imgs

efourier_minharm_imgs


# Saving the variable sets -----------------------------------------------

# Lists the various sets of variables that can be extracted from
# the analyse_objects() function.
set_var <- list(
  "results_imgs",
  "statistics_imgs",
  "object_index_imgs",
  "efourier_imgs",
  "efourier_norm_imgs",
  "efourier_error_imgs",
  "efourier_minharm_imgs"
)


# Creates the "Sheets" folder within the project directory to save
# Excel spreadsheets
if (!dir.exists("Sheets")) {
  dir.create("Sheets")
}


# Saves the variable sets in tables with the Excel extension "xlsx".
set_var |>
  map(\(x) write_xlsx(get(x), path = paste0("Sheets/", x, ".xlsx")))
```

Notes: It is important to note that some VIs in the Excel spreadsheet “object_index_imgs.xlsx” remain uncomputed due to potential indeterminacies arising from the mathematical relationships between the RGB band values, depending on the color sets in the images. Additionally, some VIs may have infinite values.

Extraction of Variable sets from Images where Object Counting Failed

The algorithm of this section permits reprocessing images that encountered failures in quantifying the objects under study (e.g., leaflets). Given the predetermined count of objects within the image and the imperative of processing each comprehensively, a higher count may imply the existence of background artifacts, such as dirt and stains, resulting in inaccuracies in

the extracted measurements. Conversely, a lower count may suggest object overlap, extreme proximity, or inadequacies in segmentation by the selected VI.

In the algorithm of the previous section, a filtering mechanism is applied to identify images that encountered failures in counting the objects, segregating their names into a distinct list named “failed_imgs_list". Subsequent reprocessing of these images requires adjustment of specific parameters within the “analyse_objects()” function for variable set extraction. For instance, alterations in VIs may be warranted in cases of segmentation failure, while setting watershed = TRUE becomes imperative to address instances of object overlap (**Figure 5**). However, it is important to note that setting “watershed = TRUE” is essential for accurately counting leaflets that are attached by petioles.

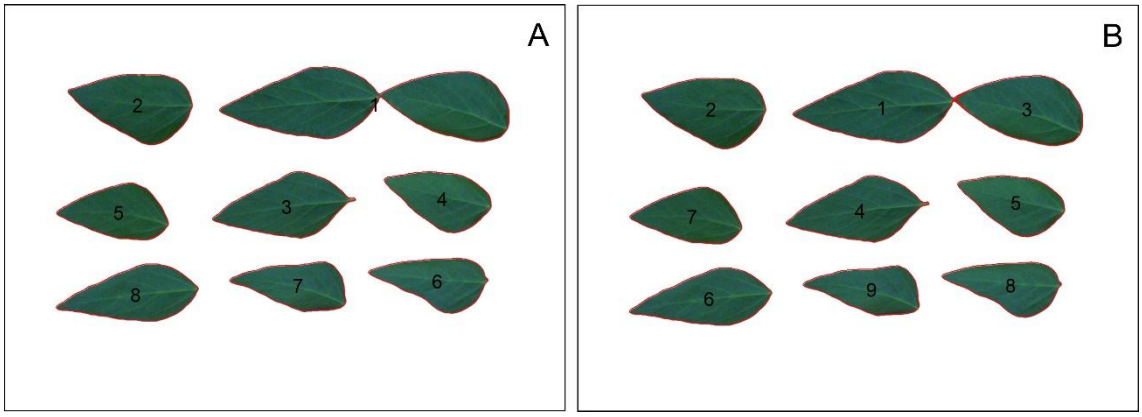

Figure 5. Failure in quantifying the objects under study by the function “analyze_object()”. In A, the parameter “watershed” of the function “analyze_object()” was set to FALSE, and in B, it was set to TRUE.

The output of reprocessed images from this new algorithm will yield an R object designated with the index “00” (i.e., “features_00”) to avert confusion with previously generated objects (i.e., “features_01”, “features_02”, “features_03”, and so forth) devoid of failures in object identification. Naturally, the algorithm below should be executed only in cases where there are failures in counting the objects of interest.

```
# Creating the script file for variable extraction (Failed Obj.) ----------

# Enter the line bellow in the RStudio Console to create the script file.
usethis::use_r("VarExtractionFailedObj")


# Packages for usage -----------------------------------------------------

library("pliman")
```

```
library("tidyverse")
library("writexl")


# Acquisition of variable sets ---------------------------------------------

# The project directory required internally for the routine established
# by the for() function.
project_path <- rstudioapi::getActiveProject()


# Available vegetative indexes
# Only "RGB", "HSB", and "CIELab" band indexes segment three-band images
# Reads VIs that can be used to segment three-band images
indexes_for_use <- readRDS("indexes_for_use.rds")


# Padding width of the numerical sequence to name the
# result files in order within the for() routine.
# Read the padding width
pad_width <- readRDS("pad_width.rds")


# If, after running the VarExtraction.R script, there are images where
# object counting failed, generating the "failed_imgs_list" object.
# Then run a line below and continue to run the script. However,
# you need to modify the arguments in the "analyse_objects()" function
# to improve object identification. In this study, simply setting
# "watershed" to TRUE was sufficient.
failed_imgs_list <- readRDS("failed_imgs_list.rds")


# If, after running the VarExtractionFailedObj.R script, there are images
# where object counting failed, generating the "new_failed_imgs_list"
# object. Then, return to this point, uncomment the line below, and rerun
# the routine with different arguments in the "analyse_objects()" function
# to improve object identification.
# failed_imgs_list <- readRDS("new_failed_imgs_list.rds")


# Routine for obtaining sets of variables.
for (i in seq_len(length(failed_imgs_list))) {
  # Item from the running list
  list_item <- names(failed_imgs_list[i])

  # This condition controls the creation of objects from the
  # analyse_objects() function. When i ≥ 1, all items in the
  # "failed_imgs_list" will be processed. If i > 1 (as long as it is ≤ the
  # size of the list), only a subset of the items will be processed.
  # Hence, if your processing unexpectedly halts (e.g., due to a power
  # outage), you can review the processed items on the list and resume
  # from where you left off, ensuring continuity in processing the
  # remaining items.
  if (i >= 1) {

    # Creates a folder for temporarily copying n images (as defined in the
    # n_imgs object above) to be processed at once.
    dir_ <- "ImagesProj/tmp"
    dir.create(dir_)

    # Copies the n images to the "tmp" folder.
    # Modify the "failed_imgs_list" object as needed.
    p <- paste0(project_path, "/ImagesProj/", failed_imgs_list[[i]])
    file.copy(p, dir_)


    # Function for extracting variable sets.
    analyze_objects(
```

```
  # When pattern = "", all images currently in the directory dir_
  # will be analyzed.
  pattern = "",
  # Directory from which the images will be analyzed.
  dir_original = dir_,
  # Directory where the processed images will be saved.
  dir_processed = "Proc_ImagesProj",
  # Determines if the processed image has a reference.
  reference = TRUE,
  # Reference area.
  reference_area = 18,
  # Index for removing only the background from the images.
  # In this work, it was determined with invert = FALSE.
  back_fore_index = "SCI",
  # Index is used to remove the background and the studied objects.
  # In this work, it was determined with invert = TRUE.
  fore_ref_index = "HUE2",
  # When objects in the image are not closely spaced, it"s preferable
  # to set watershed = FALSE. However, if they are very close or
  # slightly overlapping, it"s advisable to switch it to TRUE.
  watershed = TRUE,
  # Adds identification to the objects under study.
  marker = "id",
  # Size of the marker in the objects under study.
  marker_size = 4,
  # Filter used in image processing.
  filter = 4,
  # Filling in holes within the objects being studied. Specifying
  # fill_hull = TRUE will account for internal holes, possibly
  # created by insects in the total area of the leaflets.
  fill_hull = TRUE,
  # Executes image processing in parallel, utilizing multiple
  # processor cores simultaneously.
  parallel = TRUE,
  # Do not display the processed image (the default setting of the
  # function is TRUE).
  plot = FALSE,
  # Saves the processed images. This enables verification of the
  # image processing quality later on.
  save_image = TRUE,
  # Vector to specify the inversion of indexes. In this work, FALSE
  # was set for the "SCI" index and TRUE for the "HUE2" index.
  invert = c(FALSE, TRUE),
  # Color of the background of the processed and saved image.
  col_background = "white",
  # Outline the thickness of the objects under study in the processed
  # and saved images.
  contour_size = 2,
  # Vegetative indexes to be used as one of the variable sets.
  # This parameter of the function is defined as a vector of VIs.
  object_index = indexes_for_use,
  # Specify if texture characteristics will be computed.
  haralick = TRUE,
  # Specifies whether Fourier coefficients will be extracted.
  efourier = TRUE,
  # Defines the number of harmonics for the Fourier coefficients.
  # Objects with deeper contours need a greater number of harmonics
  nharm = 10
) -> features

# Saves the "features" object in the Features folder
saveRDS(features, file = paste0(
  "Features/features_",
  str_pad(list_item, width = pad_width, pad = "0"),
  ".rds"
))

# Removes the last object created by the analyse_objects() function.
```

```
    rm("features")

    # Removes the "tmp" directory.
    unlink(dir_, recursive = TRUE)

    # Cleans up the system memory during the routine created
    # by the for() function.
    gc()

  }
}


# Images that failed to detect the number of objects under study ----------

# CAUTION! This step should only be done if reprocessing
# images that previously failed to identify the correct number of objects
# under study has succeeded. A new filtering is also performed
# if there is a need to reprocess new failures.

# Retrieves data from images that previously failed.
"Features/features_00.rds" |>
  readRDS() -> features_00


# Images that failed to detect the number of objects under study ----------

# Lists the images where there was a failure in detecting the number of
# objects under study. In this work, the failure occurred when a number
# other than nine leaflets were identified.

# Creates a list that can substitute the "new_failed_imgs_list" in the routine
# created using the "for()" function.

new_failed_imgs_list <- list()


# Lists the images where there was no counting failure.
# In this work, the expected number is 9.
features_00$count |>
  filter(Objects != 9) -> new_failed_imgs

new_failed_imgs


# Saves a table named "new_failed_imgs", containing the identification
# of images with errors in object counting.
saveRDS(new_failed_imgs, "new_failed_imgs.rds")


# Counts the number of images that had failures.
new_failed_imgs$Image |> length()


# Stores the names of the images that failed for re-inclusion in the
# routine created by the "for()" function.
if (!is_empty(new_failed_imgs$Image)) {
  new_failed_imgs$Image |>
    sort() |>
    paste0(".jpg") ->
    new_failed_imgs_list[["0"]]

  print(new_failed_imgs_list)
}


# Saves the list of images with failures.
if (!is_empty(new_failed_imgs$Image)) {
```

```
  saveRDS(new_failed_imgs_list, "new_failed_imgs_list.rds")
}


# Below are the new data obtained after reprocessing the images.
features_00$results |>
  filter(!img %in% new_failed_imgs$Image) |>
  as_tibble() |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  results_remaning_imgs


features_00$statistics |>
  filter(!id %in% new_failed_imgs$Image) |>
  pivot_wider(values_from = value, names_from = stat) |>
  as_tibble() |>
  mutate(id = fct_relevel(id, \(x) str_sort(x, numeric = T))) |>
  arrange(id) ->
  statistics_remaining_imgs


features_00$object_index |>
  filter(!img %in% new_failed_imgs$Image) |>
  as_tibble() |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  object_index_remaning_imgs


features_00$efourier |>
  filter(!img %in% new_failed_imgs$Image) |>
  as_tibble() |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  efourier_remaning_imgs


features_00$efourier_norm |>
  filter(!img %in% new_failed_imgs$Image) |>
  as_tibble() |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  efourier_norm_remaning_imgs


features_00$efourier_error |>
  filter(!img %in% new_failed_imgs$Image) |>
  as_tibble() |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  efourier_error_remaning_imgs


features_00$efourier_minharm |>
  filter(!img %in% new_failed_imgs$Image) |>
  as_tibble() |>
  mutate(
```

```
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  efourier_minharm_remaning_imgs


# New tables with the data from the images that had failed ----------------

# The previously failed images have been reprocessed and are now accurate.

# CAUTION! This step should only be executed if there is a need to add data
# from images that previously failed to identify all objects under study
# but were correctly identified after reprocessing.
results_imgs |>
  bind_rows(results_remaning_imgs) |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  results_imgs


statistics_imgs |>
  bind_rows(statistics_remaining_imgs) |>
  mutate(
    id = fct_relevel(id, \(x) str_sort(x, numeric = T))) |>
  arrange(id) ->
  statistics_imgs


object_index_imgs |>
  bind_rows(object_index_remaning_imgs) |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  object_index_imgs


efourier_imgs |>
  bind_rows(efourier_remaning_imgs) |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  efourier_imgs


efourier_norm_imgs |>
  bind_rows(efourier_norm_remaning_imgs) |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  efourier_norm_imgs


efourier_error_imgs |>
  bind_rows(efourier_error_remaning_imgs) |>
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  efourier_error_imgs


efourier_minharm_imgs |>
  bind_rows(efourier_minharm_remaning_imgs) |>
```

```
  mutate(
    img = fct_relevel(img, \(x) str_sort(x, numeric = T)),
    id = as_factor(id)) |>
  arrange(img) ->
  efourier_minharm_imgs


# Saving the variable sets ------------------------------------------------

# Lists the various sets of variables that can be extracted from
# the analyse_objects() function.
set_var <- list(
  "results_imgs",
  "statistics_imgs",
  "object_index_imgs",
  "efourier_imgs",
  "efourier_norm_imgs",
  "efourier_error_imgs",
  "efourier_minharm_imgs"
)


# Saves the variable sets in tables with the Excel extension "xlsx".
set_var |>
  map(\(x) write_xlsx(get(x), path = paste0("Sheets/", x, ".xlsx")))
```

Notes: In cases where images exhibit excessive objects, possibly due to artifacts, you can locate these images in the "Proc_ImagesProj" folder. If adjustments to the arguments of the "analyze_objects()" function fail to address artifact counting adequately, you may need to modify the image to remove them. However, you can also perform post-processing filtering.

CONCLUSION

This method furnishes guidelines, rationales, and scripts tailored for extracting diverse sets of variables (comprising VIs, shape measures, EFDs, and texture characteristics) from RGB images, leveraging a database of soybean leaflet characterization. These procedures were systematically applied to all available images, demonstrating suitability for the objectives. Furthermore, this method offers versatility for potential employment, refinement, and adaptation across various research scenarios.

REFERENCES

BEGHIN, T. et al. Shape and Texture Based Plant Leaf Classification. 2010, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 345–353.

CORTES, D. F. M. et al. Model-assisted phenotyping by digital images in papaya breeding program. **Scientia Agricola**, v. 74, p. 294–302, 2017.

D'AGOSTINO MCGOWAN, L.; BRYAN, J. **googledrive: An Interface to Google Drive**. [S.l: s.n.]. Disponível em: <https://CRAN.R-project.org/package=googledrive>. , 2023

EKIZ, A.; ARICA, S. Detection of watermelon in RGB images via unmanned aerial vehicle by utilising texture features for predicting yield. **Pakistan Journal of Agricultural Sciences**, v. 59, n. 6, 2022.

FANG, H.; LIANG, S. Leaf Area Index Models. **Reference Module in Earth Systems and Environmental Sciences**, 2014. Acesso em: 13 out. 2022.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural Features for Image Classification. **IEEE Transactions on Systems, Man, and Cybernetics**, v. SMC-3, n. 6, p. 610–621, 1973.

KUHL, F. P.; GIARDINA, C. R. Elliptic Fourier features of a closed contour. **Computer Graphics and Image Processing**, v. 18, n. 3, p. 236–258, 1982. Disponível em: <https://www.sciencedirect.com/science/article/pii/0146664X8290034X>.

MOLLMAN, R.; ÇIFTÇI, A.; EROL, O. Variable leaf shape on short and long shoots: an elliptic Fourier analysis of *Prunus microcarpa* CA Mey. **Brazilian Journal of Botany**, v. 46, n. 1, p. 113–125, 2023.

NETO, J. C. et al. Plant species identification using Elliptic Fourier leaf shape analysis. **Computers and Electronics in Agriculture**, v. 50, n. 2, p. 121–134, 2006. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168169905001560>.

OLIVOTO, T. Lights, camera, pliman! An R package for plant image analysis. **Methods in Ecology and Evolution**, v. 13, n. 4, p. 789–798, 1 abr. 2022. Disponível em: <https://doi.org/10.1111/2041-210X.13803>.

OOMS, J. **writexl: Export Data Frames to Excel "xlsx" Format**. . [S.l: s.n.]. Disponível em: <https://CRAN.R-project.org/package=writexl>. , 2024

POSIT TEAM. **RStudio: Integrated Development Environment for R**. . Boston, MA: [s.n.]. Disponível em: <http://www.posit.co/>. , 2024

PRADEBON, L. C. et al. Selection based on the phenomenic approach and agronomic ideotic of white oat. **Agronomy Journal**, v. 116, n. 3, p. 1275–1289, 1 maio 2024. Disponível em: <https://doi.org/10.1002/agj2.21569>.

RAJU, P. P. C.; BALACHANDER, B.; NEEHARIKA, S. Comparison of Haralick Texture Features and Gray Level Run Length Matrix Features for Analyzing Textural Variation in Cotton Leaves to Identify Spot Disease. 2022, [S.l: s.n.], 2022. p. 1–17.

ROWLAND, S. D. et al. Leaf shape is a predictor of fruit quality and cultivar performance in tomato. **new phytologist**, v. 226, n. 3, p. 851–865, 2020.

TILLETT, R. D. Image analysis for agricultural processes: a review of potential opportunities. **Journal of Agricultural Engineering Research**, v. 50, p. 247–258, 1991. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0021863405800186>.

UNITED STATES DEPARTMENT OF AGRICULTURE. **International Production Assessment Dividion**. Disponível em: <https://ipad.fas.usda.gov/cropexplorer/cropview/commodityView.aspx?cropid=2222000&sel_year=2023&rankby=Production>. Acesso em: 15 maio 2024.

VISCOSI, V.; FORTINI, P. Leaf shape variation and differentiation in three sympatric white oak species revealed by elliptic Fourier analysis. **Nordic Journal of Botany**, v. 29, n. 5, p. 632–640, 1 out. 2011. Disponível em: <https://doi.org/10.1111/j.1756-1051.2011.01098.x>.

WICKHAM, H. et al. **usethis: Automate Package and Project Setup**. . [S.l: s.n.]. Disponível em: <https://CRAN.R-project.org/package=usethis>. , 2024

_____. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.

ZHAO, C. et al. Crop Phenomics: Current Status and Perspectives. **Frontiers in Plant Science**, v. 10, 2019. Disponível em: <https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2019.00714>.

CHAPTER 2

THE IMPORTANCE OF PHENOMIC DATA FOR PATTERN RECOGNITION IN SOYBEAN (*Glycine max* (L.) Merr.) LINES

ABSTRACT

Digital imaging has emerged as a transformative tool in plant characterization within breeding programs, enhancing precision, accuracy, and efficiency in phenotyping processes. Image analysis provides critical insights for interpreting agricultural phenomena, guiding breeding strategies, and optimizing resources. Soybean plants can be effectively characterized through dimensional traits and attributes extracted from images, such as area, perimeter, texture, and vegetative indices. Moreover, Haralick texture features and elliptical Fourier descriptors offer valuable metrics for morphological characterization. This study aimed to extract diverse datasets from images of soybean leaflets, encompassing traits related to shape, texture, Fourier descriptors, and vegetative indices, and to evaluate their effectiveness in discriminating and identifying distinct cultivar patterns in a germplasm bank. Results demonstrated that vegetative indices were the most frequently represented traits, while Fourier descriptors exhibited minimal redundancy. Each dataset revealed unique aspects of genotypic discrimination, identifying five distinct patterns with over 50% agreement. These findings underscore the potential of phenomic analyses to uncover genotypic patterns and identify critical traits for genetic improvement, sparking further curiosity and exploration in this field.

Keywords: vegetative indices, texture traits, shape traits, elliptical Fourier descriptors.

INTRODUCTION

Phenomics systematically studies phenotypes across diverse dimensions of an organism's biology. This field requires interdisciplinary efforts, encompassing molecular, biochemical, cellular, genetic, and morphological analyses (BILDER et al., 2009; HOULE; GOVINDARAJU; OMHOLT, 2010). Technological advancements, particularly in imaging acquisition and processing, have significantly accelerated progress in phenomics by enabling the preservation of analyzed material (WALTER; LIEBISCH; HUND, 2015) and the generation of vast datasets (DE SOUSA et al., 2015).

Image-based phenotyping stands out due to its precision, accuracy, and speed, often surpassing conventional methods susceptible to individual biases and higher costs (COBB et al., 2013; CORTES et al., 2017). Imaging technologies expand phenotypic evaluation to structural levels that were previously difficult or expensive to assess (BILDER et al., 2009; CLARK et al., 2011; HOULE; GOVINDARAJU; OMHOLT, 2010; LENK et al., 2007; MERLOT et al., 2002). Applications range from traditional methods like photography and light microscopy to cutting-edge techniques such as thermography, fluorescence, tomography, and nuclear magnetic resonance imaging (BERGER; PARENT; TESTER, 2010; HOULE; GOVINDARAJU; OMHOLT, 2010; MONTES; PAUL; MELCHINGER, 2007).

In agriculture, imaging technologies are revolutionizing decision-making processes, making them more precise and efficient in this critical domain (LI; ZHANG; HUANG, 2014; OMARI et al., 2020; TILLETT, 1991). Crop breeding, a cornerstone of agricultural development, has increasingly integrated digital imaging into its processes (CORTES et al., 2017; DIPTA et al., 2023). Significant investments have been made in high-throughput phenotyping platforms to facilitate collecting, processing, and storing image data for plant breeding (ARAUS; CAIRNS, 2014; CHAWADE et al., 2019; XU; LI, 2022).

Soybean (*Glycine max* (L.) Merr.), one of the world's most in-demand crops, has greatly benefited from image analysis, particularly in various stages of breeding (ALABI et al., 2022; ANDERSON et al., 2019; DA SILVA JUNIOR et al., 2018). Morphological characterization of vegetative parts of soybeans using image analysis is well-established (GRINBLAT et al., 2016; HE; MA; GUAN, 2022; LARESE et al., 2014; MOMIN et al., 2017). Traits such as leaf shape, texture, and color are not only helpful in distinguishing lines or cultivars but are also linked to agronomic attributes like yield potential, pest and disease resistance, and environmental adaptability (NICOTRA et al., 2008, 2011; REN; WERADUWAGE; SHARKEY, 2019; ROWLAND et al., 2020).

Once measured manually, shape traits like area, perimeter, length, and width of vegetative parts can now be accurately extracted using digital image analysis (OLIVOTO, 2022). Additionally, the shape of plant structures can be mathematically described through equations for two-dimensional contours, such as elliptical Fourier descriptors (EFDs), derived from image data (KUHL; GIARDINA, 1982; NETO et al., 2006; VISCOSI; FORTINI, 2011).

Texture and color, though visually observable, require computational metrics for quantitative analysis. Texture is quantified based on spatial dependencies in grayscale values, as demonstrated by Haralick and colleagues (HARALICK; SHANMUGAM; DINSTEIN, 1973). Colorimetric traits of plants or vegetation can be assessed using vegetative indices (VIs), calculated from image bands (e.g., RGB), which enhance sensitivity to specific traits (FANG; LIANG, 2014).

Given these advancements, this study aims to extract diverse datasets from images of soybean leaflets, focusing on traits related to shape (e.g., area, perimeter), texture, Fourier descriptors, and vegetative indices. Furthermore, it evaluates the ability of these datasets to discriminate and identify distinct patterns among cultivars in a germplasm bank, providing insights for improving genetic selection and breeding strategies.

MATERIALS AND METHODS

Soybean Cultivation

The experiment with soybean lines in the F3 generation was conducted at the experimental field of the Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUÍ), located in Ijuí – RS (Latitude: 28° 23’ 16” S; Longitude: 53° 54’ 53” W). Each line was cultivated in 2 × 0.5 m² plots, organized into augmented blocks with interspersed four control cultivars. Fifteen blocks were established, each containing approximately 100 lines, resulting in 1,449 overall.

Sowing occurred on three dates: lines 1 to 1,000 were planted on October 14, 2022; lines 1,001 to 1,300 on October 17, 2022; and lines 1,301 to 1,449 on October 18, 2022. Fertilization involved pre-application of 100 kg/block of organic fertilizer (Sulfacal), 8 tons of Biogranum (a silicon, calcium, and magnesium supplement), and 6 kg/block of a 3-21-21 (N-P-K) fertilizer applied in two equal doses of 3 kg each.

Image Acquisition and Phenotypic Data Extraction

Phenotypic analyses were based on RGB images of leaflets collected at the R5 growth stage. Between 8:00 AM and 4:00 PM, one leaf from the middle third of three plants was collected from each line. The leaves were transported to the laboratory, detached and arranged on red cardboard in three horizontal rows. A yellow reference marker with 18 cm^2 was included to convert pixel metrics into standard measurement units.

Images were captured with a digital camera positioned 50 cm above the leaflets, configured at a resolution of 1,280 × 1,024 pixels and 96 dpi. All images were systematically named. Due to a ~23.74% cultivation loss, images were taken for 1,105 lines out of the 1,449 initially planted, resulting in 9,945 observations (9 leaflets per line).

Data processing was performed using R version 4.2.3 (https://cran.r-project.org/) integrated with RStudio (https://posit.co/download/rstudio-desktop/). Phenotypic attributes were extracted using the pliman package (OLIVOTO, 2022), while data manipulation was facilitated by the tidyverse package version 2.0.0 (WICKHAM et al., 2019).

Attributes extracted from each leaflet were averaged to generate representative values for each line, reducing the 9,945 individual observations to 1,105 summarized records. A detailed description of the extracted phenotypic attributes is provided in the **Supplementary Table S1**.

Characterization of Phenomic Data

A single image can generate numerous functional attributes in various genetic studies. In this study, the soybean leaflet attributes were grouped into four distinct sets based on the type of information they provided:

1. **General shape characteristics**: area, perimeter, length, etc.;
2. **Texture characteristics**: derived from Haralick descriptors.
3. **Vegetative indices (VIs)**: calculated from RGB image bands;
4. **Elliptical Fourier descriptors**: it mathematically describes contours.

The data sets were evaluated for informational content and efficiency using the following criteria:

- **Information redundancy**: identified by Pearson correlations > 0.95.
- **Information replication**: detected by Pearson correlations equal to 1.
- **Recognition of highly correlated blocks**: visualized through correlation networks.

Discrimination Capacity Using Phenotypic Data

Statistical strategies were employed to assess the phenotypic datasets' genotypic discrimination capacity and evaluate their consistency in differentiating soybean genotypes.

A Principal Component Analysis (PCA) was first performed, a method initially proposed by Pearson (1901) and mathematically formalized by Hotelling (1933). PCA reduces the data's covariance structure by rigidly rotating the coordinate axes to align with the directions of maximum variability. It generates p principal components that are linear combinations of the original attributes, ordered by their variance magnitude.

The PCA allowed the evaluation of:

1. Genotypic differentiation through scatterplots of scores for the first two principal components.
2. Dimensionality reduction by determining the minimal number of components needed.
3. The relative contribution of each attribute to the total variability of each phenotypic set.

Subsequently, the scores from the first two principal components were used in a Procrustes analysis to assess the concordance of dissimilarity patterns across phenotypic datasets. This technique minimizes data discrepancies by applying geometric transformations such as rotation, scaling, and reflection (DRYDEN; MARDIA, 1998; GREY, 1981).

A Mantel test was also conducted to measure the association between dissimilarity matrices generated by each dataset. This test uses Pearson correlation coefficients to compare standardized Euclidean distances between genotypes (BORCARD; LEGENDRE, 2012; LEGENDRE; FORTIN, 2010; LEGENDRE; LEGENDRE, 2012). The standardized Euclidean distances were calculated as follows:

$$D_{i,i'} = \sqrt{\sum_{k=1}^{p} \frac{(x_{ik} - x_{i'k})^2}{s_{kk}}} \quad (1)$$

Where:

- x_{ik}: the value of trait k for genotype i;
- $x_{i'k}$: the value of trait k for genotype i′;
- s_{kk}: variance of trait k;
- p: total number of traits.

Scatterplots of pairwise phenomic distance matrices were generated to infer concordance graphically. Points concentrated in quadrants 1 and 3 indicated

similarity/dissimilarity agreement, while concentration in quadrants 2 and 4 suggested divergences.

Pattern Recognition in the Soybean Germplasm Bank

Pattern recognition was performed using the k-means clustering algorithm for each phenotypic dataset. This method groups observations into k clusters by assigning them to the nearest cluster centroid based on measured attributes. The algorithm iteratively refines the centroids until within-cluster similarity is maximized (BISHOP; NASRABADI, 2006).

This study used the elbow method to determine each dataset's optimal number of clusters (k). This method involved finding the more considerable distance between the perpendicular distances calculated from intermediate points (x, y), the line connecting the first (x_1, y_1) and last (x_n, y_n) points on the elbow graph (KETCHEN; SHOOK, 1996).

The line passing through the points (x_1, y_1) e (x_n, y_n) is represented by:

$$Ax + By + C = 0 \tag{2}$$

Where:

- $A = y_n - y_1$
- $B = x_1 - y_n$
- $C = x_n y_1 - x_1 y_n$

The perpendicular distance d from a point (x, y) to the line defined $Ax + By + C = 0$ is given by:

$$d = \frac{|Ax + By + C|}{\sqrt{A^2 + B^2}} \tag{3}$$

The concordance between phenotypic information sets in cluster formation was assessed after the clustering process. This procedure was done by computing the harmonic mean of the number of clusters combined with the arithmetic mean of the Jaccard indices (JACCARD, 1912). The Jaccard indices were calculated for each pair of clusters formed by the different phenotypic information sets. Thus, the concordance proportion ($P_{\mathbf{XY}}$) between two phenotypic information sets **X** e **Y** was defined as:

$$P_{\mathbf{XY}} = \frac{1}{n + m} 2 \sum_{i=1}^{n} \sum_{j=1}^{m} p_{\mathbf{XY}_{ij}} = \frac{2nm\bar{p}_{\mathbf{XY}}}{n + m} \tag{4}$$

Where:

- n and m: the number of clusters formed by X e Y, respectively;

- $p_{\mathbf{XY}_{ij}} = \frac{N_{\mathbf{X}_i \cap \mathbf{Y}_j}}{N_{\mathbf{X}_i \cup \mathbf{Y}_j}}$: the proportion of the intersection over the union of clusters $\mathbf{X}_i$ e $\mathbf{Y}_j$, commonly referred to as the Jaccard index;
- $\bar{p}_{\mathbf{XY}}$: represents the arithmetic mean of the Jaccard indexes between clusters formed from the attribute sets **X** e **Y**. It is calculated as:

$$\bar{p}_{XY} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{\mathbf{XY}_{ij}} \tag{5}$$

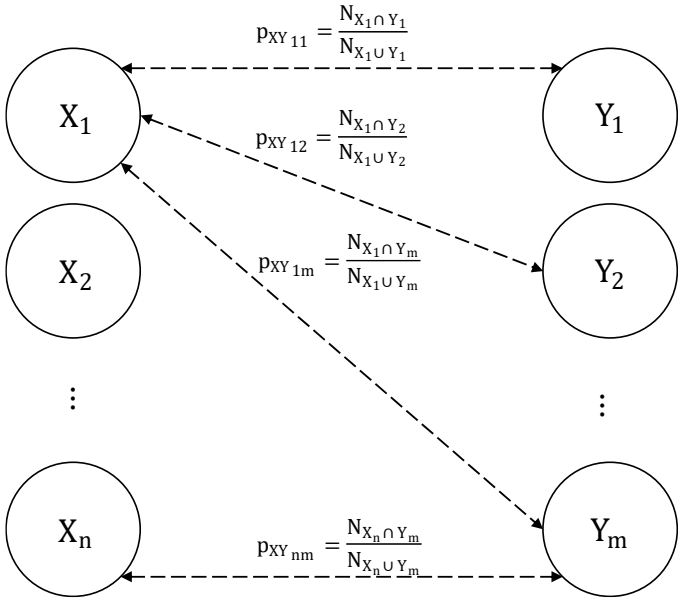

Figure 1. Schematic representation of the calculation of the Jaccard index ($p_{XY_{ij}}$) between the i-th ($i = 1, 2, \ldots, n$) cluster generated from attribute set **X** and the j-th ($j = 1, 2, \ldots, m$) cluster generated from attribute set **Y**. The bidirectional dashed arrow indicates that $p_{XY_{ij}}$ is computed in both directions, from **X** to **Y** and vice versa, resulting in two equivalent calculations.

For complete concordance between the sets in cluster formation, meaning $P_{\mathbf{XY}} = 1$, the following conditions must be met:

1) The sum of the number of clusters in **X** and **Y** must be equal twice the sum of Jaccard indexes calculated between the clusters obtained from **X** and **Y**, i.e.:

$$n + m = 2 \sum_{i=1}^{n} \sum_{j=1}^{m} p_{\mathbf{XY}_{ij}} \tag{6}$$

2) The number of clusters in **X** and **Y** must be equal ($n = m = k$):

$$k = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{\mathbf{XY}_{ij}} \quad (7)$$

3) Each cluster in X must correspond precisely to one cluster in Y ($p_{\mathbf{XY}_{IJ}} = 1$). Thus, k Jaccard indexes will equal to 1 ($p_{\mathbf{XY}_{IJ}} = 1$).

4) There must be no intersection between distinct clusters. Therefore, there will be $k^2 - k$ Jaccard indexes equal to zero ($p_{\mathbf{XY}_{IJ'}} = 0$ for $J' \neq J$).

Some correspondence exists for situations where $0 < P_{\mathbf{XY}} < 1$, but the attribute sets **X** and **Y** are not entirely equivalent.

Concordance between cluster pairs and concordance between phenotypic information sets was visualized using Sankey diagrams.

RESULTS AND DISCUSSION

Phenomic Data Characterization

This study explored the extraction of phenomic information from soybean leaflets, focusing on metrics related to shape, texture, EFDs, and vegetative indices. Quantitatively, the vegetative indices dataset was the most extensive, comprising 43 attributes, followed by the elliptical Fourier descriptors (40 attributes), shape metrics (31 attributes), and texture metrics (13 attributes). In total, 127 attributes were analyzed for pattern recognition.

While the digital image analysis yielded a relatively high number of attributes, assessing redundancy within each phenomic dataset is essential to evaluate their effectiveness in distinguishing among the evaluated units. The average attribute values of the leaflets for each soybean lineage represented these units. **Figure 2** reveals that the shape dataset exhibited the highest internal redundancy, with many attribute pairs showing strong correlations (depicted by intensely red tones). The attributes used to construct the correlation plots (**Figure 2**) are listed in Supplementary Table S1, arranged in the same order as displayed on the x-axis (left to right) and y-axis (bottom to top). Detailed descriptions of each attribute are available in the documentation for the analyze_objects() function from the pliman package (OLIVOTO, 2022).

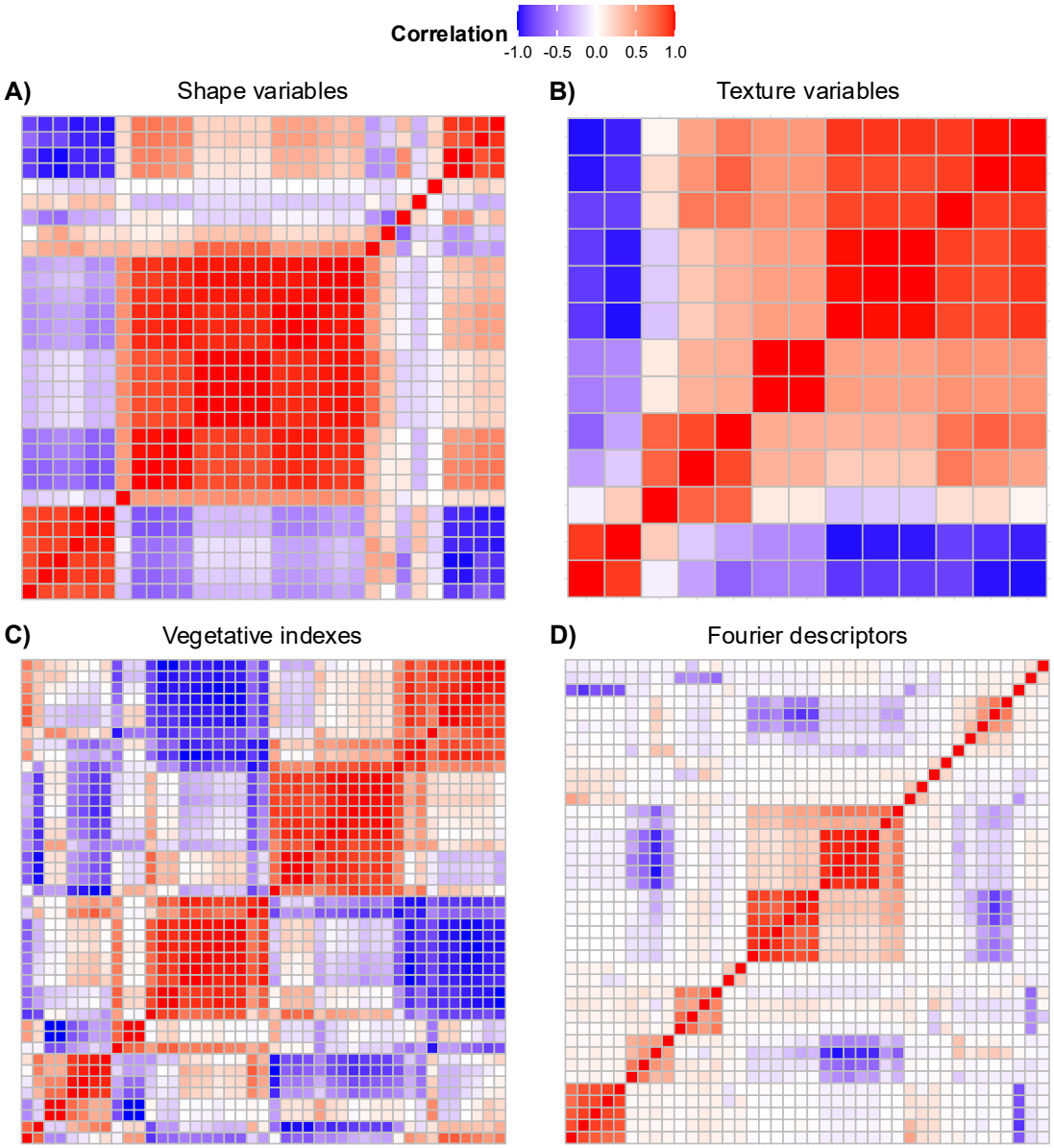

Figure 2. Colorimetric representation of Pearson correlations derived from attributes related to shape, texture, vegetative indices, and Fourier descriptors of soybean lines (*Glycine max* (L.) Merr.).

Specifically, 69 of the 465 calculated correlations in this dataset had Pearson correlation coefficients greater than 0.95 and were classified as redundant (**Table 1**). Furthermore, some shape attributes exhibited perfect correlations (Pearson = 1), such as those between average diameter and average radius and between maximum diameter and maximum radius. Despite this redundancy, size-, area-, and perimeter-related attributes offer practical criteria for differentiation, as they are easy for breeders to perceive and measure.

Conversely, the elliptical Fourier descriptors dataset showed the lowest inter-attribute associations (**Figure 2, Table 1**), suggesting that each attribute provides unique information not represented by others within the same set. Although Fourier morphological description generates four coefficients per harmonic, which may not be entirely independent (Haines et al. 2000), the limited number of highly correlated attributes observed here can be attributed to the use of average leaflet coefficients to represent each lineage in the database.

Table 1. Number and percentage of redundant attributes within each set of phenomic information. Redundant attributes were identified by Pearson correlations exceeding 0.95.

Set of features	Total of features	Total of correlations	Redundant correlations (n) *	Redundant correlations (%) *
Shape	31	465	69	14.84
Texture	13	78	5	6.41
Fourier	40	780	10	1.28
Vegetative indexes	43	903	61	6.76

* The number of replicates is also accounted for and is defined by a Pearson correlation of 1.

Genotypic Discrimination Capability of Phenomic Data

Beyond characterizing phenomic datasets and identifying redundant attributes, genotypic discrimination analysis was performed using multivariate principal component analysis (PCA). A significant cluster of lineages with high similarity patterns was concentrated near the origin of PCA scatter plots. At the same time, some dispersed points indicated distinct lineages warranting further investigation into agronomic traits (**Figure 3**).

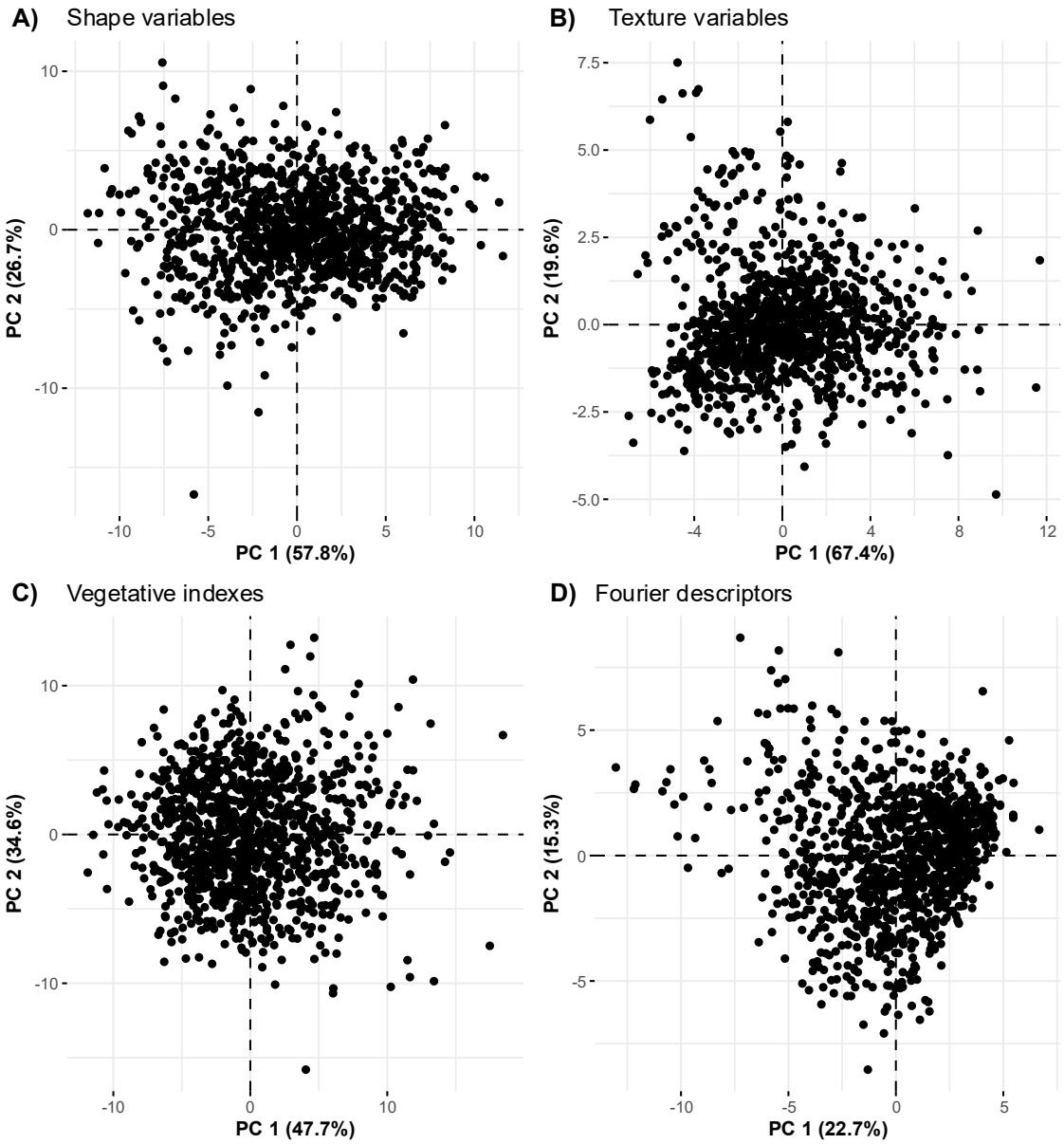

Figure 3. Scatter plot of scores from the first two principal components obtained from different phenomic information sets (Shape, Texture, Fourier, and Vegetative Indices) extracted from digital images of soybean leaflets (*Glycine max* (L.) Merr.).

PCA is advantageous for summarizing information into a few components when redundancy exists, enabling structural simplification. **Figure 4** confirms the redundancy levels observed in **Figure 2**, with approximately 80% (or more) of the variance explained by two principal components for all datasets except Fourier descriptors. This lack of redundancy highlights the unique potential of Fourier descriptors in pattern recognition or discrimination processes, though caution is advised when structurally simplifying this dataset.

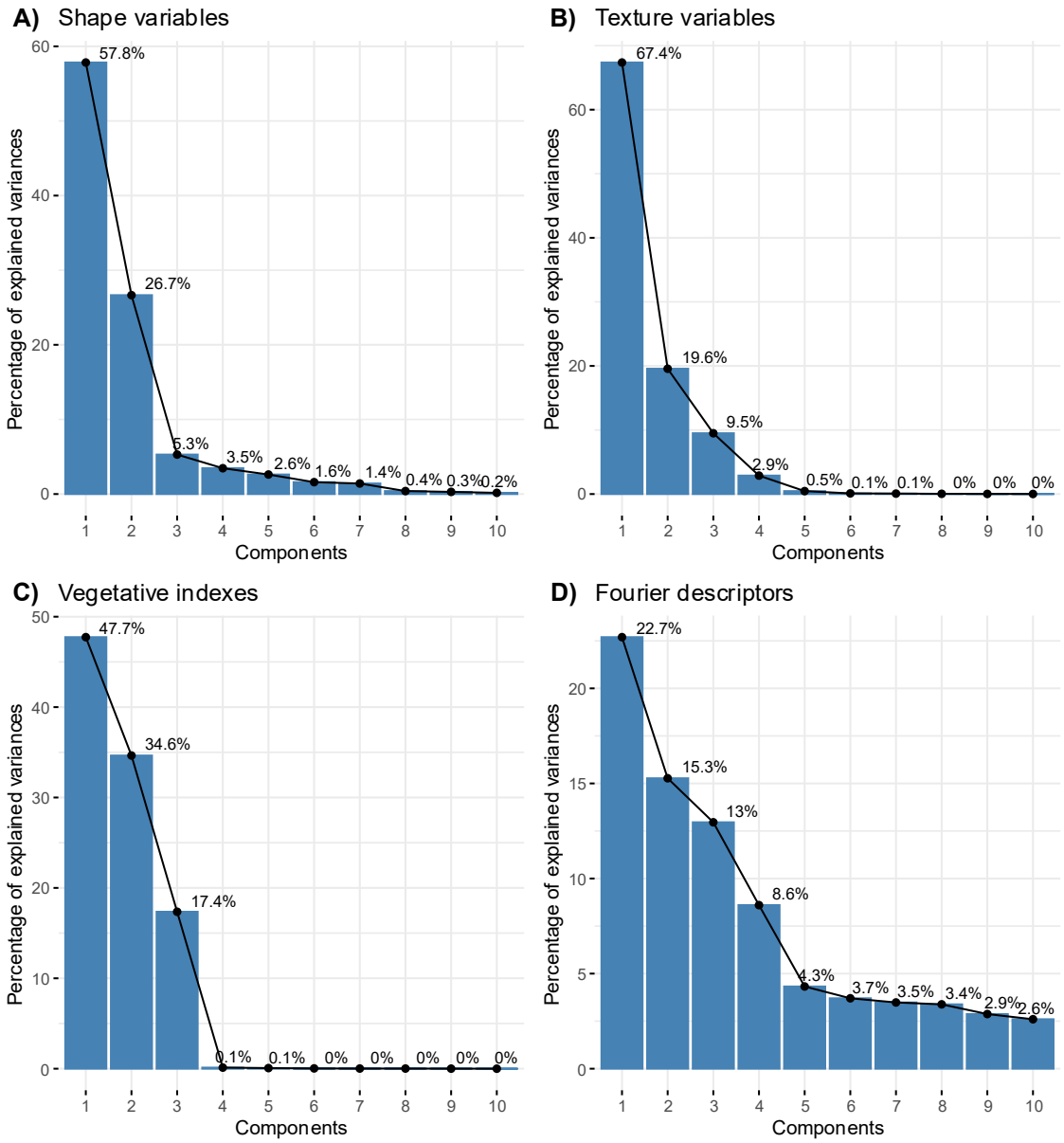

Figure 4. Percentage of variance explained by the first ten principal components derived from different phenomic information sets (Shape, Texture, Fourier, and Vegetative Indices) extracted from digital images of soybean leaflets (*Glycine max* (L.) Merr.).

Broadly, **Figure 5** displays attribute importance measures from each phenomic dataset concerning the first two principal components. Attributes with positive correlations cluster together (forming angles < 90°), while negatively correlated attributes are on opposite sides of the origin (> 90°). A color gradient at the top of each graph indicates the contribution of each attribute, with attributes close to the origin deemed less critical (KASSAMBARA, 2017). A notable number of highly correlated attributes were observed within the shape and vegetative index datasets (**Figure 5**). Nevertheless, only a few attributes in each dataset showed minimal

importance (blue tones) for forming the first two principal components, which captured the most data variance (**Figure 4**).

Dimensionality reduction using PCA for phenomic data from soybean leaflets proves valuable for statistical studies, particularly where high dimensionality presents challenges. Identifying key and less critical attributes is crucial in breeding studies, where agronomic traits differ in measurement difficulty due to labor, timing, environmental effects, and costs. Discarding redundant or invariant attributes could reduce costs and optimize resources.

In this study, image analysis depended solely on computational tools capable of extracting maximum information from refined images. Even if attribute discarding is unnecessary, **Figure 5** identifies attributes of greater interest for breeders to associate with other agronomic, morphological, or physiological traits.

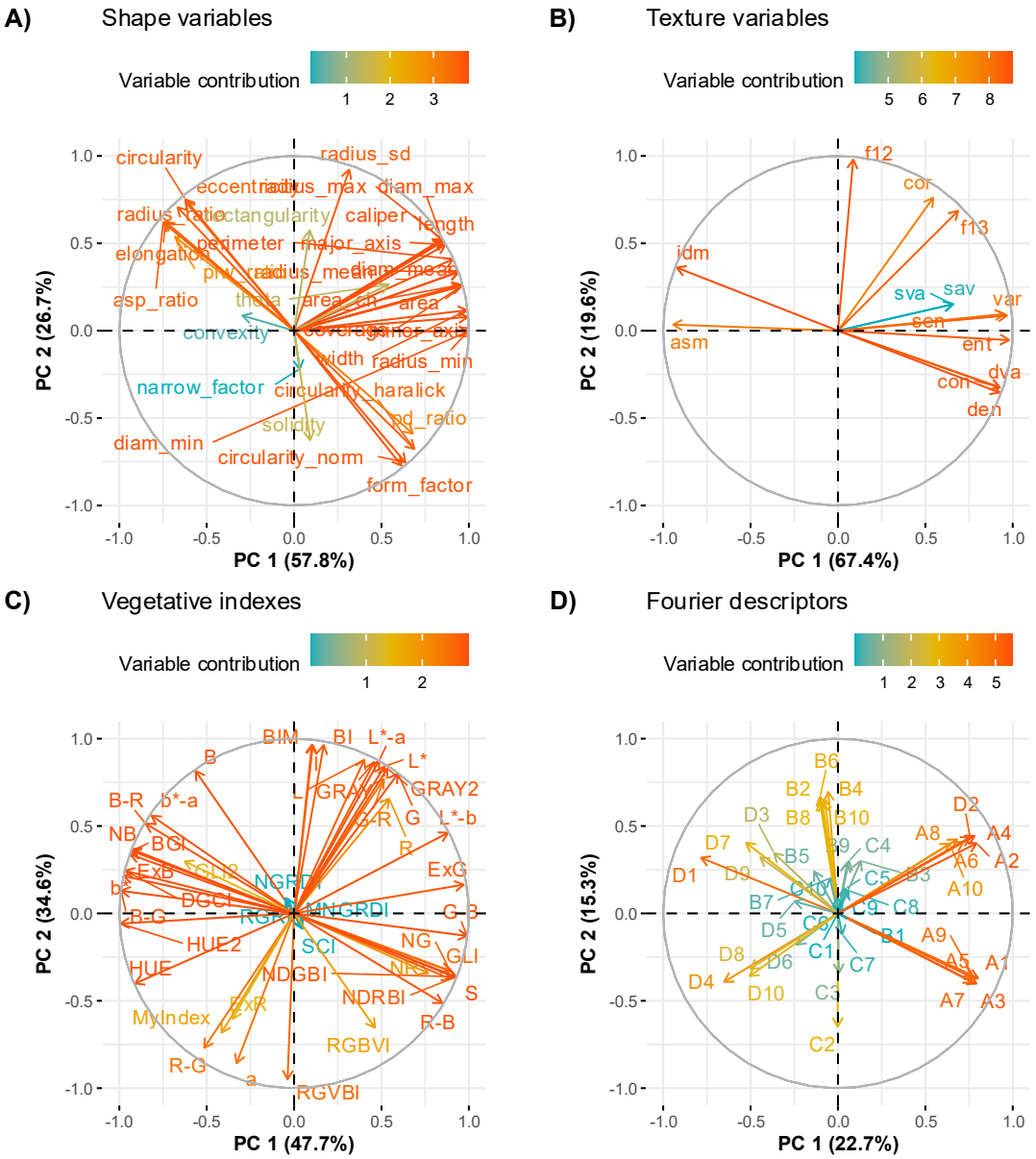

Figure 5. Correlation, representation quality, and contribution of attributes to the first principal components obtained through PCA performed on different phenomic information sets extracted from digital images of soybean leaflets (*Glycine max* (L.) Merr.).

The scores of the first two PCA components also allow graphical agreement verification between two datasets using Procrustes analysis (**Table 2**). The shape and vegetative index datasets showed the highest graphical similarity, followed by texture and vegetative indices, demonstrating significant concordance between their respective similarities and dissimilarities (**Table 2**).

Table 2. Results of the Procrustes analysis applied to the scores of the first two principal components for each pair of attribute sets (Shape, Texture, Fourier, and Vegetative Indices) extracted from digital images of soybean leaflets (*Glycine max* (L.) Merr.).

Statistics	**Pairs of feature groups**					
	Shape variables	*Shape variables*	*Shape variables*	*Texture variables*	*Texture variables*	*Fourier descriptors*
	Texture variables	*Vegetative indexes*	*Fourier descriptors*	*Vegetative indexes*	*Fourier descriptors*	*Vegetative indexes*
Sum of squares	0.9020	0.7783	0.8960	0.8143	0.9629	0.9508
Correlation	0.3130	0.4709	0.3225	0.4309	0.1926	0.2218
Significance	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010

Alternatively, the concordance among different information sets regarding genotypic dissimilarity can be assessed without structurally simplifying the data. Thus, dissimilarity matrices were generated for each phenomic dataset using Euclidean distances, and their pairwise correlations were tested with Mantel tests (**Table 3**). Consistent with Procrustes' results, except for correlation magnitude order, the highest agreement pairs were texture versus vegetative indices and shape versus vegetative indices. Notably, although **Table 3** shows significant correlations for all pairs of distance matrices, these significant results are influenced by the high degrees of freedom, as the dataset includes information from 1,105 soybean lines, resulting in 609,960 dissimilarity measures.

Table 3. Results of the Mantel test, which was applied between pairs of dissimilarity matrices derived from attribute sets (Shape, Texture, Fourier, and Vegetative Indices) extracted from digital images of soybean leaflets (*Glycine max* (L.) Merr.).

Statistics	**Pairs of feature groups**					
	Shape variables	*Shape variables*	*Shape variables*	*Texture variables*	*Texture variables*	*Fourier descriptors*
	Texture variables	*Vegetative indexes*	*Fourier descriptors*	*Vegetative indexes*	*Fourier descriptors*	*Vegetative indexes*
Mantel statistic r	0.2093	0.3091	0.2414	0.3996	0.0563	0.0487
Significance	0.0001	0.0001	0.0001	0.0001	0.0015	0.0041

A graphical inference can also be made using dissimilarity measures regarding the degree of concordance between pairs of phenomic information sets. In this context, the less elliptical distribution of points, predominance in quadrants 1 and 3 (counterclockwise order) observed in **Figure 6**, indicates a lower information overlap among the generated datasets.

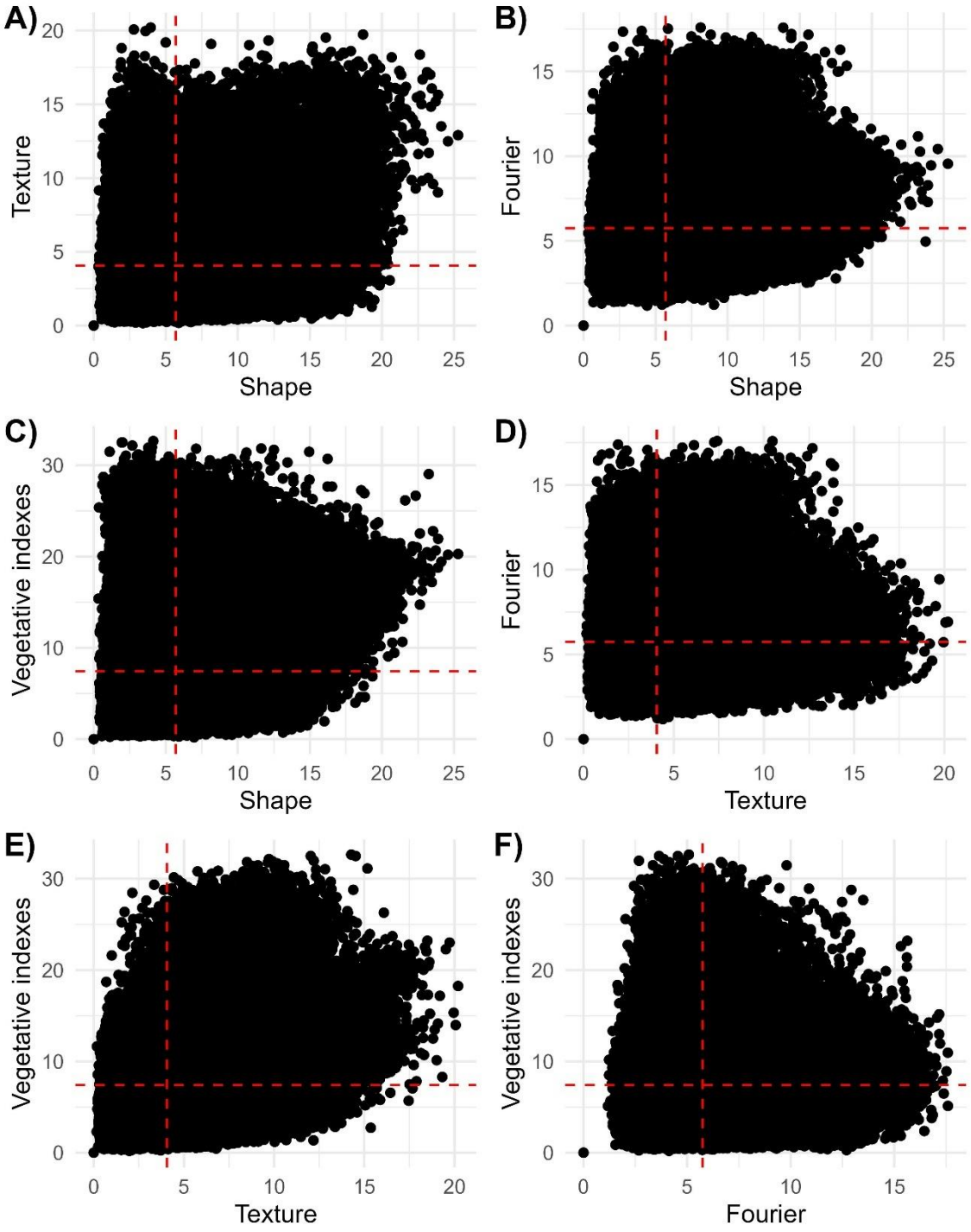

Figure 6. Relationship between dissimilarity distances for pairs of attribute sets (Shape, Texture, Fourier, and Vegetative Indices) extracted from digital images of soybean leaflets (*Glycine max* (L.) Merr.). Red dashed horizontal and vertical lines indicate the medians of dissimilarity distances for each phenomic information set.

Soybean Germplasm Pattern Recognition Using Phenomic Data

Despite their redundancy and genotypic discrimination capabilities, the phenomic datasets also have great potential for genotype pattern recognition – a critical approach in plant breeding. The goal is to identify homogeneous groups without predefined target responses (BISHOP; NASRABADI, 2006). In this study, K-means clustering identified five optimal clusters (**Figure 7**) across all datasets. Despite the coincidence, the number of clusters was mathematically determined by the most significant perpendicular distance from the intermediate (x, y) points to the line connecting the extreme points on the Elbow curve.

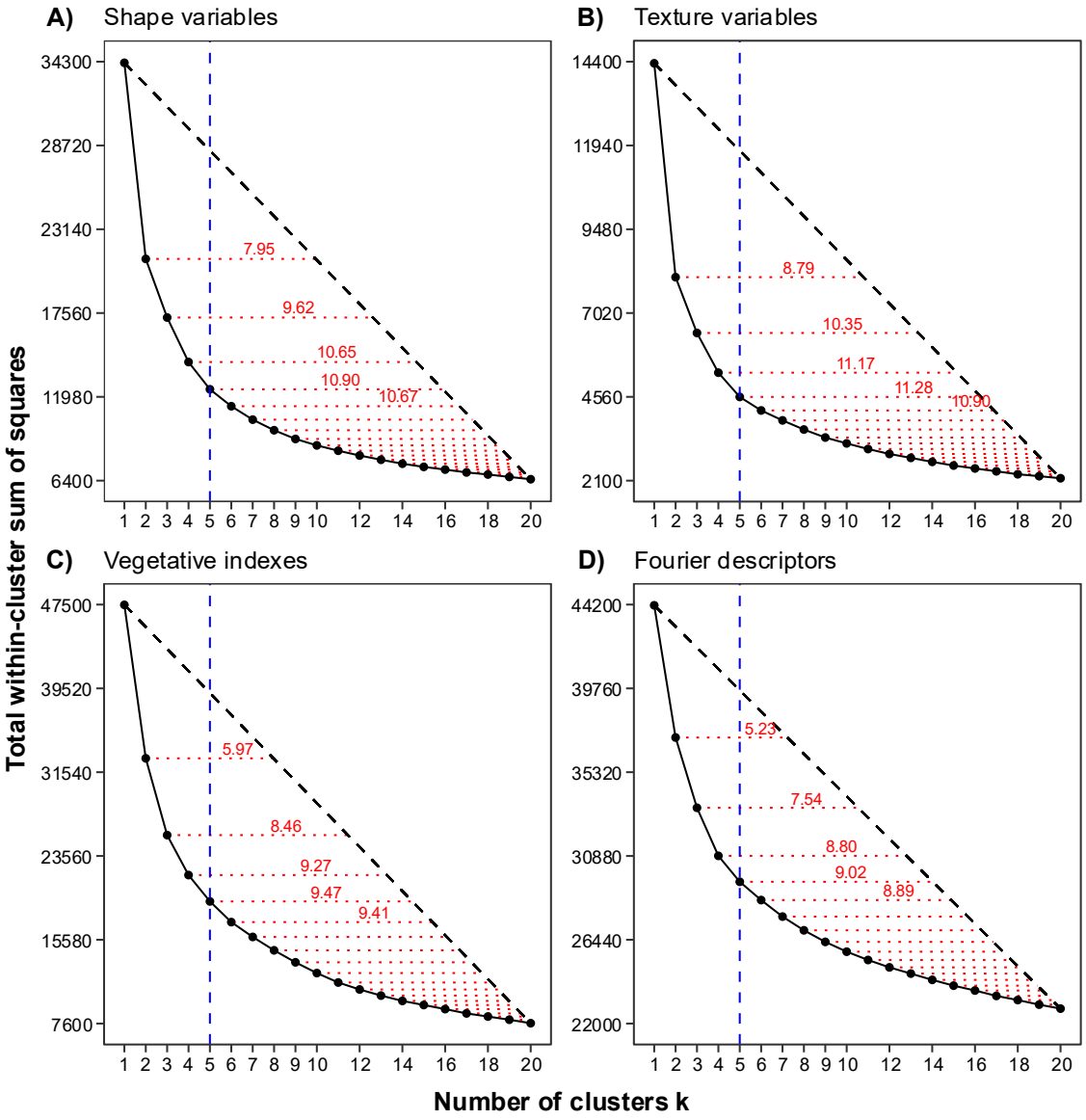

Figure 7. The elbow method determines the number of clusters used in the k-means algorithm. Red dashed lines represent the perpendicular distances from each intermediate point (x, y) to the black dashed line connecting the first (x_1, y_1) and last (x_{20}, y_{20}) points on the curve. The

vertical blue dashed line indicates the clusters associated with the most considerable perpendicular distance. **Note**: Red dashed lines are visually represented horizontally due to the compression of the y-axis relative to the x-axis in the graphs.

Cluster visualization (**Figure 8**) showed well-defined groupings for Shape and Texture, while Fourier and vegetative indices exhibited more significant overlap. **Table 4** corroborates these observations.

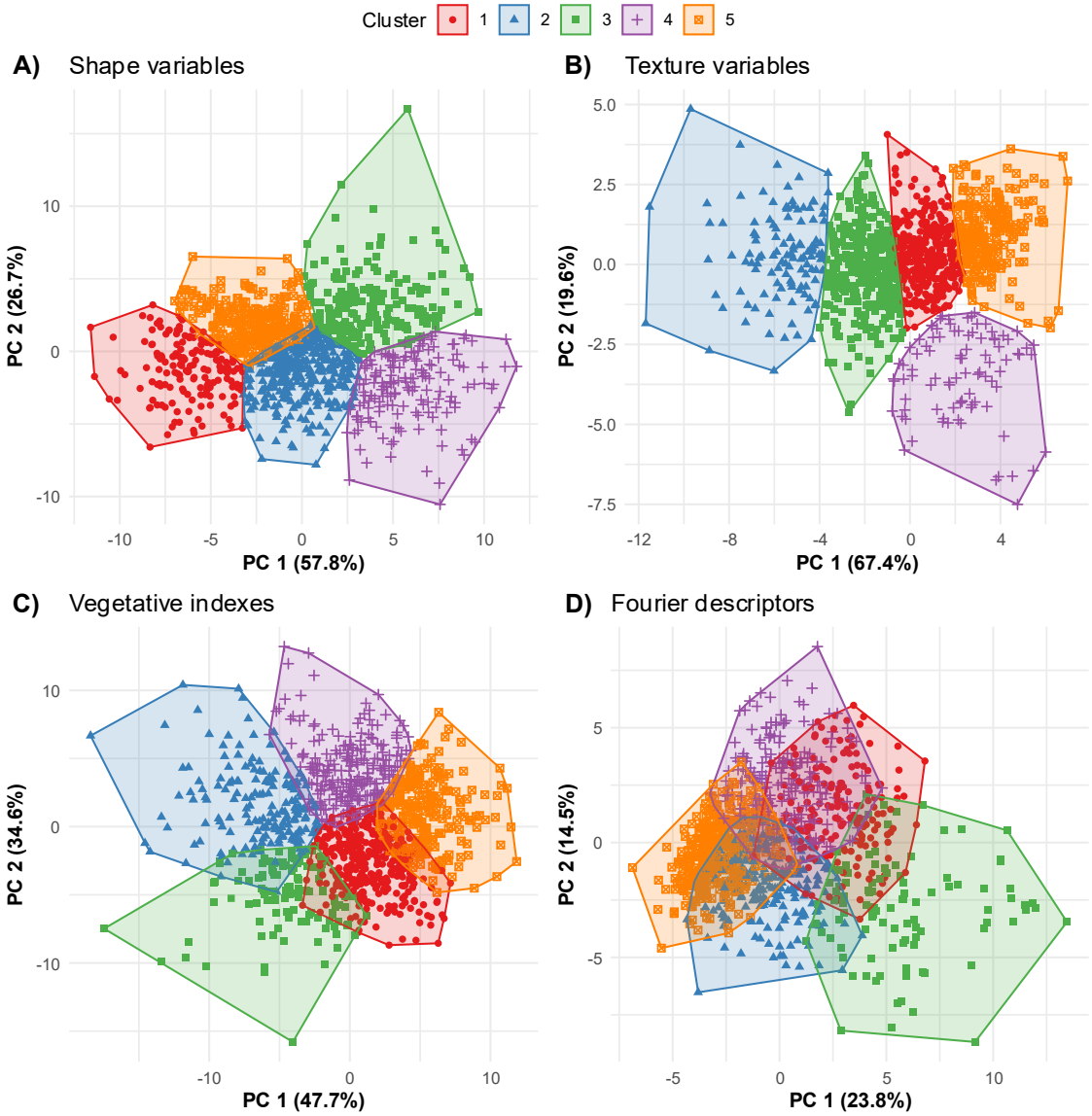

Figure 8. Clusters formed by the k-means algorithm applied to each attribute set (Shape, Texture, Fourier, and Vegetative Indices) extracted from digital images of soybean leaflets (*Glycine max* (L.) Merr.).

Table 4. Summary of k-means clustering results for each attribute set (Shape, Texture, Fourier, and Vegetative Indices) extracted from digital images of soybean leaflets (*Glycine max* (L.) Mer.r).

Set of features	SS between groups	SS within groups	Total SS	SS between groups / Total SS (%)
Shape variables	21742.15	12481.85	34224	63.53
Texture variables	9794.44	4557.56	14352	68.24
Vegetative indexes	28240.58	19231.42	47472	59.49
Fourier descriptors	14647.72	29512.28	4416	33.17

SS means the sum of squares.

Cluster sizes (**Table 5**) and Jaccard indexes (**Figure 9, Table 6**) further revealed that shape and vegetative indices had both the highest (47.00%) and lowest (0.00%) concordance values, aligning with earlier Mantel test findings.

Table 5. The number of soybean lines (*Glycine max* (L.) Merr.) assigned to each cluster identified by the k-means algorithm applied to different phenomic information sets.

Set of features	Clusters*					Total
	1	2	3	4	5	
Shape variables	167	325	200	164	249	1105
Texture variables	351	118	314	91	231	1105
Vegetative indexes	338	185	116	239	227	1105
Fourier descriptors	181	229	104	246	345	1105

* Values highlighted in red and blue indicate the largest and smallest number of lines assigned to a cluster using the k-means procedure for each set of phenomic information, respectively.

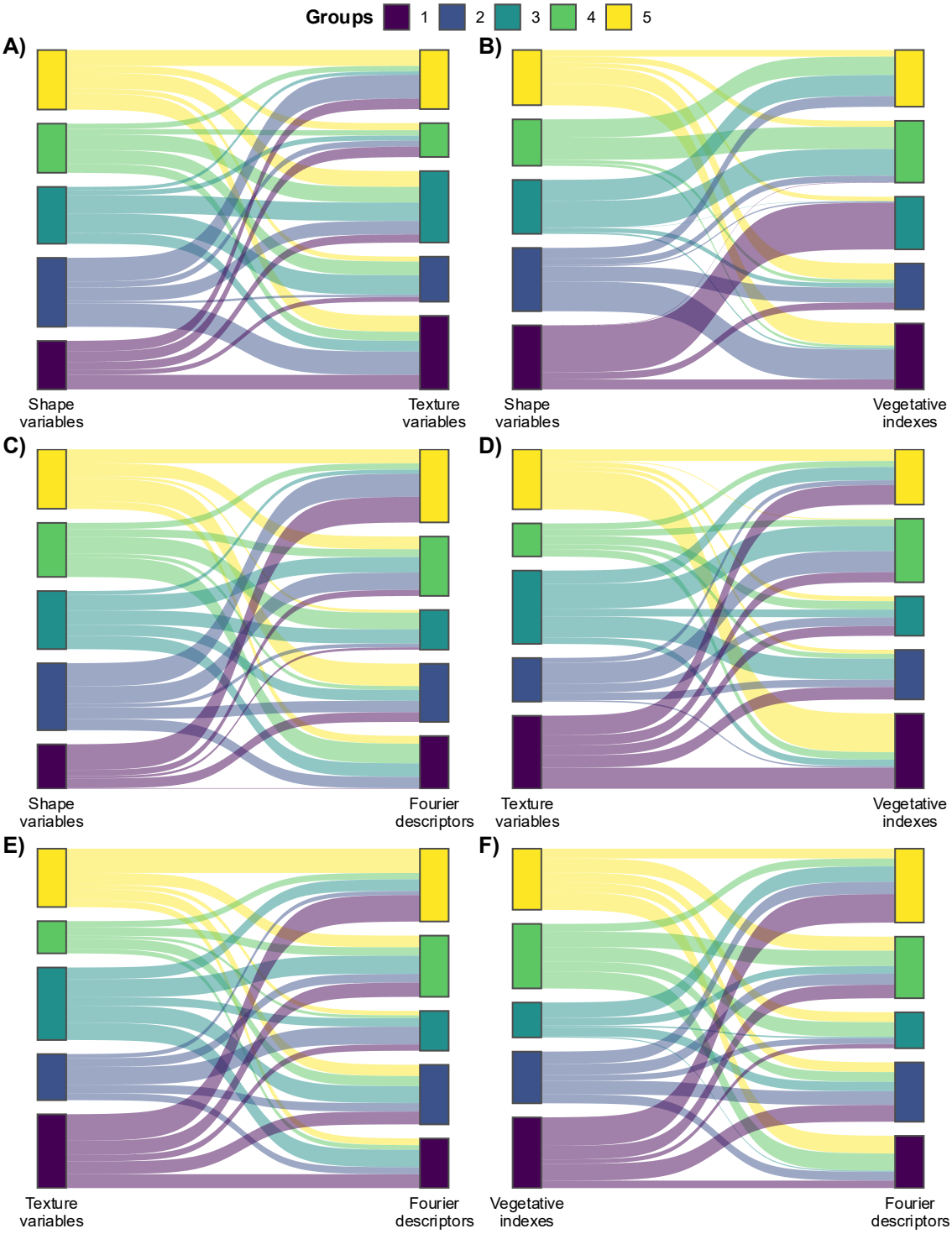

Figure 9. Jaccard index ($p_{XY_{ij}}$) between each of the five clusters formed by the k-means algorithm for every pair of the four different attribute sets (Shape, Texture, Fourier, and Vegetative Indices) extracted from soybean leaf images. The thickness of each colored band represents the value of $p_{XY_{ij}}$.

Table 6. The summary of Figure 9 shows the maximum and minimum concordance proportions ($p_{XY_{ij}}$) between clusters identified by the k-means algorithm for different phenomic information sets extracted from soybean leaflets (*Glycine max* (L.) Merr.).

Measurement	Pairs	Left cluster	Right cluster	$p_{XY_{ij}}$ (%)*
Maximum	Shape variables × Texture variables	2	5	22.20
	Shape variables × Vegetative indexes	1	3	48.95
	Shape variables × Fourier descriptors	1	5	24.88
	Texture variables × Vegetative indexes	5	1	36.45
	Texture variables × Fourier descriptors	1	5	24.29
	Vegetative indexes × Fourier descriptors	1	5	26.72
Minimum	Shape variables × Texture variables	2	2	2.07
	Shape variables × Vegetative indexes	1	5	0.00
	Shape variables × Vegetative indexes	4	3	0.00
	Shape variables × Fourier descriptors	1	1	0.58
	Texture variables × Vegetative indexes	5	4	0.86
	Texture variables × Fourier descriptors	4	3	2.63
	Vegetative indexes × Fourier descriptors	3	1	1.02

* Values highlighted in red and blue represent the highest and lowest concordance percentages between clusters formed using k-means for different pairs of phenomic information sets.

Overall, the concordance proportion (P_{XY}) in cluster formation exceeded 50% (**Table 7, Figure 10**), suggesting that any information set can be effectively used to form similar genotypic groups. However, evaluating each set's performance in predictions and classifications is crucial, where their agronomic impact can be more thoroughly explored.

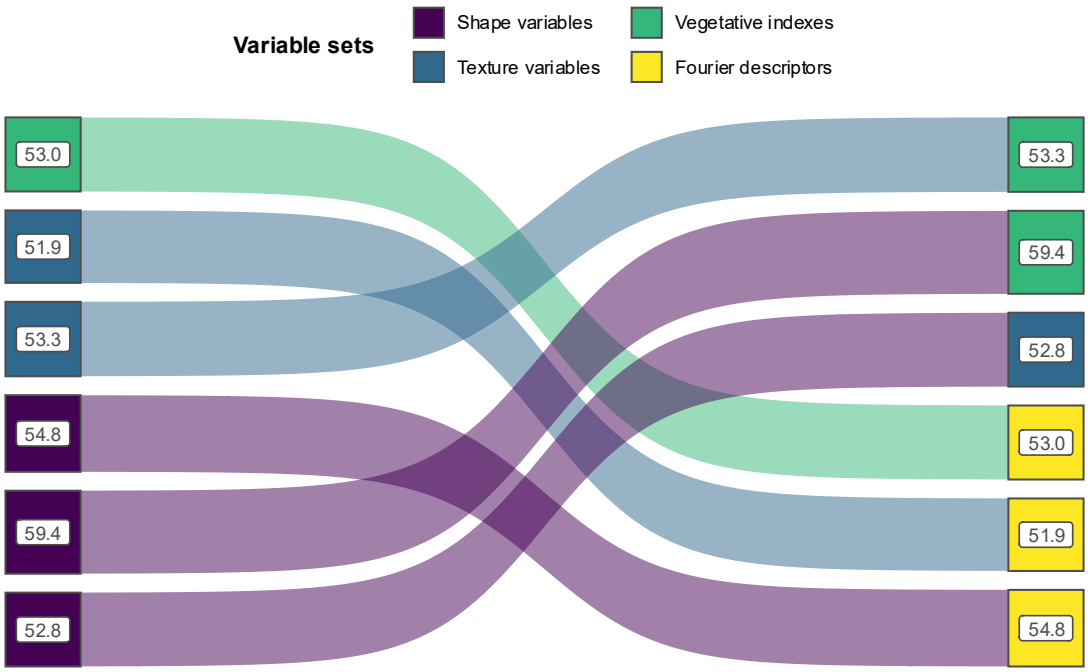

Figure 10. Concordance proportion (P_{XY}) in cluster formation is calculated using the k-means algorithm for each pair of the four distinct attribute sets (Shape, Texture, Fourier, and Vegetative Indices) extracted from soybean leaf images. The thickness of each colored band represents the value of P_{XY}.

Table 7. Concordance proportion (P_{XY}) in cluster formation by the k-means algorithm for each pair of phenomic information sets (Shape, Texture, Fourier, and Vegetative Indices) extracted from soybean leaflets. This table summarizes the P_{XY} values used to construct **Figure 10**.

Set of features		**Number of groups**		**Matching (%)**	
X	**Y**	**n**	**m**	$\bar{p}_{XY}$	P_{XY}*
Shape variables	Texture variables	5	5	10.57	52.83
Shape variables	Vegetative indexes	5	5	11.88	59.38
Shape variables	Fourier descriptors	5	5	10.96	54.80
Texture variables	Vegetative indexes	5	5	10.65	53.27
Texture variables	Fourier descriptors	5	5	10.37	51.87
Vegetative indexes	Fourier descriptors	5	5	10.60	52.99

* $P_{XY} = 2nm\bar{p}_{XY}/(n+m)$

CONCLUSIONS

This study evaluated phenomic data extracted from soybean leaflets, encompassing Shape, Texture, Fourier descriptors, and vegetative indices. Key conclusions include:

- Vegetative indices were the most abundant, while Fourier attributes exhibited the least redundancy.
- Each dataset provided distinct insights, enabling various genotypic discrimination methods.
- Five genotypic patterns were identified, with over 50% concordance across datasets.

These findings underscore the potential of phenomic analysis for genetic improvement studies. It can aid genotype pattern recognition and identify critical attributes for prediction and classification.

REFERENCES

ALABI, T. R. et al. Estimation of soybean grain yield from multispectral high-resolution UAV data with machine learning models in West Africa. **Remote Sensing Applications: Society and Environment**, v. 27, p. 100782, 2022. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2352938522000908>.

ANDERSON, E. J. et al. Soybean [*Glycine max* (L.) Merr.] breeding: History, improvement, production and future opportunities. In: AL-KHAYRI, J. M.; JAIN, S. M.; JOHNSON, D. V (Org.). **Advances in Plant Breeding Strategies: Legumes: Volume 7**. Cham: Springer International Publishing, 2019. p. 431–516. Disponível em: <https://doi.org/10.1007/978-3-030-23400-3_12>.

ARAUS, J. L.; CAIRNS, J. E. Field high-throughput phenotyping: the new crop breeding frontier. **Trends in plant science**, v. 19, n. 1, p. 52–61, 2014.

BERGER, B.; PARENT, B.; TESTER, M. High-throughput shoot imaging to study drought responses. **Journal of experimental botany**, v. 61, n. 13, p. 3519–3528, 2010.

BILDER, R. M. et al. Phenomics: the systematic study of phenotypes on a genome-wide scale. **Neuroscience**, v. 164, n. 1, p. 30–42, 2009.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. [S.l.]: Springer, 2006. v. 4.

BORCARD, D.; LEGENDRE, P. Is the Mantel correlogram powerful enough to be useful in ecological analysis? A simulation study. **Ecology**, v. 93, n. 6, p. 1473–1481, 2012.

CHAWADE, A. et al. High-throughput field-phenotyping tools for plant breeding and precision agriculture. **Agronomy**, v. 9, n. 5, p. 258, 2019.

CLARK, R. T. et al. Three-dimensional root phenotyping with a novel imaging and software platform. **Plant physiology**, v. 156, n. 2, p. 455–465, 2011.

COBB, J. N. et al. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. **Theoretical and Applied Genetics**, v. 126, p. 867–887, 2013.

CORTES, D. F. M. et al. Model-assisted phenotyping by digital images in papaya breeding program. **Scientia Agricola**, v. 74, p. 294–302, 2017.

DA SILVA JUNIOR, C. A. et al. Soybean varieties discrimination using non-imaging hyperspectral sensor. **Infrared Physics & Technology**, v. 89, p. 338–350, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1350449517308186>.

DE SOUSA, C. A. F. et al. Nova abordagem para a fenotipagem de plantas: conceitos, ferramentas e perspectivas. **Rev. Bras. Geogr. Fãsica**, v. 8, n. IV SMUD, p. 660–672, 2015.

DIPTA, B. et al. Digitalization of potato breeding program: Improving data collection and management. **Heliyon**, p. e12974, 2023.

DRYDEN, I. L.; MARDIA, K. V. Statistical shape analysis: Wiley series in probability and statistics. **New York, NY: John Wiley & Sons, Ltd**, 1998.

FANG, H.; LIANG, S. B. T.-R. M. in E. S. and E. S. Leaf Area Index Models☆. [S.l.]: Elsevier, 2014. .

GREY, D. R. Multivariate analysis, by KV Mardia, JT Kent and JM Bibby. Pp 522.£ 14· 60. 1979. ISBN 0 12 471252 5 (Academic Press). **The Mathematical Gazette**, v. 65, n. 431, p. 75–76, 1981.

GRINBLAT, G. L. et al. Deep learning for plant identification using vein morphological patterns. **Computers and Electronics in Agriculture**, v. 127, p. 418–424, 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168169916304665>.

HAINES, A. J.; CRAMPTON, J. S. Improvements to the method of Fourier shape analysis as applied in morphometric studies. **Palaeontology**, v. 43, n. 4, p. 765–783, 2000.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural Features for Image Classification. **IEEE Transactions on Systems, Man, and Cybernetics**, v. SMC-3, n. 6, p. 610–621, 1973.

HE, H.; MA, X.; GUAN, H. A calculation method of phenotypic traits of soybean pods based on image processing technology. **Ecological Informatics**, v. 69, p. 101676, 2022. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1574954122001261>.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of educational psychology**, v. 24, n. 6, p. 417, 1933.

HOULE, D.; GOVINDARAJU, D. R.; OMHOLT, S. Phenomics: the next challenge. **Nature reviews genetics**, v. 11, n. 12, p. 855–866, 2010.

JACCARD, P. The distribution of the flora in the alpine zone. **New Phytologist**, v. 11, n. 2, p. 37–50, 1 fev. 1912. Disponível em: <https://doi.org/10.1111/j.1469-8137.1912.tb05611.x>.

KASSAMBARA, A. **Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra**. [S.l.]: Sthda, 2017. v. 2.

KETCHEN, D. J.; SHOOK, C. L. The application of cluster analysis in strategic management research: an analysis and critique. **Strategic management journal**, v. 17, n. 6, p. 441–458, 1996.

KUHL, F. P.; GIARDINA, C. R. Elliptic Fourier features of a closed contour. **Computer Graphics and Image Processing**, v. 18, n. 3, p. 236–258, 1982. Disponível em: <https://www.sciencedirect.com/science/article/pii/0146664X8290034X>.

LARESE, M. G. et al. Automatic classification of legumes using leaf vein image features. **Pattern Recognition**, v. 47, n. 1, p. 158–168, 2014. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320313002641>.

LEGENDRE, P.; FORTIN, M. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. **Molecular ecology resources**, v. 10, n. 5, p. 831–844, 2010.

LEGENDRE, P.; LEGENDRE, L. **Numerical ecology**. [S.l.]: Elsevier, 2012.

LENK, S. et al. Multispectral fluorescence and reflectance imaging at the leaf level and its possible applications. **Journal of Experimental Botany**, v. 58, n. 4, p. 807–814, 2007.

LI, L.; ZHANG, Q.; HUANG, D. A review of imaging techniques for plant phenotyping. **Sensors**, v. 14, n. 11, p. 20078–20111, 2014.

MERLOT, S. et al. Use of infrared thermal imaging to isolate Arabidopsis mutants defective in stomatal regulation. **The plant journal**, v. 30, n. 5, p. 601–609, 2002.

MOMIN, M. A. et al. Machine vision-based soybean quality evaluation. **Computers and Electronics in Agriculture**, v. 140, p. 452–460, 2017. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168169916306986>.

MONTES, J. M.; PAUL, C.; MELCHINGER, A. E. Determination of chemical composition and nutritional attributes of silage corn hybrids by near-infrared spectroscopy on chopper: evaluation of traits, sample presentation systems and calibration transferability. **Plant Breeding**, v. 126, n. 5, p. 521–526, 2007.

NETO, J. C. et al. Plant species identification using Elliptic Fourier leaf shape analysis. **Computers and Electronics in Agriculture**, v. 50, n. 2, p. 121–134, 2006. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168169905001560>.

NICOTRA, A. B. et al. Leaf shape linked to photosynthetic rates and temperature optima in South African Pelargonium species. **Oecologia**, v. 154, p. 625–635, 1 fev. 2008.

_____. The evolution and functional significance of leaf shape in the angiosperms. **Functional Plant Biology**, v. 38, n. 7, p. 535–552, 2011.

OLIVOTO, T. Lights, camera, pliman! An R package for plant image analysis. **Methods in Ecology and Evolution**, v. 13, n. 4, p. 789–798, 2022.

OMARI, M. K. et al. Digital image-based plant phenotyping: a review. **Korean Journal of Agricultural Science**, v. 47, n. 1, p. 119–130, 2020.

PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin philosophical magazine and journal of science**, v. 2, n. 11, p. 559–572, 1901.

REN, T.; WERADUWAGE, S. M.; SHARKEY, T. D. Prospects for enhancing leaf photosynthetic capacity by manipulating mesophyll cell morphology. **Journal of Experimental Botany**, v. 70, n. 4, p. 1153–1165, 2019.

ROWLAND, S. D. et al. Leaf shape is a predictor of fruit quality and cultivar performance in tomato. **new phytologist**, v. 226, n. 3, p. 851–865, 2020.

TILLETT, R. D. Image analysis for agricultural processes: a review of potential opportunities. **Journal of Agricultural Engineering Research**, v. 50, p. 247–258, 1991. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0021863405800186>.

VISCOSI, V.; FORTINI, P. Leaf shape variation and differentiation in three sympatric white oak species revealed by elliptic Fourier analysis. **Nordic Journal of Botany**, v. 29, n. 5, p. 632–640, 1 out. 2011. Disponível em: <https://doi.org/10.1111/j.1756-1051.2011.01098.x>.

WALTER, A.; LIEBISCH, F.; HUND, A. Plant phenotyping: from bean weighing to image analysis. **Plant Methods**, v. 11, n. 1, p. 14, 2015. Disponível em: <https://doi.org/10.1186/s13007-015-0056-8>.

WICKHAM, H. et al. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.

XU, R.; LI, C. A review of high-throughput field phenotyping systems: Focusing on ground robots. **Plant Phenomics**, v. 2022, 2022.

SUPPLEMENTARY MATERIAL

Supplementary Table S1. Attributes from the different phenomic information sets used to construct the correlation plots shown in Figure 2. The attributes are organized in the same order as displayed on the graph axes: along the x-axis (from left to right) and the y-axis (from bottom to top).

Shape variables	Texture variables	Vegetative indexes	Fourier descriptors
plw_ratio	asm	RGBVI	B2
radius_ratio	idm	RGVBI	B4
circularity	f12	SCI	B6
eccentricity	cor	RGRI	B8
asp_ratio	f13	ExR	B10
elongation	sav	MyIndex	D3
theta	sva	R-G	D7
radius_min	den	a	D1
diam_min	con	GLI2	D9
minor_axis	dva	NGRDI	B7
width	var	MNGRDI	B9
length	sen	B	B3
caliper	ent	B-R	B5
major_axis		b*-a	C6
radius_max		B-G	C10
diam_max		b*	A6
coverage		DGCI	A8
area		NB	A10
area_ch		BGI	D2
diam_mean		ExB	A2
perimeter		HUE	A4
radius_mean		HUE2	A9
radius_sd		G-R	A1
rectangularity		I	A3
solidity		BI	A5
convexity		BIM	A7
narrow_factor		R	B1
form_factor		L	C1
circularity_norm		GRAY	C4
pd_ratio		L*-a	C8
circularity_haralick		G	C5
		GRAY2	C9
		L*	D5
		L*-b	D6
		G-B	D10
		ExG	D4
		NR	D8
		NG	C2
		GLI	C3
		NDGBI	C7
		S	
		NDRBI	
		R-B	

CHAPTER 3

CLASSIFICATION OF SOYBEAN CULTIVARS USING MACHINE LEARNING AND DIGITAL PHENOTYPING OF LEAFLET IMAGES

ABSTRACT

Soybean is a globally significant crop essential for human and animal nutrition and biofuel. Adequate germplasm characterization depends on quantifying genetic variability and establishing criteria to distinguish, standardize, and stabilize cultivars. Traditional morphological descriptors often need to meet the growing demand for precision in classification, driving the search for novel phenotypic attributes. This study employed digital image analysis and machine learning to evaluate features related to leaflet shape, texture, vegetative indices, and elliptical Fourier descriptors (EFDs) for soybean cultivar classification. The Random Forest (RF) algorithm was used as the classification method, with strategies implemented to address dataset imbalances and improve representation among classes. Results revealed significant morphological variability among genotypes, with maternal germplasm exhibiting the most discriminatory power. Image-derived data enhanced classification efficiency, with leaflet shape and texture emerging as the most relevant attributes. This research underscores the potential of combining digital phenotyping and machine learning to advance soybean cultivar classification and genetic improvement programs.

Keywords: *Glycine max* L., random forest, data imbalance, morphological attributes, Haralick textures, elliptical Fourier descriptors.

INTRODUCTION

Soybean (*Glycine max* L.) is a cornerstone of global agriculture, playing a vital role in human and animal nutrition and biofuel production. Its importance extends across sectors, serving as a primary source of protein and nutrients for human consumption, an essential input for livestock feed, and a critical raw material for renewable energy industries (ANDERSON et al., 2019; MUHAMMAD et al., 2024). Expanding cultivation into new regions has highlighted the need for genetic breeding to increase productivity and address biodiversity and sustainability challenges in the production chain (ALEMU et al., 2024; SAEED et al., 2024).

The foundation of genetic breeding relies on quantifying genetic variability and establishing clear criteria for germplasm characterization (SINGH et al., 2024; YADAVA et al., 2022). These efforts optimize the use of genetic resources and ensure cultivars meet distinctiveness, uniformity, and stability (DUS) requirements mandated for legal protection. Among DUS requirements, distinctiveness ensures that a cultivar is distinguishable from any other previously registered (BRASIL, 1997).

Traditionally, morphological descriptors have been extensively used to distinguish soybean cultivars and assess their agronomic performance. However, they often fail to meet the demands of modern classification processes (ANDERSON et al., 2019). Thus, expanding the list of morphological traits evaluated is critical, particularly traits related to leaflet morphology, which hold significant potential for discrimination (NOGUEIRA et al., 2008; SHILPASHREE et al., 2021).

Leaflet characteristics, such as size and shape, are strongly associated with a plant's growth potential and light capture efficiency, critical for photosynthesis. Larger leaflets can increase biomass, while morphological adaptations may enhance tolerance to environmental stresses like drought and heat (NICOTRA et al., 2008, 2011; REN; WERADUWAGE; SHARKEY, 2019; ROWLAND et al., 2020). Therefore, measuring leaflet traits provides breeders valuable information for selecting resilient and productive cultivars (NOGUEIRA et al., 2008; SHILPASHREE et al., 2021). In addition, when combined with other agronomic traits, these measurements can increase selection efficiency.

Digital image analysis is an innovative tool for capturing a broader range of phenotypic information, including size and shape, as well as colorimetric and texture attributes (AMARAL et al., 2024; RAJU; BALACHANDER; NEEHARIKA, 2022; VISCOSI; FORTINI, 2011). In this context, this study aims to evaluate the potential of soybean cultivar discrimination based on leaflet characteristics using image analysis and machine learning techniques. Three

fundamental methodological issues are considered in classification studies: genetic material, which reflects the inherent variability within individuals or populations that may enhance or hinder discrimination; information utilized, encompassing the range of attributes selected to optimize discrimination accuracy; and biometric procedures, involving the methods applied to identify genotype similarity patterns for effective discrimination.

Classification studies focus on developing functions or algorithms capable of categorizing individuals based on measurable characteristics while minimizing classification errors. Once a model's effectiveness is validated, these functions can be employed to classify new individuals with unknown origins (LIAW; WIENER, 2015; SANTANA et al., 2023). Although linear and quadratic discriminant functions are traditionally popular, machine learning techniques, such as decision trees and their advanced variants, have shown significant potential. These methods excel in capturing nonlinear relationships and are less sensitive to issues like data distribution assumptions, missing data, and outliers.

This study adopts the Random Forest (RF) algorithm for its precision, robustness, and ability to manage complex, high-dimensional datasets. RF constructs an ensemble of decision trees, with final classifications determined by the majority vote across all trees (BREIMAN, 2001; PRASAD; IVERSON; LIAW, 2006). While highly effective, RF can be adversely affected by imbalanced datasets, where majority classes dominate and bias the results (WANG et al., 2021). For instance, consider two scenarios: one with equal representation (50 genotypes each of large and small leaflets) and another with a skewed distribution (95 genotypes of large leaflets and only 5 of small leaflets). The latter scenario highlights the critical need for balanced data partitioning and the selection of appropriate evaluation metrics to prevent classification biases.

In this context, the present study seeks to advance genetic improvement efforts by evaluating the potential of leaflet images for discriminating against soybean genotypes. Furthermore, it proposes robust biometric strategies to address data imbalance issues effectively. By leveraging the Random Forest algorithm, this research aims to provide practical and methodological insights into cultivar differentiation, ultimately enhancing precision and efficiency in the selection process.

MATERIAL AND METHODS

Soybean cultivars

An experiment involving soybean lines in the F3 generation was established at the experimental field of the Regional University of the Northwest of the State of Rio Grande do Sul (UNIJUÍ), located in Ijuí, RS, Brazil (Latitude: 28º 23’ 16” S; Longitude: 53º 54’ 53” W). The plots for each line covered an area of 2 × 0.5 m² and were organized in augmented blocks, interspersed with control plots. One thousand four hundred forty-nine lines were sown, distributed across 15 blocks (~100 lines per block) in the experimental area. The sowing was carried out on three different dates: on October 14, 2022, lines 1 to 1000 were planted; on October 17, 2022, lines 1001 to 1300; and on October 18, 2022, lines 1301 to 1449. Prior to the experiment’s implementation, the soil was fertilized with 100 kg/block of organic fertilizer (Sulfacal), 8 tons of Biogranum (a mixture of silicon, calcium, and magnesium), and two applications of 3 kg/block of 3-21-21 fertilizer (N-P-K).

Classification Study Scenarios

Aspects Related to Genetic Material and Phenotypic Information

For the classification study, various scenarios were considered, reflecting the specific characteristics of the genetic material (maternal, paternal, and hybrid combinations) and the type of information analyzed (e.g., shape, texture, vegetative indices, Fourier descriptors, and the complete set of features). A detailed examination of these aspects is presented below.

Between 8:00 a.m. and 4:00 p.m., one leaf from the middle third of three plants per soybean line (F3 generation) at the R5 growth stage was collected for RGB image capture. The images were obtained in the laboratory under ambient temperature (22 ºC) and diffuse artificial lighting. The leaflets from each line (i.e., nine leaflets per line) were detached and arranged in three horizontal rows on red cardboard. An 18 cm² yellow color reference was placed on the cardboard to calibrate metrics derived from the leaflets during image analysis. The leaflets were then photographed at a resolution of 1280 × 1024 pixels and 96 dpi (dots per inch) using a digital camera positioned 50 cm above the cardboard. Due to an approximate 23.74% loss in cultivation, 1,105 lines were photographed out of the 1,449 initially planted, resulting in 9,945 observations (9 leaflets × 1,105 lines).

Phenotypic data were extracted from the soybean leaflet images using the “pliman” package (version 2.1.0) (OLIVOTO, 2022) within the R software environment (version 4.2.3) (https://cran.r-project.org/), integrated with the RStudio development interface

(https://posit.co/download/rstudio-desktop/). The extracted datasets were processed and organized using the "tidyverse" package (version 2.0.0) (WICKHAM et al., 2019).

The extracted attributes were grouped into four distinct datasets based on the type of information they represented:

1. **General Shape:** Contained 31 attributes.
2. **Haralick Textures:** Included 13 attributes derived from texture analysis.
3. **Vegetative Indices (VIs):** Comprised of 43 attributes calculated as the average VIs of all pixels forming each leaflet.
4. **Elliptical Fourier Descriptors (EFDs):** Comprised of 40 attributes capturing detailed shape information.

The four datasets were subsequently merged into a comprehensive dataset containing 127 attributes. This unified dataset was further refined by averaging the values of the nine leaflets per line, yielding 1,105 averaged observations, with each observation representing a single soybean line.

Additionally, the genealogy of each soybean line was well-documented, allowing information about the maternal and paternal parents and their hybrid combinations to be incorporated into the dataset. In this context, "recombinants" refer to the parental pairs that gave rise to the lines, regardless of their roles as male or female parents in the crosses.

Aspects Related to the Database

For the classification study, various scenarios were evaluated to address imbalances in the representativeness of genealogical information within the database (maternal, paternal, and hybrid combinations). Data from 26 maternal parents, 24 paternal parents, and 31 recombinants were analyzed in one scenario. In another scenario, the dataset was restricted to only four maternal parents, five paternal parents, and four recombinants. The implications of these imbalances in the database observations are discussed below.

The classification procedure was implemented using k-fold cross-validation. This process began with the randomization of the soybean lines, followed by their systematic division into k distinct groups within each label, corresponding in this study to a specific parent (maternal or paternal) or a recombinant. The procedure sequentially combined $k - 1$ groups from each label to train the model, while the remaining group from each label was used for testing. This process was iterated k times ensuring each line participated in the validation phase exactly once.

While k-fold cross-validation is widely used, it imposes a critical requirement. Each label must have at least k observations to ensure the procedure can be executed appropriately. Since the soybean breeding program that provided the data for this study resulted in fewer than k lines for given parents (maternal or paternal) or hybrids, the initial dataset (containing 1,105 lines) needed to be filtered. This filtering process ensured that every parent (maternal or paternal) or recombinant had at least k-associated lines. Furthermore, an additional restriction was applied: the database was further filtered to require each parent or recombinant to have $15 \times k$ associated lines. Although this restriction reduced the number of parents or hybrids in the database, it ensured that those remaining were associated with a more significant number of lines. This balance of observations among ancestors tends to strengthen classification models and improve their accuracy (TANG; HENDERSON; GARDNER, 2021), as the performance of the RF algorithm heavily relies on data partitioning.

With k set to 5 for the cross-validation process, the first restriction reduced the database from 33 to 26 maternal parents, 35 to 24 paternal parents, and 42 to 31 recombinants. The second restriction further narrowed the dataset to four maternal parents, five paternal parents, and four recombinants. Supplementary Tables S1, S2, and S3 provide a detailed breakdown of the number of lines retained for each filtering step, categorized by maternal, paternal, or hybrid origins**123.**

Biometric Approach in Classification Study

This study employed the machine learning algorithm Random Forest (RF) to classify soybean lines (F3 generation) based on their ancestry – maternal parents, paternal parents, or recombinants. RF constructs an ensemble of decision trees using resampling techniques applied to both observations and predictors in the dataset. This approach enhances predictive performance by reducing variance, lowering error rates, and mitigating overfitting (BREIMAN, 2001; PRASAD; IVERSON; LIAW, 2006). The RF models were implemented using the randomForest() function from the "randomForest" package (versão 4.7-1.2) (LIAW; WIENER, 2002, 2014), available in the R environment (R CORE TEAM, 2024).

The RF models were constructed using four datasets containing phenotypic information from soybean leaflets, focusing on shape, texture, vegetative indices (VIs), and Fourier descriptors. An additional, comprehensive dataset was created by combining all attributes from the previous datasets. Each dataset contained a specific number (n_i) of variables, where $i = 1, 2, 3, 4, \text{and } 5$. The attributes within each phenotypic dataset were selected iteratively during

algorithm cycles (loops), which determined the objects to be passed as arguments to the randomForest() function. Similarly, observations were filtered according to the restrictions imposed by the k-fold cross-validation process during these cycles.

The algorithm implementation included five repetitions of the k-fold cross-validation procedure (as previously described). Each repetition involved a new random shuffle of the lines within their respective ancestry groups (maternal parents, paternal parents, or recombinants), ensuring a robust and diverse model performance evaluation.

Several model parameters were adjusted to optimize classification accuracy. Predictor sampling was conducted with and without replacement, as specified by the replace parameter in the randomForest() function. The number of variables randomly sampled as candidates for each tree varied according to a numerical sequence $\{S_i\}$, which depended on the number of attributes (n_i) in the phenotypic datasets. This sequence comprised $z-1$ terms ($j = 1, 2, 3, \ldots, z-1$), defined by the formula:

$$s_j = x + \frac{y-x}{z-1} \times (j-1) \quad (1)$$

Where:

s_j: the j-th value in the sequence;

x: 10% of the number of attributes in the i-th dataset ($x = 0.1 \times n_i$);

y: the total number of attributes in the i-th dataset ($y = n_i$);

z: the total number of elements in the sequence, defined as $z = 5$.

The default value used by the randomForest() function (the square root of the number of predictors) was included in each sequence $\{S_i\}$, generating an extended sequence $\{S_i'\}$ with z elements. This extended sequence $\{S_i'\}$ was ordered as $\{S_{i_{ordered}}'\}$, and its values were rounded upward to the nearest integer, following the rule:

$$\left\lceil s_{i_{ordered_l}}' \right\rceil = \min\left\{ n \in \mathbb{Z}^+ \middle| n \geq s_{i_{ordered_l}}' \right\} \quad (2)$$

where $l = 1, 2, 3, \ldots, z$ denotes the position of the term in the ordered sequence $\{S_{i_{ordered}}'\}$.

Additionally, the maximum number of terminal nodes in each tree (controlled by the "maxnodes" parameter) varied as 1%, 2.5%, 5%, 7.5%, 10%, 17.5%, 25%, 50%, 75%, and 90% of the number of observations in the training dataset, which the k-fold cross-validation process had restricted. These values were also rounded according to the rule in Equation 2.

Figure 1 illustrates the flow of objects used to construct the RF models, which include, in order, ancestry of the lines, observation filtering, phenotypic information, data partitioning, repetitions/shuffling, number of predictors, sampling method, and maximum number of nodes.

The algorithm flow was implemented using functions from the "purr" package (WICKHAM; HENRY, 2023) and the "furrr" package (VAUGHAN; DANCHO, 2022). The "furrr" package enabled parallel processing of the loop, controlling the iterations for the maximum number of nodes (lower layer of **Figure 1**), significantly accelerating model execution.

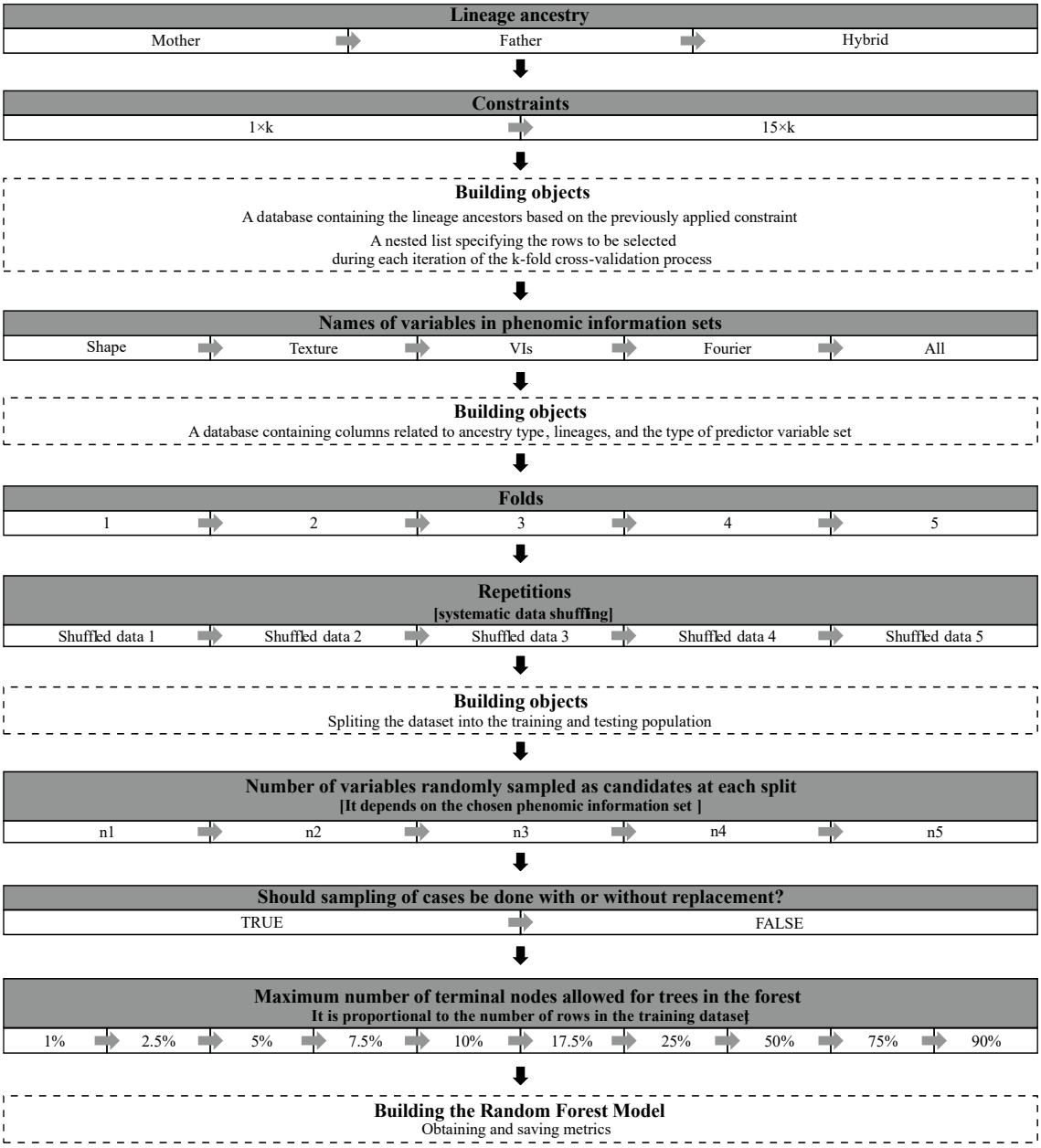

Figure 1. Diagram of the object flow for training random forest models. The diagram illustrates the flow of selected objects used to train the Random Forest models. In order, the selected objects include ancestry of the lines, observation filtering, phenotypic information, data partitioning, repetition/shuffling, number of predictors, sampling method, and maximum number of nodes. Each iterative cycle is displayed as a table with a gray-highlighted header,

which controls selecting one object type at a time. The objects required for subsequent iterations are obtained between cycles, represented by dotted rectangles. Gray arrows indicate the direction of selection within each object type, while black arrows show the progression of the selected object types.

Metrics Used to Calculate Classification Efficiency

Three metrics were employed to assess the classification efficiency of the RF models: sensitivity, specificity, and balanced accuracy. For multiclass classification problems, these metrics are computed using a one-vs-all approach based on confusion matrices (Figure 2) that display counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For each class ($u = 1, 2, 3, \ldots, v$), the model is evaluated on its ability to distinguish that class from all others. Additionally, the average of these metrics provides a global perspective on the models' discriminatory power, assigning equal weight to each class and ensuring that majority classes do not overshadow minority classes.

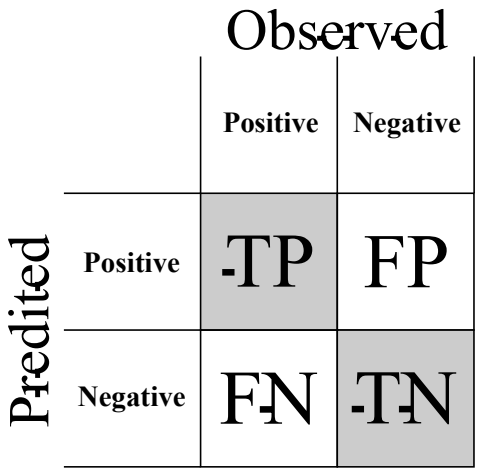

Figure 2. Confusion matrix for evaluating classification model performance. The table compares predicted and observed classes, presenting values for true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Below are the formulas and definitions of the metrics used to evaluate the RF models in this study:

- **Sensitivity or True Positive Rate (TPR):** The proportion of true positives correctly identified by the model.

$$\text{Sensitivity}_u = \frac{\text{TP}_u}{\text{TP}_u + \text{FN}_u} \quad (3)$$

$$\text{Mean sensitivity} = \frac{1}{v}\sum_{u=1}^{v}\frac{TP_u}{TP_u + FN_u} \quad (4)$$

- **Specificity or True Negative Rate (TNR):** The proportion of true negatives correctly identified by the model.

$$\text{Specificity}_u = \frac{TN_u}{TN_u + FP_u} \quad (5)$$

$$\text{Mean specificity} = \frac{1}{v}\sum_{u=1}^{v}\frac{TN_u}{TN_u + FP_u} \quad (6)$$

- **Balanced Accuracy:** The average of sensitivity and specificity, providing an adjusted measure for handling class imbalances.

$$\text{Balanced accuracy}_u = \frac{1}{2}\left(\frac{TP_u}{TP_u + FN_u} + \frac{TN_u}{TN_u + FP_u}\right) \quad (7)$$

$$\text{Mean balanced accuracy} = \frac{1}{v}\sum_{u=1}^{v}\frac{1}{2}\left(\frac{TP_u}{TP_u + FN_u} + \frac{TN_u}{TN_u + FP_u}\right) \quad (8)$$

Classification Methods

Once the classifiers are established, determining the class to which a specific line belongs is essential. During the training phase (or classifier adjustment), correct and incorrect classification rates measure the technique's efficiency. However, a particular line often exhibits varying degrees of membership across different classes, making the choice of an appropriate classification method critical.

Soybean lines were classified according to their ancestry using two distinct methods. The first method utilized the standard classification rule based on the Maximum a posteriori (MAP) criterion. The second method employed a decision threshold, determined using the Receiver Operating Characteristic (ROC) curve, to balance TPR and the false negative rate (FNR) and optimize classification. Within the second method, two approaches (PERKINS; SCHISTERMAN, 2006; ROBIN et al., 2011) were applied:

1. The top-left corner method identifies the point closest to (0,1) on the ROC curve.
2. The Youden index-based method balances sensitivity and specificity.

1) Standard Classification Method – MAP Rule

In the MAP rule, a class is assigned based on the highest posterior probability estimated by the model. For a sample $\mathbf{x}_q$, the predicted class $\hat{y}_q$ is given by:

$$\hat{y}_q = \arg \max_{c \in \{C1,C2,\ldots,C_v\}} \hat{P}(y = c|\mathbf{x}_q) \quad (9)$$

Where:

- $\mathbf{x}_q$ is the feature vector of the q-th line ($q = 1, 2, 3, \ldots, r$);
- $\hat{y}_q$ is the predicted ancestry of the q-th line;
- c represents one of the possible ancestries (C_u) to which $\mathbf{x}_q$ may belong ($c \in \{C1, C2, \ldots, C_v\}$ for $u = 1, 2, 3, \ldots, v$);
- $\hat{P}(y = c|\mathbf{x}_q)$ é a probabilidade predita pelo modelo de $\mathbf{x}_q$ pertencer à uma determinada ascendência c;
- $\max_{c \in \{C1,C2,\ldots,C_v\}} \hat{P}(y = c|\mathbf{x}_q)$ is the posterior probability predicted by the model for $\mathbf{x}_q$ to belong to ancestry c;
- arg max represents the c value which maximizes $\hat{P}(y = c|\mathbf{x}_q)$.

2) Optimal Threshold-Based Classification Method

For this method, the predicted class $\hat{y}_q$ for a sample $\mathbf{x}_q$ is given by:

$$\hat{y}_q = \arg \max_{c \in \{C1,C2,\ldots,C_v\}} [\hat{P}(y = c|\mathbf{x}_q) - t_c] \quad (10)$$

Where t_c represents the optimized threshold for each class or ancestry.

2.1) Top-Left Corner Method on the ROC Curve

This method identifies the point on the ROC curve closest to (0,1), minimizing the distance between TPR and FPR (false positive rate). The optimal threshold t_c is calculated as:

$$t_c = \arg \min_t \sqrt{(1 - \text{Sensitivity})^2 + (1 - \text{Specificity})^2} \quad (11)$$

2.2) Youden Index Method

This method maximizes the difference between sensitivity and FPR, balancing sensitivity and specificity. The Youden index (J) is defined as (YOUDEN, 1950):

$$J = \text{Sensitivity} - (1 - \text{Specificity}) \quad (12)$$

The optimal threshold t_c is given by:

$$t_c = \arg \max_t [\text{Sensitivity} - (1 - \text{Specificity})] \quad (13)$$

Random Forest Model Selection

Although many RF models were generated by combining different arguments for the randomForest() function inputs, not all were expected to perform effectively as ideal classifiers. Therefore, two strategies for model selection based on the mean balanced accuracy (MBA) were adopted: **model selection before summarization (BS)** and **model selection after summarization (AS)** of this metric. The MBA was chosen because it is more suitable for imbalanced datasets and offers straightforward interpretability.

Model Selection Before MBA Summarization (BS)

1. The best model with a specific configuration of arguments was selected based on the highest MBA value obtained from the validation set for each fold within each repetition.
2. The mean evaluation metrics (MBA, sensitivity, and specificity) were then calculated for each repetition using the metric values from the folds in the validation set.
3. Finally, estimates of the mean and standard deviation of the metrics (calculated from the repetition averages) were obtained for the validation dataset.

Model Selection After MBA Summarization (AS)

1. Initially, models with specific argument configurations had their metrics (MBA, sensitivity, and specificity) summarized by averaging the values across each repetition (i.e., using the metrics from the folds).
2. These metrics were further summarized by calculating their overall mean and standard deviation.
3. Finally, the best argument configuration was selected based on the highest MBA mean obtained from the validation set.

In the cross-validation process, which utilized five folds and five repetitions, the model selection before summarization (BS) strategy resulted in 25 models trained with different combinations of arguments, while the model selection after summarization (AS) strategy resulted in 25 models trained with the same argument combinations.

In both approaches, model selection was performed either by including the different phenotypic datasets as part of the arguments defining the model structure or by excluding these datasets from the argument set. In the first case, the summarized evaluation metrics (mean and standard deviation) were obtained independently of the datasets. In the second case, the metrics differed between the datasets.

Finally, graphical perspectives of the model selection results were illustrated using the "ggplot2" (versão 3.5.1) (WICKHAM, 2016), "ggtext" (versão 0.1.2) (WILKE; WIERNIK, 2022), and "ggh4x" (versão 0.2.8.9000) (VAN DEN BRAND, 2024) packages in R. Additionally, confusion matrices were created with the "cvms" (versão 1.5.2) to represent the results of the models selected with the $15 \times k$ restriction on the number of lines.

RESULTS AND DISCUSSION

Discrimination and classification techniques are crucial in genetic improvement, but various factors significantly influence their results. Among these, the specific genetic characteristics of the populations being evaluated (CRUZ; FERREIRA; PESSONI, 2011), the degree of imbalance in the dataset (WANG et al., 2021), the choice of the classification model, and the metric used to evaluate the approach's effectiveness stand out (KOKLU; CINAR; TASPINAR, 2021).

Figure 3 demonstrates model selection results, considering the inclusion of different phenotypic information datasets to configure the models and the variations incorporated into other parameters.

Regarding the particularities of the evaluated genetic set, the maternal parents group exhibited greater genetic variability, which led to better performance in metrics indicating classification effectiveness. The sensitivity level for maternal parents was significantly higher than for paternal parents, and this advantage was also evident when compared to the recombinant group, whose performance was considered satisfactory.

Classification for maternal ancestors outperformed paternal and recombinant ancestors, particularly under the $15 \times k$ restriction. In the $1 \times k$ restriction, the low MBA (~0.55) in ancestry classifications was more associated with an increase in the FNR than the FPR. Conversely, a better balance between FPR and FNR was observed when applying the $15 \times k$ restriction.

The positive impact of applying restrictions was evident from a methodological perspective, including classification method, effectiveness metrics, and dataset restrictions to mitigate class representation effects. In this study, restricting the dataset yielded various benefits, including reducing imbalance, providing a more equitable distribution of lineage representation across classes, and decreasing the number of classes within each genetic group (maternal parent, paternal parent, and hybrids). Methods like random undersampling and oversampling, which balance class observations by reducing samples in majority classes or

replicating samples in minority classes, respectively, have significantly improved the performance of decision tree-based machine learning methods such as Random Forest (TANG; HENDERSON; GARDNER, 2021).

In general, although not in all cases, classification using the Youden index-based method was more advantageous than other methods, as evidenced by sensitivity and MBA metrics. Additionally, model selection performed before summarizing the MBA showed slightly better results than after summarizing the MBA, especially under the $15 \times k$ restriction, as reflected in MBA and average sensitivity metrics across all ancestries. A slight increase in average sensitivity was also observed when BS selection was applied to the $1 \times k$ restriction dataset for all ancestries.

Notably, the phenotypic dataset comprising all variables formed the configuration of most models selected before MBA summarization (BS). Furthermore, only datasets containing all variables were included in models selected after MBA summarization (AS) (**Supplementary Table S4**).

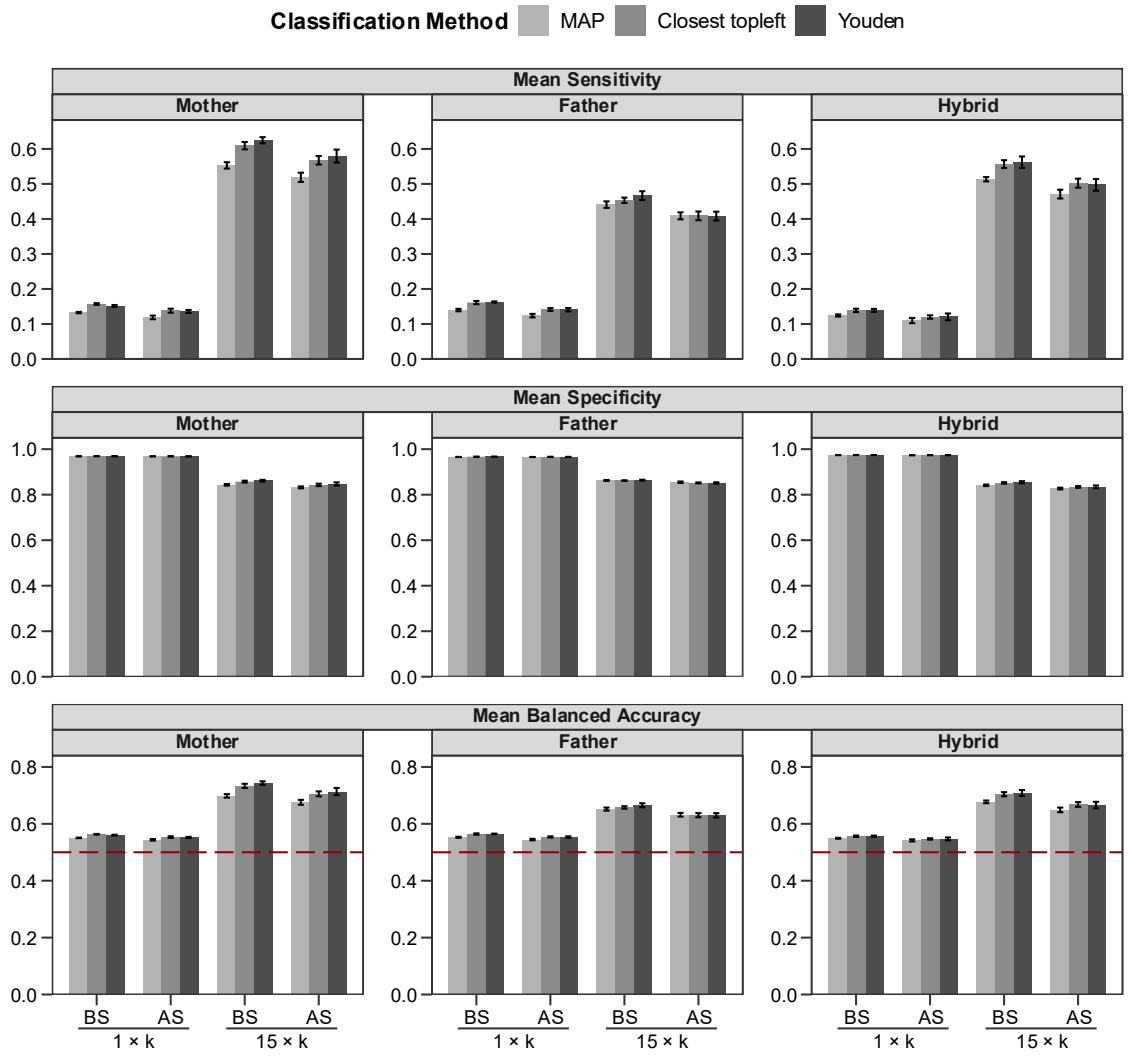

Figure 3. Average sensitivity (true positive rate), specificity (true negative rate), and mean balanced accuracy (MBA) obtained from Random Forest classifications of the paternal, maternal, and recombinant ancestries of soybean lines (*Glycine max* L.) using the Maximum a posteriori (MAP) method, Youden index, and top-left corner method. On the x-axis, 1 × k, and 15 × k indicate the restrictions applied to the number of lines in the dataset due to the k-fold cross-validation procedure. Also, on the x-axis, "BS" and "AS" indicate that model selection was performed before and after MBA summarization. The dashed horizontal red line indicates an MBA of 0.5

Information about soybean genotype leaflets is widely recognized for its association with photosynthetic efficiency, which directly influences productivity, adaptability, and resistance to pathogens and abiotic factors. However, such data has often been limited to leaflet size and shape. In this context, image-based analysis can significantly enhance morphological

studies by providing additional details on texture and contour. These improvements lead to more precise characterizations and more effective discrimination of lineages.

Therefore, it is essential to consider the information presented in **Figure 4**, which shows model selection results while accounting for the exclusion of different phenotypic datasets during model configuration. This figure highlights the predominance of selected models that included all variables when these were part of the model configuration definition (**Supplementary Table S4**). Consistently, **Figure 4** demonstrates that using a general dataset comprising all phenotypic datasets achieves higher classification accuracy regardless of ancestry type, model selection criteria, or applied restrictions.

Suppose analyses using subsets of information need to be prioritized. In that case, datasets containing shape or texture variables should receive greater attention due to their enhanced impact on metrics that quantify classification efficiency.

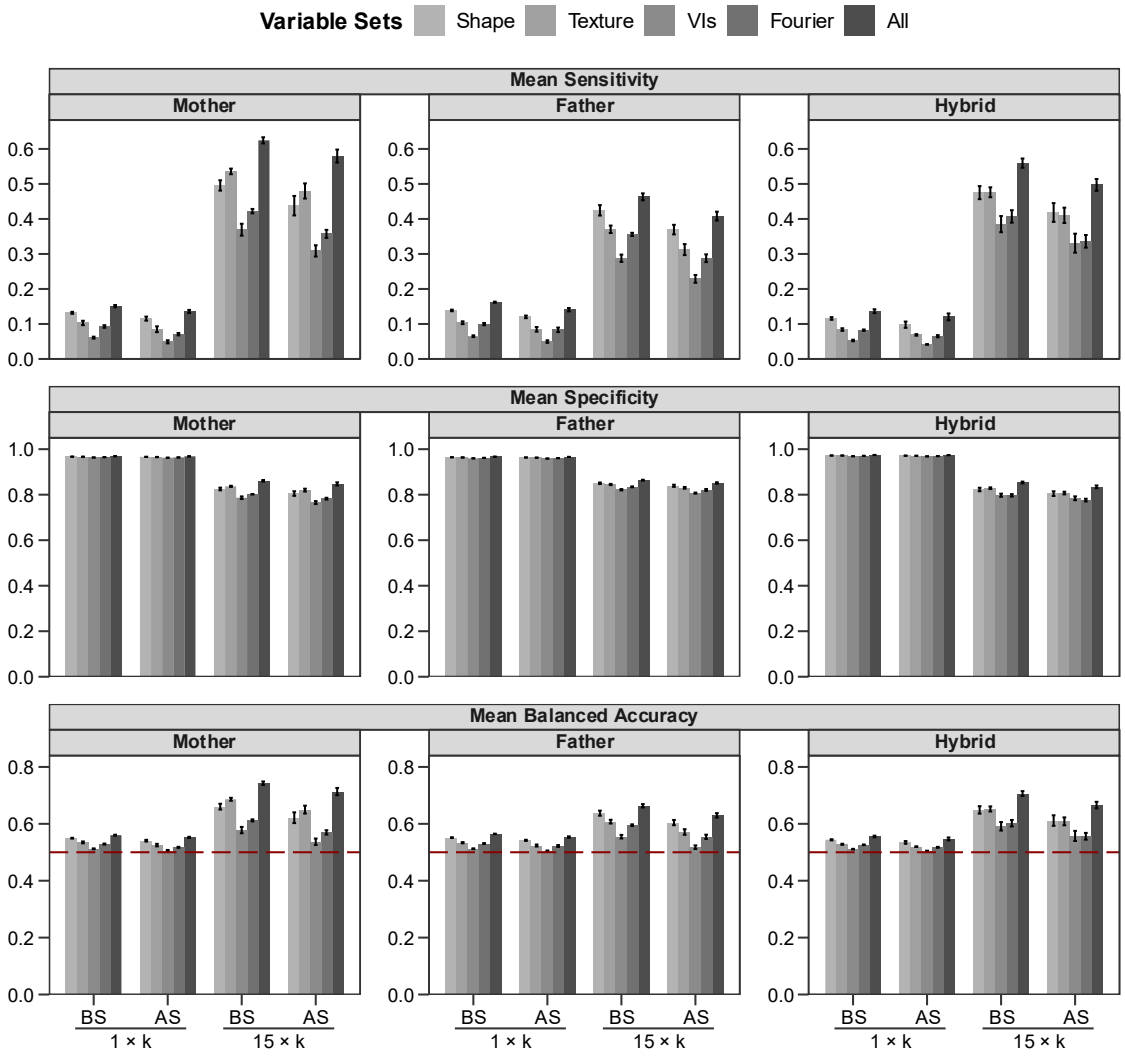

Figure 4. Average sensitivity (true positive rate), specificity (true negative rate), and mean balanced accuracy (MBA) obtained from Random Forest classifications of the paternal, maternal, and recombinant ancestries of soybean lines (*Glycine max* L.) for each phenotypic dataset used. On the x-axis, 1 × k, and 15 × k indicate the restrictions applied to the number of lines in the dataset due to the k-fold cross-validation procedure. Also, on the x-axis, "BS" and "AS" indicate that model selection was performed before and after MBA summarization. The Youden index was chosen for this graph as it generally demonstrated superior performance compared to other classification methods tested. The dashed horizontal red line indicates an MBA of 0.5

When conducting classification studies on a group of lineages, the expectation is to identify representative sets of well-characterized genotypes that can be discriminated with relative ease. These genotypes exhibit typical configurations of size, shape, color (as indicated

by vegetative indices), and texture, distinguishing them from others (BRASIL, 1997). However, lineages also represent a more homogeneous group of genotypes, where distinguishability is reduced even when broader information and more sophisticated biometric approaches are applied.

Figures 5, **6**, and **7** illustrate that only a few ancestries (maternal, paternal, or recombinant) achieved a balanced accuracy of close to 70% in the selected models. Furthermore, for ancestries with a balanced accuracy of around 50% and specificity of 100%, the models could correctly classify all negative examples (non-target ancestries) but consistently failed to identify positive examples (target ancestry). This classification indicates that the selected models effectively distinguish non-target ancestries but face significant challenges in accurately recognizing examples belonging to the target ancestry.

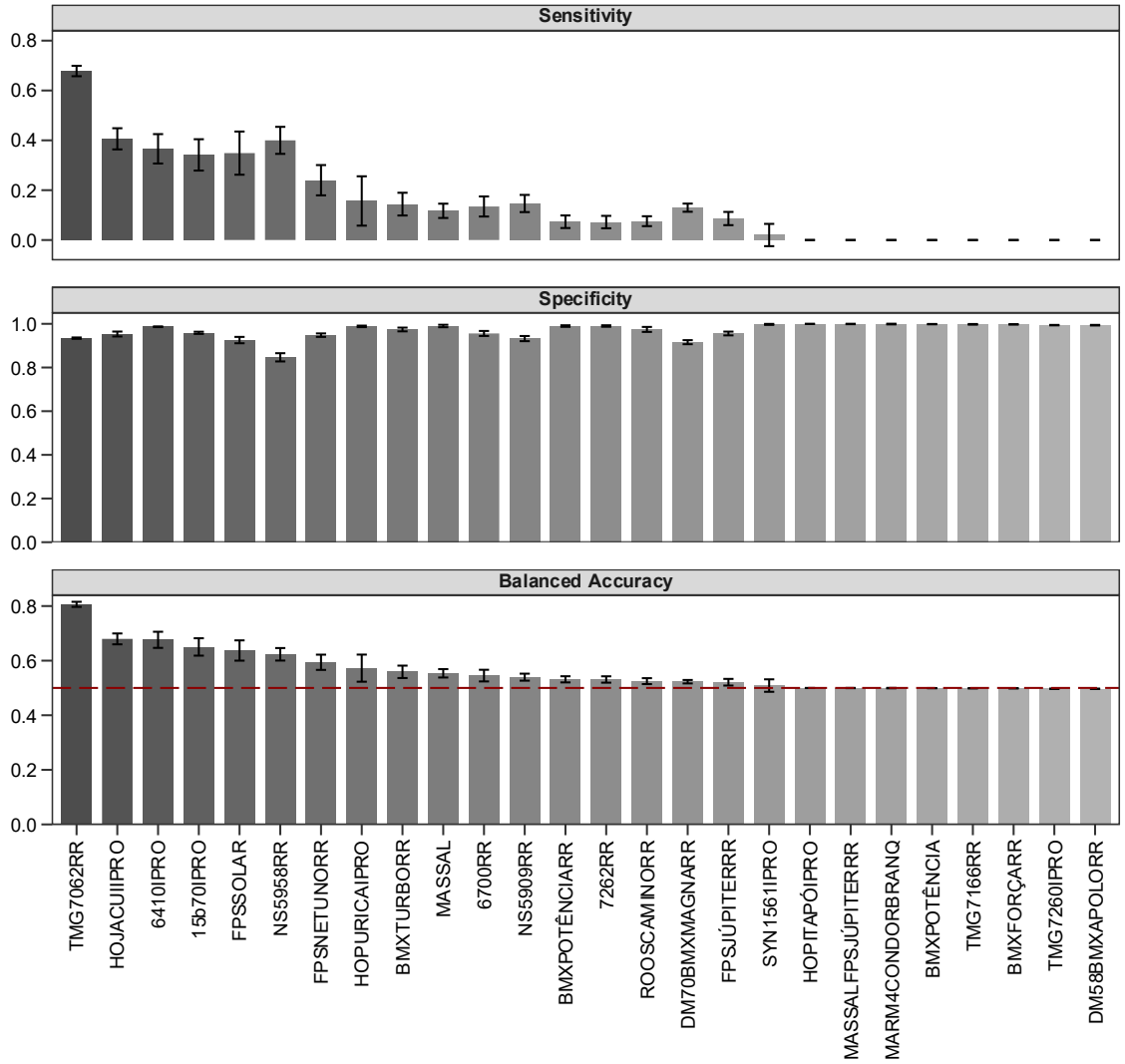

Figure 5. Average sensitivity (true positive rate), specificity (true negative rate), and mean balanced accuracy (MBA) obtained for each maternal ancestor of soybean lines (*Glycine max* L.) using the 1 × k restriction dataset with Random Forest models. The Youden index and the selection of models performed before MBA summarization (BS) were chosen to compose the data presented in this graph, as both showed superior performance compared to other classification methods and model selection criteria tested. The dashed horizontal red line indicates a balanced accuracy of 0.5.

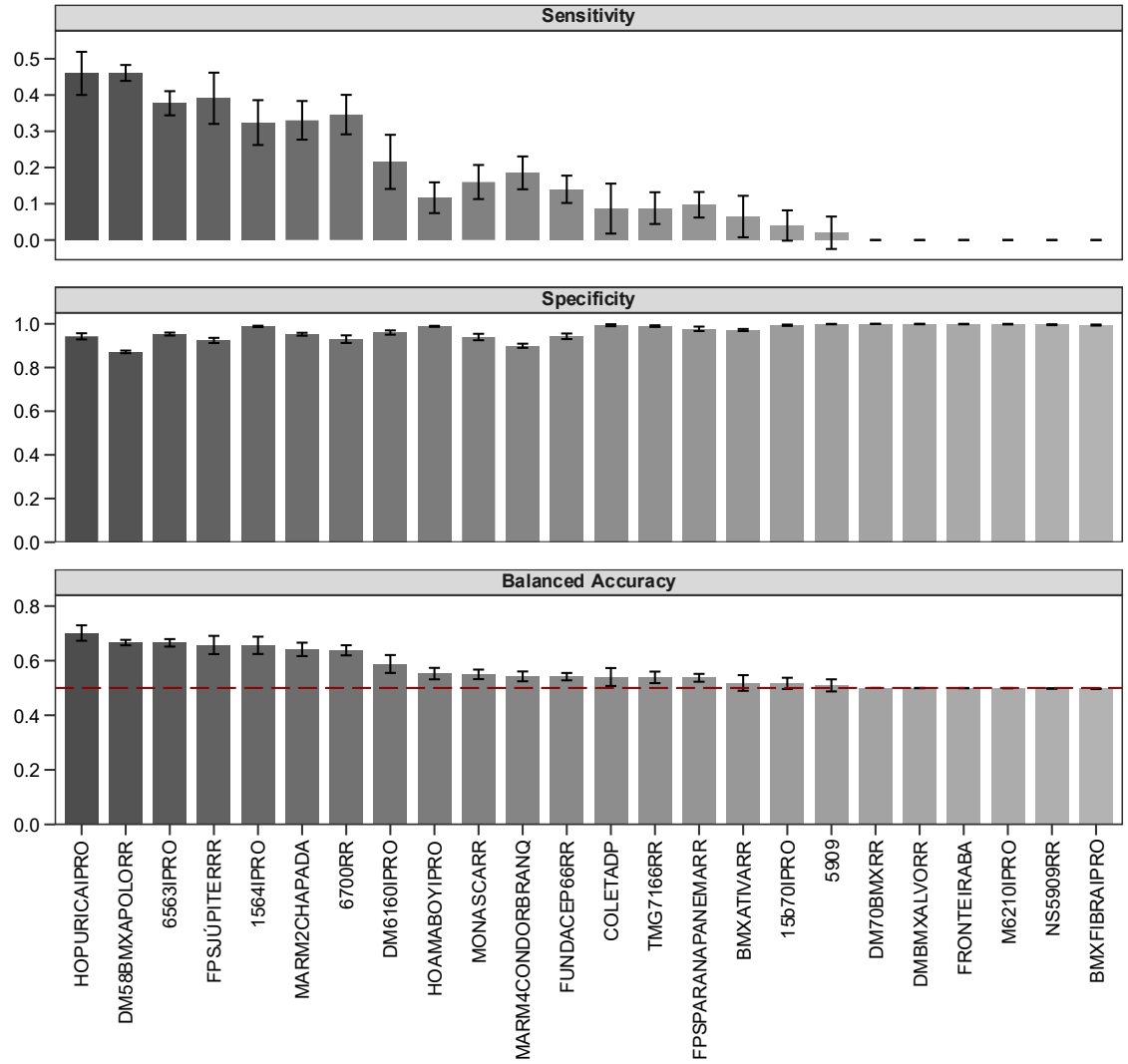

Figure 6. Average sensitivity (true positive rate), specificity (true negative rate), and mean balanced accuracy (MBA) obtained for each paternal ancestor of soybean lines (*Glycine max* L.) using the 1 × k restriction dataset with Random Forest models. The Youden index and the selection of models performed before MBA summarization (BS) were chosen to compose the data presented in this graph, as both showed superior performance compared to other classification methods and model selection criteria tested. The dashed horizontal red line indicates a balanced accuracy of 0.5.

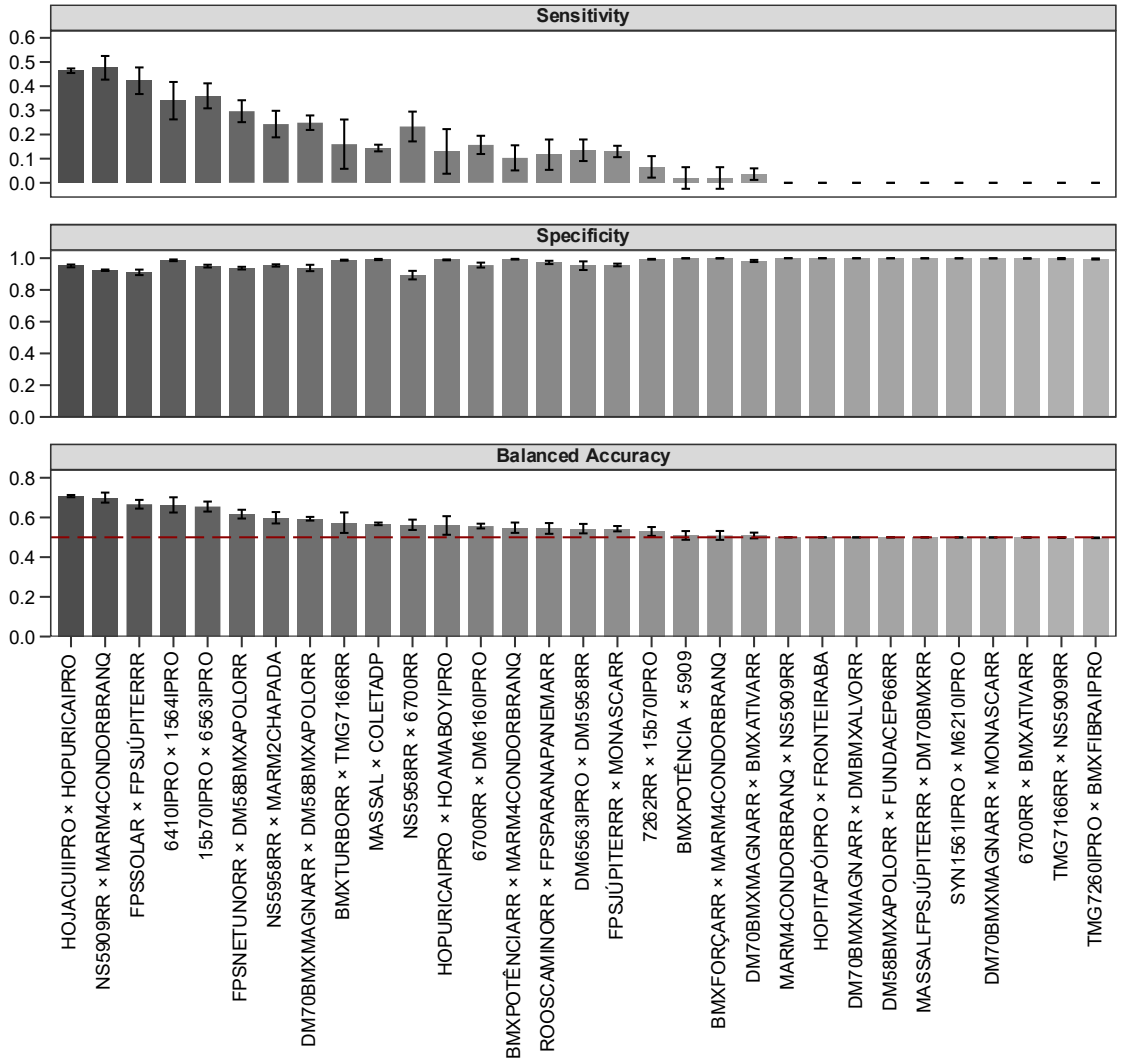

Figure 7. Average sensitivity (true positive rate), specificity (true negative rate), and mean balanced accuracy (MBA) obtained for each recombinant ancestor of soybean lines (*Glycine max* L.) using the 1×k restriction dataset with Random Forest models. The Youden index and the selection of models performed before MBA summarization (BS) were chosen to compose the data presented in this graph, as both showed superior performance compared to other classification methods and model selection criteria tested. The dashed horizontal red line indicates a balanced accuracy of 0.5.

The distinguishability of the genotypes observed and discussed earlier may have been affected by the low heterogeneity of the studied genotypes and the degree of imbalance in the dataset. In Random Forest's voting-based classification process, such low representation can lead to misclassifications. In this study, a scenario with reduced imbalance was analyzed by applying restrictions to the dataset, involving fewer genotypes per ancestry. The results of this analysis are presented in **Figure 8**.

The findings in **Figure 8** indicate that reducing the number of ancestries to those with greater representation in terms of the number of associated lines – resulting in a more balanced distribution of lines per ancestry – improved classification performance for the remaining ancestries. However, the increase in balanced accuracy should be interpreted cautiously, as the metric values for each class were derived from a much larger number of negative observations in the one-vs-all approach based on simplified confusion matrices.

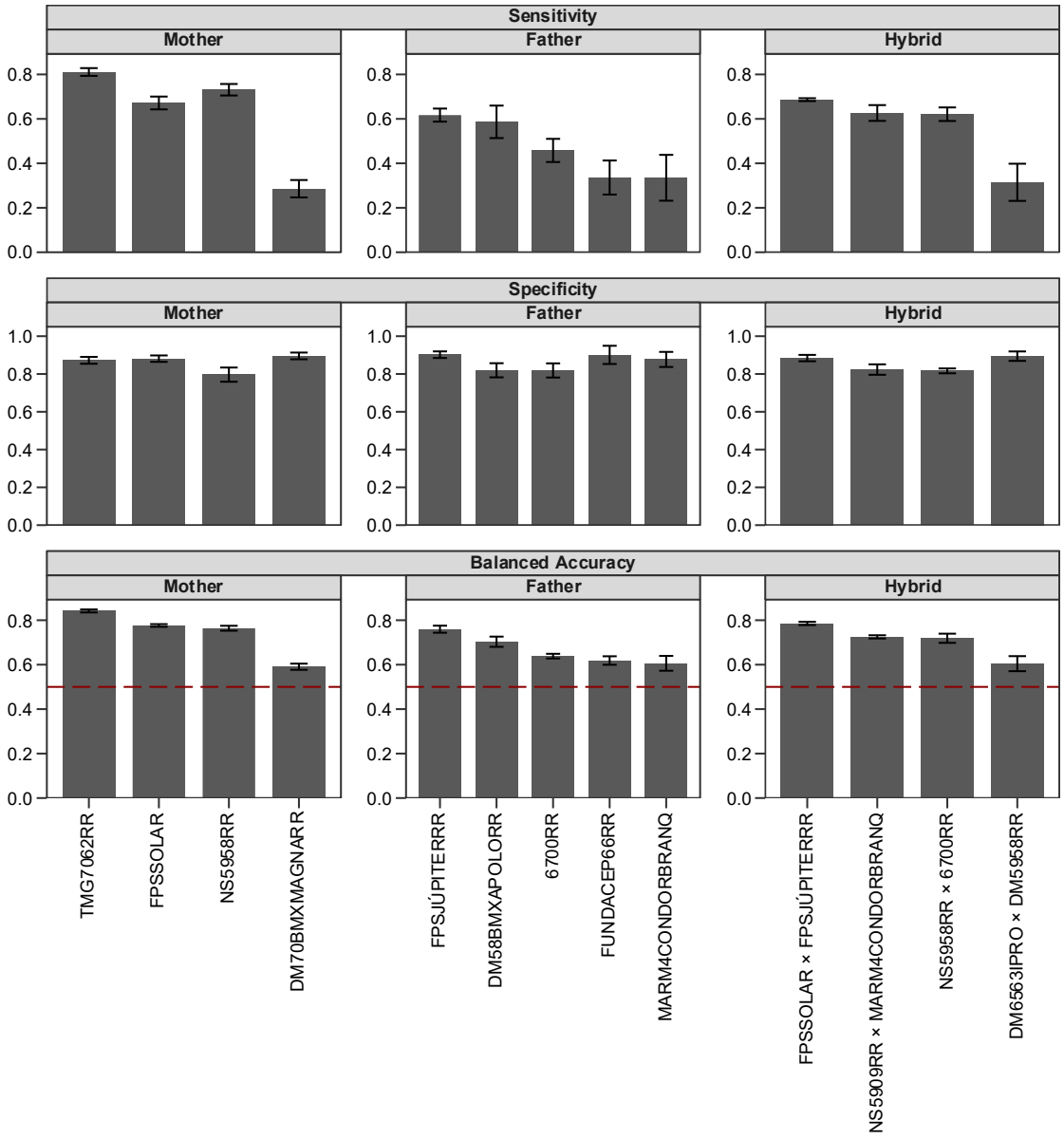

Figure 8. Average sensitivity (true positive rate), specificity (true negative rate), and mean balanced accuracy (MBA) for each maternal, paternal, or recombinant ancestor of soybean lines (*Glycine max* L.) obtained using the 15 × k restriction dataset with Random Forest models. The Youden index and the selection of models performed before MBA summarization (BS) were

chosen to compose the data presented in this graph, as both showed superior performance compared to other classification methods and model selection criteria tested. The dashed horizontal red line indicates a balanced accuracy of 0.5.

Finally, **Figures 9**, **10**, and **11** provide a detailed view of the classification of maternal, paternal, and recombinant ancestries using the Youden index to validate the best Random Forest models selected before MBA summarization. The maternal, paternal, and recombinant ancestries that achieved the highest accuracy rates were TMG7062RR, FPSJÚPITERRR, and FPSSOLAR × FPSJÚPITERRR, respectively. Conversely, the maternal, paternal, and recombinant ancestries that achieved the highest error rates were DM70BMXMAGNARR, FUNDACEP66RR, MARM4CONDORBRANQ, and DM6563IPRO × DM5958RR.

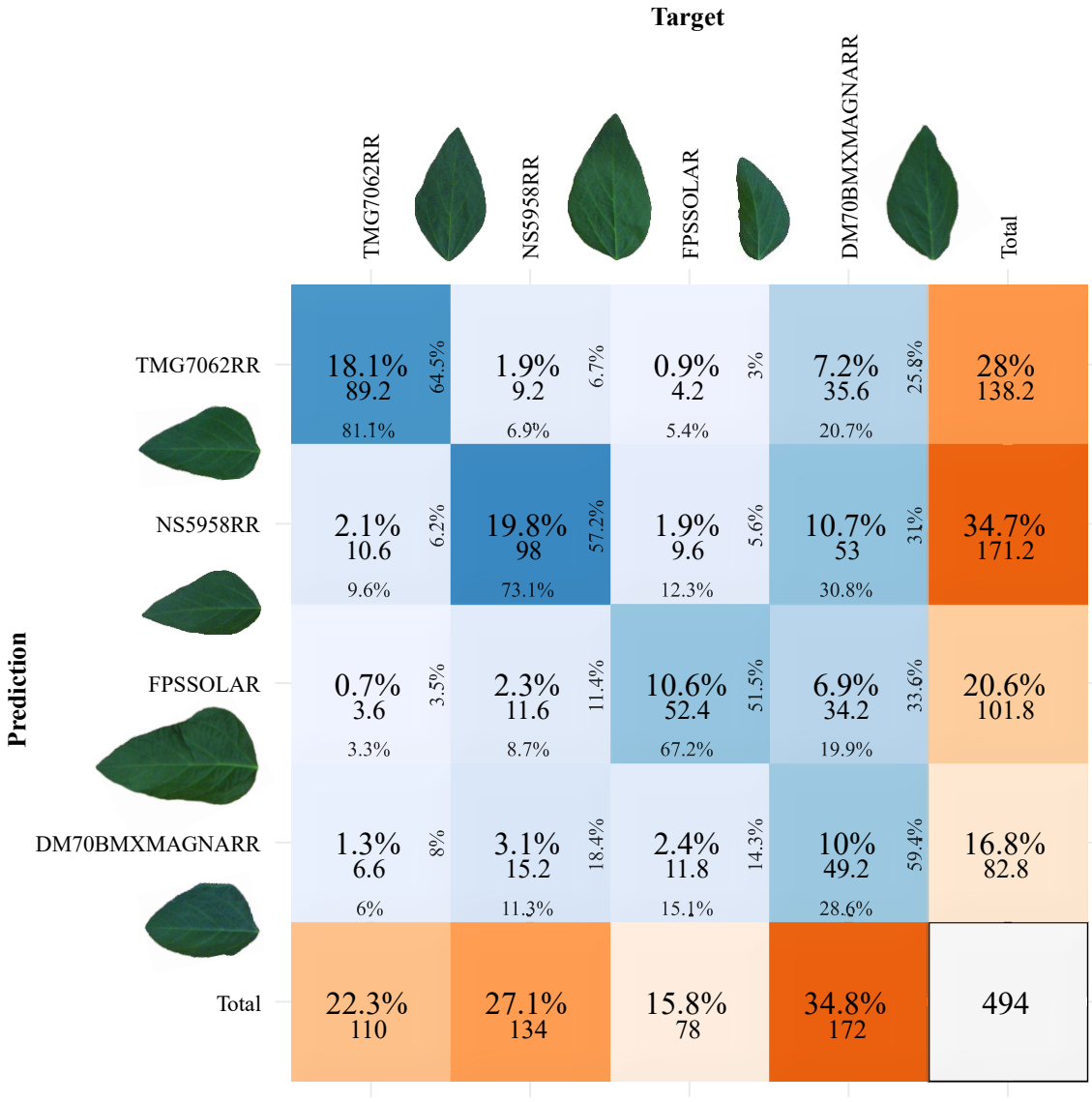

Figure 9. Confusion matrix presenting the validation of Random Forest models for classifying maternal ancestors of soybean lines (*Glycine max* L., Merr.) using the 15 × k restriction dataset. The Youden index and the selection of models performed before MBA summarization (BS) were chosen to compose the data for this figure, as both showed superior performance compared to other classification methods and selection criteria tested. The leaflets were randomly selected to represent the lines associated with each ancestor, positioned to the right or below the names of maternal ancestors (on the x-axis above and the y-axis to the left, respectively).

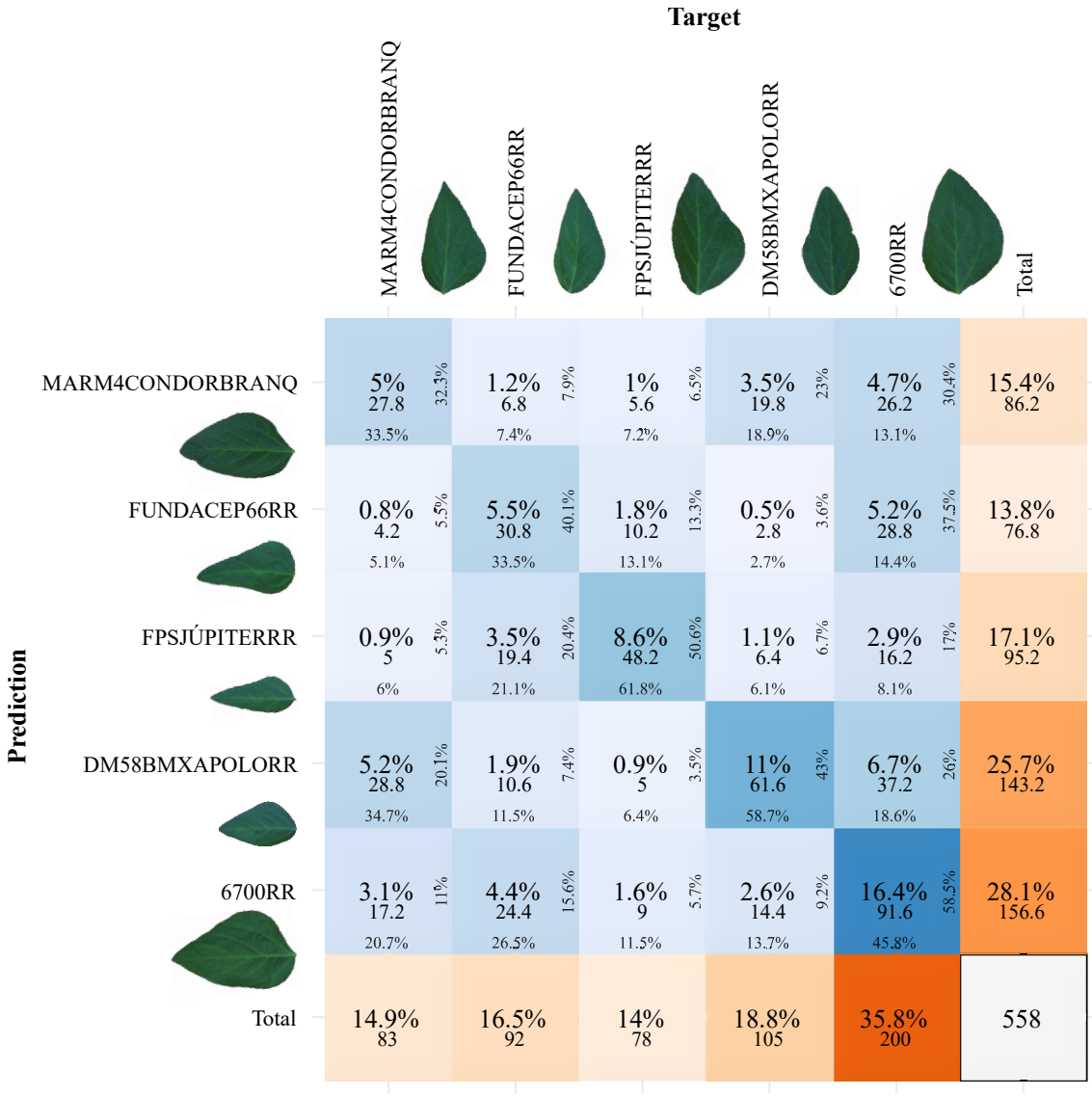

Figure 10. Confusion matrix presenting the validation of Random Forest models for classifying paternal ancestors of soybean lines (*Glycine max* L.) using the 15 × k restriction dataset. The Youden index and the selection of models performed before MBA summarization (BS) were chosen to compose the data for this figure, as both showed superior performance compared to other classification methods and selection criteria tested. The leaflets, positioned to the right or below the names of paternal ancestors (on the x-axis above and the y-axis to the left, respectively), were randomly selected to represent the lines associated with each ancestor.

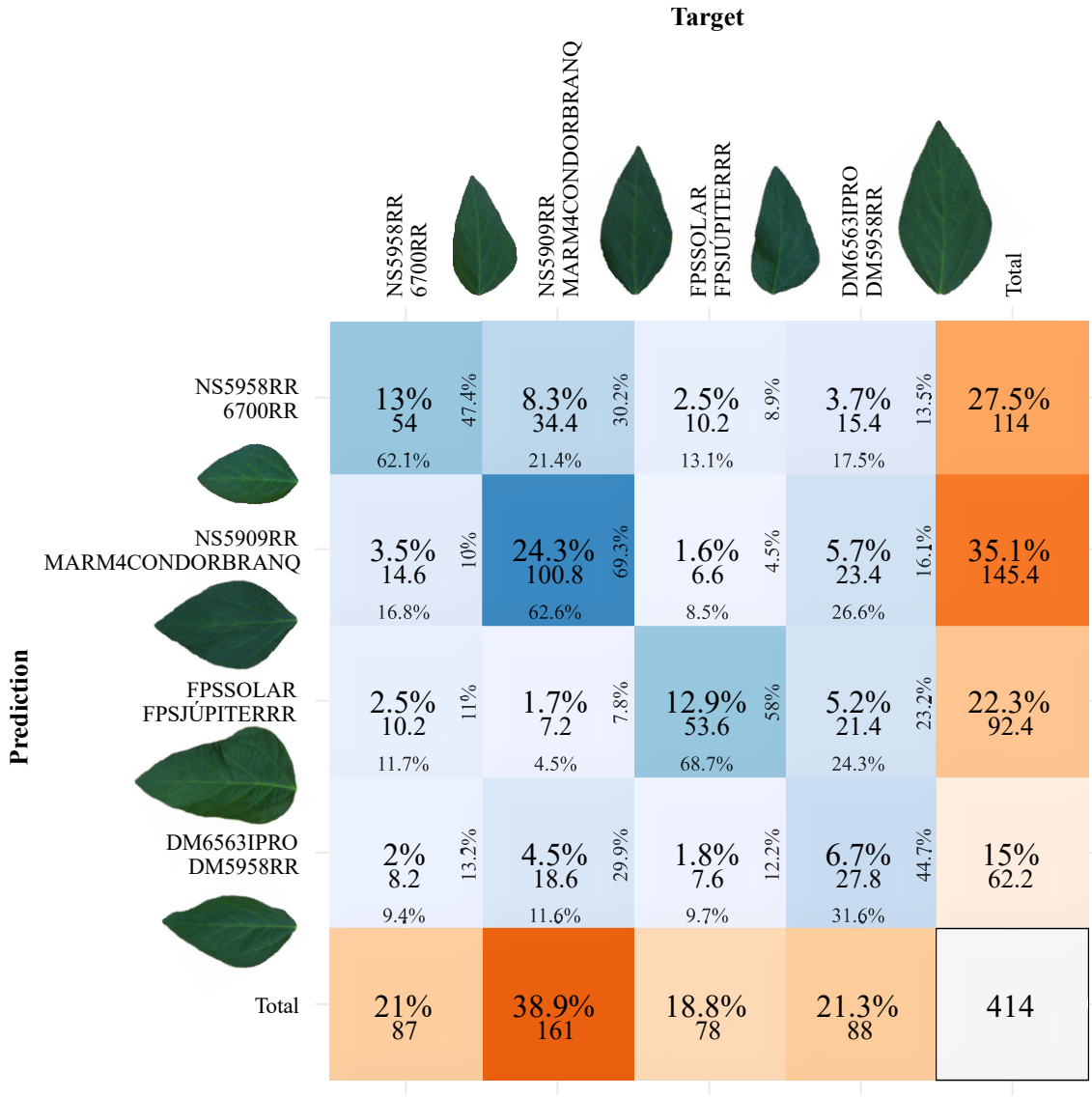

Figure 11. Confusion matrix presenting the validation of Random Forest models for classifying recombinant ancestors of soybean lines (*Glycine max* L.) using the 15 × k restriction dataset. The Youden index and the selection of models performed before MBA summarization (BS) were chosen to compose the data for this figure, as both showed superior performance compared to other classification methods and selection criteria tested. The leaflets, positioned to the right or below the names of recombinant ancestors (on the x-axis above and the y-axis to the left), were randomly selected to represent the lines associated with each ancestor.

CONCLUSIONS

- There is variability among genotypes that could be effectively discriminated and classified using the Random Forest machine learning algorithm.
- The germplasm used as maternal parents for the studied lineages demonstrated better discrimination compared to other ancestries.
- The degree of imbalance in the dataset significantly affects the representativeness of genotypic classes. Restricting the dataset to ensure more equitable representation yielded more favorable discrimination results.
- Comprehensive information extracted from images enhances discrimination efficacy. However, for leaflet-based discrimination, data on shape and texture should be prioritized.

REFERENCES

ALEMU, A. et al. Genomic selection in plant breeding: Key factors shaping two decades of progress. **Molecular Plant**, v. 17, n. 4, p. 552–578, 1 abr. 2024. Disponível em: <https://doi.org/10.1016/j.molp.2024.03.007>.

AMARAL, L. R. et al. Remote sensing imagery to predict soybean yield: a case study of vegetation indices contribution. **Precision Agriculture**, v. 25, n. 5, p. 2375–2393, 2024.

ANDERSON, E. J. et al. Soybean [*Glycine max* (L.) Merr.] breeding: History, improvement, production and future opportunities. In: AL-KHAYRI, J. M.; JAIN, S. M.; JOHNSON, D. V (Org.). . **Advances in Plant Breeding Strategies: Legumes: Volume 7**. Cham: Springer International Publishing, 2019. p. 431–516. Disponível em: <https://doi.org/10.1007/978-3-030-23400-3_12>.

BRASIL. **Lei nº 9.456, de 25 de Abril de 1997. Institui a Lei de Proteção de Cultivares e dá outras providências**. **Diário Oficial da União**. [S.l: s.n.]. , 1997

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética**. [S.l: s.n.], 2011. v. 620.

KOKLU, M.; CINAR, I.; TASPINAR, Y. S. Classification of rice varieties with deep learning methods. **Computers and Electronics in Agriculture**, v. 187, p. 106285, 2021. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168169921003021>.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>.

LIAW, A.; WIENER, M. Package “randomForest”: Breiman and Cutler’s random forests for classification and regression. **R Development Core Team**, v. 4, p. 6–10, 1 jan. 2014.

LIAW, A.; WIENER, M. randomForest: Breiman and Cutler's random forests for classification and regression. **R package version**, v. 4, p. 14, 2015.

MUHAMMAD, H. et al. Global impact of soybean production: A review. **Asian Journal of Biochemistry, Genetics and Molecular Biology**, v. 16, n. 2, p. 12–20, jan. 2024. Disponível em: <https://journalajbgmb.com/index.php/AJBGMB/article/view/357>.

NICOTRA, A. B. et al. Leaf shape linked to photosynthetic rates and temperature optima in South African Pelargonium species. **Oecologia**, v. 154, p. 625–635, 1 fev. 2008.

_____. The evolution and functional significance of leaf shape in the angiosperms. **Functional Plant Biology**, v. 38, n. 7, p. 535–552, 2011.

NOGUEIRA, A. P. O. et al. Novas características para diferenciação de cultivares de soja pela análise discriminante. **Ciência Rural**, v. 38, p. 2427–2433, 2008.

OLIVOTO, T. Lights, camera, pliman! An R package for plant image analysis. **Methods in Ecology and Evolution**, v. 13, n. 4, p. 789–798, 2022.

PERKINS, N. J.; SCHISTERMAN, E. F. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. **American Journal of Epidemiology**, v. 163, n. 7, p. 670–675, 1 abr. 2006. Disponível em: <https://doi.org/10.1093/aje/kwj063>.

PRASAD, A. M.; IVERSON, L. R.; LIAW, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. **Ecosystems**, v. 9, n. 2, p. 181–199, 2006.

R CORE TEAM. **R: A Language and environment for statistical computing**. . Vienna, Austria: [s.n.]. Disponível em: <https://www.R-project.org/>. , 2024

RAJU, P. P. C.; BALACHANDER, B.; NEEHARIKA, S. Comparison of Haralick Texture Features and Gray Level Run Length Matrix Features for Analyzing Textural Variation in Cotton Leaves to Identify Spot Disease. 2022, [S.l: s.n.], 2022. p. 1–17.

REN, T.; WERADUWAGE, S. M.; SHARKEY, T. D. Prospects for enhancing leaf photosynthetic capacity by manipulating mesophyll cell morphology. **Journal of Experimental Botany**, v. 70, n. 4, p. 1153–1165, 2019.

ROBIN, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. **BMC Bioinformatics**, v. 12, n. 1, p. 77, 2011. Disponível em: <https://doi.org/10.1186/1471-2105-12-77>.

ROWLAND, S. D. et al. Leaf shape is a predictor of fruit quality and cultivar performance in tomato. **new phytologist**, v. 226, n. 3, p. 851–865, 2020.

SAEED, A. et al. Genetic evaluation and breeding strategies under water deficit environment to develop the drought tolerant wheat germplasm. **Polish Journal of Environmental Studies**, 2024. Disponível em: <https://doi.org/10.15244/pjoes/188056>.

SANTANA, D. C. et al. Machine learning in the classification of soybean genotypes for primary macronutrients' content using UAV–multispectral sensor. **Remote Sensing**, v. 15, n. 5, p. 1457, 2023.

SHILPASHREE, N. et al. Morphological characterization, variability and diversity among vegetable soybean (Glycine max L.) genotypes. **Plants**, v. 10, n. 4, p. 671, 2021.

SINGH, S. et al. Assessing genetic variability in taramira (*Eruca sativa* Mill.) germplasm for enhanced breeding strategies. **International Journal of Economic Plants**, v. 11, n. Feb, 1, p. 18–25, 2024.

TANG, J.; HENDERSON, A.; GARDNER, P. Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced data sets. **Analyst**, v. 146, n. 19, p. 5880–5891, 2021.

VAN DEN BRAND, T. **ggh4x: Hacks for "ggplot2"**. [S.l: s.n.]. Disponível em: <https://github.com/teunbrand/ggh4x>. , 2024

VAUGHAN, D.; DANCHO, M. **furrr: Apply Mapping Functions in Parallel using Futures**. [S.l: s.n.]. Disponível em: <https://CRAN.R-project.org/package=furrr>. 2022

VISCOSI, V.; FORTINI, P. Leaf shape variation and differentiation in three sympatric white oak species revealed by elliptic Fourier analysis. **Nordic Journal of Botany**, v. 29, n. 5, p. 632–640, 1 out. 2011. Disponível em: <https://doi.org/10.1111/j.1756-1051.2011.01098.x>.

WANG, L. et al. Review of classification methods on unbalanced data sets. **Ieee Access**, v. 9, p. 64606–64628, 2021.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. [S.l.]: Springer-Verlag New York, 2016. Disponível em: <https://ggplot2.tidyverse.org>.

_____. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.

WICKHAM, H.; HENRY, L. **purrr: Functional Programming Tools**. . [S.l: s.n.]. Disponível em: <https://CRAN.R-project.org/package=purrr>. , 2023

WILKE, C. O.; WIERNIK, B. M. **ggtext: Improved text rendering support for "ggplot2"**. [S.l: s.n.]. Disponível em: <https://CRAN.R-project.org/package=ggtext>. , 2022

YADAVA, D. K. et al. **Fundamentals of field crop breeding**. [S.l.]: Springer, 2022.

YOUDEN, W. J. Index for rating diagnostic tests. **Cancer**, v. 3, n. 1, p. 32–35, 1 jan. 1950. Disponível em: <https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO>.

SUPPLEMENTARY MATERIAL

Supplementary Table S1. Utility of paternal parents in the dataset based on the $1 \times k$ and $15 \times k$ restrictions imposed for performing k-fold cross-validation using Random Forest models.

Count	Father	Number of lines	Utility in the First Restriction (Minimum of 5 Lines)	Utility in the Second Restriction (Minimum of 75 Lines)
1	6700RR	200	Yes	Yes
2	DM58BMXAPOLORR	105	Yes	Yes
3	FUNDACEP66RR	92	Yes	Yes
4	MARM4CONDORBRANQ	83	Yes	Yes
5	FPSJÚPITERRR	78	Yes	Yes
6	6563IPRO	67	Yes	No
7	HOPURICAIPRO	59	Yes	No
8	MONASCARR	55	Yes	No
9	DM6160IPRO	49	Yes	No
10	MARM2CHAPADA	47	Yes	No
11	FPSPARANAPANEMARR	44	Yes	No
12	BMXATIVARR	34	Yes	No
13	COLETADP	26	Yes	No
14	TMG7166RR	25	Yes	No
15	1564IPRO	24	Yes	No
16	15b70IPRO	21	Yes	No
17	HOAMABOYIPRO	18	Yes	No
18	NS5909RR	15	Yes	No
19	BMXFIBRAIPRO	12	Yes	No
20	M6210IPRO	8	Yes	No
21	5909	7	Yes	No
22	FRONTEIRABA	5	Yes	No
23	DMBMXALVORR	5	Yes	No
24	DM70BMXRR	5	Yes	No
25	COLETACHAPADALL	4	No	No
26	DM5958RR	3	No	No
27	COLETATENENTEPORTELA	3	No	No
28	FPSURANORR	2	No	No
29	MARM3VIDENTEDUTRA	2	No	No
30	ROTA54IPRORR	2	No	No
31	TMG	1	No	No
32	COLETAVICENTEDUTRA	1	No	No
33	M5892	1	No	No
34	NS5958RR	1	No	No
35	FUNDACEP65RR	1	No	No

Supplementary Table S2. Utility of maternal parents in the dataset based on the $1 \times k$ and $15 \times k$ restrictions imposed for performing k-fold cross-validation using Random Forest models.

Count	Mother	Number of lines	Utility in the First Restriction (Minimum of 5 Lines)	Utility in the Second Restriction (Minimum of 75 Lines)
1	DM70BMXMAGNARR	172	Yes	Yes
2	NS5958RR	134	Yes	Yes
3	TMG7062RR	110	Yes	Yes
4	FPSSOLAR	78	Yes	Yes
5	15b70IPRO	65	Yes	No
6	FPSNETUNORR	60	Yes	No
7	HOJACUIIPRO	59	Yes	No
8	6700RR	57	Yes	No
9	NS5909RR	51	Yes	No
10	FPSJÚPITERRR	49	Yes	No
11	ROOSCAMINORR	44	Yes	No
12	MASSAL	34	Yes	No
13	BMXTURBORR	25	Yes	No
14	6410IPRO	24	Yes	No
15	7262RR	21	Yes	No
16	BMXPOTÊNCIARR	18	Yes	No
17	HOPURICAIPRO	18	Yes	No
18	DM58BMXAPOLORR	16	Yes	No
19	TMG7260IPRO	12	Yes	No
20	TMG7166RR	10	Yes	No
21	SYN1561IPRO	8	Yes	No
22	BMXPOTÊNCIA	7	Yes	No
23	BMXFORÇARR	7	Yes	No
24	MARM4CONDORBRANQ	5	Yes	No
25	HOPITAPÓIPRO	5	Yes	No
26	MASSALFPSJÚPITERRR	5	Yes	No
27	DM6563IPRO	3	No	No
28	M6210IPRO	2	No	No
29	4823RR	2	No	No
30	M3	1	No	No
31	NS6209	1	No	No
32	TMG7262RR	1	No	No
33	FPSPARANAPANEMARR	1	No	No

Supplementary Table S3. Utility of recombinants in the dataset based on the $1 \times k$ and $15 \times k$ restrictions imposed for performing k-fold cross-validation using Random Forest models.

Count	Recombinants	Number of lines	Utility in the First Restriction (Minimum of 5 Lines)	Utility in the Second Restriction (Minimum of 75 Lines)
1	NS5909RR × MARM4CONDORBRANQ	161	Yes	Yes
2	DM6563IPRO × DM5958RR	88	Yes	Yes
3	NS5958RR × 6700RR	87	Yes	Yes
4	FPSSOLAR × FPSJÚPITERRR	78	Yes	Yes
5	15b70IPRO × 6563IPRO	65	Yes	No
6	HOJACUIIPRO × HOPURICAIPRO	59	Yes	No
7	FPSNETUNORR × DM58BMXAPOLORR	57	Yes	No
8	6700RR × DM6160IPRO	56	Yes	No
9	FPSJÚPITERRR × MONASCARR	49	Yes	No
10	DM70BMXMAGNARR × DM58BMXAPOLORR	48	Yes	No
11	NS5958RR × MARM2CHAPADA	47	Yes	No
12	ROOSCAMINORR × FPSPARANAPANEMARR	46	Yes	No
13	DM70BMXMAGNARR × BMXATIVARR	28	Yes	No
14	MASSAL × COLETADP	26	Yes	No
15	BMXTURBORR × TMG7166RR	25	Yes	No
16	6410IPRO × 1564IPRO	24	Yes	No
17	7262RR × 15b70IPRO	21	Yes	No
18	BMXPOTÊNCIARR × MARM4CONDORBRANQ	18	Yes	No
19	HOPURICAIPRO × HOAMABOYIPRO	18	Yes	No
20	TMG7260IPRO × BMXFIBRAIPRO	12	Yes	No
21	TMG7166RR × NS5909RR	10	Yes	No
22	SYN1561IPRO × M6210IPRO	8	Yes	No
23	BMXPOTÊNCIA × 5909	7	Yes	No
24	BMXFORÇARR × MARM4CONDORBRANQ	7	Yes	No
25	DM58BMXAPOLORR × FUNDACEP66RR	7	Yes	No
26	DM70BMXMAGNARR × MONASCARR	6	Yes	No
27	6700RR × BMXATIVARR	6	Yes	No
28	MARM4CONDORBRANQ × NS5909RR	5	Yes	No
29	HOPITAPÓIPRO × FRONTEIRABA	5	Yes	No
30	DM70BMXMAGNARR × DMBMXALVORR	5	Yes	No
31	MASSALFPSJÚPITERRR × DM70BMXRR	5	Yes	No
32	MASSAL × COLETACHAPADALL	4	No	No
33	MASSAL × COLETATENENTEPORTELA	3	No	No
34	FPSNETUNORR × 6700RR	3	No	No
35	DM58BMXAPOLORR × FPSURANORR	2	No	No
36	6700RR × MARM3VIDENTEDUTRA	2	No	No
37	4823RR × ROTA54IPRORR	2	No	No
38	M3 × TMG	1	No	No
39	MASSAL × COLETAVICENTEDUTRA	1	No	No
40	NS6209 × M5892	1	No	No
41	TMG7262RR × NS5958RR	1	No	No
42	FPSPARANAPANEMARR × FUNDACEP65RR	1	No	No

Supplementary Table S4. Counting of the different variable sets used to configure the Random Forest models selected before the summarization (BS) of the mean balanced accuracy (MBA).

Model Selection*	Ancestry	Restriction	Classification Method	Variable sets	N**
BS	Maternal	001	MAP	Shape	6
BS	Maternal	001	MAP	Texture	1
BS	Maternal	001	MAP	All	18
BS	Maternal	001	Youden	Shape	2
BS	Maternal	001	Youden	All	23
BS	Maternal	001	Closest top left	Shape	3
BS	Maternal	001	Closest top left	All	22
BS	Maternal	015	MAP	Shape	2
BS	Maternal	015	MAP	Texture	1
BS	Maternal	015	MAP	All	22
BS	Maternal	015	Youden	Texture	1
BS	Maternal	015	Youden	All	24
BS	Maternal	015	Closest top left	Texture	3
BS	Maternal	015	Closest top left	All	22
BS	Paternal	001	MAP	Shape	7
BS	Paternal	001	MAP	All	18
BS	Paternal	001	Youden	Shape	1
BS	Paternal	001	Youden	All	24
BS	Paternal	001	Closest top left	Shape	4
BS	Paternal	001	Closest top left	All	21
BS	Paternal	015	MAP	Shape	4
BS	Paternal	015	MAP	Texture	1
BS	Paternal	015	MAP	All	20
BS	Paternal	015	Youden	Shape	3
BS	Paternal	015	Youden	Texture	1
BS	Paternal	015	Youden	All	21
BS	Paternal	015	Closest top left	Shape	3
BS	Paternal	015	Closest top left	Texture	1
BS	Paternal	015	Closest top left	All	21
BS	Recombinants	001	MAP	Shape	6
BS	Recombinants	001	MAP	All	19
BS	Recombinants	001	Youden	Shape	4
BS	Recombinants	001	Youden	All	21
BS	Recombinants	001	Closest top left	Shape	5
BS	Recombinants	001	Closest top left	All	20
BS	Recombinants	015	MAP	Shape	1
BS	Recombinants	015	MAP	Texture	3
BS	Recombinants	015	MAP	All	21
BS	Recombinants	015	Youden	Shape	1
BS	Recombinants	015	Youden	Texture	2
BS	Recombinants	015	Youden	All	22
BS	Recombinants	015	Closest top left	Texture	2
BS	Recombinants	015	Closest top left	All	23

* Only the datasets formed by "All" variables were used to configure the models selected after summarization (AS) of the mean balanced accuracy (MBA). ** The variable sets formed by vegetative indices (VIs) and Fourier descriptors did not configure any of the best models selected.

GENERAL CONCLUSIONS

This study highlights the power and versatility of digital phenotyping and machine learning in soybean leaflet analysis, offering practical contributions to genotypic discrimination and plant breeding.

The methods established in Chapter 1 provided comprehensive guidelines, rationales, and scripts for extracting diverse phenotypic traits from RGB images including vegetative indices (VIs), shape measures, Fourier elliptical descriptors (EFDs), and texture characteristics. These procedures, applied systematically to all available images, demonstrated robustness, flexibility, and potential for adaptation across various research contexts.

In Chapter 2, phenomic data analysis revealed distinct contributions of each dataset. Fourier descriptors exhibited the least redundancy among all datasets. The identification of five genotypic patterns, with more than 50% concordance across datasets, reinforces the value of phenomic analysis in genetic improvement studies, aiding in genotype pattern recognition and the identification of critical predictive and classificatory attributes.

Chapter 3 demonstrated the efficacy of the Random Forest machine learning algorithm for genotype discrimination and classification, especially when the dataset was balanced to address representational biases. The germplasm used as maternal parents for the studied lineages exhibited superior discriminatory potential. In addition, shape and texture attributes contribute to the best leaflet-based genotypic differentiation.

These findings collectively underscore the transformative role of computational tools and machine learning in advancing phenotypic studies, providing a robust framework for future research in genotype identification, cultivar characterization, and plant breeding.