Anthocyanins estimation in homogeneous bean landrace (*Phaseolus vulgaris* L.) using probabilistic representation and convolutional neural networks

José Luis Morales-Reyes,[1] Héctor-Gabriel Acosta-Mesa,[1] Elia Nora Aquino Bolaños,[2] Socorro Herrera Meza,[3] Aldo Márquez Grajales[1]

[1]Artificial Intelligence Research Institute; [2]Centre for Food Research and Development; [3]Institute of Psychological Research, University of Veracruz, Xalapa, Veracruz, Mexico

Abstract

Studying chemical components in food of natural origin allows us to understand their nutritional contents. However, nowadays, this analysis is performed using invasive methods that destroy the sample under study. These methods are also expensive and time-consuming. Computer vision is a non-invasive alternative to determine the nutritional contents through digital image processing to obtain the colour properties. This work employed a probability mass function (PMF) in colour spaces HSI (hue, saturation, intensity) and CIE L*a*b* (International Commission on Illumination) as inputs for a convolutional neural network (CNN) to estimate the anthocyanin contents in landraces of homogeneous colour. This proposal is called AnthEstNet (Anthocyanins Estimation Net). Before applying the CNN, a methodology was used to take digital images of the bean samples and extract their colourimetric properties represented by PMF. AnthEstNet was compared against regression methods and artificial neural networks (ANN) with different characterisation in the same colour spaces. The performance was measured using precision metrics. Results suggest that AnthEstNet presented a behaviour statistically equivalent to the invasive method results (pH differential method). For probabilistic representation in channels H and S, AnthEstNet obtained a precision value of 87.68% with a standard deviation of 10.95 in the test set of samples. As to root mean square error (RMSE) and R^2, this configuration was 0.49 and 0.94, respectively. On the other hand, AnthEstNet, with probabilistic representations on channels a* and b* of the CIE L*a*b* colour model, reached a precision value of 87.49% with a standard deviation of 11.84, an RMSE value of 0.51, and an R^2 value of 0.93.

Correspondence: Héctor-Gabriel Acosta-Mesa, Artificial Intelligence Research Institute, University of Veracruz, Campus Sur, Calle Paseo 112, Lote 2, Colonia Nueva Xalapa, C.P. 91097, Xalapa, Veracruz, Mexico. E-mail: heacosta@uv.mx

Key words: anthocyanin estimation, bean landrace, colour distribution, probabilistic representation, machine learning.

Acknowledgements: the first author acknowledges the National Council of Science and Technology (CONACyT) of Mexico for granting support for the realisation of this investigation through scholarship 712056. We also thank Dr. José Luis Chavez Servia for the samples of bean landraces used in this work.

Received: 8 April 2022.
Accepted: 28 September 2022.

Journal of Agricultural Engineering 2023; LIV:1421
doi:10.4081/jae.2023.1421

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Introduction

Chemical components studied in food of natural origin provide knowledge about their nutritional contents. For example, colour is a property used to evaluate the maturity degree of a fruit or vegetable. Furthermore, food colour determines nutrition impact and disease prevention (Horbowicz *et al*., 2008; Steinmetz and Potter, 1996).

Food analysis has focused on a compound present in flowers, leaves, fruits, vegetables, and seeds called anthocyanins (Singh and Singh, 2018). Anthocyanins belong to the flavonoid group responsible for red, purple, and blue colouration (Garzón, 2008). Moreover, this compound has benefits for health. For example, its antioxidant activity favours diabetes control and reduces coronary heart disease risks (Ataie-Jafari *et al*., 2008). Furthermore, anthocyanins prevent cancer cell proliferation (Grimes *et al*., 2018), cholesterol reduction (Farrell *et al*., 2015), and kidney stone prevention (Nirumand *et al*., 2018). Also, anthocyanins have anti-inflammatory properties (Bowen-Forbes *et al*., 2010; Hidalgo *et al*., 2012).

The quantification of anthocyanins is regularly obtained by a laboratory procedure (Wrolstad, 1993). However, this procedure is a complex and destructive process that consumes time, human, economic, and chemical resources. For example, Aquino-Bolaños *et al*. (2016) report the quantification of anthocyanins from 26 common bean landraces. First, they manually separated the seed coat, adding solvents to the extracts to generate a homogeneous mixture. Moreover, the pH differential method was employed to obtain the anthocyanin content. As a result, they obtained concentrations between 0.04 and 9.07 *mg* (*C3G*)/*g* dry weight.

Other non-invasive techniques for the quantification of anthocyanins currently employed include computational vision methods. For example, Yoshioka *et al*. (2013) estimated the contents of anthocyanins in digital images of strawberries from different sites.

This estimation was performed by computing the average values of each channel in the colour spaces RGB, HSL, and CIE L*a*b* and applying multiple regression over these average values. Results suggest that the average values are relevant for estimating anthocyanins, reaching a value for the determination coefficient of 0.978 for RGB colour space, 0.948 for HSL, and 0.980 for CIE L*a*b*.

Also, Taghadomi-Saberi *et al*. (2014) estimated the antioxidant activity and the anthocyanins amount contained in six ripening states of cherry present in 420 cherry specimens grouped in 42 sets. Three colour spaces (RGB, HSV, and CIE L*a*b*) were used for the digital image processing stage, and each digital image was characterised using 11 features (statistical mean, statistical variance, asymmetry, kurtosis, chromaticity, among others). Furthermore, several ANNs were implemented for the estimation task, selecting the one with the higher performance (a correlation coefficient of 0.98). Finally, the quantification results were evaluated by mean absolute error, reaching an error rate of 0.13.

On the other hand, Fernandes *et al*. (2015) estimated the anthocyanins present in 240 bunches of grapes. Six grapes were randomly selected from the samples to be captured in digital images in which the reflectance spectrum (reflected intensity percentage by the sample) was computed. Due to the high dimensionality of such a spectrum, a reduction dimensionality was performed using the principal component analysis (PCA), obtaining 14 components for the description of the reflectance spectrum. Based on these components, an ANN was performed to estimate anthocyanins obtaining a coefficient range of 0.87 and 0.95.

Similarly, Chen *et al*. (2015) determined the contents of anthocyanins using digital images with 50 grapes and a 928-1695 nm wavelength range. Furthermore, the average value of the interest regions in each image was computed for the analysis, and the support vector regression was employed as a predictive model using this average as the predictor variable. The obtained anthocyanins coefficient of determination was 0.9414.

Another research on the contents of anthocyanins is presented by Del Valle *et al*. (2018), where anthocyanins of six species of flowers were estimated in petals, stems, pedicels, and calyxes. For the data representation, they used the colour average of the tissues in each colour channel of the RGB colour space and some indexes as inputs for the simple linear regression algorithm. The reported determination coefficients by the regression models were between 0.63 and 0.94.

Chen *et al*. (2020) estimated the contents of anthocyanins in purple lettuce leaves. First, they captured digital images of each specimen and used the RGB, HSV, and CIE L*a*b* colour spaces. Once digital images were obtained, the average value of each channel in each colour space (R, G, B, H, S, V, L*, a*, and b*) was calculated to characterise them and perform the digital image processing. The colour features V, R-B, B/(R+G), 2R(G-B)/(G+B) y L*+a*+b* were employed to train the multiple linear regression models, support vector machines (SVM) and random forest method. Results suggest that the random forest method was the most competitive, with a determination coefficient of 0.5709 for the testing set.

Finally, Zhang *et al*. (2020) report the estimation of anthocyanins in black goji berries. Digital images were acquired using near-infrared hyperspectral imaging (NIR-HSI). For the spectra extraction, was employed the wavelet transform using Daubechies 8 for pre-processing the pixel-wise spectra in each black goji fruit. The successive projections algorithm and the competitive adaptive reweighted sampling were used for selecting the wavelengths. Moreover, principal component analysis and wavelet transform were used to extract features. Three supervised learning approaches were employed for anthocyanin estimation, partial least squares regression, SVM, and CNN. The performance between models showed similarity. However, no significant differences were obtained. The reported determination coefficients by the CNN models were between 0.63 and 0.94. Most of the research on the estimation of anthocyanins focuses on apples, strawberries, cherries, grapes, flowers, and purple lettuce leaves. However, to the best of our knowledge, no one has studied the estimation of anthocyanins in bean landraces. In México, many bean landraces are preserved by different ethnicities and cultivated for the private consumption of farmers, which is the researchers' main interest to analyse their nutritional components (Chávez-Servia *et al*., 2016). The main contribution of this paper is the colour characterisation as a two-dimensional PMF and how it can be used as the primary input to the CNN. Furthermore, our proposal is a robust, innovative, faster, cheaper, and non-invasive method for anthocyanin estimation that considers the entire image colour distributions of a little-explored food such as beans.

Consequently, the particular objectives of this work are described as follows. First, select the most suitable statistical or artificial intelligence techniques for estimating anthocyanins used in other domains of the estimation of anthocyanins. Finally, to compare the estimation obtained for the selected technique against the estimation computed by the pH differential method to analyse if our proposal is an alternative to this destructive, complex, and time-consuming technique, and if so, how accurate this artificial intelligence method is.

Materials and Methods

Bean landraces

This work employed 40 different colour homogeneous bean landraces (*Phaseolus vulgaris* L.). Therefore, the number of available examples was subject to the number of bean landraces available by the producer. In addition, the time required to quantify anthocyanins by the traditional invasive laboratory method increases as the number of samples increases. Consequently, the sample size was considered sufficient to estimate anthocyanins.

The bean landraces were recollected from several municipalities of Oaxaca, México. Each sample contains 60 grams of healthy and clean grains. Moreover, the colour groups used were: 18 black bean landraces, nine red bean landraces, eight yellow bean landraces, four white bean landraces, and one brown bean landrace. Due to the invasive and destructive chemical procedure, digital images of the bean grains were captured before applying the chemical methods for the estimation of anthocyanins.

Monomeric anthocyanins quantification

Anthocyanins were measured in the seed coats through the method described by Xu *et al*. (2007). First, the beans were soaked in distilled water for 12 hours to remove the seed coats. After, three grams were crushed using a mixture of 25 mL of acetone/water/acetic acid (70:29.5:0.5, v/v/v). Such a mixture was homogenised for 20 s (Wisetis Homogenizer, HG-15-A, 110 v; DAIHAN-brand, Gang-won, Korea). Therefore, the crushed material was centrifuged at 4000 rpm for 20 min (Hettich Centrifuge, Universal 32R, Tuttlingen, Germany) by removing the supernatant and repeating this procedure in the residues under the same conditions. Finally, the monomeric anthocyanin content was evaluated, pooling each supernatant of the crushed material. Moreover, the anthocyanin content was determined using the pH differential method. Two extract dilutions were made: potassium chloride buffer at pH 1.0 and sodium acetate buffer at pH 4.5. Subsequently, the absorption spectrum was obtained between 460 and 710 nm (Spectrophotometer UV-1800, Shimadzu, Kyoto, Japan) to determine maximum absorbance. Finally, the monomeric anthocyanin concentration was computed according to the equation proposed by Giusti and Wrolstad (2001). Results were expressed as mg cyanidin-3-glucoside per gram of dry weight [*mg* (*C*3*G*)/*g*].

Figure 1 shows, in ascending order, the anthocyanins concentration values obtained from the laboratory results. Moreover, it is essential to highlight the relationship between bean colouration and the amount of anthocyanin content.

Illumination environment and image reproduction workflow

An image capture workflow aims to obtain a digital representation as close as possible to the colour of the seeds of a landrace. The controlled illumination environment and colour image reproduction workflow proposed by Morales-Reyes *et al*. (2021) were employed to estimate anthocyanins in landraces. It follows a similar procedure reported by Korytkowski and Olejnik-Krugly (2017) to reproduce the colour image in computer vision.

This prototype consists of an aluminium box of 68 cm long × 68 cm wide × 60 cm tall and eight bulbs of fluorescent daylight balanced light of 45 watts with a correlated colour temperature equal to 6500K used for the photography illumination. Figure 2a shows the structure and components of the capture prototype. A hole was made at the top of the prototype (Figure 2c) to place the camera lens for capturing images.

Also, this prototype contains a diffusion box to decrease the specular reflexes and shadows among seeds. This diffusion box is an internal cube 38 cm long × 38 cm wide × 45 cm tall lined with a translucid white cloth leaving uncovered the box top for visualisation of the samples (Figure 2b). For measuring the relative spectral power distribution inside the diffusion box, was used the Sciencetech Monochromator 9057.

The colour image reproduction workflow used to define the final image digitisation process is expressed in Figure 3 and explained as follows:

- *Image capture*: SONY ILCE 3500 digital camera was used for image acquisition with exposure settings set at shutter speed 1/60, focal length 50 mm, ISO 100, and aperture f/8.0. The white balance was a custom set in the camera by putting the X-Rite ColourChecker Passport white chart as a reference in front of the digital camera. The X-Rite ColourChecker Passport classic chart with 24 patches was used as a reference for creating a custom camera profile.
- *RAW image processing (ARW 12-bits)*: The RAW image of reference was acquired with a resolution of 5456×3632 pixels. Darktable software was used to process raw images to create the standard ICC profile and assign sRGB as a color space. The images were saved in TIFF format, suitable for saving images without compression.
- *Custom ICC profile calculation*: The ColorChecker Camera Calibration v2.0 - X-Rite was employed to create the custom ICC profile, including a colourimetric label AToB1 for the transformation from RGB to PCS (Profile Connection Space), providing information to convert images from one colour space to another.
- *Custom ICC profile assignment*: Matlab software replaced the standard ICC profile with a custom ICC profile and converted the colour space sRGB to CIE L*a*b*.
- *Colour Accuracy ΔE^*_{ab}*: The colour accuracy is computed using an average and maximum delta E. The formula for determining the delta E value is CIE 1976 L*a*b*.

At the same time, the Minolta CM-2600d spectrophotometer (Konica Minolta Sensing Inc., Japan) was employed to measure the 24 patches of the ColourChecker Passport classic chart using an illuminant D50 observer angle of 10°. Each patch was measured six times and averaged the values of parameters L*, a*, and b*. The spectrophotometer was used as a validation test to verify that

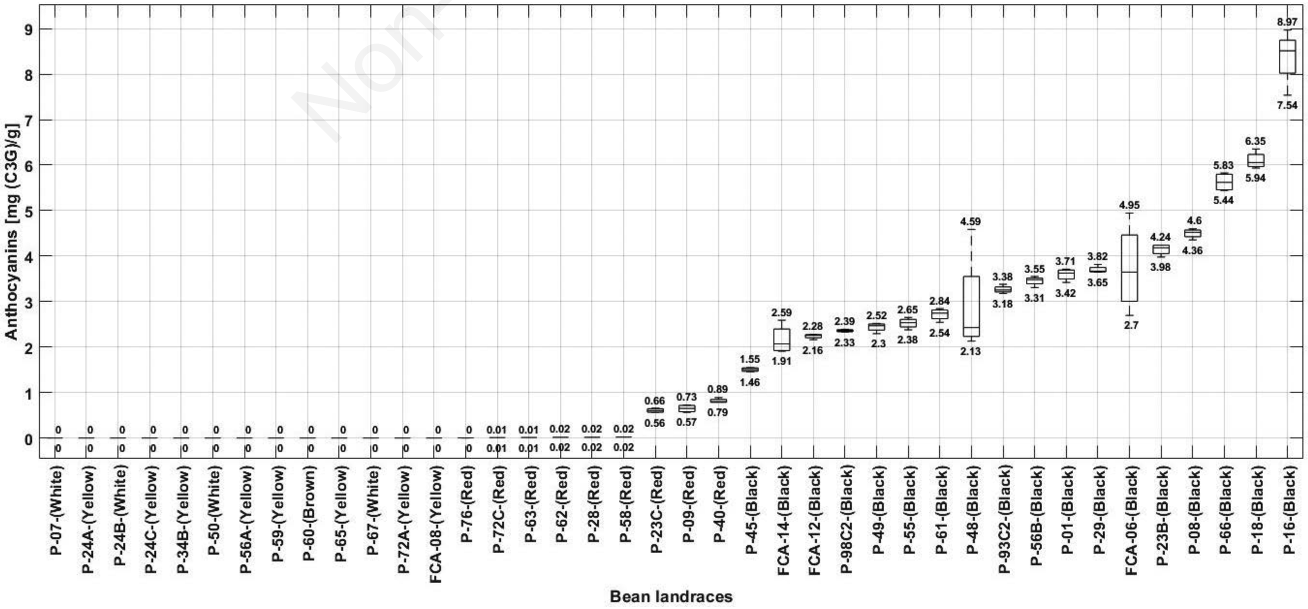


Figure 1. Samples of bean landraces sorted by average anthocyanins concentration.

the colour space conversions were accurate. Then, using a custom ICC profile, the RGB image values were converted to CIE 1976 L*a*b*. Finally, the colour differences ΔE_{ab} were computed between the spectral measurements and a pixel colour of the ColourChecker classic chart image when the camera calibration profile was applied. As a result, the colour accuracy achieved by CIE L*a*b*(ΔE^*_{ab}) was average =5.5 and max =11.8.

Image acquisition

The image acquisition process consisted of samples of the bean landraces. Such samples were located in a movable tray controlled by the prototype's front side. Also, the tray surface is covered with a blue background to contrast the seed landraces. Each captured image contains 60 grams of seeds with a proper separation to avoid occlusions and shadows. Figure 4 shows examples of the image captured by the prototype. Finally, the captured images were processed using the workflow explained above.

Image segmentation

Once each bean landrace is captured, extracting each seed area from the digital image is next. For this purpose, the growing region segmentation algorithm was employed. This method groups the neighbour pixels concerning a predefined similarity measure (Gonzalez and Woods, 2002; Tang, 2010). The first step of this method is to find pixel seeds as start points and then merge all pixels with similar seed properties. These merged pixels are now the new seed pixel. The process stops when no more pixels satisfy the similarity criterion.

As the background colour was homogeneous, it was employed as the similarity criterion for the segmentation algorithm. The three-colour channels of the background contrast colour were used for the initialisation of the seed algorithm, and as a similarity metric between a pixel and its neighbour has employed the colour accuracy, CIE L*a*b* 1976 (ΔE^*_{ab}). Eq. 1 computes the colour accuracy of CIE L*a*b* by the distance between the colour channels (L, a, b) of the sample (s) and the colour channels (L, a, b) of the reference image (r).

$$\Delta E^*_{ab} = \sqrt{(L_s - L_r)^2 + (a_s - a_r)^2 + (b_s - b_r)^2} \qquad (1)$$

Colour characterisation approaches

Several authors use colour spaces to estimate the quantification of anthocyanins (Yoshioka *et al*., 2013, Chen *et al*., 2015, 2020), like RGB, HSI, HSV, HSL, and CIE L*a*b*. However, it is desired to explore the chromaticity channels. For this reason, the RGB colour space is not appropriate to separate the chromaticity channels of luminosity.

Consequently, other colour spaces are appropriate to decouple intensity from chromaticity. For example, the CIE L*a*b* colour space is used in laboratory equipment for colour metrics due to its closeness to human visual perception. HSI (hue, saturation, intensity) colour space is also closer to human perception colour, calculated from RGB colour space. These colour characterisation forms (HSI, CIE L*a*b*) were used to evaluate which is adequate to develop a non-invasive anthocyanins estimation method and explore the chromaticity channels. For sRGB colour space, the colour image reproduction workflow described above is employed for transforming it to CIE L*a*b* colour space.

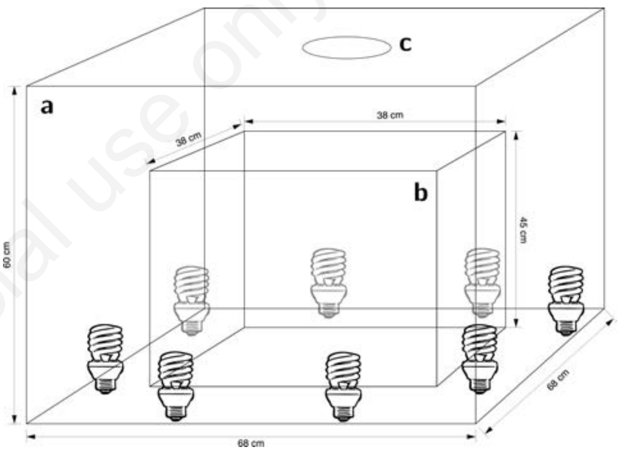


Figure 2. Image acquisition prototype: a- external box made of aluminium in which the bulbs of fluorescent daylight balanced light are connected; b- diffusion box used to decrease the specular reflections and shadows in the samples; c- hole for the camera lens.

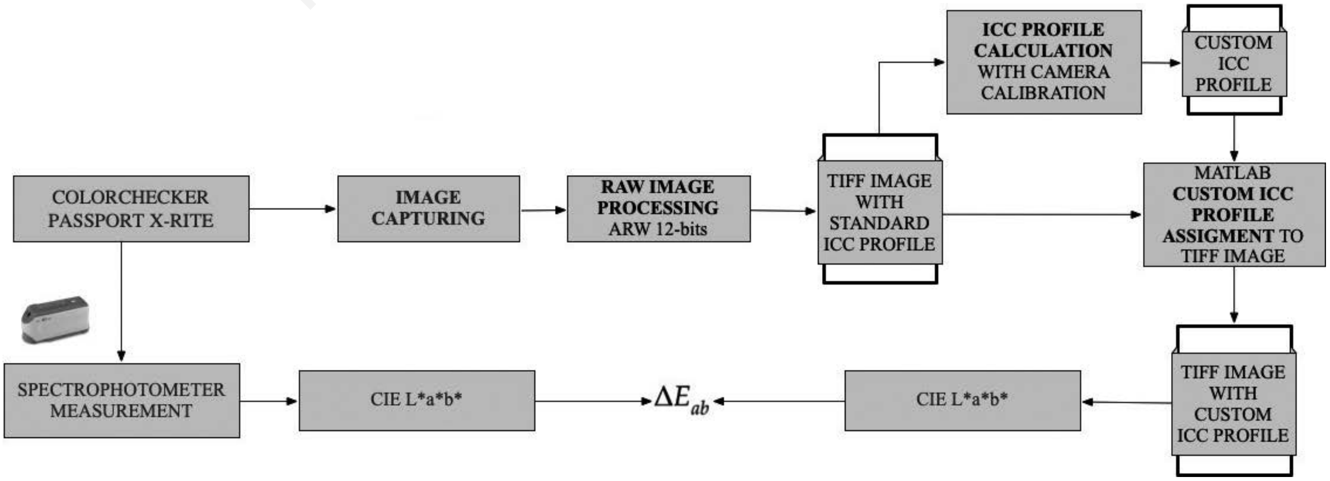

ICC PROFILE CALCULATION WITH CAMERA CALIBRATION
CUSTOM ICC PROFILE
COLORCHECKER PASSPORT X-RITE
IMAGE CAPTURING
RAW IMAGE PROCESSING ARW 12-bits
TIFF IMAGE WITH STANDARD ICC PROFILE
MATLAB CUSTOM ICC PROFILE ASSIGMENT TO TIFF IMAGE
SPECTROPHOTOMETER MEASUREMENT
CIE L*a*b*
ΔEab
CIE L*a*b*
TIFF IMAGE WITH CUSTOM ICC PROFILE

Figure 3. Representation of the workflow used to reproduce a colour image.

Averaging colour characterisation

The most straightforward colour characterisation approach is the average method. First, this technique extracts the colour channels L*, a*, and b* from 'landraces' segmented seeds. After that, the average value from each channel ($\bar{L}, \bar{a}, \bar{b}$) is obtained by Eq. 2, where i represents a pixel from the interest regions. In the same way, this procedure was applied to the HSI colour space.

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i \, , \bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i \, , \bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i \tag{2}$$

The colour average is a simple characterisation of the colour expressed as a three-dimensional vector for each bean landrace.

Probability mass function

Another more robust colour characterisation uses the joint probability mass function (PMF). For example, the conjunction of three-color channels in colour digital images conforms to the final colouration. Therefore, the pixel value occurrence frequency in a colour image is counted as the joint probability of each channel value in such pixel.

The joint PMF ($f_{X,Y}$) of two discrete random variables (X, Y) is computed as Eq. 3 (Pishro-Nik, 2016). Thus, $f_{X,Y}$ can be interpreted as the simultaneous probability of x and y values.

$$f_{X,Y} = P(X = x, Y = y) \tag{3}$$

The joint probability of the chromaticity channels $f_{i,j}$ is given by Eq. 4, where $o_{i,j}$ is the occurrence frequency of the pixel values inside a bin, and n is the number of occurrence frequencies ($o_{i,j}$).

$$f_{i,j} = \frac{o_{ij}}{n^2} \tag{4}$$

Under this representation, it is possible to form a two-dimensional histogram using the chromaticity channels of HSI and CIE L*a*b* colour space. The colour histogram bins number are computed by discretising the chromaticity channels (H, S for HSI space and a*, b* for CIE L*a*b*) into 256 values (2^8) because of the domain of a* and b* channels inside [–128, 127]. Consequently, the two-dimensional histograms are expressed in the interval [0, 255] by adding the lower boundary absolute value of the channel domain interval to each histogram value. The total occurrence frequency is obtained by counting the number of pixels belonging to a bin. This work represents the colourimetric properties representation of common bean landraces by colour histograms as joint probabilities. Moreover, this characterisation adequately represents the total colour distribution of the seeds set, not just an average. The histograms are represented as a matrix of dimension 256×256 (Figure 5).

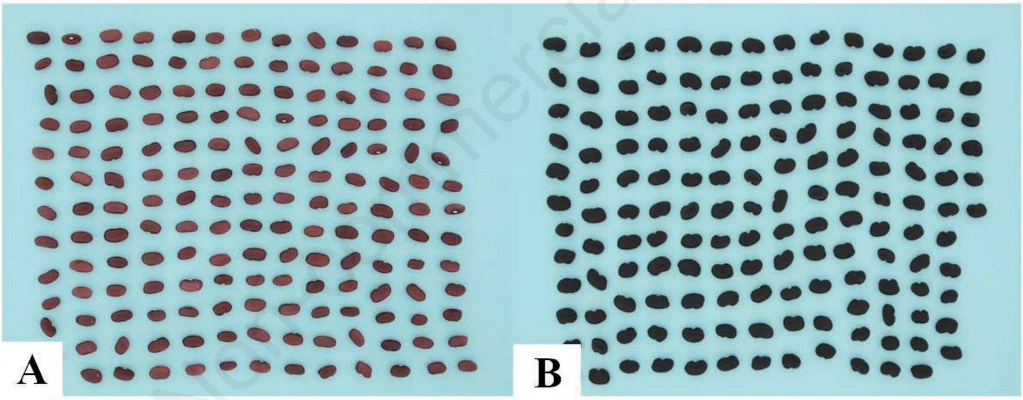


Figure 4. Digital image examples of beans landraces were captured through the acquisition prototype: A) 60 grams of red coat landraces; B) 60 grams of black coat landraces.

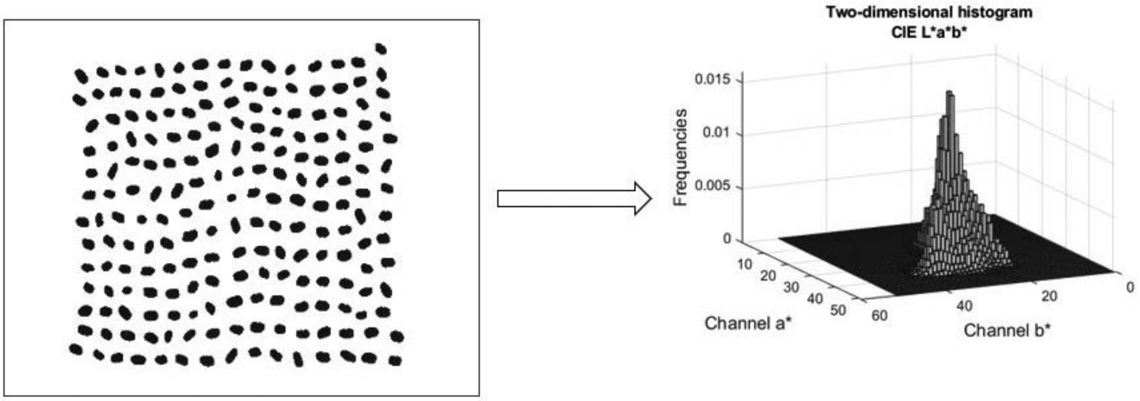


Figure 5. The colour distribution is represented as a joint probability mass function. The histogram is just a graph of a probability mass function.

Principal component analysis

Principal component analysis (PCA) is a statistical method widely used for data dimensionality reduction. PCA extracts the main data characteristics by linear transforming the original variables into orthogonal and uncorrelated variables called principal components (PCs) (Zhang *et al.*, 2020). This technique was employed to get the PCs from the colour histograms (PMF) expressed as vectors of 65,536 (256×256) positions.

PCA was used to compute the first 30 components that captured 98% of the variance. Then those components were used to project each PMF on those bases to have a compressed representation of 30 coefficients.

Evaluation measure

In estimation tasks for continuous values, several metrics are employed to compare the predicted value against the actual value. Notably, this work uses a precision metric based on the mean absolute percentage error (MAPE) measure.

The MAPE is the average absolute error between the predicted and current model values (Paul, 2000). This error is obtained by subtracting the actual value from the predicted value. Afterward, the error is converted in percentage, allowing a better understanding of the resulting value. Eq. 5 expresses the MAPE computation (Goodwin and Lawton, 1999). MAPE lower values are preferred.

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100}{n} \qquad (5)$$

The MAPE metric was integrated to know how close the anthocyanin estimation value is to the anthocyanin concentration value reported by the chemical method. This measure is easy to understand because it provides the error in terms of percentages.

On the other hand, the following formula is integrated to know how close the estimated value is in terms of precision to the actual value (Eq. 6).

$$Precision = 100 - MAPE \qquad (6)$$

Therefore, a higher precision value represents a better estimation of anthocyanin.

Dataset splitting

In this research, 40 homogeneous colour bean landraces samples were used. Each bean landrace was randomly divided into four subsets for training, validation, and testing to allocate 160 examples. In the case of regression methods, two splitting sets were done, one with 70% of the seeds (training set) and the other with 30% (testing set). Due to the requirements of these methods, three splitting sets were performed according to ANN and CNN algorithms, the first with 70% of the seeds (training set) and the others with 15% each (validation and testing set). Figure 6 shows an example of how each bean landrace was divided.

Anthocyanins estimation net (AnthEstNet)

A CNN is a deep neural network employed for image recognition, classification, and estimation, among other tasks (Kim, 2017). The main advantage of this modern approach is receiving a set of matrices as input, which is suitable for the colour histograms (PMF). Our proposal is a CNN designed and applied to estimate the anthocyanins in homogeneous-colour bean landraces. Our CNN, called *AnthEstNet* (anthocyanins estimation net), contains six convolutional and max-pooling layers selected experimentally. Figure 7 shows the basic structure of the AnthEstNet.

AnthEstNet receives a 255×255 input matrix corresponding to the colour histogram dimensions. The AnthEstNet's convolution layer contains 16 filters of 7×7, and the rest contains 18 filters of the same dimensions. Moreover, all convolutional layers contain a [1,1] stride and padding with [2,2,2,2]. They perform batch normalisation and the activation function ReLU. Furthermore, the pooling layers use the max strategy with dimensions of 2×2 and a [2,2] stride, the positions where the kernel will move. AnthEstNet contains a fully connected layer and a regression layer to estimate anthocyanins. The overall AnthEstNet workflow is described in Figure 8. This figure shows the integration of each aforementioned stage, as well as the proposed flow to estimate anthocyanins in bean landraces. Moreover, the AnthEstNet code is available at https://github.com/JLMR-Code-Creator/AnthEstNet.

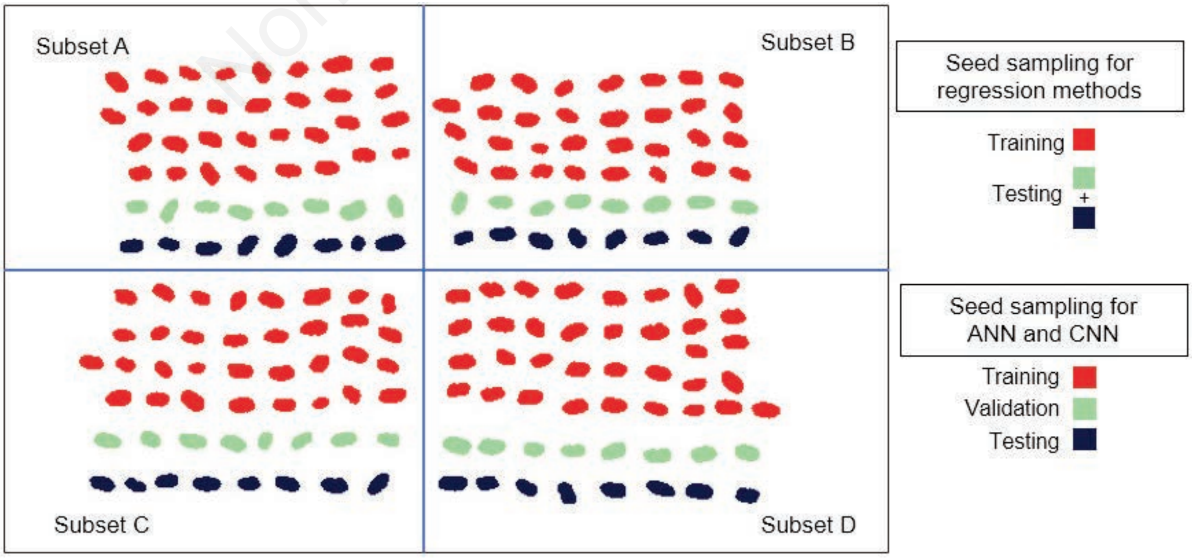


Figure 6. Example of the splitting process of seeds for creating the training, validation, and testing sets from bean landraces. ANN, artificial neural networks; CNN, convolutional neural network.

Experimental design

This research seeks a non-invasive method for the estimation of anthocyanins in homogeneous colour beans. Therefore, the experiments described in this section were designed to reach the objectives described previously.

Experiment 1 - Analysis of colour characterisation and statistical techniques

The first experiment performed is the comparison of statistical and artificial intelligence techniques. Regression models (linear, quadratic, pure quadratic, and regression with interactions) and ANN were employed for this comparison. The first one was evaluated using averaging colour characterisation and with different numbers of principal components mentioned before. Meanwhile, ANN was evaluated using a reduced histogram form as the primary input, compressed by the PCA technique (Taghadomi-Saberi *et al*., 2014; Fernandes *et al*., 2015). Next, several neural network architectures were evaluated by changing the number of neurons in the hidden layer.

Each method was performed using the MATLAB (R2021a) software (fitlm for regression and fitnet for ANN). The selected colour characterisations were the averaging colour characterisation and PCA. Finally, the precision metric was the evaluation measure used for selecting the model with the higher performance. Moreover, 20 executions were performed by each method applying a different random dataset split in each execution. The average precision values and their standard deviation were stored for each execution.

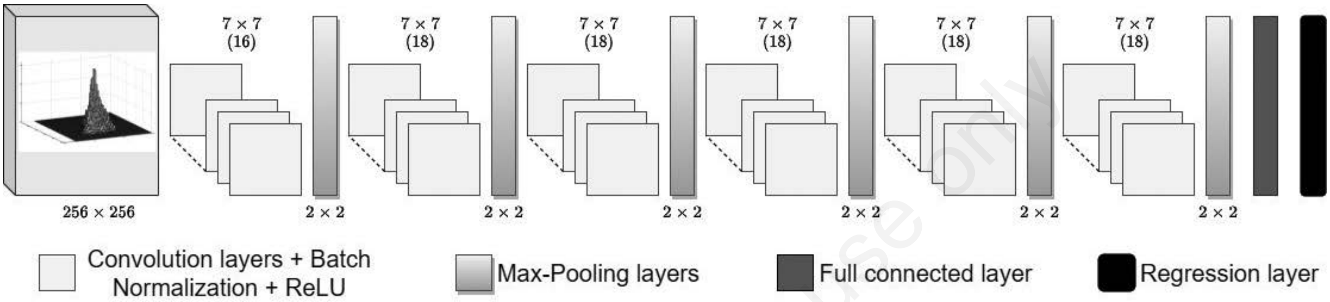


Figure 7. The architecture of AnthEstNet for anthocyanins estimation.

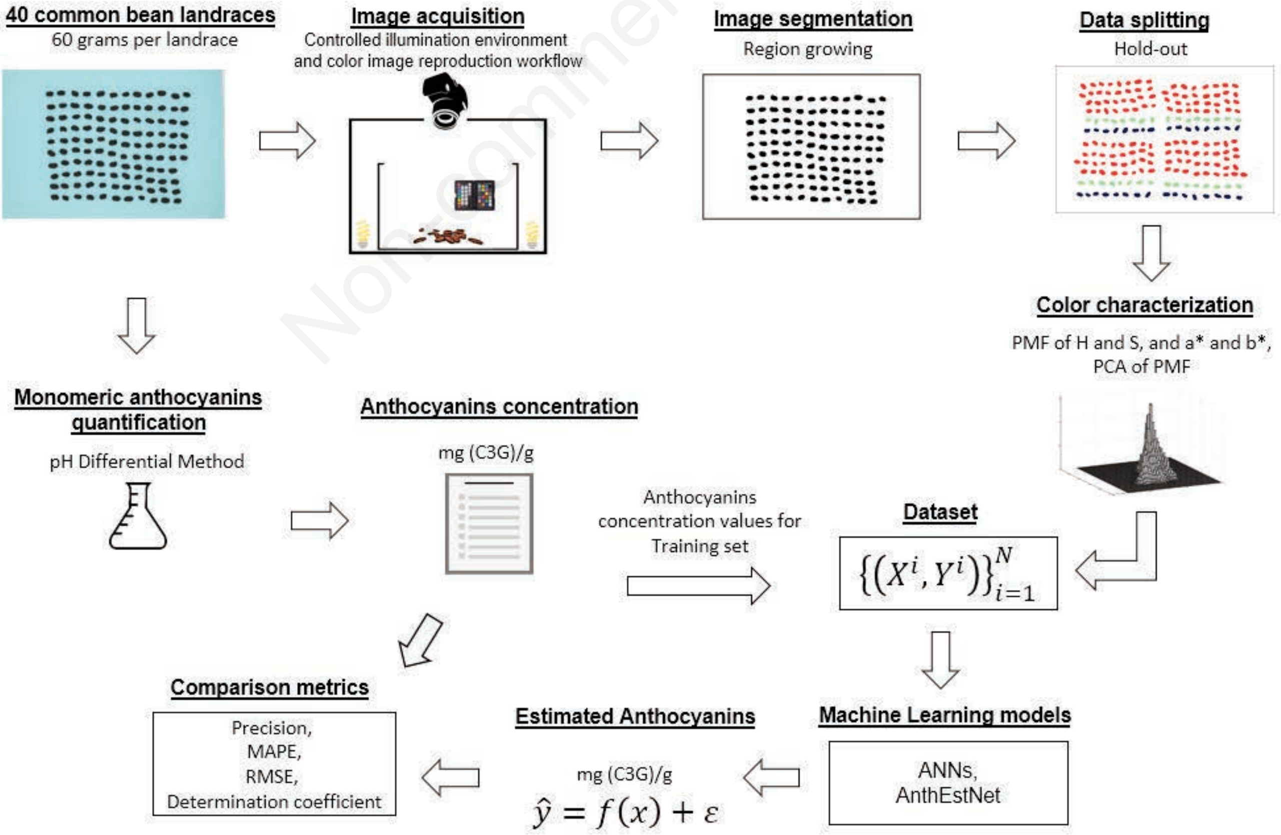

40 common bean landraces
60 grams per landrace
Image acquisition
Controlled illumination environment and color image reproduction workflow
Image segmentation
Region growing
Data splitting
Hold-out
Color characterization
PMF of H and S, and a* and b*, PCA of PMF
Monomeric anthocyanins quantification
pH Differential Method
Anthocyanins concentration
mg (C3G)/g
Anthocyanins concentration values for Training set
Dataset
$\{(X^i, Y^i)\}_{i=1}^{N}$
Comparison metrics
Precision, MAPE, RMSE, Determination coefficient
Estimated Anthocyanins
mg (C3G)/g
$\hat{y} = f(x) + \varepsilon$
Machine Learning models
ANNs, AnthEstNet

Figure 8. AnthEstNet overall workflow for estimation of anthocyanins.

Experiment 2 - Anthocyanin estimation obtained by AnthEstNet

This experiment was designed to compare our proposal (AnthEstNet) and the best models from the previous comparison. Therefore, our proposal was executed under the same conditions as the previously described methods. Furthermore, AnthEstNet was trained with 250 epochs, an initial learning rate of 0.001, and a function for updating the weights in the backpropagation process called stochastic gradient descent with momentum, which minimises the loss function by small steps in gradient direction in each algorithm iteration.

Concerning statistical comparisons, a normality test (Shapiro-Wilk test) was carried out to select suitable statistical techniques. Then, once normality distribution was confirmed, the parametric test one-way analysis of variance (ANOVA) was selected, applied with a 95%-confidence, and validated using the post hoc test called the Tukey's test. Furthermore, a correlation analysis of standard deviations was performed to investigate the relationship between the dispersion of the four chemical estimations of anthocyanins with the pH differential method and the proposed approaches. For this purpose, the execution with the best precision value was selected for the proposed techniques. The analysis aims to know the effect of variations between the reported dispersions of the anthocyanin estimation obtained by the pH differential method in each bean landraces and the standard deviation reached by ANN and AnthEstNet approaches. Like estimation analysis, the Shapiro-Wilk test was applied to observe if the standard deviation values follow a normal distribution. However, the normality test revealed that the standard deviation did not present a normal distribution. Therefore, the non-parametric test Spearman test was employed for the correlation analysis.

Results

Analysis of colour characterisation and statistical methods

Table 1 shows each execution's mean and standard deviation values in all evaluation measures for regression models and ANN, respectively, grouped by the colour space.

Anthocyanin estimation obtained by AnthEstNet

Table 2 details the results of comparing AnthEstNet against the best models found with ANN in each colour space.

Table 1. Results of the regression models and the artificial neural networks in each colour space.

Colour space		HSI				CIE L*a*b			
Model		Regression (interactions)	Regression (linear)	ANN (3:36:1)	ANN (28:3:1)	Regression (quadratic)	Regression (pure quadratic)	ANN (3:10:1)	ANN (6:21:1)
Colour characterisation technique		Averaging	PCA (10 components) 90.00% variance of data	Averaging	PCA (28 components) 98.31% variance of data	Averaging	PCA (17 components) 98.75% variance of data	Averaging	PCA (6 components) 90.71% variance of data
Training set	Precision	66.29%±29.05	85.84%±14.68	90.35%±11.10	89.71%±10.25	76.37%±28.89	89.46%±10.77	89.45%±11.07	89.52%±11.78
	RMSE	1.17	0.73	0.51	0.56	1.19	0.61	0.49	0.58
	R^2	0.69	0.88	0.93	0.92	0.69	0.93	0.96	0.92
Validation set	Precision	-	-	83.79%±20.89	87.04%±13.75	-	-	86.69%±14.04	86.73%±13.89
	RMSE	-	-	0.63	0.64	-	-	0.58	0.66
	R^2	-	-	0.90	0.90	-	-	0.92	0.89
Testing set	Precision	66.11%±30.05	85.61%±14.82	83.03%±24.35	86.83%±14.11	76.22%±28.89	85.06%±14.58	86.55%±14.41	87.04%±13.80
	RMSE	1.17	0.74	0.65	0.65	1.19	0.70	0.60	0.67
	R^2	0.69	0.88	0.90	0.90	0.69	0.90	0.91	0.89

HSI, hue, saturation, intensity; ANN, artificial neural networks; PCA, principal component analysis; RMSE, root mean square error. The ANN name depends on the number of input, hidden, and output layers separated by two points.

Table 2. Comparison of AnthEstNet against artificial neural networks models.

Colour space		HSI		CIE L*a*b	
Models		ANN (28:3:1)	AnthEstNet	ANN (6:21:1)	AnthEstNet
Colour characterisation technique		PCA (28 components) 98.31% variance of data	PMF of h and s	PCA (6 components) 90.71% variance of data	PMF of a* and b*
Training set	Precision	89.71%±10.25	93.83%±6.39	89.52%±11.78	93.52%±7.75
	RMSE	0.56	0.32	0.58	0.39
	R^2	0.928	0.97	0.92	0.96
Validation set	Precision	87.04%±13.75	87.93%±10.66	86.731%±13.89	87.51%±11.81
	RMSE	0.64	0.48	0.66	0.52
	R^2	0.90	0.94	0.89	0.93
Testing set	Precision	86.83%±14.11	87.68%±10.95	87.04%±13.80	87.49%±11.84
	RMSE	0.65	0.49	0.67	0.51
	R^2	0.90	0.94	0.89	0.93

HSI, hue, saturation, intensity; ANN, artificial neural networks; PCA, principal component analysis; PMF, probability mass function; RMSE, root mean square error.

Figures 9 and 10 show the comparison results between the anthocyanins estimation reached by ANN with PCA of PMF, AnthEstNet with PMF, and the anthocyanins estimation reported by the laboratory procedure pH differential method (Figure 1) in both colour spaces.

Figure 11A shows the statistical results for precision measure where no significant statistical difference among all compared methods is presented with a P-value of 0.06. As a result, the standard deviation of the 20 executions was considered to analyse the data variance. Since standard deviation values also present a nor-

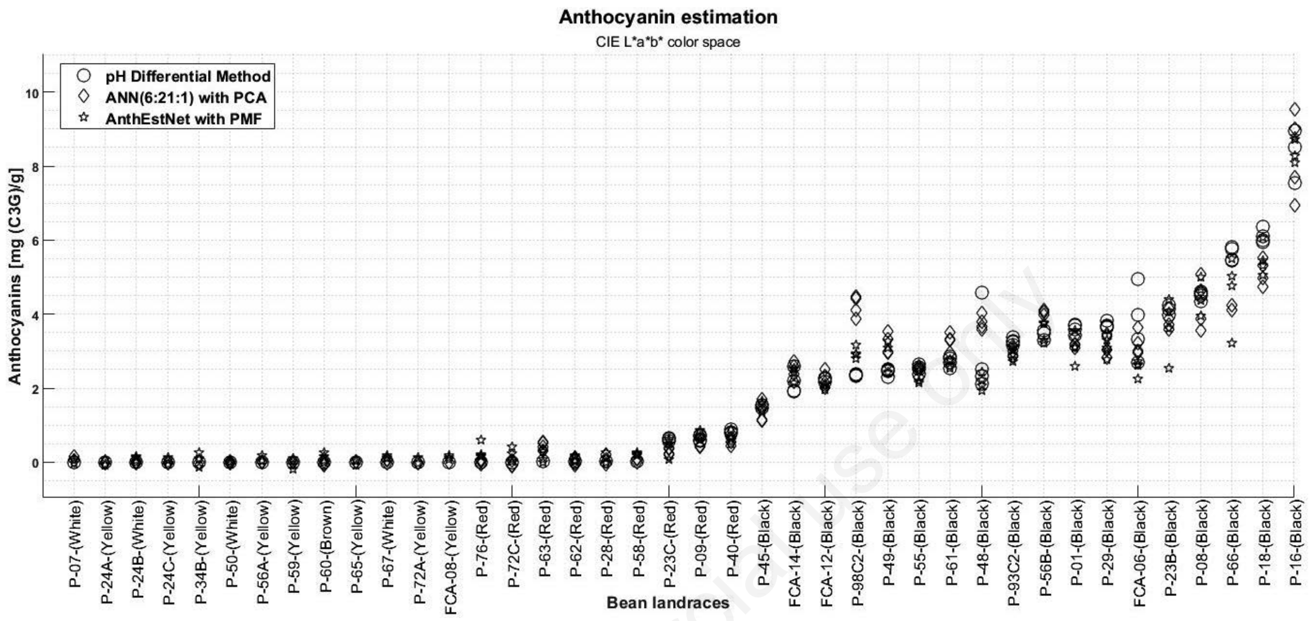


Figure 9. Estimation results comparison reached by each approach against the laboratory procedure results in CIE L*a*b colour space. ANN, artificial neural networks; PCA, principal component analysis; PMF, probability mass function.

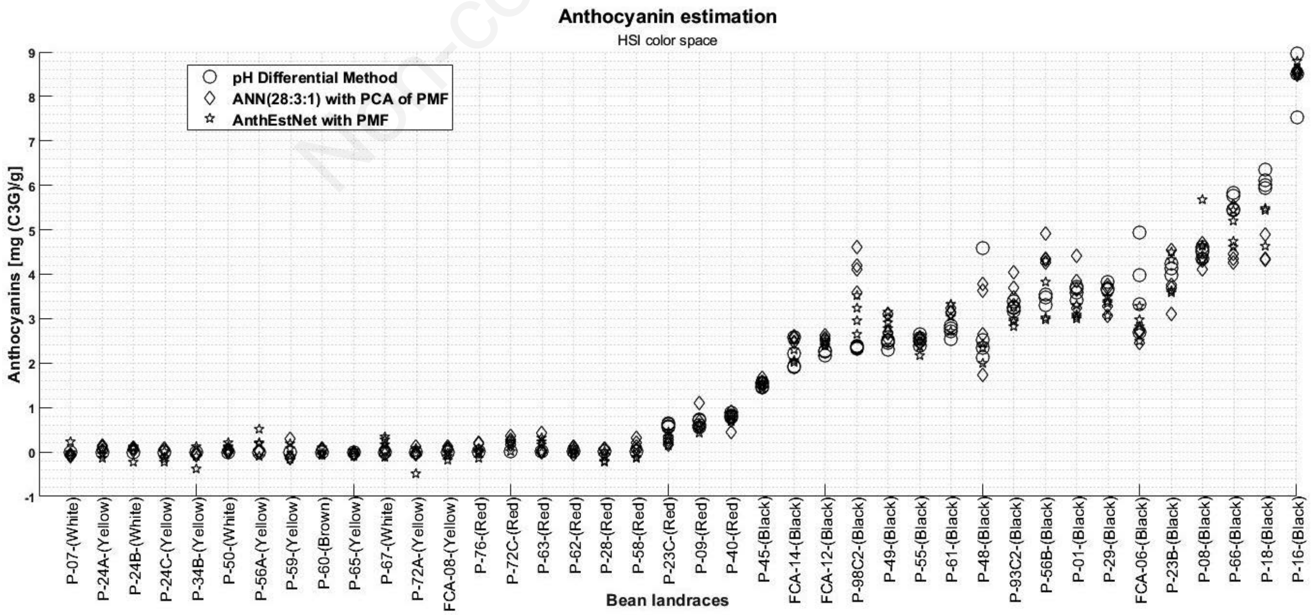


Figure 10. Estimation results comparison reached by each approach against the laboratory procedure results in hue, saturation, intensity (HSI) colour space. ANN, artificial neural networks; PCA, principal component analysis; PMF, probability mass function.

mal distribution, ANOVA and Tukey's tests were applied for the statistical analysis. Figure 11B presents the statistical results on standard deviation values where AnthEstNet groups present significant statistical differences compared to ANN groups with a P-value of 1.73301×10^{-10}.

Regarding correlation analysis, Figure 12 shows the standard deviation of estimation values reported by five methods in each of the 40 bean landraces. Besides, Table 3 shows the correlation and P-values between the pH differential method and the computational methods proposed.

Discussion

Different approaches and colour characterisations were explored to know their potential in anthocyanin estimation in landraces. As we can observe in Table 1, the averaging colour characterisation and the regression models reached a lower anthocyanin estimation precision in the testing set with a precision value of 66.11%±30.05 and 76.22%±28.89 for the HSI and CIE L*a*b*, respectively. For ANN, the precision value obtained was 83.03% ±24.35 in the HSI colour space and 86.55%±14.41 for CIE L*a*b* colour space. As a result, the ANN outperforms the regression

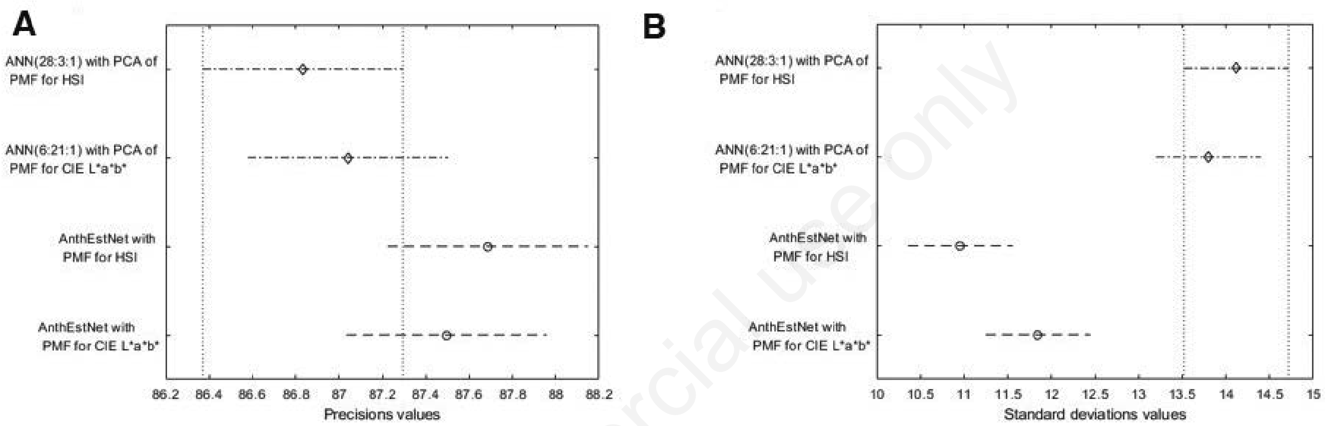
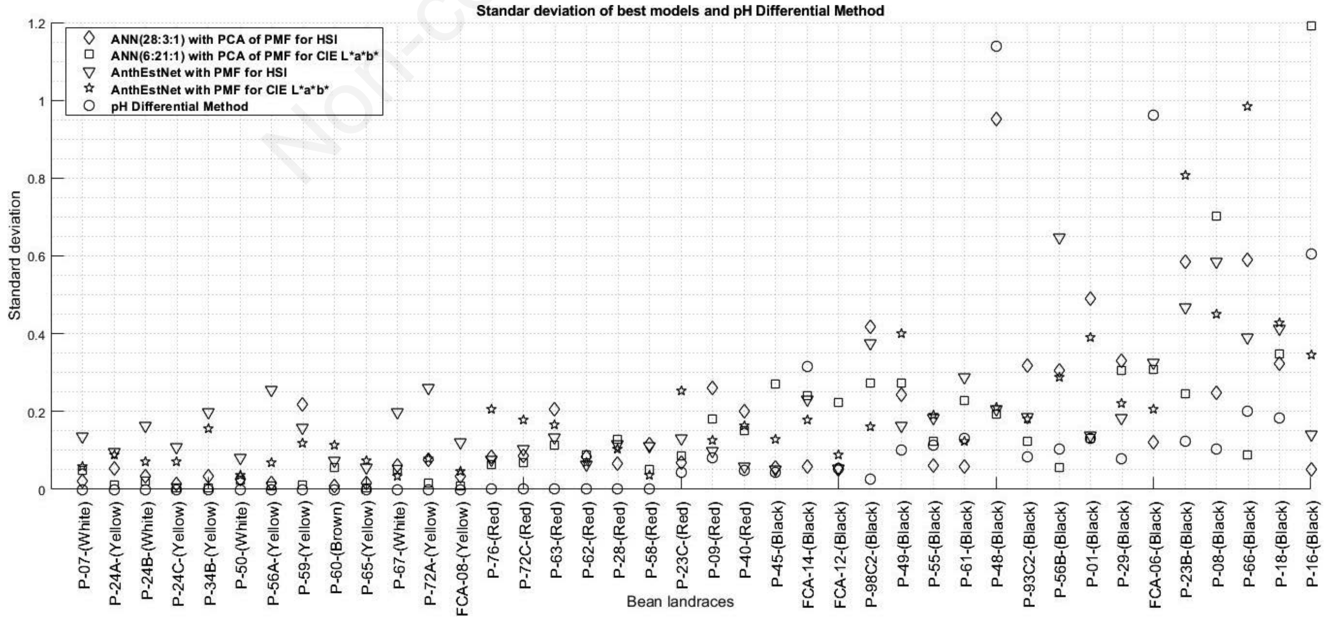


Figure 11. ANOVA statistical results for precision values (**A**) and standard deviation (**B**). ANN, artificial neural networks; PCA, principal component analysis; PMF, probability mass function; HSI, hue, saturation, intensity.



ANN, artificial neural networks; PCA, principal component analysis; PMF, probability mass function; HSI, hue, saturation, intensity.

models reaching a lower anthocyanin estimation error in the testing set. On the other hand, the characterisation of colour by the PMF compressed with the PCA method and the regression model reduced the estimation error of anthocyanins, reaching a precision value of 85.61%±14.82 using ten components (explaining 90% of the PMF information) in the HSI colour space. Regarding the CIE L*a*b space colour, PCA obtained a precision value of 85.06%±14.58 with 17 components, explaining 98.75% of the data variance. It is essential to mention that PCA characterisation provides more information than averaging colour characterisation. The performance of ANN with PCA characterisation decreases the estimation error in HSI colour space, obtaining a precision value of 86.83%±14.11 with 28 components and 87.04%±13.80 with six components (explaining the 90.71% of data variance) for CIE L*a*b.

ANN presents a more competitive behaviour in both colour spaces than the regression models because ANN presents a better adjustment to the data. Moreover, using the gradient descent and backpropagation methods in ANN significantly minimizes the estimation error regarding the regression models. However, different characterisation techniques got the best values in each colour space due to the several ANN architectures employed. Results suggest that the colour histogram's (PMF) PCA method obtains the lower estimation error concerning the regression models. Furthermore, since a colour histogram is a complex structure created from a matrix with PMF, it provides more information about the colour properties related to the number of anthocyanins present in the bean landraces, allowing lower values of estimation errors. Consequently, this structure was employed as input in our proposal to improve the anthocyanin estimation.

On the other hand, Table 3 shows that the AnthEstNet model using the chromaticity colour histogram of the colour space used as input reached a precision value of 87.68±10.95 and 87.49±11.84 in the CIE L*a*b colour space. The precision mean values were statistically equivalent in all compared experiments. However, the standard deviation values obtained by the AnthEstNets were minor compared to those obtained by the ANN models in both colour spaces. The ANOVA statistical test analysed the differences between precision and standard deviation values, and the post hoc test called the Tukey test. Regarding precision values, statistical tests confirm that they are equivalent among all compared methods, but in the standard deviation values, the statistical test presents significant statistical differences, confirming that AnthEstNet presents a more robust behavior than ANN models.

Moreover, since AnthEstNet uses colour histograms as colour characterisation, it does not require a compression process, allowing the integration of more colorimetric information in the estimation process. Finally, AnthEstNet can capture the spatial relationships of the 2D histogram (PMF).

According to Table 3, the standard deviation values of compared models' estimations and pH differential method present a positive correlation, which means if the pH differential method estimation presents a change, it also will be reflected by AnthEstNet estimation. As a result, AnthEstNet presents a similar behaviour in estimation and data variance with invasive-chemical methods.

Although the estimations reached were closer to the pH differential method results in most bean landraces, none of its anthocyanin estimations are close to the actual value in the P-98C2 sample black colour, because AnthEstNet employs the chromaticity channels (a* and b*) of CIE L*a*b* to estimate the anthocyanins concentration. This situation is presented in channels H and S of the HSI colour space. However, another detected problem was the similarity of chromaticity in several samples, causing estimation errors by not having more information to identify different samples. The results showed that AnthEstNet could potentially be used to estimate total anthocyanins in bean landraces.

Conclusions

This work estimated anthocyanins in bean landraces using conventional and deep learning-based techniques. The averaging colour characterisations and PCA of probabilistic distribution with regression techniques and ANN were implemented to analyse the estimation with conventional and widely used techniques for the estimation of anthocyanins. Finally, a CNN, using a set of colourimetric probabilistic characterisations as input, was compared against the methods mentioned above to verify if the estimation could be improved using the deep learning approach.

Results suggest that AnthEstNet reached a competitive estimation of anthocyanins regarding the pH differential method, demonstrating that using CNN is viable as a non-invasive alternative for anthocyanins estimation in bean landraces.

The colour characterisation using a probabilistic distribution of the bean landraces colour provides essential information about the colour patterns related to the anthocyanin concentration, a crucial characteristic for minimising the estimation error. Moreover, the probabilistic distribution can represent the colour distribution of a 20-megapixel image of 8-bit colour depth. Consequently, the CNN with PMF as input is a competitive tool able to reach anthocyanin estimations closer to the laboratory procedure in most of the beans landraces samples, representing a promising method to get estimations quickly, similar to the non-invasive method with an average precision of 87.68% (±10.95) and 87.49% (±11.84) in the colour spaces HSI and CIE L*a*b*.

The colour concentration relationship was explored in this research by considering the chromaticity channels. Consequently, three aspects can be explored to improve the AnthEstNet results in future work: i) to analyse the contribution of the luminosity channel as a part of the colour characterisation to decrease the estimation error; ii) AnthEstNet was used to estimate anthocyanins in bean landraces of homogeneous colour. Therefore, a study of greater complexity, such as anthocyanin estimation in heteroge-

Table 3. Correlation results among the standard deviation of anthocyanin estimations models. P letter represents the P-value reached by the Spearman test.

Correlation of standard deviation of estimationsp	ANN (28:3:1) with PCA of PMF for HSI	PCA of the PMF for CIE L*a*b* and ANN (6:21:1)	AnthEstNet with PMF for HSI	AnthEstNet with PMF for CIE L*a*b*
PH differential method	0.6309 P: 0.0000	0.8236 P: 0.0000	0.4418 P: 0.0043	0.7970 P: 0.0000

ANN, artificial neural networks; PCA, principal component analysis; PMF, probability mass function; HSI, hue, saturation, intensity.

neous coloured bean landraces, can be introduced; iii) building a convolutional neural network is a process that involves several parameters. For this reason, we propose to use methods to find the appropriate architecture for the estimation of anthocyanins task incorporating the Neuroevolution approach.

References

Aquino-Bolaños E.N., García-Díaz Y.D., Chavez-Servia J.L., Carrillo-Rodríguez J.C., Vera-Guzmán A.M., Heredia-García, E. (2016). Anthocyanins, polyphenols, flavonoids and antioxidant activity in common bean (Phaseolus vulgaris L.) landraces. Emirates J. Food Agric. 581-8.

Ataie-Jafari A., Hosseini S., Karimi F., Pajouhi M. 2008. Effects of sour cherry juice on blood glucose and some cardiovascular risk factors improvements in diabetic women: a pilot study. 38:355-60.

Bowen-Forbes C.S., Zhang Y., Nair M.G. 2010. Anthocyanin content, antioxidant, anti-inflammatory and anticancer properties of blackberry and raspberry fruits. J. Food Compos. Analysis 23:554-60.

Chai T., Draxler R.R. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7:1247-50.

Chávez-Servia J.L., Heredia-García E., Mayek-Pérez N., Aquino-Bolaños E.N., Hernández-Delgado S., Carrillo-Rodríguez J.C., Gill-Langarica H.R., Vera-Guzmán A.M. 2016. Diversity of common bean (Phaseolus vulgaris L.) landraces and the nutritional value of their grains. In: A.K. Goyal (Ed.), Grain legumes. IntechOpen. Availble from: https://doi.org/10.5772/63439

Chen S., Zhang F., Ning J., Liu X., Zhang Z., Yang S. 2015. Predicting the anthocyanin content of wine grapes by NIR hyperspectral imaging. Food Chem. 172:788-93.

Chen Y., Zheng L., Wang M., Wu M., Gao W. 2020. Prediction of chlorophyll and anthocyanin contents in purple lettuce based on image processing 2020 ASABE Annual International Virtual Meeting, St. Joseph, MI, USA.

del Valle J.C., Gallardo-López A., Buide M.L., Whittall J.B., Narbona E. 2018. Digital photography provides a fast, reliable, and noninvasive method to estimate anthocyanin pigment concentration in reproductive and vegetative plant tissues. Ecol. Evol. 8:3064-76.

Farrell N., Norris G., Lee S.G., Chun O.K., Blesso C.N. 2015. Anthocyanin-rich black elderberry extract improves markers of HDL function and reduces aortic cholesterol in hyperlipidemic mice [10.1039/C4FO01036A]. Food Funct. 6:1278-87.

Fernandes A.M., Franco C., Mendes-Ferreira A., Mendes-Faia A., Costa P.L.D., Melo-Pinto P. 2015. Brix, pH and anthocyanin content determination in whole Port wine grape berries by hyperspectral imaging and neural networks. Comput. Electron. Agric. 115:88-96.

Garzón G.A. 2008. Las antocianinas como colorantes naturales y compuestos bioactivos: revisión. Acta Biol. Colomb. 13:27-36.

Giusti M.M., Wrolstad R.E. 2001. Characterization and measurement of anthocyanins by UV-visible spectroscopy. Curr Protocols Food Analyt. Chem. 00:F1.2.1-F1.2.13.

Gonzalez R.C., Woods R.E. 2002. Digital image processing. Prentice Hall Upper Saddle River, NJ, USA.

Goodwin P., Lawton R. 1999. On the asymmetry of the symmetric MAPE. Int. J. Forecast. 15:405-8.

Grimes K.L., Stuart C.M., McCarthy J.J., Kaur B., Cantu E. J., Forester S.C. 2018. Enhancing the cancer cell growth inhibitory effects of table grape anthocyanins. J. Food Sci. 83:2369-74.

Hidalgo M., Martin-Santamaria S., Recio I., Sanchez-Moreno C., de Pascual-Teresa B., Rimbach G., de Pascual-Teresa S.J.G. 2012. Potential anti-inflammatory, anti-adhesive, anti/estrogenic, and angiotensin-converting enzyme inhibitory activities of anthocyanins and their gut metabolites. Nutrition 7:295-306.

Horbowicz M., Kosson R., Grzesiuk A., Dębski H. 2008. Anthocyanins of Fruits and vegetables - their occurrence. Analy. Role Human Nutr. 68:5.

Kim P. 2017. Convolutional neural network. pp. 121-147 in MATLAB deep learning. Springer.

Korytkowski P., Olejnik-Krugly A. 2017. Precise capture of colors in cultural heritage digitization. Color Res. Appl. 42:333-6.

Morales-Reyes J.L., Acosta-Mesa H.G., Aquino-Bolaños E.N., Herrera-Meza S., Cruz-Ramírez N., Chávez-Servia J.L., 2021. Classification of bean (Phaseolus vulgaris L.) landraces with heterogeneous seed color using a probabilistic representation 2021 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 2021, pp. 1-7, doi: 0.1109/ROPEC53248.2021.9668106.

Nirumand M.C., Hajialyani M., Rahimi R., Farzaei M.H., Zingue S., Nabavi S.M., Bishayee A. 2018. Dietary plants for the prevention and management of kidney stones: preclinical and clinical evidence and molecular mechanisms. Int. J. Mol. Sci. 19:765.

Paul M.S. 2000. MAPE (mean absolute percentage error). In: P.M. Swamidass (Ed.), Encyclopedia of production and manufacturing management. Springer US, pp. 462-462.

Pishro-Nik H. 2016. Introduction to probability, statistics, and random processes. Kappa Research LLC, 2014, available from: https://www.probabilitycourse.com

Singh B., Singh S. 2018. Advances in postharvest technologies of vegetable crops. Apple Academic Press.

Steinmetz K.A., Potter J.D. 1996. Vegetables, fruit, and cancer prevention: a review. J. Am. Diet. Assoc. 96:1027-39.

Taghadomi-Saberi S., Omid M., Emam-Djomeh Z., Ahmadi H. 2014. Evaluating the potential of artificial neural network and neuro-fuzzy techniques for estimating antioxidant activity and anthocyanin content of sweet cherry during ripening by using image processing. J. Sci. Food Agricult. 94:95-101.

Tang J. 2010. A color image segmentation algorithm based on region growing. 2nd International Conference on Computer Engineering and Technology, Chengdu, China, 2010, pp. V6-634-V6-637, doi: 10.1109/ICCET.2010.5486012.

Wrolstad R.E. 1993. Color and pigment analyses in fruit products. Agricultural Experiment Station, Oregon State University. Station Bulletin, 624.

Xu B.J., Yuan S.H., Chang S.K.C. 2007. Comparative analyses of phenolic composition, antioxidant capacity, and color of cool season legumes and other selected food legumes. J. Food Sci. 72:S167-77.

Yoshioka Y., Nakayama M., Noguchi Y., Horie H. 2013. Use of image analysis to estimate anthocyanin and UV-excited fluorescent phenolic compound levels in strawberry fruit. Breed Sci. 63:211-7.

Zhang C., Wu W., Zhou L., Cheng H., Ye X., He Y. 2020. Developing deep learning based regression approaches for determination of chemical compositions in dry black goji berries (Lycium ruthenicum Murr.) using near-infrared hyperspectral imaging. Food Chem. 319:126536.