Mathematical and Computational Applications

MDPI

Article

Classification of Common Bean Landraces of Three Species Using a Neuroevolutionary Approach with Probabilistic Color Characterization

José-Luis Morales-Reyes [1,*], Elia-Nora Aquino-Bolaños [1], Héctor-Gabriel Acosta-Mesa [2], Nancy Pérez-Castro [3] and José-Luis Chavez-Servia [4]

[1] Centre for Food Research and Development, University of Veracruz, Xalapa 91190, Mexico; eliaquino@uv.mx
[2] Artificial Intelligence Research Institute, University of Veracruz, Xalapa 91097, Mexico; heacosta@uv.mx
[3] Agroengineering Institute, University of Papaloapan, Loma Bonita 68400, Mexico; nperez@unpa.edu.mx
[4] National Polytechnic Institute, Hornos 1003, Santa Cruz Xoxocotlán, Oaxaca 71230, Mexico; jchavezs@ipn.mx
* Correspondence: jluismorey@hotmail.com

Abstract: The common bean is a widely cultivated food source. Many domesticated species of common bean varieties, known as landraces, are cultivated in Mexico by local farmers, exhibiting various colorations and seed mixtures as part of agricultural practices. In this work, we propose a methodology for classifying bean landrace samples using three two-dimensional histograms with data in the CIE L*a*b* color space while additionally integrating chroma (C*) and hue (h°) to develop a new proposal from histograms, employing deep learning for the classification task. The results indicate that utilizing three histograms based on L*, C*, and h° brings an average accuracy of 85.74 ± 2.37 compared to three histograms using L*, a*, and b*, which reported an average accuracy of 82.22 ± 2.84. In conclusion, the new color characterization approach presents a viable solution for classifying common bean landraces of both homogeneous and heterogeneous colors.

Keywords: common bean landraces; classification; CNN; histograms; heterogenous color; neuroevolution

check for updates

Academic Editor: Gianluigi Rozza

Received: 15 April 2025
Revised: 13 June 2025
Accepted: 17 June 2025
Published: 19 June 2025

Citation: Morales-Reyes, J.-L.; Aquino-Bolaños, E.-N.; Acosta-Mesa, H.-G.; Pérez-Castro, N.; Chavez-Servia, J.-L. Classification of Common Bean Landraces of Three Species Using a Neuroevolutionary Approach with Probabilistic Color Characterization. *Math. Comput. Appl.* **2025**, *30*, 66. https://doi.org/10.3390/mca30030066

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The common bean (*Phaseolus vulgaris* L.) is an essential crop for food security and the rural economy in Mexico and worldwide. According to information from the Agri-Food and Fisheries Information Service (SIAP) [1], after a sharp drop in production in 2023 due to drought conditions, a significant recovery was observed in 2024, reaching nearly 856,000 tons. This rebound highlights the need to value and conserve the genetic diversity of this legume, especially through local varieties such as the common bean landrace, the lima bean (*Phaseolus lunatus* L.), and the ayocote bean (*Phaseolus coccineus* L.), which represent a biocultural heritage of great importance. They are significant global crops, and due to their high protein and carbohydrate content, they constitute essential food sources in human nutrition [2].

One of the most remarkable characteristics of these species is the diversity of seed colors, which has aesthetic implications and correlates with nutritional and functional properties. The seeds of *Phaseolus vulgaris* L. exhibit different hues, from white and yellow to black and red, which influence their commercial acceptance and culinary applications. Similarly, *Phaseolus lunatus* L. seeds typically exhibit white, cream, or green coloration with mottled patterns [3].

Common bean landrace species are prized for their drought tolerance and adaptability to warm environments, rendering them well suited for water-scarce regions [4]. Thus, the task of classifying *Phaseolus* species is crucial for biodiversity preservation and agricultural advancement, as it enables the selection of resilient and nutritious varieties [5].

Traditionally, human experts have classified *Phaseolus* species using morphological methods, which rely on visible characteristics that can vary due to environmental factors. Additionally, molecular analyses have been employed, but these are often costly, require specialized equipment, and can be invasive. Faced with these disadvantages, using digital images to assess seed color is emerging as a non-destructive, rapid, and accessible alternative. Although this technique also faces challenges related to color variability and lighting conditions, recent research has demonstrated its effectiveness and accuracy in *Phaseolus* seed classification, highlighting the potential of artificial intelligence to automate the process.

A review of the existing literature, with a special emphasis on the incorporation of artificial intelligence techniques, provides insight into the advances and limitations of image-based classification methods for Phaseolus species and other seed species.

The literature review was organized according to the number of feature classes extracted. Table 1 presents these and the respective descriptors employed in the proposals derived from the reviewed articles. Studies focus on classifying beans and other seeds. They analyze features like color, shape, texture, and size. The approaches are described from three classes to one class.

Table 1. Features and descriptors found in reviewed works.

Feature Class	Descriptors (Identifier)
Size (S)	Area (S1), Perimeter (S2), Diameter (S3), Major and Minor Axis Lengths (S4), Convexity (S5), Extent (S6), Solidity (S7), Roundness (S8), Biomass (S9), Height (S10), Width (S11), Weight (S12), Thickness (S13), Entropy (S14).
Shape (H)	Shape Factors (SF1, SF2, SF3, SF4) (H1), Compactness (H2), Roundness (H3), Solidity (H4), Shape Index (H5), Projected Volume (H6), Elliptic Fourier Descriptors (EFDs) (H7), Seed Form (H8).
Color (C)	Mean (C1), Variance (C2), Skewness (C3), Kurtosis (C4) (in RGB, HSV, L*a*b*, or HIS Color Space), 2D Histograms using PMF (C5), Predominant and Secondary Colors (C6), Brightness (C7), 3D Histograms and Color Information Reduction via Gaussian Mixture Model (GMM) (C8).
Texture (T)	Scale-Invariant Feature Transform (SIFT) (T1), Speeded-Up Robust Features (SURF) (T2), Root Scale-Invariant Feature Transform (RootSIFT) (T3), Haralick-Contrast (T4), Correlation (T5), Energy (T6), Homogeneity (T7), Histogram of Curvature (HoC) (T8), Bag-of-Visual Words (BoW) (T9), LBP-Contrast (T10), Correlation (T11), Energy (T12), Homogeneity (T13), Entropy (T14), Gabor Filter-Mean (T15), Standard Deviation (T16), Gray Level Co-occurrence Matrix (GLCM)-Dissimilarity (T17), Correlation (T18), Contrast (T19), Homogeneity (T20), Angular Second Moment (ASM) (T21).

Several works have explored three-class feature extraction using different types of descriptors. Lo Bianco et al. [6] analyzed seed images to classify 67 types of Italian beans based on size, shape, and texture features. They extracted 138 features from digital images and applied stepwise Linear Discriminant Analysis (LDA) on morphological characteristics. The results showed identification accuracies between 94.3% and 99.7% for different color categories. Furthermore, the system successfully identified the cropping region with over

88% accuracy and differentiated batches of seeds from the same variety grown under different practices with 100% accuracy. However, the authors noted the need for further validation in varying growing conditions and the expansion of the feature set to improve classification performance.

A more advanced approach was presented by Yi et al. [7], who used a multi-kernel SVM to classify seeds based on shape, texture, and color. Their approach achieved 97% accuracy, surpassing the 90% expectation. This study highlighted the importance of texture features in classification.

Garcia et al. [8] developed a computer vision and machine learning system to classify beans. They created a protocol using a Nikon D3300 camera (Nikon Corporation, Tokyo, Japan) and a photo studio box, which improved image quality and reduced shadows. The method employed Otsu thresholding, distance transform, and watershed techniques. A Random Forest algorithm trained on seven attributes achieved an accuracy of 98.5% and an F1-score of 0.98. While the model performed well overall, it struggled slightly with Cargamanto Rojo and MBC 46 due to their visual similarity. The system outperformed contemporary methods, although it was limited to six bean varieties.

The most common approach in digital image seed classification studies is combining two kinds of features. Specifically, five of the seven reviewed papers employ color together with size, shape, or texture, highlighting its importance in this type of analysis, while the other two focus on extracting size and shape.

The work of Kılıc et al. [9] was among the earliest proposals, where they developed a computer vision system that could measure and classify bean samples with an accuracy of 90.6%. The computer vision system was able to identify high-quality white beans 99.3% of the time, outperforming human inspection due to its superior performance, reproducibility, and objectivity. However, factors such as sample size, placement, misclassification, and optimization need to be considered to achieve more accurate size measurements.

Ozktan et al. [10] presented a methodology that employs image processing and machine learning to predict and differentiate 20 common bean genotypes based on their physical attributes, capturing images of the bean seeds and measuring their color parameters. Four machine learning algorithms were applied to predict seed mass. The Random Forest (RF) algorithm predicted bean mass accurately, and the first two principal components explained 73% of the variation in the physical attributes. Hierarchical clustering grouped similar genotypes effectively.

Studies have also explored seed classification by focusing solely on extracting color and texture features, as seen in the works of Mendoza et al. [11], Eric et al. [12], and Djoulde et al. [13].

Mendoza et al. [11] developed a machine vision system for inspecting and predicting the color and quality of canned beans. They extracted color and texture features from images of beans and brine, calibrating RGB values to CIE color spaces. The system showed a strong predictability, with a 0.937 correlation for color and 0.871 for appearance. It correctly grouped 89.7% of the samples.

Similarly, Eric et al. [12] focused on improving cocoa bean classification by extracting color and texture features. An Artificial Neural Network model was trained, and its parameters were fine-tuned using Bayesian optimization. L1L2 regularization, early stopping, and reduced model complexity were applied, resulting in an accuracy of 85.36%.

In contrast, Djoulde et al. [13] used color filter array (CFA) images and an SVM model to classify Penja pepper seeds, achieving 87% accuracy. Images were captured using a custom lightbox and converted to 16-bit grayscale. A broader set of texture features was extracted. It was also found that raw CFA images outperformed demosaiced images in classification accuracy.

Some studies have focused solely on extracting size and shape features for seed classification. For example, Koklu and Özcan [14] used a computer vision system with machine learning to classify seven types of dry beans, extracting 12 shape features from 13,611 images. The SVM model achieved 93.13% accuracy because the color was not distinctive. The major axis length, shape, and suture axis were key contributors, but excluding the suture axis is a limitation. Similarly, Mendigoria et al. [15] employed machine learning to predict and classify dry bean shapes, extracting 16 shape-related features. Their findings indicate that the KNN7 model attained the highest accuracy (93.69%), while GPR performed best in predicting solidity, and RT performed best for roundness and compactness. Both studies demonstrate the effectiveness of size and shape descriptors for bean classification, although differences are observed in the feature sets and modeling approaches.

Several studies have focused exclusively on extracting color features for classification tasks. Early, the work of Nasirahmadi and Behroozi-Khazaei [16] utilized a machine vision system and an Artificial Neural Network (ANN) to classify ten types of beans based on their color attributes, achieving an overall classification accuracy of over 96%. However, due to shared color characteristics, lower accuracy was observed for certain bean types, such as Sarab1 and KS21108. In comparison, Morales-Reyes et al. [17] introduced a computer vision system based on color histograms to classify common bean landraces, achieving an average classification accuracy of 80%. Their method outperformed other color-based techniques, such as spectrophotometer measurements (68.24%) and color averages (CIE L*a*b* at 53.80% and RGB at 44.44%). The study demonstrated that color histograms, particularly 2D and 3D versions, provided a more accurate representation of the color distributions in beans.

Morales-Reyes et al. [18] continued their research with further advancements. They developed a methodology to estimate anthocyanin concentrations in common beans using digital image processing combined with a neuroevolutionary approach. This method incorporated a region-growing algorithm for image segmentation and optimized a Convolutional Neural Network (CNN) using Deep Genetic Algorithm (DeepGA). The results showed over 93% precision in estimating anthocyanin concentrations, particularly when three two-dimensional histograms were used and the luminosity channel was incorporated. Subsequently, the authors [19] refined the technique for estimating anthocyanins, highlighting the challenges of estimating anthocyanin concentrations in beans with heterogeneous coloration. They found that combining 2D histograms and the luminosity channel enhanced precision, offering a promising approach for more accurate estimation in diverse bean landraces.

Sonmez et al. [20] also employed machine learning for classification, but in this case, wheat varieties and hybrids. The study achieved over 99% accuracy using reduced color features extracted from images in different color spaces, although the Naive Bayes algorithm performed somewhat lower (~96%). The results demonstrated the potential of color-based classification for grain identification, though the focus was on wheat rather than beans.

Recently, López-Lobato et al. [21] used a data set of 40 bean landraces from rural Oaxaca, Mexico. Each landrace was represented by a 60 g sample, and its color distribution was captured as a 3D histogram in the CIE L*a*b* color space, resulting in 400 histograms, 10 histograms by sample. The researchers proposed a classification approach using Gaussian mixture models fitted to the histogram data, with the Gini Index algorithm determining the number of Gaussian components. The means of these components were then used to identify primary colors, and classification was performed using the K-Nearest Neighbors algorithm. While the method achieved 100% accuracy in broad color classification, its performance declined to 53.33–80% when distinguishing specific local varieties, particularly black-colored landraces.

Table 2 categorizes and details key aspects of the studies conducted, highlighting the features, descriptors, type of seed, machine learning algorithms, color space, and image acquisition of each work.

Table 2. Detailed description of articles revised.

Authors	Features	Descriptors	Type of Seed	ML Algorithms	Color Space	Image Acquisition
Lo Bianco et al. [6]	S, H, T	C1, S12, S14, T1, T4, T5, T6, T7	Beans (*Phaseolus vulgaris* L.)	LDA and SVM	Gray scale	Flatbed scanner
Yi et al. [7]	C, H, T	C5, C2, C3, C1, H8, H9, T1, T2, T3	Different seed species	NB and SVM	CIE L*a*b*	Motorized multi-purpose zoom microscope
Garcia et al. [8]	C, H, S	H8, C6, C7, S1, S2, S3, S4	Beans (*Phaseolus vulgaris* L.)	RF	RGB	Photo-eBox Plus (PeP)
Kılıc et al. [9]	S, C	C1, C2, C3, C4	Beans	ANN	RGB	Custom-built light black box image capturing
Ozaktan et al. [10]	H, C	S1, S2, S3, S4, S5, S6, S7, H1, H2, S8, H3, H4	Common bean genotypes from Türkiye	MLP, RF, SVR, and KNN	CIE L*a*b*	Controlled illumination and a digital camera
Mendoza et al. [11]	C, T	C1, C2, C3, C4, T4, T5, T6, T7	Bean (*Phaseolus vulgaris* L.)	Non-linear SVM	RGB, CIE L*a*b*, HSV, and gray scale	Controlled lighting system
Eric et al. [12]		C1, C2, C3, C4, C5, C6, T8, T4, T5, T6	Cocoa beans	SVM, RF, DT, NB, and ANN	RGB	Not mentioned
Djoulde et al. [13]		S6, S1, S10, S12, S11, S7, T10, T4, T5, T6, T7, T15, T16, T17, T18, T19, T20, T21	Pepper seeds	KNN, SGD, SVM, and RF	RGB	Custom-built light box
Koklu and Ozkan [14]	S, H	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, H1, H2, H3, H4	Turkish beans	SVM, DT, MLP, and KNN	RGB, gray scale	Custom-built illumination box
Mendigoria et al. [15]		S1, S2, S3, S4, S5, S6, S7, S8, S9, H1, H2, H3, H4, H5, H6, H7	Bean (*Phaseolus vulgaris* L.)	LDA, KNN, DT, and NB	CIE L*a*b*	The Prosilica GT2000C camera
Nasirahmadi and Behroozi-Khazaei [16]	C	C1	Iranian bean	MLP-ANN	RGB	Controlled lighting
Morales-Reyes et al. [17–19]		C5	Bean (*Phaseolus vulgaris* L.)	KNN	RGB, HIS, and CIE L*a*b*	Custom-built light and diffuser box
Sönmez et al. [20]		C1, C2, C3, C4	Durum wheat seed	ANN, SVM, KNN, DT, RF, and NB	RGB, HSV, CIE L*a* b*, and YCrCb	Custom-built light black box
López-Lobato et al. [21]		C8	Bean (*Phaseolus vulgaris* L.)	KNN	CIE L*a* b*	NA

Acronyms: Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Artificial Neural Network (ANN), Multilayer Perceptron (MLP), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Non-linear Support Vector Machine (Non-linear SVM), Decision Tree (DT), and Stochastic Gradient Descent (SGD). The different color spaces considered are red, green, and blue (RGB); CIE 1976 (L*, a*, b*) color space (CIE L*a*b*); hue, saturation, and value (HSV); hue, intensity, and saturation (HIS); luminance and chrominance components (YCrCb).

- Color characterization of each bean landrace using three two-dimensional histograms with the data of each bean landrace in CIE L*a*b* color space.
- The second color characterization proposal uses the same color characterization with the same data in CIE L*a*b* color space; our contribution is using chroma and hue instead of a* and b* to characterize the color.

Mexico is the origin and diversification hub for numerous common bean landraces, which are maintained by various ethnic groups and cultivated and harvested for private consumption by farmers. The researchers primarily focus on creating a classification system for common bean landraces across different species [22].

Most research focuses on grading individual seeds of different bean varieties by characterizing their color, including averages and other values of color, and complementing this with other values of texture and shape characteristics of the seed. However, there is evidence that considers the color distribution of a set of seeds, which are characterized by three two-dimensional histograms, to estimate anthocyanin concentration using a regression task [19]. In this sense, this research work aims to classify common bean landraces into homogeneous and heterogeneous colorations across three different bean species. Due to the complexity and challenges presented by working with landraces of heterogeneous coloration, the use of color characterization using three two-dimensional histograms is proposed. The paper presents the following:

This work also incorporates a neuroevolutionary approach, as a CNN architecture is desired for the custom classification task.

The structure of this document is as follows: Section 2 describes the methodology for assigning labels to each landrace and classifying common bean landraces with heterogeneous and homogeneous coloration, utilizing two-color characterizations through deep learning and neuroevolutionary approaches. Next, Section 3 presents the results obtained from its implementation. Section 4 discusses the findings based on the results reported in this work. Finally, Section 5 presents the conclusions.

2. Materials and Methods

2.1. Workflow of the Study

In this work, the classification of common bean landraces with homogeneous and heterogeneous coloration is performed through digital image processing, as shown in the overall workflow in Figure 1. A series of steps were performed as follows:

1. Gather the set of bean landraces, and obtain a representative sample of each represented by 60 g.
2. Acquire digital photographic images of each bean landrace sample, then apply color calibration using an image reproduction workflow.
3. In this work, it is required to characterize the color of a set of seeds; therefore, the application of an algorithm for the separation of the regions of interest is required.
4. A physical characteristic of the study of this work is color, and a process was defined to identify the color of each bean landrace to facilitate the complex task and reduce its subjectivity when reviewing 60 g of seeds.
5. To characterize the color, we propose considering the color distribution of a set of seeds, using the joint probability distribution for two color characterizations.
6. Grouping of local bean varieties by labels, which were assigned based on the two principal predominant colorations in each sample; moreover, the total samples were randomly separated into 50% of the samples by each color group to create training and test data.

7. To develop a Convolutional Neural Network (CNN) architecture tailored to the classification of common beans, a key step in creating deep learning architectures for a specific task is to use a neuroevolutionary approach to evolve CNNs.
8. Finally, the results were compared using CNNs and two-color characterizations for the classification task of common bean landraces.

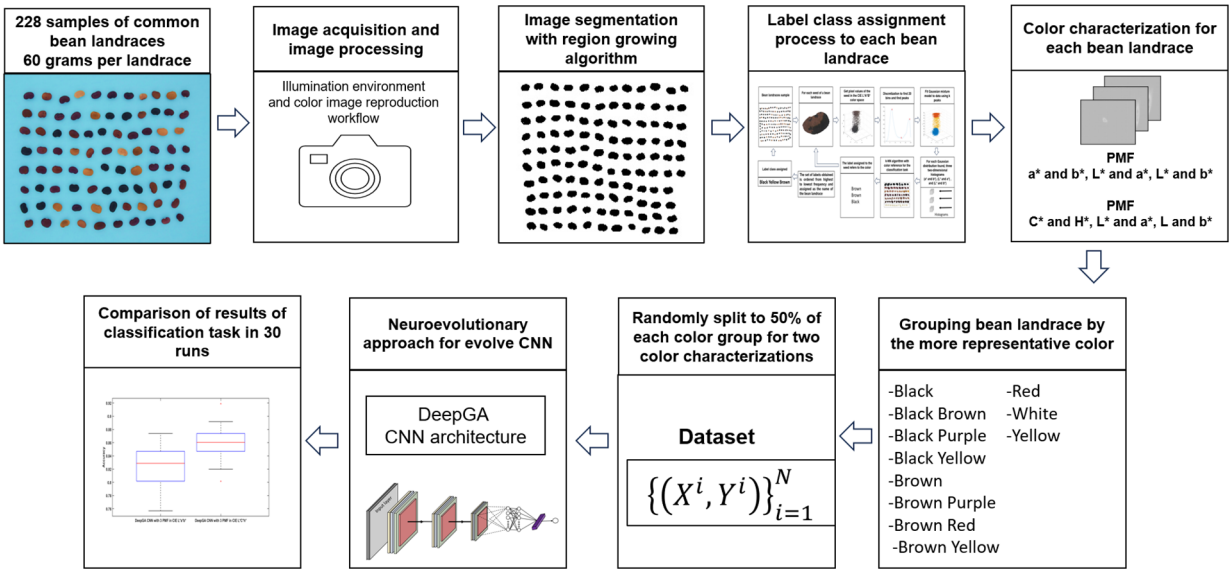


Figure 1. The workflow illustrates the procedure for classifying common bean landraces with heterogeneous and homogeneous coloration.

2.2. Common Bean Landraces

The common bean landraces used in this work were collected from several municipalities in Oaxaca, Mexico. These landraces correspond to three domesticated species (*Phaseolus vulgaris* L., *Phaseolus lunatus* L., and *Phaseolus coccineus* L.). The samples consisted of various colorations, ranging from homogeneous to heterogeneous, and included a mixture of differently colored or variegated seeds. The colors observed in the bean landraces employed in this work included seed coats of red, black, yellow, brown, pink, purple, and white. This color variability in heterogeneous landraces ensures a representative sample of each landrace, represented with a weight of 60 g. Figure 2 shows examples of the common bean landraces.

2.3. Acquisition System and Image Segmentation

The controlled illumination environment for this image acquisition process and protocol involved positioning a sample of bean landraces with individually spaced seeds to reduce occlusions and shadows, along with the optimal image reproduction workflow detailed in [17,18]. In the aluminum box, the camera lens was placed into the hole at the top of the light diffusion box, with regulated lighting and standardized settings, which were crucial for consistent photographic results. Figure 3 displays the workflow, which is summarized below:

1. The image acquisition under standardized conditions using a color reference chart and SONY digital camera model ILCE 3500 with a focal length of 50 mm and exposure configuration: shutter speed 1/60, ISO 100, and Aperture f/8.0.
2. Process the RAW image with the software Darktable 4.4.2 to preserve color data and convert it to a TIFF format with a standard ICC profile. Next, assign the custom ICC profile created with the software calibration tool to the processed TIFF images.

3. Shown in Figure 2, colored paper was used as a contrasting background to differentiate the seed hues. A blue background was employed to accentuate the seed landraces.
4. The RAW images were processed using the same software to create a TIFF image with the standard ICC profile. MATLAB 2023b software was utilized to change the standard ICC to a custom ICC profile and convert sRGB to the CIE L*a*b* color space.

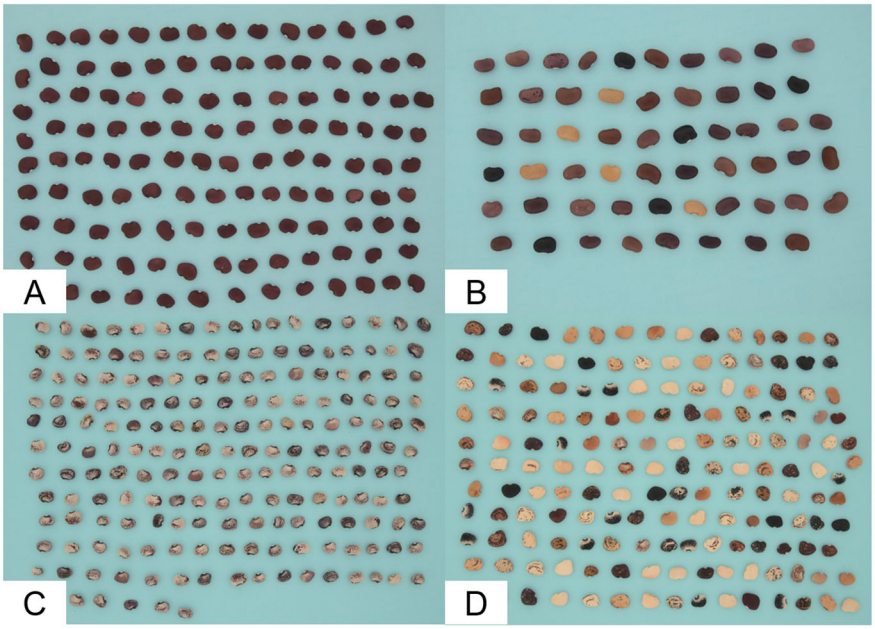


Figure 2. Four cases of coloration were observed in common bean landraces: (**A**) a sample of seeds with homogeneous coloration; (**B**) a sample of a mixture of seeds with different colorations, where each seed has a homogeneous coloration; (**C**) a sample of variegated seed coloration; (**D**) a sample of a mixture of seeds with homogeneous colorations and seeds with variegated coloration.

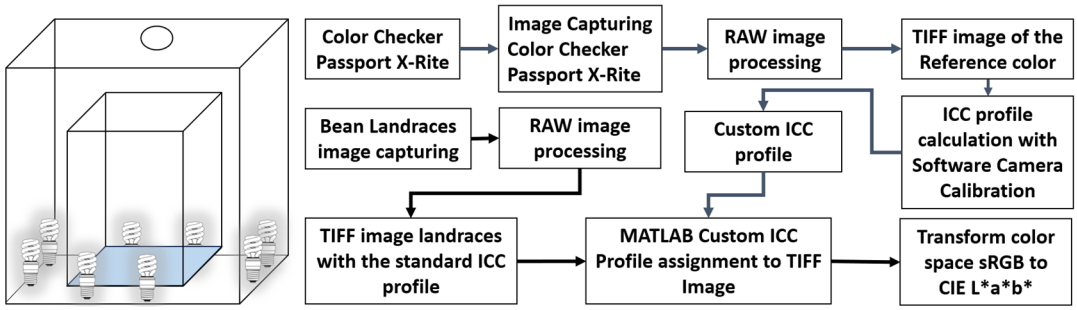


Figure 3. A light-controlled environment is crucial for standardizing image acquisition. The workflow shows the process for image reproduction.

A region-growing algorithm was applied to color images to obtain the regions of interest. Due to the background contrast used during image acquisition, it was defined as an area to select the seed pixel of the algorithm. The algorithm operates by grouping pixels based on a similarity criterion. It starts with the definition of a seed pixel from the image background. The Euclidean distance was used as a similarity measure between neighboring pixels. The appropriate metric to measure color differences using CIE L*a*b* coordinates was Delta E (CIE 1976) defined in Equation (1).

$$\Delta E = \sqrt{\left(L_{np} - L_{sp}\right)^2 + \left(a_{np} - a_{sp}\right)^2 + \left(b_{np} - b_{sp}\right)^2} \quad (1)$$

where ΔE is a color distance computed value of L*a*b*, (L, a, b) of seed pixel (sp), and (L, a, b) of neighboring pixel (np). A standardized metric for quantifying the difference between two colors. The lower the ΔE value, the greater the similarity to the seed pixel.

The result is a binary matrix with equal dimensions to the processed image. In the binary image, the regions of interest are marked with a value of 0, while the region of no interest are marked with a value of 1.

2.4. Objective Color Measurement

Color analysis is often a critical consideration when determining efficacy classification by color. The CIE L*a*b* color space is used by color measurement instruments that report a set of Cartesian coordinates based on the color measurement of a sample. These values represent the color within a three-dimensional space, but it is crucial to understand that they do not convey details of hue and chroma. Chroma quantifies the degree of deviation from gray toward pure chromatic color, whereas hue denotes specific color types, including red, blue, purple, yellow, and others. As shown in Figure 4, the L* channel refers to luminance, with a range of values between 0 and 100, corresponding to black and white, respectively. For any measured color, a* and b* position the color on a rectangular coordinate grid that is perpendicular to the L* axis. The color at the grid's origin (a* = 0, b* = 0) is achromatic (gray). a* positive indicates a purplish-red hue on the horizontal axis, while a* negative indicates a bluish-green hue. On the vertical axis, b* positive denotes yellow, and b* negative is blue.

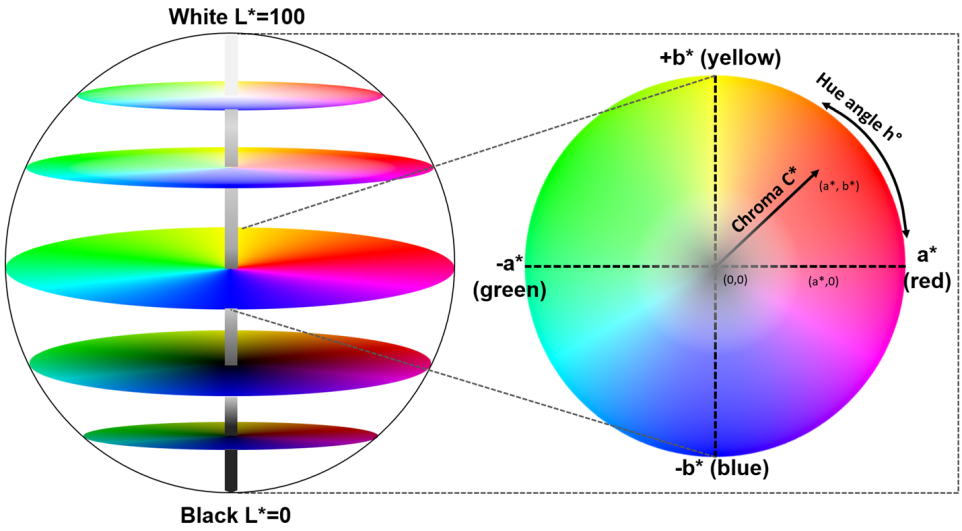


Figure 4. CIE L*a*b* color space is represented by axes L*, a*, and b*, displaying chroma in a color wheel and hue angle.

The chroma C* is calculated as $\sqrt{(a^{*2} + b^{*2})}$ and represents the hypotenuse of a right triangle created by joining the points (0, 0), (a*, 0), and (a*, b*). The hue angle is defined as the angle between the hypotenuse and 0° on the a* (blue-green/red-purple) axis; h is calculated as the arc tangent of b*/a*.

The arc tangent, however, assumes positive values in the first and third quadrants and negative values in the second and fourth quadrants. For correct interpretation, h° must remain positive between 0° and 360° of the color screen; the following pseudocode (Table 3) shows the algorithm for calculating hue [23].

Table 3. Pseudocode for calculating chroma and hue.

Input L a b Value
C = sqrt ((a × a) + (b × b))
Theta = (atan (b/a)/6.2832) × 360
If a > 0 and b >= 0 then h = theta;
If a < 0 and b >= 0 then h = 180 + theta;
If a < 0 and b < 0 then h = 180 + theta;
If a > 0 and b < 0 then h = 360 + theta;
return C, h

2.5. Color Characterization Using a Probability Mass Function

In the present study, the color characterization of bean landraces was performed using the CIE L*a*b* color space, which have been used in previous similar studies because they approximate human visual perception. This space allows for the separation of chromaticity from luminosity, facilitating a more detailed examination. In this sense, a probability mass function (PMF) is employed to describe the color distribution within a set of seeds. Specifically, a joint PMF is used to model the relationship between two discrete random variables, X and Y, whose joint probability distribution is expressed as $f(X,Y) = P(X = x, Y = y)$, where $f(X,Y)$ represents the probability of the occurrence of x and y values, subject to standard probability conditions as follows:

1. $f(x,y) \geq 0 \, for \, all(x,y)$,
2. $\Sigma x \Sigma y f(x,y) = 1$,
3. $P(X = x, Y = y) = f(x,y)$

Each pixel value in two channels is counted position by position, and the frequency of occurrence of the pixel value is given by Equation (2) to calculate the joint probability of a pair of channels.

$$pmf_{i,j} = \frac{p_{ij}}{n} \tag{2}$$

The overall procedure applied to characterize the color in the CIE L*a*b* color spaces is detailed as follows:

(a) *Color space transformation:* Convert images to CIE L*a*b*.

(b) *First color characterization:* The values of L*, a*, and b* were modified.

   a. Adjust the a* and b* values within the range of [−128 to 127] by adding the absolute value of the lower limit to ensure non-negative representations.
   b. For L*, normalize the values and scale them to the range of [0–255]. Three two-dimensional histograms were generated based on joint probability mass functions (PMFs) for color characterization. Specifically, PMFs were created for (a*, b*), (L*, a*), and (L*, b*) with dimensions of 256 × 256 × 3.
   c. Three two-dimensional histograms were used to characterize the color of each seed that forms the color palette and to characterize the color of each bean landrace integrated into this work.

(c) *Second color characterization:* The values of L* conserved, a*, and b* used to calculate chroma and hue.

   a. Chroma (C*) is calculated by the square root of the sum of a* and b* squares.
   b. Hue (h°) is performed by the arc tangent of b divided by a, and the result is converted to degrees. The procedure is reported in Table 3.
   c. CIE L*C*h°: Three two-dimensional histograms created by PMFs (C*, h°), (L*, C*), (L*, h°), and the color characterization as a 360 × 360 × 3 matrix, preserving joint distributions.

2.6. Gaussian Mixture Model or Gaussian Mixture

In this work, the color distributions of single seeds or all seeds belonging to a sample bean landrace are analyzed by obtaining cloud points in the CIE L*a*b* color space. The Gaussian mixture model possesses several important analytical properties, and it is a fundamental unsupervised learning algorithm in clustering and probability density estimation when the data are unlabeled [24]. To separate the color of variegated seeds, a Gaussian mixture was integrated. A Gaussian mixture is defined by Equation (3).

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3}$$

where

$\boldsymbol{x}$: represents a d-dimensional data vector.
K: number of Gaussian distributions (components) in the mixture.
$\boldsymbol{\mu}_k$: centroid (mean vector) of the k-th component.
$\boldsymbol{\Sigma}_k$: matrix describing the shape and orientation of the k-th component.
$N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: Gaussian density called a component of the mixture with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$.
$\boldsymbol{\pi}_k$: mixing coefficient, (priori probability that a data belongs to the component).
The parameter $\boldsymbol{\pi}_k$ is subject to the restrictions of Equations (4) and (5).

$$\sum_{k=1}^{K} \pi_k = 1 \tag{4}$$

$$0 \leq \pi_k \leq 1 \tag{5}$$

The main characteristic of any mixture model is its structure, which comprises marginal, conditional, and joint distributions. As a mixture model is directed to solve density estimation problems, the final goal is to identify a distribution over the data vectors $\boldsymbol{x}$, denoted as $p(\boldsymbol{x})$. For this purpose, the joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$ can be considered as in Equation (6).

$$p(x) = \sum_{k=1}^{K} p(x, z = k) \tag{6}$$

So, the proposed distribution is the marginal distribution of the joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$.

The joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$ can be defined in terms of a marginal distribution $p(\boldsymbol{z})$ and a conditional distribution $p(\boldsymbol{x}|\boldsymbol{z})$, as in Equation (7).

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) \tag{7}$$

Regarding the Gaussian mixture model, $\boldsymbol{z}$ is the latent variable indicating the cluster $(\boldsymbol{z} = k)$, and $\boldsymbol{x}$ represents the observed data.

The joint distribution $p(\boldsymbol{x}, \boldsymbol{z} = k)$ is factored as Equation (8).

$$p(\boldsymbol{x}, \boldsymbol{z} = k) = p(\boldsymbol{z} = k)p(\boldsymbol{x} \mid \boldsymbol{z} = k) \tag{8}$$

where

$p(\boldsymbol{z} = k) = \pi_k$ (cluster weight, k).
$p(\boldsymbol{x}|\boldsymbol{z} = k) = N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ (normal distribution of the cluster k).

The conditional distribution of D-dimensional vector $\boldsymbol{x}$ given a particular value of $\boldsymbol{z} = k$ is a multivariate Gaussian distribution and can be written in the form (Equation (9)).

$$N(\boldsymbol{x}|\boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma_k}|^{1/2}} exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) \right\} \tag{9}$$

where $\boldsymbol{\mu}_k$ is a D-dimensional mean vector, $\boldsymbol{\Sigma}_k$ is a $D \times D$ covariance matrix, and $|\boldsymbol{\Sigma}_k|$ denotes the determinant of $\boldsymbol{\Sigma}_k$.

So, with the Gaussian characterization displayed in (8), the structure of the distribution $p(\boldsymbol{x})$ can be written as in Equation (10).

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x},\ \boldsymbol{z} = k) = \sum_{k=1}^{K} p(\boldsymbol{z} = k) p(\boldsymbol{x} \mid \boldsymbol{z} = k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{10}$$

Suppose that the data points $\{\boldsymbol{x}_i\}_{i=1}^{M}$ are drawn independently of the distribution. In that case, we can express the Gaussian mixture model with the likelihood function shown in (11).

$$\mathcal{L}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{x}) = \prod_{i=1}^{M} \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_\mathrm{k},\ \boldsymbol{\Sigma}_\mathrm{k}) \tag{11}$$

And the logarithm of the likelihood function is given in Equation (12).

$$ln p(\boldsymbol{X}|\boldsymbol{\pi},\ \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{12}$$

A notable issue with applying the maximum likelihood framework to Gaussian mixture models is the occurrence of singularities. As a result, maximizing the log likelihood function poses challenges, since these singularities are persistent and arise when one of the Gaussian components "collapses" at a particular data point. To avoid the singularities, we can employ appropriate heuristics, such as detecting when a Gaussian component is collapsing and resetting its mean to a randomly chosen value, while also adjusting its covariance to some meaningful value, and then continuing with the optimization.

The Expectation-Maximization algorithm (EM algorithm) is a sophisticated method for finding maximum likelihood solutions for models with latent variables. The model parameters were estimated using the Expectation-Maximization (EM) algorithm, which iteratively maximizes the likelihood of the observed data. The EM algorithm alternates between two steps: the Expectation step (E-step) and the Maximization step (M-step), as described in Algorithm 1.

The EM algorithm is repeated until convergence, defined as a negligible change in the log likelihood between successive iterations. The EM algorithm requires significantly more iterations to reach convergence compared to the *K*-means algorithm, and each cycle requires substantially more computation. It is therefore common to run the K-means algorithm to find a suitable initialization for a Gaussian mixture model, which is subsequently adapted using the EM algorithm. The model is implemented using MATLAB's *fitgmdist* function.

Algorithm 1: Expectation-Maximization for Gaussian Mixture Models

1. Initialize parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, π_k for each $k = 1, \ldots, K$
2. Compute initial log-likelihood:
$$lnp(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi) = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$
3. **Repeat**
   a. **E-step:** Compute responsibilities
   $$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$
   b. **M-step:** Update parameters
   $$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$
   $$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x_n}$$
   $$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_\mathrm{n} - \boldsymbol{\mu}_k)(\boldsymbol{x}_\mathrm{n} - \boldsymbol{\mu}_k)^T$$
   $$\pi_k = \frac{N_k}{N}$$
   c. **Recompute log-likelihood:**
   $$lnp(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi) = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$
4. **Until:** the convergence of parameters or log-likelihood

2.7. Label Class Assignment to Common Bean Landraces

Since color is a physical characteristic related to chemical composition, the common bean landraces used in this work were processed to assign labels to each bean landrace and then described by color. The label class assignment procedure used in this methodology was part of the color quantification described in the work by [25]. Figure 5 illustrates the workflow used for assigning labels to each one.

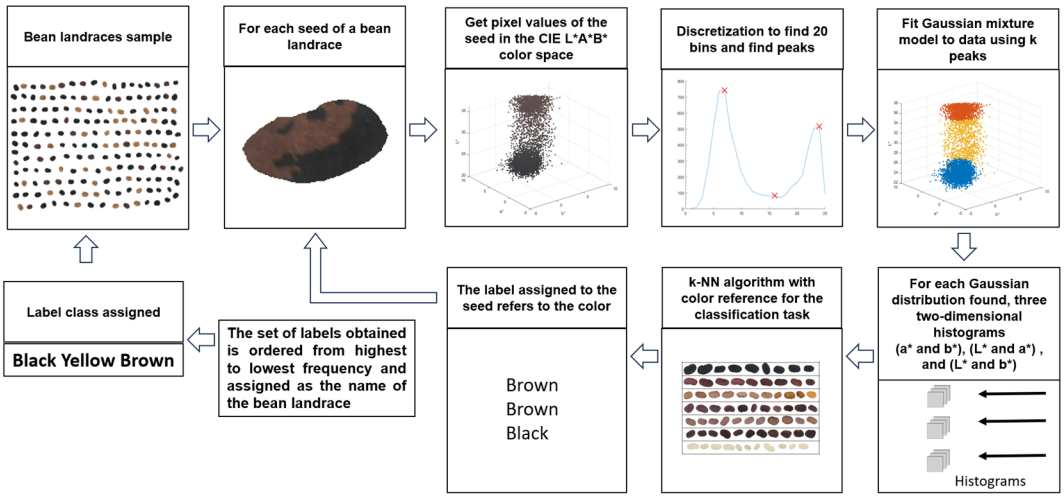


Figure 5. A workflow that visually illustrates the methodology for assigning class labels to each bean landrace.

Due to the complexity of manually assigning a label to each bean landrace weighing 60 g, the task is intricate for samples composed of a mixture of seeds with homogeneous colorations and seeds with variegated colorations and can be quite subjective. We es-

tablished a procedure to assign a label generated through a solution that includes color characterization and a machine learning algorithm, as shown below:

1. For each bean landrace, the binary image obtained from segmentation was used to apply digital image processing and to analyze each region of interest for its color distribution. For this purpose, connected components were used.
2. Discretize the data into 20 bins and search for histogram peaks. In the MATLAB program 2023b, the function used was *findpeaks*.
3. Given the point cloud, a Gaussian mixture distribution model (GMModel) with k components was fitted to the data using the *fitgmdist* function. The parameter value for *k* corresponds to the number of identified peaks.
4. A triple two-dimensional histogram was created for each component to characterize the colors of the seeds and obtain labels that describe their color.
5. To identify colors and assign labels, it was essential to define a color palette based on the hues found in native bean varieties. It was noted that uniform seed colorations also appear in variegated seeds. Consequently, a palette featuring seeds of uniform colors was created, and a classification algorithm was implemented to leverage the knowledge base established and conduct similarity searches based on color. A machine learning algorithm was used to generate and assign labels to each seed, achieving this goal.
6. The labels derived from analyzing each seed of every local variety of common bean were organized into color categories based on their presence in the sample, arranged from highest to lowest frequency. For bean varieties exhibiting heterogeneous coloration or a mixture of seeds in different colors, the assigned label combined the two most representative labels in the sample, reflecting the most frequently occurring colorations.

Color Palette Reference and K-Nearest Neighbors Algorithm

Manually labeling a 60 g sample representing a heterogeneous coloration of beans is quite challenging and to minimize subjectivity in this task, it is necessary to develop an automated process to assign a label to each seed per bean landrace, so that the set of labels obtained will allow each bean landrace to be described by its color. For this purpose, color references were developed using information derived from seeds of homogeneous coloration due to their similar shades. A data set was compiled with the color information of selected seeds from various bean landraces. In this way, with the support of an expert, a set of seed samples was selected from the different images to create six color groups. The chosen seed color was characterized using PMFs composed of channels (a*, b*), (L*, a*), and (L*, b*). For each seed, three two-dimensional histograms were then generated to create the color palette.

Figure 6 shows a set of seeds for each type of coloration detected in the samples of this work.

The K-Nearest Neighbors (k-NN) is a supervised learning algorithm that predicts the category of new observations based on the majority class of K-Nearest Neighbors. Its operation relies on the locations of points in multidimensional space using a distance metric; it calculates the distance from the new observation to all the points in the training set. The K-Nearest Neighbors are selected, and the highest frequency class is assigned to the unknown observation through a voting process. The cityblock distance is suitable for comparing histograms [26] (see Equation (13)).

$$d_T(\boldsymbol{a}, \boldsymbol{b}) = \sum_{j=1}^{n} |a_j - b_j| \tag{13}$$

where $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$, $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$, $\boldsymbol{a}$ and $\boldsymbol{b}$ are histograms, and n-points represent the value frequencies.

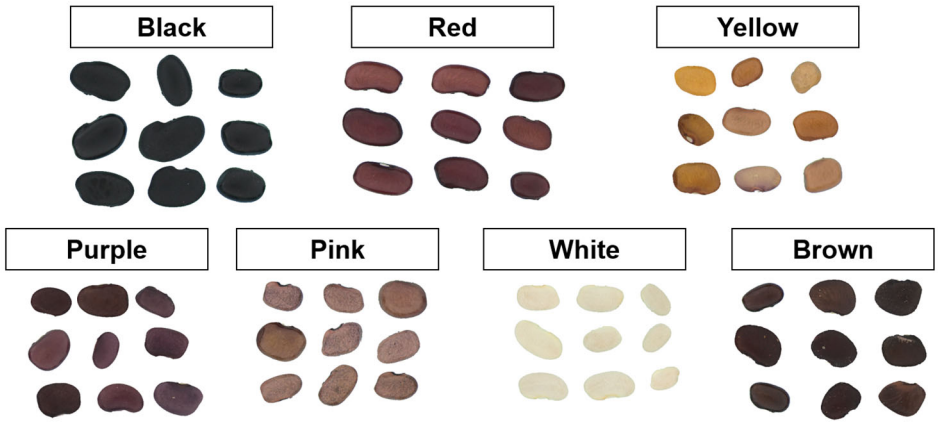


Figure 6. Seven color groups for the knowledge base; the seed for color reference represents training data for the machine learning algorithm k-NN.

The *k*-NN method stores a training data set instead of undergoing a training phase. When an unknown observation is received, classification or prediction calculations take place. To identify its neighbors and assign them a label, the distance between the query point and the base points is calculated using a metric.

A *k*-NN classification model is represented in this work using the *fitcknn* function from the MATLAB program 2023b, and the training data set used was the color reference represented by a set of three two-dimensional histograms.

In the process of assigning a class to each bean landrace, to ensure the correct assignment of labels to the seed, steps were performed as shown in Figure 5. In summary, the first step was to obtain the color distribution of the seed, and discretization of the data to find peaks of frequency. The number of peaks was used for the value of k components of a Fit Gaussian mixture model to data using a k-component Gaussian mixture model; for each component corresponding to a new observation in the *k*-NN model, the parameter *k* was set to 9, 21, 31, 41, and 51 neighbors; from the acquired labels, the label with the highest frequency was assigned. This same process was performed for each seed.

Once the process of assigning labels to all seeds was completed and given the colorimetric variability of the heterogeneously colored bean landraces, the next step was to order the set of labels and group from highest to lowest frequency to assign the two labels with the highest frequency as class labels. To correspond to coloration for maximal presence in the sample, it is important to note, some colors that make up the palette may not appear on the labels due to their minimal presence, as is the case with pink seeds. For the case of the homogeneously colored bean landrace, a single label was assigned due to its uniform coloration.

2.8. Data Splitting

A total of 228 samples of common bean landraces with different shades were considered. This included homogeneously colored bean landraces with uniformly colored seeds and heterogeneously colored bean landraces with mixed seeds of various shades. As a result, the data set consists of 12 color groups, as shown in Table 4. In this preliminary assessment, the designated label refers to the two colors most prominently featured in the bean landrace of mixed coloration. However, these colors may not be reflected in the class labels of common bean landraces due to their limited occurrence.

Table 4. The number of samples for each color group used in this paper.

Group Color	Number of Samples	Group Color	Number of Samples
Black	78	Black Purple	11
Brown	5	Black Yellow	6
Red	31	Brown Purple	10
White	8	Brown Red	6
Yellow	16	Brown Yellow	11
Black Brown	39	Purple Yellow	7

The process involved random sampling from each color group of bean landraces with a 50% sample separation to generate training and test data.

Subsequently, two color characterizations were created from each separation: the first conformed to three two-dimensional histograms using CIE L*a*b* color space data, and the second was represented by three two-dimensional histograms employing CIE L*C*h° data.

2.9. Convolutional Neural Networks

The color characterization is represented by histograms that are $256 \times 256 \times 3$ matrices, and CNNs are suitable for receiving this type of matrix. The basic structure of the CNN consists of convolutional layers that implement filters (kernels) used for the convolution process on images, generating feature maps. An image can be defined as a two-dimensional function $f(x,\ y)$, where x and y are spatial coordinates; the amplitude f of a coordinate pair is referred to as the level of the image. The image contains M rows and N columns, and the value of the origin coordinate is $(x,\ y) = (0,\ 0)$; the following coordinate is represented by $(x,\ y) = (0,\ 1)$, and the filter is a grid of discrete numbers called convolutional kernel and defined as $w(x,\ y)$; to define the size of an mxn kernel, the following expressions result from applying $m = 2a + 1$ and $n = 2b + 1$, let a and b be non-negative integers [27].

The operation between the image and the kernel simply involves moving them from one point to another on the image. At each point $(x,\ y)$, the response of the kernel at that location is calculated as the sum of the products of the kernel coefficients and the corresponding image pixels in the area covered by the convolution kernel, as expressed in Equation (14).

$$R = w_1 z_1 + w_2 z_2 + \ldots + w_{mn} z_{mn} = \sum_{i=1}^{mn} w_i z_i \tag{14}$$

where R of an $m \times n$ kernel at any point $(x,\ y)$, w represents the kernel coefficients, z are the values of the gray levels of the image, and mn is the total number of kernel coefficients.

The process of convolution of an image f of size $M \times N$ with a kernel of size $m \times n$ is given by the following Equation (15) and graphically represented in Figure 7.

$$(w * f)(x,y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(s,t) f(x-s, y-t) \tag{15}$$

where $x = 0,\ 1,\ 2,\ \ldots,\ M-1, y = 0,\ 1,\ 2,\ \ldots,\ N-1$, and $(w\ *\ f)(x,y)$ is the output of a complete process of convolution of an image.

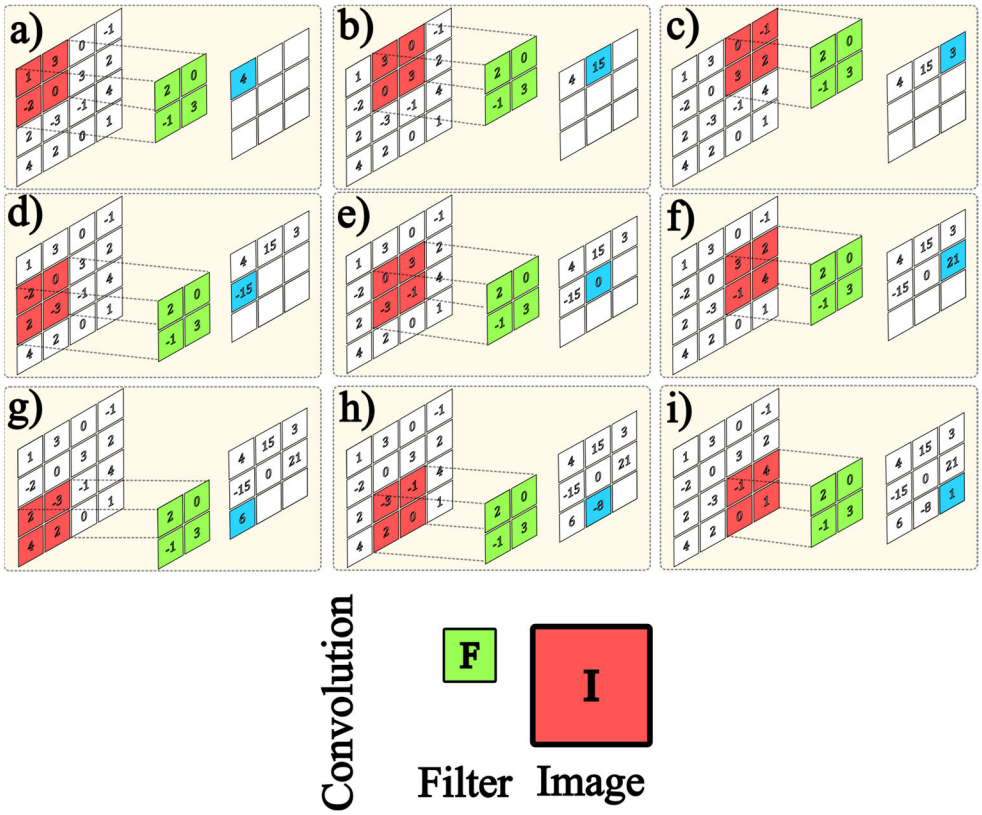


Figure 7. An example of the convolution process involves moving the kernel through the image, from position (**a**–**i**), to extract feature maps taken of [28].

The convolution has several properties. The convolution is as follows:

- Commutative: $w * f = f * w$
- Associative: $f * (w1 * w2) = (f * w1) * w2)$
- Distributive: $f * (w1 + w2) = (f * w1) + (f * w2)$

The convolutional layer implements kernels that are used for the convolution process on the images, generating the aforementioned feature maps, and the pooling layers reduce the dimensions of feature maps generated in the convolutional layers. Like convolutional layers, reduction layers utilize filters to derive a new value from the neighboring pixels. Two strategies, mean reduction and maximum reduction, are commonly used to calculate the value from the filters. The average reduction obtains the average value of all pixels within the filter, and the max-reduction layer obtains the maximum value [29].

The batch normalization algorithm is a technique for transforming input values of each layer into values with zero mean and constant standard deviation [30].

In a fully connected layer, every unit is linked to all units in the preceding layer. Typically, in a CNN, these layers are positioned near the end of the architecture. Fully connected layers convert the resulting feature maps from the reduction layer into a vector representing an ANN [29].

The ANN comprises a collection of basic units known as artificial neurons, linked by weights. These neurons take inputs from their peers and pass along their outputs to others in the form of a simple scalar value. A layer in a neural network consists of a cluster of neurons operating at the same depth or level. These networks utilize basis functions structured similarly to Equation (16), meaning each basis function is a non-linear function

derived from a linear combination of the inputs, with the coefficients in this combination being adaptable parameters [24].

$$y(\boldsymbol{x}, \boldsymbol{w}) = f\left(\sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x})\right) \tag{16}$$

where w_j represent the coefficients, ϕ_j are the parameters, $\boldsymbol{x}$ is a input vector, $\boldsymbol{w}$ correspond all weight and bias parameters grouped into a vector, and $f(\cdot)$ is a non-linear activation function. This leads to the basic neural network model, which can be described as a series of functional transformations for the first layer of the network, with input variables $x_i, \ldots, x_D$, expressed in Equation (17).

$$a_j = \sum_{i=1}^{D} w_{ji} x_j + w_{j0} \tag{17}$$

where w_{ji} are the weights, w_{j0} as biases, and a_j are known as activations.

Each activation is then transformed using a differentiable, non-linear activation function $h(\cdot)$ (see Equation (18)).

$$z_j = h(a_j) \tag{18}$$

where z_j represents the hidden units; the non-linear functions $h(\cdot)$ are generally chosen to be functions such as the ReLU function for to give output unit activations. For a second layer of the network expressed in Equation (19).

$$a_k = \sum_{j=1}^{M} w_{kj} z_j + w_{k0} \tag{19}$$

where $j = 1, \ldots, M$ correspond to M linear combination, $k = 1, \ldots, K$, K is the total number of outputs, w_{kj} are the weights second layer. The output unit activations (a_k) are transformed with a suitable activation function to produce a set of network outputs y_k. Therefore, each output unit's activation is transformed using the ReLU function (see Equation (20)).

$$y_k = \sigma(a_k) \tag{20}$$

where $\sigma(a_k)$ is a ReLU function and mathematically defined in Equation (21).

$$\sigma(a_k) = max(0,\ a_k) \tag{21}$$

where a_k is the input to the neuron, the function returns a_k if a_k is greater than 0.

If a_k is less than or equal to 0, the function returns 0.

Finally, the overall network function for the ReLU function output unit activation takes the form of Equation (22).

$$y_k(\boldsymbol{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj} z_j\, h\left(\sum_{i=1}^{D} w_{ji} x_i + w_{j0}\right) + w_{k0}\right) \tag{22}$$

The ANN model is a set of input variables $\{x_i\}$ to a set of output variables $\{y_k\}$ controlled by a vector $\boldsymbol{w}$ of adjustable parameters.

2.10. Neuroevolutionary and DeepGA Algorithm

Neuroevolution, also known as Evolutionary Neural Networks, is a machine learning technique that uses evolutionary algorithms to optimize Artificial Neural Networks, including their weights, architectures, and hyperparameters [31].

One of the most well-known types of evolutionary algorithms, extensively applied to neural network optimization, is the Genetic Algorithm (GA) [18]. GA is a class of search algorithms inspired by natural selection, where the fittest individuals have a higher probability of survival and reproduction. These algorithms evolve a population of candidate solutions, each represented as a set of genes, utilizing selection, mutation, and crossover operators to create improved solutions over time [19]. In the context of neuroevolution, the genes typically represent the connection weights and biases of the neural network architecture. The Genetic Algorithm then explores this parameter space to find an optimal set of weights that maximizes the neural network's performance on a given task.

DeepGA is a neuroevolution algorithm that utilizes a hybrid coding approach to represent Convolutional Neural Networks (CNNs) at two levels [20]. The first level consists of convolutional blocks, each serving as a layer with various filters and sizes. These layers feature a stride of 1, zero padding, batch normalization, and a rectified linear unit (ReLU) activation function. Additionally, this level includes a fully connected block to encode the number of neurons. The second level employs binary coding to describe the connections between layers: 1 indicates a connection, while 0 signifies no connection. This approach leads to a CNN architecture with fewer layers. DeepGA has been optimized [32] and utilized to optimize the design of neural networks for various image classification and regression tasks, including breast cancer diagnosis [21], vehicle make and model recognition [22], steering angle estimation [23], and anthocyanin level estimation in common bean landraces [14].

2.11. Metric Performance

We assessed the Convolutional Neural Network using the optimal architecture obtained from DeepGA. The results showcased the network's accuracy in classifying common bean landraces.

The accuracy is calculated as the ratio of the total number of correct predictions made by the algorithm to the total number of data points (see Equation (23)).

$$Accuracy = \frac{correctly\ classified\ images}{total\ images} \tag{23}$$

2.12. Experiment Design

The experiment was designed to achieve the above-mentioned objectives by comparing classification results obtained using two color characterizations and deep learning. A total of 30 runs were conducted using the common bean landrace data set, each involving a random separation of the data to create training and testing sets for the two different color characterizations. Two data sets corresponding to these color characterizations were employed to develop the neuroevolved architectures using DeepGA. Table 5 presents the values assigned to the hyperparameters of the DeepGA setup for the classification task, particularly to ensure consistent conditions across the experiments.

Table 5. Hyperparameter values for DeepGA in the classification task for evolving CNNs. All other parameters retained their original settings values.

Hyperparameters of Evolutionary Algorithm	Value	Hyperparameters of CNN	Value
Population size	15	Epochs	50
Generations number	50	Learning rate	0.001
Crossover rate	0.7	Optimization method	ADAM
Mutation rate	0.5	Loss function	Cross-entropy

CNN Architecture Optimized Through DeepGA

Two architectures were obtained with DeepGA. The first architecture, depicted in Figure 8, consists of three convolutional layers with batch normalization and ReLU activation. All convolutional layers incorporate 3×3 average pooling with a stride of 2×2. The first convolutional layer contains eight filters of 2×2, the second convolutional layer has 16 filters of 3×3, and the third convolutional layer has eight filters of 2×2. Finally, the architecture was complemented with two fully connected layers, each with 128 and 12 neurons, and the first fully connected layer included a ReLU activation.

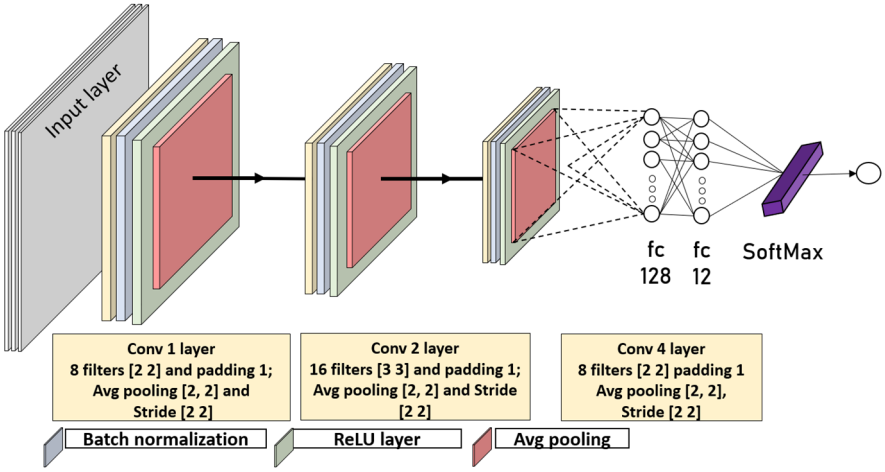


Figure 8. The CNN architecture evolved by DeepGA, utilizing color characterization represented by three two-dimensional histograms: (a* and b*), (L* and a*), and (L* and b*).

The second evolved architecture is illustrated in Figure 9. The solution reported by DeepGA consisted of five convolutional layers and four fully connected layers, all of which contained batch normalization and ReLU layers. The first convolutional layer has 16 filters of 4×4; the second one has eight filters of 2×2; the third one has eight filters of 3×3; the fourth one has 128 filters of 3×3; and the fifth one has eight filters of 3×3. The architecture includes four skip connections. The max pooling from the first convolutional layer has three connections leading to input convolutional layers 3, 4, and 5. The average pooling of the second convolutional layer connects to the input of the fourth convolutional layer, followed by four fully connected layers with 16, 128, 256, and 12 neurons.

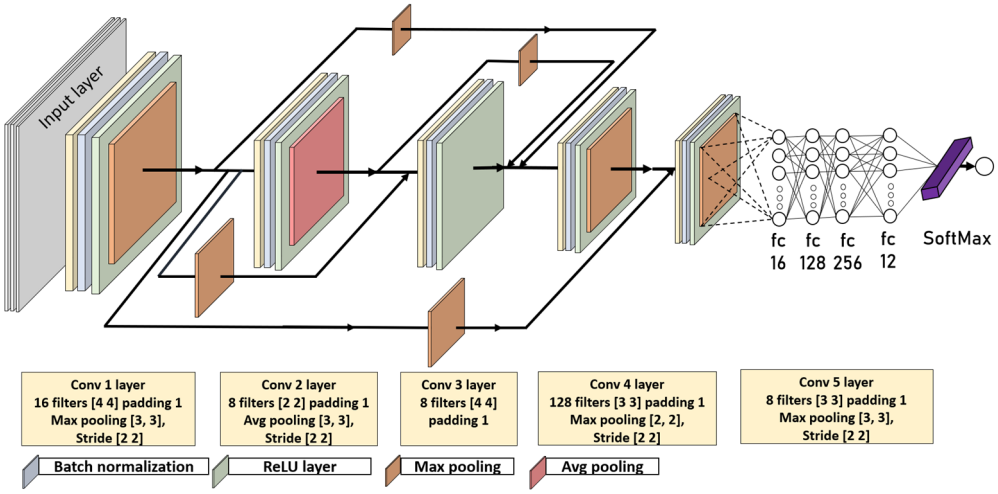


Figure 9. CNN architecture evolved by DeepGA using the color characterization conformed with three two-dimensional histograms (C* and h°), (L* and C*), (L* and h°).

Two architectures of DeepGA were evaluated to compare classification results based on the color characterizations of three two-dimensional CIE L*a*b* histograms and the contributions of three two-dimensional CIE L*C*h° histograms. Each architecture was trained for 500 epochs on each data set over 30 runs.

A normality test, specifically the Shapiro–Wilk test, was conducted to analyze the differences in classification results using two color characterizations and deep learning. In this normality test, a 95% confidence level was established to accept the null hypothesis of normality, indicating that the results follow a normal distribution.

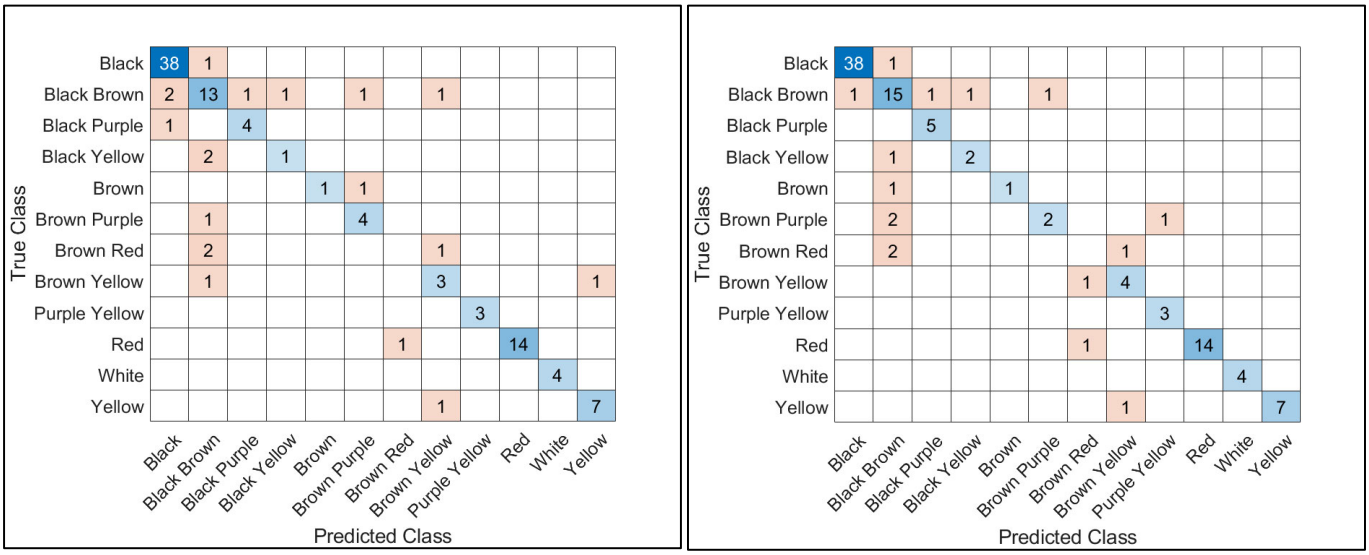
3. Results

The classification results using the architectures obtained with DeepGA and two-color characterizations are displayed in Table 6.

Table 6. Results from 30 runs of classifying bean landraces of various colors are presented. The maximum, minimum, and mean values obtained in two color characterizations using the architectures reported by the DeepGA algorithm.

Models		DeepGA Architecture	DeepGA Architecture
Color characterization technique		PMF of a* and b* PMF of L* and a* PMF of L* and b*	PMF of C* and h° PMF of L* and C* PMF of L* and h°
Accuracy in 30 runs	Max	87.39	91.89
	Min	75.68	80.18
	Mean	82.22 ± 2.84	85.74 ± 2.37

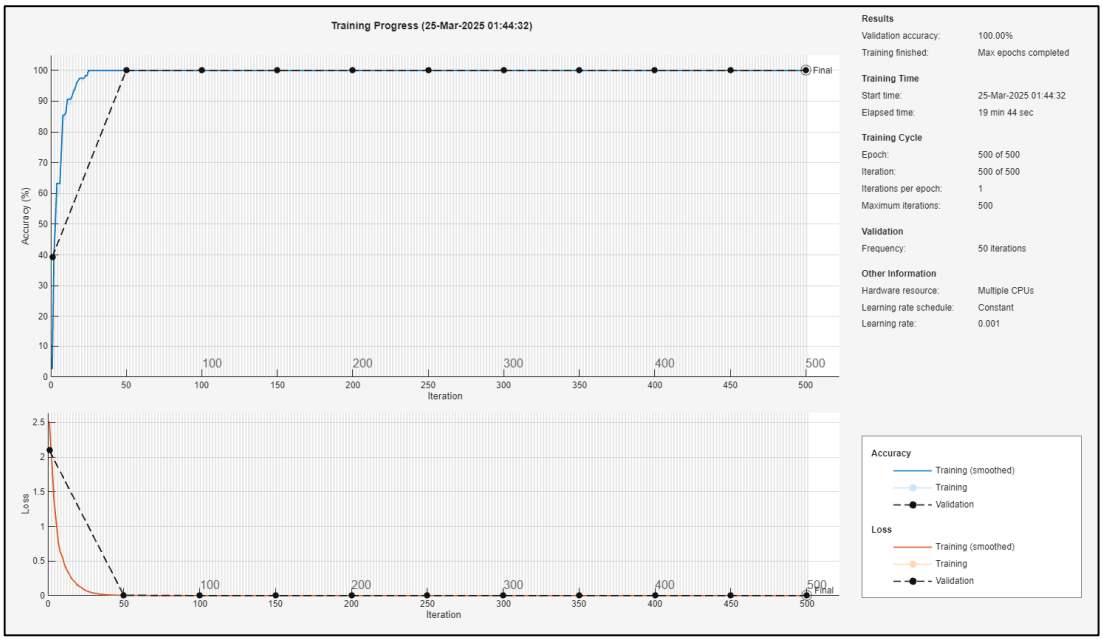
The outcomes from the trials, which exhibited accuracies close to the mean for each color characterization, were used to depict the confusion matrix. Figure 10 shows two confusion matrices.



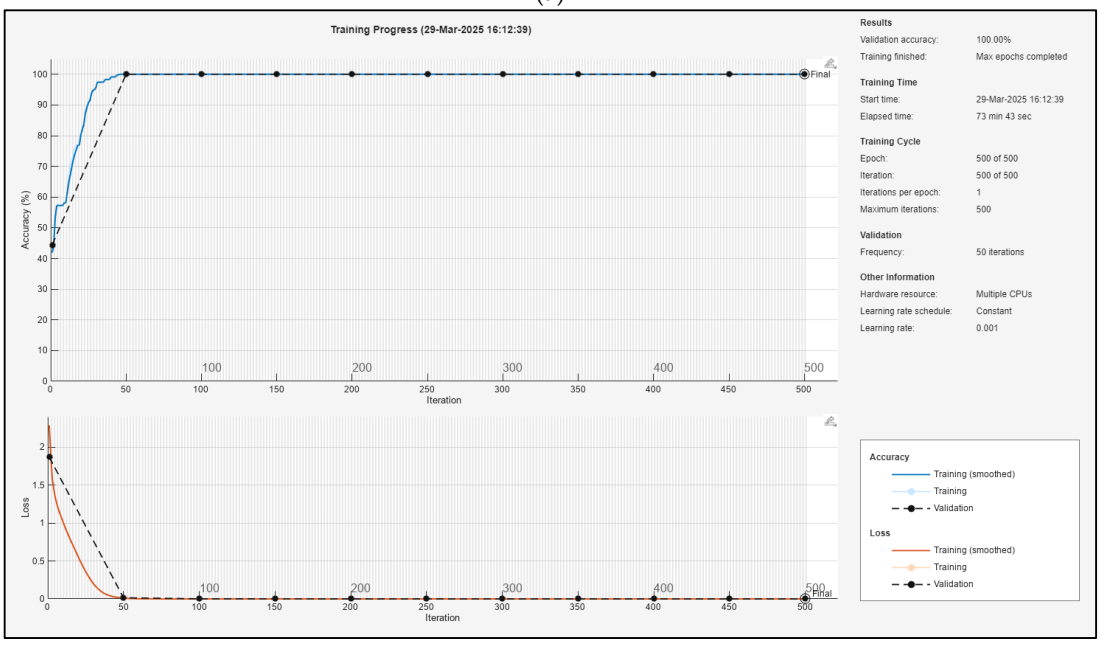
(**a**) (**b**)

Figure 10. Confusion matrix of runs exhibiting accuracy comparable to the average of the 30 tests across 12 classes: (**a**) Confusion matrix showing classification results from the DeepGA evolved architecture, with color characterization represented in three two-dimensional histograms of (a* and b*), (L* and a*), (L* and b*). (**b**) Confusion matrix displaying results achieved using color characterization from three histograms of (C* and h°), (L* and C*), (L* and h°), alongside their corresponding architecture provided by DeepGA.

Figure 11 shows the convergence plots of the training process of the CNN in the runs with accuracy close to the mean accuracy reported in Table 3.



(**a**)



(**b**)

Figure 11. Convergence plots of the training process for the architectures generated by DeepGA feature two color characterizations: (**a**) The process of architecture, with color characterization illustrated in three two-dimensional histograms of (a* and b*), (L* and a*), (L* and b*). (**b**) The process utilizing color characterization from three histograms of (C* and h°), (L* and C*), (L* and h°), alongside their corresponding architectures provided by DeepGA.

Statistical Results

When comparing the performance of models trained using the CIE L*a*b* and CIE L*C*h° color representations, we found that the accuracy distributions for both followed a normal distribution. The independent samples t-test revealed a statistically significant difference between the two groups (p-value = 2.6409×10^{-6}).

Boxplot analysis revealed differences in the dispersion of the results (see Figure 12). The model based on CIE L*C*h° had a more compact box, suggesting lower variability in accuracy across runs. Two outlying observations were noted, situated on either side of the main distribution, and may be related to variability during the training phase. In contrast, the model using the CIE L*a*b* representation exhibited greater central dispersion, as evidenced by a longer box and extended whiskers. No outliers were detected, but the overall variability was higher than the L*C*h–based model. These findings suggest that the model trained with CIE L*C*h° achieved higher average accuracies and demonstrated more stable behavior.

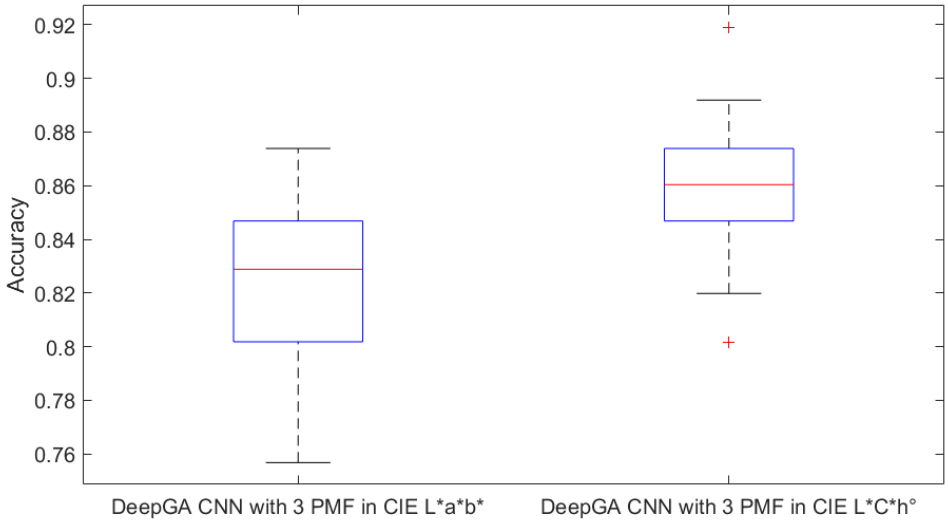


Figure 12. A box plot illustrating the comparative results of homogeneous and heterogeneous coloration bean landraces classification tasks.

4. Discussion

Mexico has hundreds of domesticated and wild bean landraces in various colors. Implementing an automatic classification solution reduces subjectivity in color organization. Furthermore, the colors of bean landraces are linked to bioactive compounds, enabling us to analyze this relationship.

Several studies classify bean varieties using averages of color, texture, or morphological properties to characterize seeds [8,11,18]. In this work, due to the complexity of labeling portions of 60 g by each bean landrace, which contain approximately 100 to 200 seeds depending on the size of the seeds, the complexity increase per the variability of their heterogeneous colors, it was necessary to define a methodology for this task; additionally, we explore the color characterization of a set of seeds, and in that sense, we have examined the use of PMF in two color characterizations.

The methodology used to assign class labels that describe each landrace by coloration allowed us to automatically organize the color groups. This was necessary to create the database, rather than relying on manual assignment again, as reported in previous work for a small number of seeds (20 seeds) [17]. The challenge involved heterogeneous color landraces, which consisted of a mixture of seeds with various color patterns. Additionally, a second challenge arose with variegated color seeds due to their lack of a single-color pattern.

The DeepGA algorithm is essential because it provides an evolved network architecture. In this work, two CNN architectures were required to perform the classification process of bean landraces. For this reason, the same values of the hyperparameters were retained. Unlike other cases where the architectures reported by DeepGA are identical [19],

this work has different reported architectures. As shown in Figure 9, a more complex architecture was developed for new color characterization, demonstrating the complexity of the work.

The results obtained from the architecture created by DeepGA, using the color characterization of three two-dimensional histograms (constructed with the L*, a*, and b* channels), reported a mean accuracy of 82.22 $\pm$ 2.84, a maximum accuracy of 87.39, and a minimum accuracy of 75.68, which are lower than those reported by the second architecture evolved by DeepGA and our proposed color characterization of three two-dimensional histograms with the L*, C*, and h$^\circ$ information, which achieved a mean accuracy of 85.74 $\pm$ 2.37, a maximum accuracy of 91.89, and a minimum accuracy of 80.18.

Based on the results, the information provided by chroma and hue (C* and h$^\circ$) enhances our understanding, as chroma indicates the color purity that varies from the center to the periphery of the CIE L*a*b* color circle. In contrast, hue indicates the angle of the color on the circumference. Therefore, this information is crucial for the deep learning model to accurately identify the colors of the test set samples.

In Figure 10, we observe the confusion matrices from the runs close to the mean of each color characterization. In Figure 10a, it is evident that the classification errors occur in landraces with heterogeneous colors. For example, the bean landraces labeled as Black Brown exhibit a higher classification error, which is an important factor contributing to their color variability. Although the dominant colors are black and brown, these landraces also contain seeds of other colorations, making classification complicated due to the differing colorimetric patterns. The values of the channels in CIE L*a*b* share the same scale, which can be a disadvantage for the similarity of values in landraces that exhibit different colorimetric patterns. In contrast, these values are on a different scale when using chroma and hue. This distinction aids in differentiating between classes, as illustrated in Figure 10b, where a reduction in classification error is observed.

The difference in accuracy reported by both classification models is significant, as illustrated in Figure 12, where notable differences in the results of the two-color characterizations are observed. Therefore, this new proposal accommodates significant differences, making it a suitable solution for classifying bean landraces of both homogeneous and heterogeneous colorations.

5. Conclusions

Initially, images of each bean landrace were obtained, followed by a color calibration procedure to ensure consistent results and prevent chromatic deviations. The definition of a color palette and the application of the k-NN algorithm enabled the accurate assignment of labels to the set of bean landraces used in this study, thereby reducing the subjectivity inherent in the manual classification of bean landraces. The development of a method to correctly separate the colors of variegated seeds allowed the assignment of labels based on the number of colors present.

The color characterization determines the results; for that purpose, two color characterizations were compared to know the classification results of common bean landraces. In that sense, the DeepGA algorithm was used to have custom architectures. The results showed a significant difference. Therefore, the color characterization represented by three two-dimensional histograms (C* and h$^\circ$), (L* and C*), (L* and h$^\circ$), and the neuroevolved CNN architecture with DeepGA offered higher accuracy.

In this work, we analyzed the classification of homogeneous and heterogeneous colored bean landraces, considering a reduced number of heterogeneous samples. As future work, we propose the following: i. A greater number of samples of bean landraces with heterogeneous coloration is needed to better represent the diversity of colors present

in different bean species. ii. To improve the label assignment method, the total color distribution present in a sample of bean landraces should be considered to identify the different colors rather than each seed individually.

Author Contributions: Conceptualization, J.-L.M.-R. and E.-N.A.-B.; methodology, J.-L.M.-R. and E.-N.A.-B.; software, J.-L.M.-R.; validation, E.-N.A.-B. and H.-G.A.-M.; formal analysis, J.-L.M.-R.; investigation, J.-L.M.-R.; resources, E.-N.A.-B. and J.-L.C.-S.; data curation, J.-L.M.-R. and E.-N.A.-B.; writing—original draft preparation, J.-L.M.-R. and N.P.-C.; writing—review and editing, E.-N.A.-B. and H.-G.A.-M.; visualization, J.-L.M.-R. and N.P.-C.; supervision, E.-N.A.-B. and H.-G.A.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data sets employed in this article are not readily available since they are part of an ongoing study. Requests for access to the data sets should be directed to eli-aquino@uv.mx and heacosta@uv.mx.

Acknowledgments: The first author acknowledges the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) of Mexico for granting support for the realization of this investigation through scholarship 712056 awarded for postdoctoral studies at the Centre for Food Research and Development in the University of Veracruz.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Servicio de Información Agroalimentaria y Pesquera (SIAP). Expectativas Agroalimentarias Agosto 2024. Available online: https://www.gob.mx/cms/uploads/attachment/file/940838/Agosto_2024.pdf (accessed on 10 April 2025).
2. Smith, M.R.; Dinglasan, E.; Veneklaas, E.; Polania, J.; Rao, I.M.; Beebe, S.E.; Merchant, A. Effect of Drought and Low P on Yield and Nutritional Content in Common Bean. *Front. Plant Sci.* **2022**, *13*, 814325. [CrossRef] [PubMed]
3. Aquino-Bolaños, E.N.; Garzón-García, A.K.; Alba-Jiménez, J.E.; Chávez-Servia, J.L.; Vera-Guzmán, A.M.; Carrillo-Rodríguez, J.C.; Santos-Basurto, M.A. Physicochemical Characterization and Functional Potential of *Phaseolus vulgaris* L. and *Phaseolus Coccineus* L. Landrace Green Beans. *Agronomy* **2021**, *11*, 803. [CrossRef]
4. López, C.M.; Pineda, M.; Alamillo, J.M. Differential Regulation of Drought Responses in Two *Phaseolus vulgaris* Genotypes. *Plants* **2020**, *9*, 1815. [CrossRef] [PubMed]
5. Polania, J.A.; Chater, C.C.C.; Covarrubias, A.A.; Rao, I.M. Phaseolus Species Responses and Tolerance to Drought. In *The Plant Family Fabaceae*; Hasanuzzaman, M., Araújo, S., Gill, S.S., Eds.; Springer: Singapore, 2020; pp. 319–336, ISBN 978-981-15-4751-5.
6. Bianco, M.L.; Grillo, O.; Cremonini, R.; Sarigu, M.; Venora, V. Characterisation of Italian Bean Landraces ("*Phaseolus vulgaris*" L.) Using Seed Image Analysis and Texture Descriptors. *Aust. J. Crop Sci.* **2015**, *9*, 1022–1034.
7. Yi, X.; Eramian, M.; Wang, R.; Neufeld, E. Identification of Morphologically Similar Seeds Using Multi-Kernel Learning. In Proceedings of the 2014 Canadian Conference on Computer and Robot Vision, Montreal, QC, Canada, 6–9 May 2014; IEEE: New York, NY, USA, 2014; pp. 143–150.
8. Garcia, M.; Chaves, D.; Trujillo, M. An Automatic Bean Classification System Based on Visual Features to Assist the Seed Breeding Process. In *Trends and Advancements of Image Processing and Its Applications*; Johri, P., Diván, M.J., Khanam, R., Marciszack, M., Will, A., Eds.; EAI/Springer Innovations in Communication and Computing; Springer International Publishing: Cham, Switzerland, 2022; pp. 165–176, ISBN 978-3-030-75944-5.
9. Kılıç, K.; Boyacı, İ.H.; Köksel, H.; Küsmenoğlu, İ. A Classification System for Beans Using Computer Vision System and Artificial Neural Networks. *J. Food Eng.* **2007**, *78*, 897–904. [CrossRef]
10. Ozaktan, H.; Çetin, N.; Uzun, S.; Uzun, O.; Ciftci, C.Y. Prediction of Mass and Discrimination of Common Bean by Machine Learning Approaches. *Environ. Dev. Sustain.* **2023**, *26*, 18139–18160. [CrossRef]
11. Mendoza, F.A.; Kelly, J.D.; Cichy, K.A. Automated Prediction of Sensory Scores for Color and Appearance in Canned Black Beans (*Phaseolus vulgaris* L.) Using Machine Vision. *Int. J. Food Prop.* **2017**, *20*, 83–99. [CrossRef]
12. Eric, O.; Gyening, R.-M.O.M.; Appiah, O.; Takyi, K.; Appiahene, P. Cocoa Beans Classification Using Enhanced Image Feature Extraction Techniques and a Regularized Artificial Neural Network Model. *Eng. Appl. Artif. Intell.* **2023**, *125*, 106736. [CrossRef]
13. Djoulde, K.; Ousman, B.; Hamadjam, A.; Bitjoka, L.; Tchiegang, C. Classification of Pepper Seeds by Machine Learning Using Color Filter Array Images. *J. Imaging* **2024**, *10*, 41. [CrossRef] [PubMed]

14. Koklu, M.; Ozkan, I.A. Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques. *Comput. Electron. Agric.* **2020**, *174*, 105507. [CrossRef]
15. Mendigoria, C.H.; Concepcion, R.; Dadios, E.; Aquino, H.; Alaias, O.J.; Sybingco, E.; Bandala, A.; Vicerra, R.R.; Cuello, J. Seed Architectural Phenes Prediction and Variety Classification of Dry Beans (*Phaseolus vulgaris*) Using Machine Learning Algorithms. In Proceedings of the 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), Bangalore, India, 30 September–2 October 2021; IEEE: New York, NY, USA, 2021; pp. 1–6.
16. Nasirahmadi, A.; Behroozi-Khazaei, N. Identification of Bean Varieties According to Color Features Using Artificial Neural Network. *Span. J. Agric. Res.* **2013**, *11*, 670–677. [CrossRef]
17. Morales-Reyes, J.L.; Acosta-Mesa, H.G.; Aquino-Bolanos, E.N.; Herrera-Meza, S.; Cruz-Ramirez, N.; Chávez-Servia, J.L. Classification of Bean (*Phaseolus vulgaris* L.) Landraces with Heterogeneous Seed Color Using a Probabilistic Representation. In Proceedings of the 2021 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 10–12 November 2021; IEEE: New York, NY, USA, 2021; pp. 1–7.
18. Morales-Reyes, J.L.; Acosta-Mesa, H.-G.; Aquino-Bolaños, E.-N.; Herrera Meza, S.; Márquez Grajales, A. Anthocyanins Estimation in Homogeneous Bean Landrace (*Phaseolus vulgaris* L.) Using Probabilistic Representation and Convolutional Neural Networks. *J. Agric. Eng.* **2023**, *54*. [CrossRef]
19. Morales-Reyes, J.-L.; Aquino-Bolaños, E.-N.; Acosta-Mesa, H.-G.; Márquez-Grajales, A. Estimation of Anthocyanins in Heterogeneous and Homogeneous Bean Landraces Using Probabilistic Colorimetric Representation with a Neuroevolutionary Approach. *MCA* **2024**, *29*, 68. [CrossRef]
20. Sönmez, M.E.; Sabancı, K.; Aydın, N. Classification of Wheat Rootstock and Their Hybrids According to Color Features by Machine Learning Algorithms. *Int. J. Appl. Math. Electron. Comput.* **2022**, *10*, 39–48. [CrossRef]
21. López-Lobato, A.-L.; Avendaño-Garrido, M.-L.; Acosta-Mesa, H.-G.; Morales-Reyes, J.-L.; Aquino-Bolaños, E.-N. Bean Landraces Color Identification Through Image Analysis and Gaussian Mixture Model. In *Advances in Computational Intelligence. MICAI 2024 International Workshops*; Martínez-Villaseñor, L., Ochoa-Ruiz, G., Montes Rivera, M., Barrón-Estrada, M.L., Acosta-Mesa, H.G., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2025; Volume 15465, pp. 112–124, ISBN 978-3-031-83881-1.
22. Aquino-Bolaños, E.; García-Díaz, Y.; Chávez-Servia, J.; Carrillo-Rodríguez, J.; Vera-Guzman, A.; Heredia-Garcia, E. Anthocyanins, Polyphenols, Flavonoids and Antioxidant Activity in Common Bean (*Phaseolus vulgaris* L.) Landraces. *Emir. J. Food Agric.* **2016**, *28*, 581. [CrossRef]
23. McGuire, R.G. Reporting of Objective Color Measurements. *HortSci* **1992**, *27*, 1254–1255. [CrossRef]
24. Bishop, C.M. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer: New York, USA, 2006; ISBN 978-0-387-31073-2.
25. Morales-Reyes, J.-L.; Aquino-Bolaños, E.-N.; Acosta-Mesa, H.-G. Color Quantification in Common Bean Landraces Using a Supervised Learning Technique. In *Advances in Computational Intelligence. MICAI 2024 International Workshops*; Martínez-Villaseñor, L., Ochoa-Ruiz, G., Montes Rivera, M., Barrón-Estrada, M.L., Acosta-Mesa, H.G., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2025; Volume 15465, pp. 167–178, ISBN 978-3-031-83881-1.
26. Gustavo, E.A.P.A.B.; Diego, F.S. How K-Nearest Neighbor Parameters Affect Its Performance. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b83da505b60e32469152b986cbd6199842403b11 (accessed on 15 April 2025).
27. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 4th ed.; Pearson India: Uttar Pradesh, India, 2018; ISBN 978-93-5306-298-9.
28. Khan, S.; Rahmani, H.; Shah, S.A.A.; Bennamoun, M. *A Guide to Convolutional Neural Networks for Computer Vision*; Synthesis Lectures on Computer Vision; Morgan & Claypool Publishers: San Rafael, CA, USA, 2018; ISBN 978-1-68173-021-9.
29. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep Learning for Visual Understanding: A Review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
30. Bjorck, J.; Gomes, C.; Selman, B.; Weinberger, K.Q. Understanding Batch Normalization. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31, pp. 7705–7716.
31. Galvan, E.; Mooney, P. Neuroevolution in Deep Neural Networks: Current Trends and Future Challenges. *IEEE Trans. Artif. Intell.* **2021**, *2*, 476–493. [CrossRef]
32. Barradas-Palmeros, J.-A.; López-Herrera, C.-A.; Acosta-Mesa, H.-G.; Mezura-Montes, E. Efficient Neural Architecture Search: Computational Cost Reduction Mechanisms in DeepGA. In *Advances in Computational Intelligence. MICAI 2024 International Workshops*; Martínez-Villaseñor, L., Ochoa-Ruiz, G., Montes Rivera, M., Barrón-Estrada, M.L., Acosta-Mesa, H.G., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, Switzerland, 2025; Volume 15465, pp. 125–134, ISBN 978-3-031-83881-1.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.