Journal of Food Engineering xxx (xxxx) 111850

Contents lists available at ScienceDirect

Journal of Food Engineering

journal homepage: www.elsevier.com/locate/jfoodeng

Machine learning for the identification of colour cues to estimate quality parameters of rocket leaves

Michela Palumbo [a, b, *], Maria Cefola [a], Bernardo Pace [a], Giancarlo Colelli [b], Giovanni Attolico [c]

[a] Institute of Sciences of Food Production, National Research Council of Italy (ISPA-CNR), C/o CS-DAT, Via Michele Protano, 71121, Foggia, Italy
[b] Department of Agriculture, Food, Natural Resources and Engineering (DAFNE), University of Foggia, Via Napoli 25, 71122, Foggia, Italy
[c] Institute on Intelligent Industrial Systems and Technologies for Advanced Manufacturing, CNR-National Research Council of Italy (STIIMA-CNR) Via G. Amendola, 122/O, 70126, Bari, Italy

ARTICLE INFO

Keywords:
Machine learning
Colour features identification
Chlorophyll and ammonia prediction
Fast and interpretable algorithms

ABSTRACT

Computer Vision Systems (CVSs) have proved to be a powerful tool to evaluate the quality of agricultural products in a non-destructive, contactless, sustainable and objective way. Machine learning techniques have proved to simplify the development of CVS and to provide better performance and greater flexibility in matching the requirements of different products and environmental characteristics, but they are often computationally complex and difficult to be understood by humans. It is desirable to develop methods that exploit the benefits of learning and generate simple and fast solutions that are also interpretable by humans. The approach described in this paper analyses a previously developed and effective machine learning model to extract the information useful to develop computationally light and easily understandable algorithms that evaluate the characteristics of interest on rocket leaves. A Random Forest model previously developed to classify visual quality and to estimate chlorophyll and ammonia contents in rocket leaves has been studied to identify a small set of visual characteristics (colours) that correlate with relevant properties of the product. These visual characteristics have been used as input for several simple, fast and easily understandable algorithms that classify visual quality (QL) and estimate chlorophyll and ammonia contents with lower computational complexities compared to the original Random Forest model. Results obtained by these methods are shown and compared with the ones provided by the original Random Forest model. All the algorithms provided a good separation between marketable and non-marketable samples. They required from 1ms to 22 ms to classify a new sample instead of the 25 ms of the original Random Forest model. Additionally, two methods provided good prediction of chlorophyll (R^2_v = 0.70) and ammonia (R^2_v = 0.72) contents requiring only 3 ms and 1 ms respectively.

1. Introduction

Recently, researchers have focused their attention on contactless, non-destructive, rapid, accurate and more sustainable techniques to objectively assess sensory and compositional quality of fruit and vegetables. Nevertheless, although these non-invasive methods offer significant advantages compared to analytical and destructive analyses, they cannot completely replace them. They are complementary, enabling lower time and cost, continuous and reliable monitoring and reduction of impact on environment along the supply chain (Chaudhry et al., 2020). Computer Vision Systems (CVSs) represent an innovative and contactless non-destructive technology suitable for in-line grading and quality assessment of fruit and vegetables (Fan et al., 2020). The integration of machine learning techniques inside CVSs has proved to make them more effective, more flexible e easier to be designed and configured for specific tasks on different products. A serious limitation of machine learning methodologies is their difficulty to be understood by humans. Unfortunately, th most powerful models are too complex to be evaluated and interpreted and operate as black-box: ensemble methods and deep learning are just two examples of very successful tools that are difficult to be explained. Instead, there is an increasing need of explainable machine learning in most application domains, for several reasons ranging from reliability assessment up to legal questions depending on their use (Lisboa et al., 2023). The growing research on eXplainable Artificial Intelligence (XAI) aims to explain the behaviour of complex methodologies (Wang et al., 2023). However, even when these ap-

* Corresponding author. Institute of Sciences of Food Production, National Research Council (CNR), c/o CS-DAT, Via Michele Protano, 71121, Foggia, Italy.
E-mail address: michela.palumbo@ispa.cnr.it (M. Palumbo).

https://doi.org/10.1016/j.jfoodeng.2023.111850
Received 25 May 2023; Received in revised form 31 October 2023; Accepted 19 November 2023
0260-8774/© 20XX

proaches succeed in explaining the behaviour of intelligent agents, they do not simplify their complexity. Sometimes, time constraints of specific applications can make acceptable a small loss in accuracy while most of the best performing machine learning approaches are not suitable for tuning the trade-off between effectiveness and complexity.

This research described in this paper exploited machine learning to extract information useful to design simple and fully understandable algorithms for quality assessment or characteristics estimation in agricultural products. The goal was not to investigate the best machine learning approach to solve the tasks of interest. Instead, we started from a machine learning methodology that in previous experiments had proved to be proficient in evaluating quality and in estimating internal properties of rocket leaves and used it to identify effective colour features. Then these features were used to develop simple and fully interpretable algorithms to accomplish the desired tasks, accepting limited losses of performance to achieve significative reductions in terms of computational complexity. In our experiments, we moved from a previous work (Palumbo et al., 2022), in which a machine learning model based on the Random Forest methodology had been used to solve a classification problem (assessment of quality level of rocket leaves) and two regression problems (estimation of chlorophyll and ammonia content of rocket leaves). That system worked successfully on both packaged and unpackaged products. The Random Forest approach is based on the hypothesis that the combination of several weak learners, each configured as a decision tree, can provide more powerful and robust results in classification and regression (Breiman, 1984). The Random Forest methodology has proved to be successful to achieve several goals in agriculture: to solve organic and conventional discrimination problems (Natarajan and Ponnusamy, 2023), to distinguish healthy and infected vegetables crops (Ghule et al., 2020), to predict freshness index in fruit and vegetables (Gokhale et al., 2023). Each tree is trained on the same task of the complete forest using random subsets of the available training samples and of the considered features. The training of each tree selects the most relevant features among the available ones contained in the current random set to achieve the required results: therefore, the analysis of the resulting forest is expected to provide useful hints about the relevance of each feature. Random Forest models are generally quite efficient and effective, but their significant conceptual complexity may prevent their application in some operational contexts. Moreover, they are hard to be read and interpreted by humans. Our research was not intended to compare the performance of Random Forest with other machine learning techniques: we tried, instead, to overcome the limitations of this approach while exploiting as much as possible the contribution of learning.

Therefore, we started from the results previously achieved on packaged and unpackaged rocket leaves in Palumbo et al. (2022) to move further. An analysis of the features used by the Random Forest trees was used to identify a compact yet efficient set of colour cues that can be used by simpler classifiers and regressors. Due to the nature of the original universe of features (frequencies of colour occurrence in the product), this subset can be associated to colour regions that are informative about the nature of the product. Each region provides a single measure that becomes a single feature for new simpler classifiers and regressors. That simplifies the computational load. The structure of the classifiers and regressors used is very simple and easily interpretable. In fact, each specific colour region can be correlated by humans to specific characteristics of the products, whose changes, during the storage, depends on well-known chemical or physical processes. Therefore, this colours hints provide a sound and objective base to describe quality marker parameters of the product, whose textual and iconographic descriptions become more robust.

The two innovative goals achieved by this paper are: i) to use a sound and effective machine learning model to automatically identify objective and sound colour regions in the ap-plane of the CIELab color space that effectively correlate to properties of interest of rocket leaves; ii) to show simple and easily interpretable algorithms to separate marketable from non-marketable rocket leaves and to estimate their ammonia and chlorophyll contents.

2. Materials and methods

2.1. Identification and preliminary selection of clusters

The experiments used the same images of packaged and unpackaged rocket leaves acquired for Palumbo et al. (2022). They were chosen to make significant the comparison between the performance of the simpler algorithms with the Random Forest model used in that previous paper.

The Random Forest model described in Palumbo et al. (2022) was analysed to identify the colour features more relevant for classification and regression. That resulting small set of relevant features provided the input for the simple and understandable algorithms described in this paper. The vocabulary of features of the Random Forest was composed by the elements of colours histogram (percentage of presence of a specific colour) in the *ab*-plane of the CIELab colour space. The relevance of each colour feature was estimated by considering its use to generate splits in the trees and its correlation with other features (Loh, 2002; 2002). Then, features were sorted by decreasing importance. The n most important features (with n empirically set to 36) were selected providing a set of 36 colours (Fig. 1). In this way, the machine learning model was used only to select the colours more informative to accomplish the tasks of interest. Those features purposely do not contain the L component because past experiments have shown this channel too sensitive to illumination levels and to uneven distribution of light across the scene.

Some of the selected colours were isolated in the ab-plane of the CIELab space while others were close to each other. A hierarchical clustering approach was used to group the 36 colours: the resulting clusters are shown in Fig. 2 using both rectangular (A) and polar (B) representations (Hastie et al., 2009). The clusters composed by one or two isolated colours have been considered less able to identify colour regions of interest for our work and discarded.

Three clusters, whose colours are shown in Fig. 3, collected a larger number of elements, namely the cluster 4 (C4) which contains 10 colours, the cluster 5 (C5) with 11 colours, and the cluster 9 (C9) with 6 colours. A compact group of relevant colours has been considered able to suggest a region of the ab-plane with a significant relationship with the characteristics of interest. The rectangles defined by the minimum and maximum values of the two components of the colours of each cluster are shown in Fig. 4 in both rectangular and polar representations.

The proposed approach associates a region of the *ab*-plane to each of the more populated clusters. A single feature is extracted from each region. This reduces the number of features and simplify the processing required to their evaluation, that is evaluating the percentages of pixels of each image whose colour belongs to each region.

2.2. Methodologies for the definition of the colour region corresponding to each cluster

A first statistical analysis (data not shown), done using a one-way ANOVA, pointed out the greater relevance of C5 and C9 compared to C4 which was therefore discarded in the following experiments. Only in two features, the frequencies of colours belonging to the C5 and C9 respectively, were evaluated and used to build the simple and understandable algorithms that are presented in the paper. Each cluster was composed by a set of sparse points in the ab-plane: it was necessary to transform these points in a continuous connected region to which to assign the colours of pixels of each image. Several possibilities to define this transform were analysed and compared. It is relevant to note that the points belonging to C5 were well separated by any other clus-

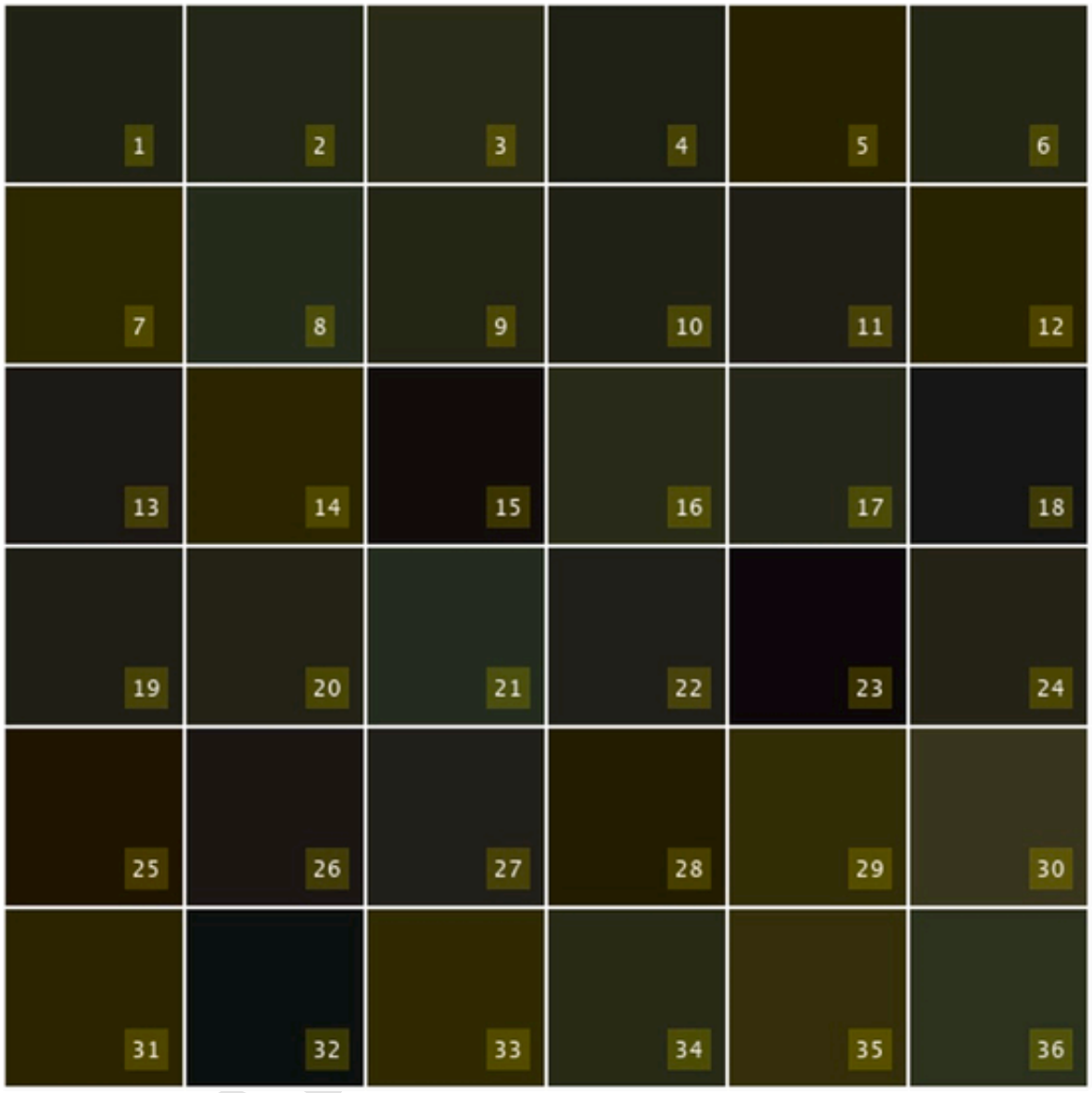


Fig. 1. The colours in the squares (from 1 to 36) are the most important features, resulting from the analysis of the Random Forest model, in descending order of importance.

ter both in the polar and rectangular representations of colours. On the contrary, the points belonging to C9 were very close to the cluster C4 in both the representations. Therefore, to correctly assign colours to C9 seemed feasible in the polar space but much more challenging in the rectangular space. For this reason, it was decided to apply each method to both the representations to verify their efficacy in transforming points into continuous connected regions suitable for the desired tasks. Each method also determines the extension and the shape of the region associated to each cluster: this choice affects the efficiency both in terms of computational complexity and flexibility.

Let us now describe in detail all the methods compared in this paper. Let us define (a_i, b_i) with $i = 1, \ldots, 36$, as the rectangular coordinates of the relevant colours identified in the *ab*-plane by the analysis of the Random Forest model. Every point in the plane can be expressed also using the polar representation:

$$\begin{aligned} distance_i &= \sqrt{a_i^2 + b_i^2} \\ angle_i &= arctg\left(\frac{b_i}{a_i}\right) \end{aligned} \tag{1}$$

For simplicity, all the formulas in the rest of the paper will be written only once using the symbols *x* and *y*. It is intended that they can be applied to both the representations (rectangular and polar), providing results that just need different interpretations. In the rectangular representation, *x* stands for *a* and *y* for *b* while in the polar representation *x* stands for *angle* and *y* for *distance*. Therefore, every method that we will describe has been applied into both the coordinates systems to compare their performance in characterizing the colour regions of interest. Let it denote:

$$min_x^k = \min\{x_i | i \in cluster\ k\}$$
$$max_x^k = \max\{x_i | i \in cluster\ k\} \tag{2}$$

$$min_y^k = \min\{y_i | i \in cluster\ k\}$$
$$max_y^k = \max\{y_i | i \in cluster\ k\} \tag{3}$$

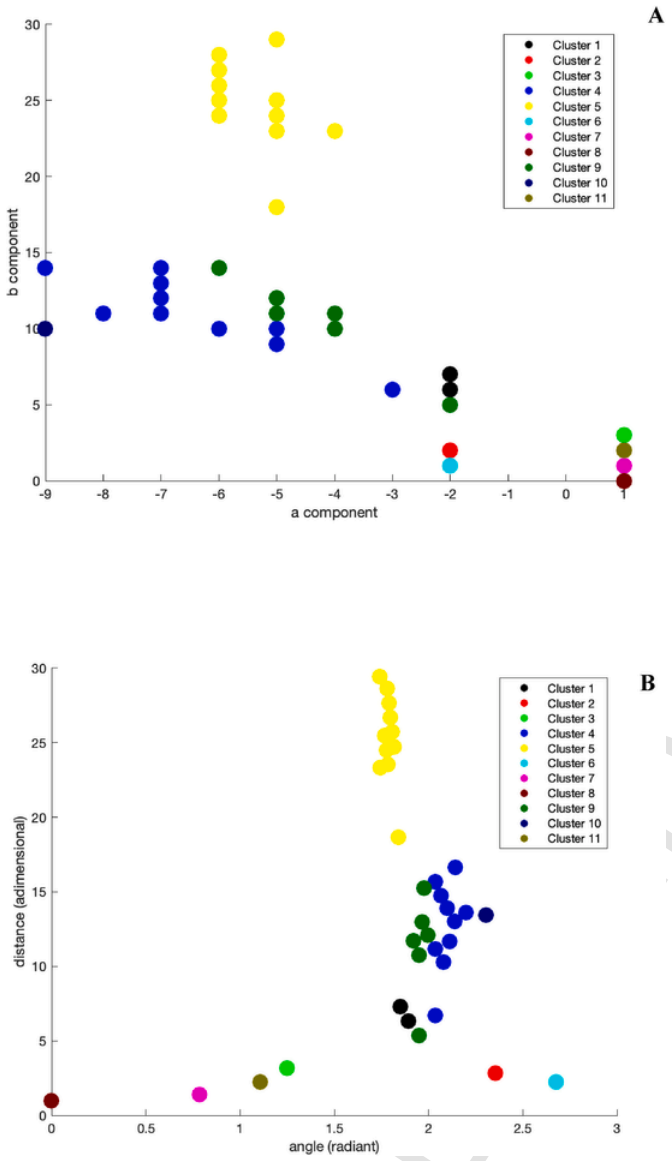


Fig. 2. The clusters of the relevant colours as identified by the analysis of the Random Forest model. In A, the abscissa and ordinate axes represents respectively the *a* and *b* components of the colours in the rectangular representation of the *ab*-plane in the CIELab color space. In B, the same points are represented in polar coordinates, where the ordinate represents the distance of the point from the origin (which is an achromatic point): this distance is evaluated in the ab-plane of the CIELab color space and it is adimensional. The abscissa represents the angle (in radiant) compared to the line having *b* equal to 0. In the latter representation, the elements of each cluster lie inside a narrow region of the x axis.

The method 1 (M1) associates to each cluster a rectangular region in the *ab*-plane whose limits are the minimum and maximum values of each component of the colours, out of the 36, belonging to that cluster. The upper-left and bottom-right corners of the resulting rectangular region have coordinates respectively:

$$\left(\min_x^k, \max_y^k\right) \qquad \left(\max_x^k, \min_y^k\right) \tag{4}$$

All the pixels whose colours belong to the rectangular region corresponding to the cluster *k*, are assigned to that cluster.

Fig. 4 shows that this method separates the clusters C4 and C9 in the polar representation but not in the rectangular representation in which the rectangular regions corresponding to C4 and C9 largely overlap.

The method 2 (M2) represents each region using the central point of the cluster, that is the point whose coordinates are the mean values between the minimum and the maximum of *x* and *y*. The central point of the cluster *k* is:

$$\left(\frac{\min_x^k + \max_x^k}{2}, \frac{\min_y^k + \max_y^k}{2}\right) \tag{5}$$

Each colour of an image is counted in the cluster whose central point is closest to it.

The method 3 (M3) represents each region using the centroid of the colours belonging to the cluster. The coordinates of the centroid of the cluster k are:

$$\left(\frac{\sum_i x_i}{n_k}, \frac{\sum_i y_i}{n_k}\right) \tag{6}$$

where $i \in cluster\ k$ and n_k is the number of colours, out of the 36, associated to the cluster k.

Each colour of an image is counted in the cluster whose centroid is closest to it.

The method 3 M3 has been applied in two different variants. In the first one (M3a), only the centroids of C5 and C9 have been considered when looking for the proper cluster for a colour in the image: each pixel was assigned to the cluster (out of these two) whose representative point (centroid) was closest to the colour of the pixel. In the second one (M3b), all the centroid points of all the clusters have been considered when looking for the proper cluster for a colour in the image: each pixel was assigned to the cluster (out of all the 11 clusters) whose representative point (centroid) was closest to the colour of the pixel. Even in this last case, only the values corresponding to C5 and C9 were considered for further processing. The difference between the two variants is that in the first variant (a) all the pixels are assigned either to C5 or to C9, according to which one is the closest. In the second case (b), a significative number of pixels were assigned to other clusters (different from C5 or C9) and were not considered in further processing. In geometrical terms, variant (b) reduces the size and modifies the shape of the regions of the *ab*-plane assigned to each of the two clusters of interest (C5 and C9).

The method 4 (M4) assigns each pixel to the colour, out of the 36, which is the closest in the *ab*-plane. Then, all the pixels associated to colours belonging to the same cluster are cumulated to evaluate the number of pixels belonging to that cluster. This last method enables a finer definition of the shape of the colour region associated to each cluster.

It is important to remember that each method has been applied using both the polar and the rectangular representations. The results have been compared to verify the expressivity of each representation compared to the tasks at hand. All the methods reported above were applied on the images of unpackaged and packaged samples of rocket leaves acquired by the CVS in Palumbo et al. (2022). The features corresponding to the different clusters were normalized: they were divided by the total number of foreground pixels in the image.

2.3. Statistical analysis

The values corresponding to the most relevant clusters (C5 and C9) were subjected to a one-way ANOVA analysis to find significant relationships with the quality level (QL) scores of rocket leaves reported in Palumbo et al. (2022).

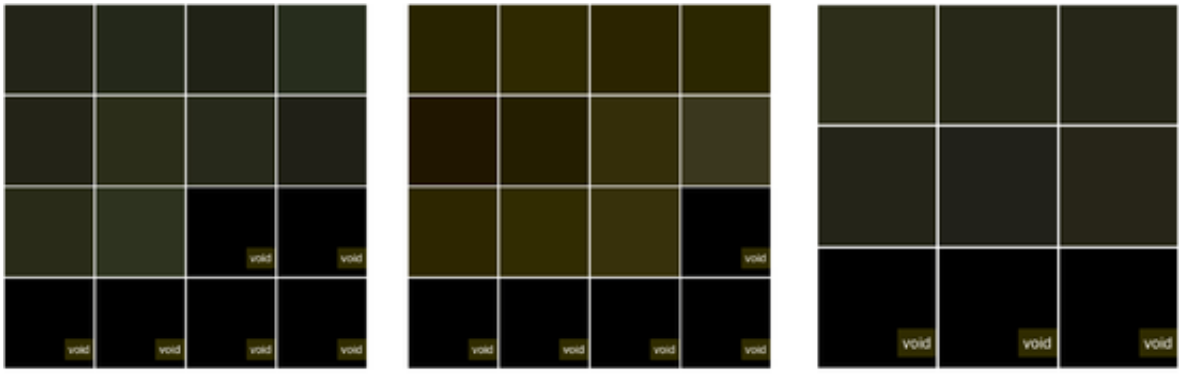


Fig. 3. Colours belonging to the most relevant clusters 4, 5 and 9. Please note that the black cells marked as *void*, are empty and purposely introduced in the image to simplify its construction.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The mean values were separated using the Student-Newman-Keuls (SNK) test and Statgraphics Centurion (version 18.1.12, Warrenton, Virginia, USA) was used for statistical analyses.

Principal component analysis (PCA) was performed by the software Statistica (version 6.0, StatSoft, Inc., Tulsa, OK, USA), using as variables the values of the C5 and C9 obtained by all the methods described above, in both the polar or rectangular versions (Method 1, Method 2, Method 3a, Method 3b, Method 4) and the chemical data of total chlorophyll and ammonia content previously reported (Palumbo et al., 2022). While, as the case, data were mediated in two visual quality group: 5-4-3 (marketable) and 2-1 (unmarketable).

Significant correlations were highlighted between each method and the chemical data (total chlorophyll or ammonia content) reported in Palumbo et al. (2022). In particular, the correlation matrices based on the Pearson correlation coefficient were explored by an heatmap and the level $p = 0.05$ was assumed significant for the correlation coefficients. Data analysis was carried out using the software Statistica (version 6.0, StatSoft, Inc., Tulsa, OK, USA). Moreover, a partial least square regression (PLSR) analysis was carried out to predict the total chlorophyll or ammonia content using The Unscrambler X software (CAMO AS, Oslo, Norway).

3. Results and discussions

3.1. Selection of methods associated to rocket leaves marketability

Significant relationships among the values of C5 and C9 evaluated using the 4 methods and the QL scores attributed to rocket leaves during the cold storage (Palumbo et al., 2021) are reported in Table 1.

Results from the one-way ANOVA highlights that all the compared methods can separate marketable samples (QL5, QL4 and QL3) from non-marketable ones (QL2 and QL1). This information is normally sufficient in most commercial applications where the QL3 represents the limit of marketability. It is interesting to note that the method M2, when applied in the polar representation, is able to separate all the QLs. Moreover, the methods require different times to assign a quality level to a sample: they allow the designer of a Computer Vision System to choose the appropriate trade-off between accuracy and computation time. The Random Forest model analysed to identify the 36 most relevant colours required about 25 ms to classify a sample. The methods derived by the relevant colours allow different reduction of computation time and are easily understandable by humans. The method M1 takes only 1 ms to evaluate a sample. The method M2 requires 3 ms for classifying a sample. The methods M3, regardless its version, requires 11 ms. M3a, applied in the polar representation, is able to separate all the QLs but not the QL5 and QL4 (corresponding to very good and good product, respectively). The method M4 involves a computation time of 22 ms.

All the methods showed a general reduction in the value associated to C9 (associated to green nuances), going from QL5 to QL1. The senescence produced an increase in the value of C5, associated to yellow pigments. While a Random Forest model behaves as a black-box, the methods based on such phenomena are clearly interpretable by humans: in rocket leaves, the reduction of green pigments and the simultaneous increase of yellow ones during the cold storage is due to biological degradation of chlorophyll (Cefola and Pace, 2015; Cefola et al., 2010; Watkins, 2006), as also described by Palumbo et al. (2022). Indeed, in Fig. 5A, data of total chlorophyll and ammonia content reported in Palumbo et al. (2022) are presented. The total chlorophyll content of rocket leaves showed a significant reduction (42.4 %) during the storage.

Additionally, postharvest chlorophyll breakdown may contribute to ammonia accumulation in vegetable tissues (Amodio et al., 2018) which is highly correlated to hue angle variations (related to leaves yellowing) in rocket leaves stored at 10 °C, already demonstrated by Palumbo et al. (2022) and Mastrandrea et al. (2016). As for ammonia content, at harvest samples showed very low values (6.11 ± 2.42 μg NH_4^+/g of fresh weight), but a significant increase was recorded at the end of storage (132.51 ± 8.67 μg NH_4^+/g of fresh weight) (Fig. 5B). High levels of ammonia may cause tissue damage with visible senescence effects, influencing the overall quality of the product.

Because of their strict relation to the senescence of the product, both chlorophyll and ammonia content may be considered objective markers for quality loss of rocket leaves (Palumbo et al., 2022).

These qualitative associations between greenish and yellowish contents of rocket leaves are known but not easy to be translated into an exact description of the corresponding green and yellow regions of the colour space. The proposed approach exploits the Machine Learning technique to quantitatively identify the regions of the colour space that must be observed to monitor the quality of the product.

These results are clearly visible also by looking at the score plot obtained by PCA analysis, that uses as variables the values of the C5 and C9 obtained by the 4 methods and the chlorophyll and ammonia data (Fig. 6).

The first and the second components accounted for 66.2 % and 20.4 % of the total variance respectively, displaying a different distribution of marketable and non-marketable samples in the PCA quadrants (Fig. 6A): the formers were mostly clustered at the left side, while the non-marketable ones at the right side along the first component.

In the PCA score scatter plot, all the methods in the higher-left and lower-left quadrants showed a significant correlation with the chlorophyll content of rocket leaves which presented negative component 1

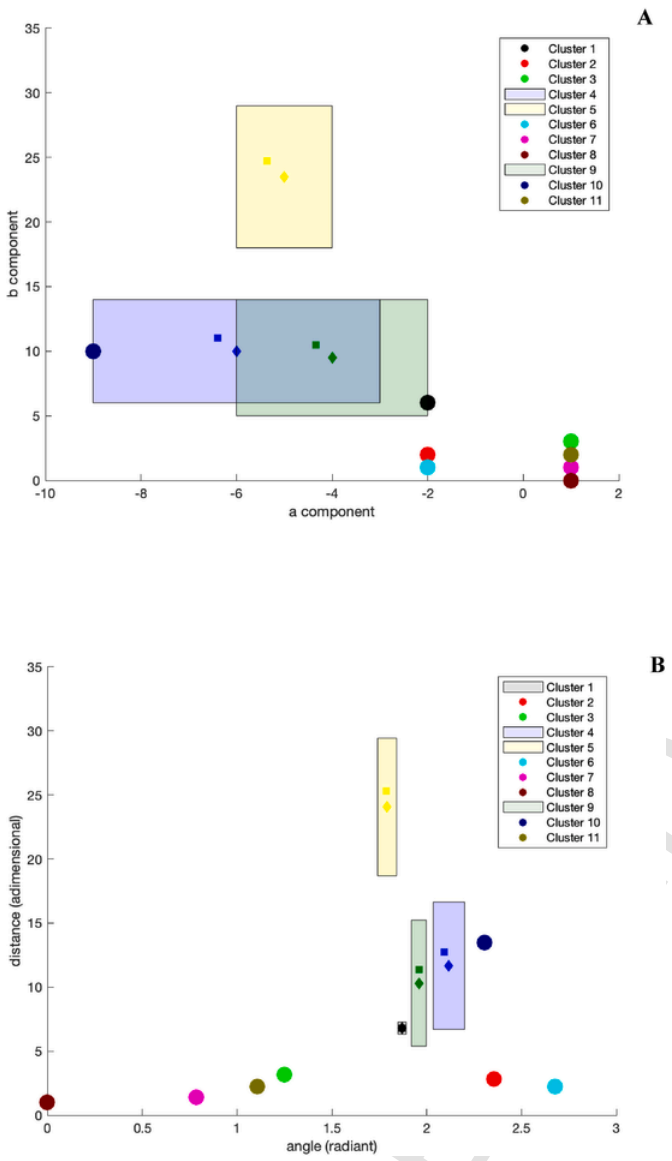


Fig. 4. The figure shows the rectangular regions defined by the minimum and maximum values of the colours belonging to clusters 4, 5 and 9 in the rectangular (A) and polar representation (B). In both figures, the symbol of diamond represents the mean point of each region while the square represents the centroid of the same region. In the polar representation the regions related to different clusters have an empty intersection while they largely overlap in the rectangular representation.

and 2 and was found in the lower-left quadrant of the score plot; on the other hand, the ammonia content, that is placed in the lower-right quadrant of the score plot, was correlated significantly to the methods placed in the same quadrant (Fig. 6B).

The relationships between clusters values of the 4 methods and the chemical attributes (chlorophyll and ammonia content) were explored by the heatmap reported in Table 2: the methods M2-C9-P, M2-C9-R and M3a-C9-R, in the higher-left quadrants, and M3a-C9-P and M3b-C9-P, in the lower-left quadrant, showed higher correlations with the chlorophyll content than the others in the same quadrants, while the ammonia content was highly correlated to the methods M1-C5-R, M4-C5-R and M4-C9-R. Even for what concern the correlation with chlorophyll content, M2 e M3 exhibit a high correlation at a very low computational cost (3 ms). The differences between C5 and C9 do not appear to be relevant. Instead, for what concern ammonia, the two clusters provide very different performances. C9 seems to poorly correlate with ammonia while C5 is much more effective in estimating this property of the product. In particular, the presence of C5 evaluated using the M1, applied to the rectangular representation, exhibits a very high correlation with ammonia content. The C5 is characterized by a strong separation from the other ones in the *ab*-plane of the CIELab colour space. The corresponding colour region defined by the M1 and the rectangular representation remains well separated and have a relevant extension in the colour plane that could explain this high correlation.

Fig. 7 shows the values estimated by the CVS against the values measured in the laboratory for chlorophyll (A) and ammonia (B) content on unpackaged and packaged rocket leaves using as variables of the models the values provided by the methods M1-C5-R and M3b-C9-P. The accuracy of predictions, expressed in Pearson's correlation coefficient, were interesting ($r = 0.82$ for chlorophyll and $r = 0.89$ for ammonia). They are similar to those obtained by the Random Forest model developed in Palumbo et al., (2022) (about $r = 0.85$ and 0.91 for chlorophyll and ammonia, respectively). The advantage of the methods proposed in this paper is that they are computationally simpler and faster than the Random Forest model.

In literature, there are several examples vision systems integrating machine learning models that proved to simplify the development of CVS and to provide better performance and greater flexibility. Dange et al. (2023) proposed a novel approach to predict the quality of grape using a CNN-based model and considering colour as an additional indicator. By evaluating the size of grapes in images acquired by a computer vision system, they predicted the quality of grapes at different stages of harvesting. The accuracy range of grape quality prediction by the CNN-model was among 90% and 92%, demonstrating a good level of performance. Additionally, the authors compared the performances of the model with different classifiers, showing the importance of selecting appropriate algorithm based on specific performance criteria. Ismail and Malik (2022) proposed a machine vision system based on deep learning techniques to offer a non-destructive and cost-effective solution for automating the freshness and appearance of apples and bananas. They compared the performance of several deep learning models (ResNet, DenseNet, MobileNetV2, NASNet and EfficientNet) and the average accuracy of the system was about 99.2% and 98.6% using EfficientNet model for apples and bananas test sets, respectively. Torkashvand et al. (2017) evaluated a novel methodology about the extrapolative ability of multiple linear regressions (MLR) and artificial neural networks (ANN) to estimate firmness of kiwifruit together with nutrients concentrations such as nitrogen, potassium, calcium along with magnesium. As results, the MLR model predicted firmness of fruit with a higher accuracy than the ANN model, but when they added the application of the N/Ca ratio as input dataset in the ANN model, it enhanced the prediction firmness of fruit compared to the MLR model.

All the methodologies used in these studies are computationally complex and difficult to be understood by humans. This paper proposes simpler methodologies for the construction of predictive algorithms that are easily understandable by operators, at the cost of light losses in performances. Further research is needed about the best way to exploit the contribution of machine learning methodologies in realizing real-world systems that provide useful performance in industrial systems while remaining fully understandable by humans.

3.2. Chlorophyll and ammonia content prediction

The methods that reported the highest correlations with chlorophyll and ammonia contents were used to build two PLS models to predict these two quality markers of rocket leaves (Table 3).

Results showed good prediction of chlorophyll (Model 1) and ammonia (Model 2) content by using as predictors the values of the methods M3b-C9-P and M1-C5-R, respectively. In detail, the Model 1 with R^2 of 74 % in calibration and 70 % in validation was obtained for chlorophyll content. Higher performances were obtained with the Model 2 to

Table 1
Relationships between the quality levels of packaged and unpackaged rocket leaves and the studied clusters 5 (C5) and 9 (C9) obtained by 4 different methods (M1, M2, M3a, M3b, M4) in both polar and rectangular representations. The values represent, for each method, the mean of presence (in terms of percentage, %) of image colours belonging to the studied clusters.

Method	Representation	Clusters	Quality Level										*P*-value
			5		4		3[a]		2		1		
			very good		good		fair		poor		very poor		
			Presence percentage (%)										
M1	**polar**	5	0.00053	c	0.00075	c	0.00089	c	0.00684	b	0.01638	a	****
		9	0.12830	c	0.16343	a	0.12125	c	0.16444	a	0.14662	bc	****
	rectangular	5	0.00074	c	0.00103	c	0.00131	c	0.00916	b	0.01927	a	****
		9	0.39882	b	0.44795	a	0.37086	c	0.37985	bc	0.34676	d	****
M2	**polar**	5	0.20719	e	0.22294	d	0.23866	c	0.33814	b	0.37348	a	****
		9	0.79280	a	0.77705	b	0.76133	c	0.66185	d	0.62652	e	****
	rectangular	5	0.15652	d	0.17772	c	0.18929	c	0.29372	b	0.33330	a	****
		9	0.84347	a	0.82227	b	0.81076	b	0.70627	c	0.66669	d	****
M3a	**polar**	5	0.16327	d	0.17472	d	0.18858	c	0.27137	b	0.30264	a	****
		9	0.83672	a	0.82527	a	0.81141	b	0.72863	c	0.69735	d	****
	rectangular	5	0.12073	d	0.13578	c	0.14621	c	0.23119	b	0.26607	a	****
		9	0.87926	a	0.86421	b	0.85378	b	0.76880	c	0.73392	d	****
M3b	**polar**	5	0.17793	d	0.18900	d	0.20609	c	0.29534	b	0.32366	a	****
		9	0.22341	a	0.19918	b	0.19192	b	0.14904	c	0.13697	c	****
	rectangular	5	0.12913	d	0.14746	c	0.16002	c	0.26392	b	0.30112	a	****
		9	0.16967	b	0.18999	a	0.14444	c	0.18574	a	0.19132	a	****
M4	**polar**	5	0.18734	e	0.20148	d	0.21638	c	0.30907	b	0.34268	a	****
		9	0.20359	bc	0.22448	a	0.19776	c	0.20940	b	0.20097	c	****
	rectangular	5	0.09040	d	0.11072	c	0.12132	c	0.24132	b	0.29544	a	****
		9	0.22129	c	0.29558	b	0.23595	c	0.36851	a	0.37268	a	****

For each quality level, the mean values followed by different letters (a, b, c, d, e) are significantly different (P-value <0.05) according to Student-Newman-Keuls (SNK) test.
Significance: **** = significant at P-value ≤ 0.0001.
[a] Limit of marketability.

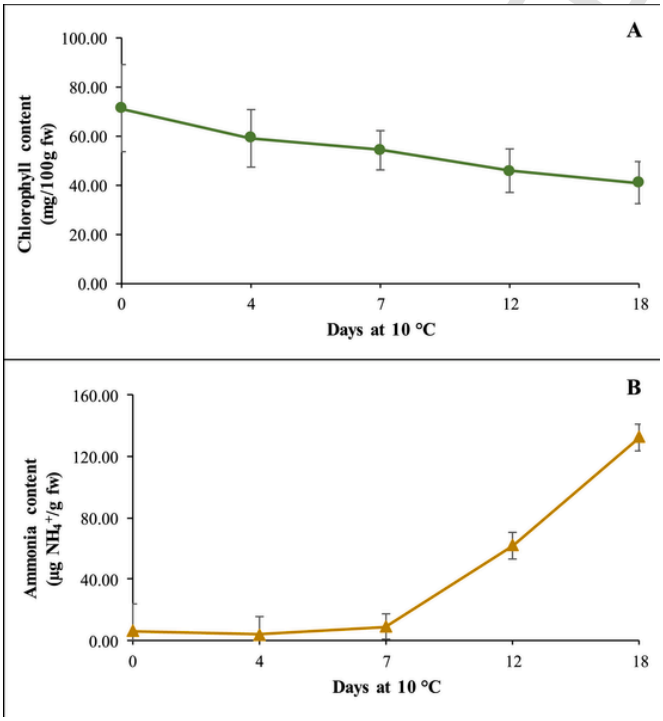


Fig. 5. Changes in chlorophyll (A) and ammonia (B) contents of rocket leaves during 18 days of storage at 10 °C. Each data is the mean value of 60 samples ± standard deviation.

predict the ammonia content (R^2 of 0.83 and 0.72 in calibration and validation, respectively). Similar performances were achieved by the prediction reported in Palumbo et al. (2022), in which Random Forest model was used (R^2_v = 0.77 and 0.80 for packaged and unpackaged samples, respectively). Hendrawan et al. (2023) measured the chlorophyll content of Moringa leaves using machine vision and an optimized artificial neural network with high performances in validation (R = 0.97). Higher performances were obtained by Cavallo et al., (2017), in which the combination of CVS and a Random Forest model provided a very interesting predictive model (R^2_v = 0.90). In our work, M3b-C9-P provides lower performances for chlorophyll prediction than Hendrawan et al. (2023), Palumbo et al. (2022) and Cavallo et al., (2017), but the methodology adopted provided simpler algorithms, easily interpretable by humans, and a lower computational speed (about 3 ms against the more than 20 ms of the random forest model). Additionally, while no relevant correlation was identified in Palumbo et al. (2022) for ammonia content, often used as another senescence indicator in leafy vegetables, the novel approach allowed to obtain a significant prediction of this parameter by M1-C5-R, the simplest model that has a computational time of 1 ms. There are no research works about the prediction of ammonia content in rocket leaves through the analysis of images acquired by a CVS to compare to our results. The possibility to evaluate this senescence parameter in a consistent and objective way, may be useful in the food processing industry to monitor the quality and shelf life of rocket leaves. In the future, once the methodology has been explored by other works, it would be useful to have a direct comparison of all results.

4. Conclusions

The present research addresses two limits of machine learning models: their computational complexity and their being not understandable by humans. The most effective machine learning method-

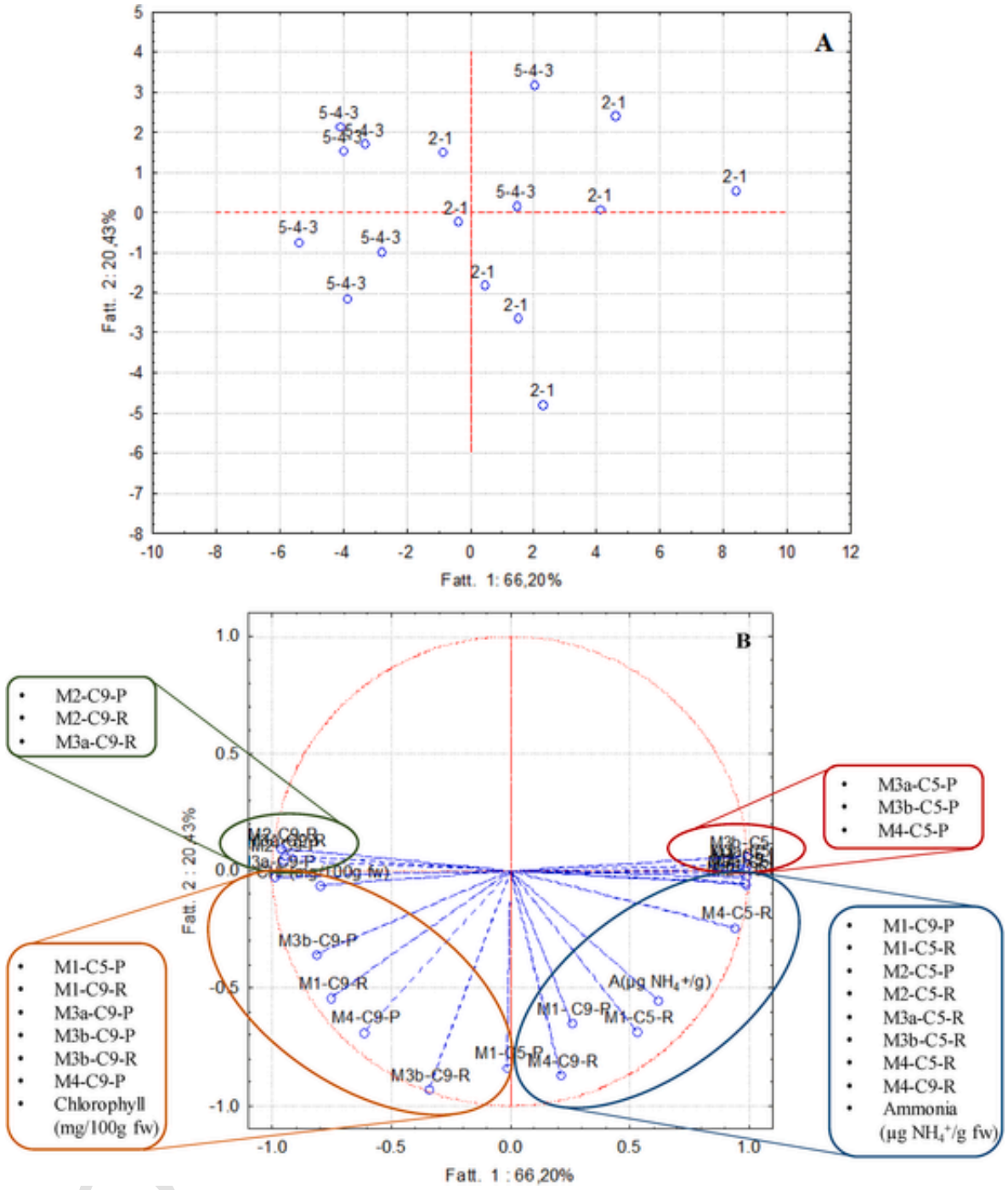


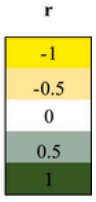
Fig. 6. PCA loading plot (A) and score scatter plot (B) carried out on the values of clusters 5 (C5) and 9 (C9) in the polar (P) or rectangular (R) representation obtained by the 4 methods adopted (M1, M2, M3a, M3b, M4). The abscissa and ordinate axes represent respectively the first and the second component with different percentage of the total variance. In A, each quadrant of the PCA represents the different distribution of marketable (5-4-3) and non-marketable (2-1) samples; in B, correlations among chlorophyll and ammonia contents and all the methods adopted (M1, M2, M3a, M3b, M4) are displayed.

ologies (Random Forest models or Convolutional Neural Networks are just a couple of the most popular examples) behaves as black-box and it is impossible to interpret their behaviour and to explicit the reasons that lead to their decisions. This lack of transparency is a serious weakness that justify why relevant research (under the general label of eXplanable Artificial Intelligence) is currently working on methodologies that make these models comprehensible. The approach proposed in this paper explores the possibility of extracting and explicating information hidden into an effective machine learning model developed to classify visual quality and to estimate internal properties of rocket leaves. This information provided useful hints to develop methods fully understandable by humans that achieve the same results with lower computational costs. An effective Random Forest model, developed in previous experiments to classify visual quality and to estimate chlorophyll and ammonia contents in rocket leaves, has been analysed to point out a set of relevant colours that correlate with the characteristics of interest. These colours have been used to automatically identify regions in the *ab*-plane of the CIELab colour space that can be exploited by simple and easily understandable methods that accomplish the same classification and estimations

Significance: ns = not significant; * significant for P ≤ 0.05; ** significant for P ≤ 0.01; *** significant for P ≤ 0.001; **** significant for P ≤ 0.0001.

Table 2
Significance: ns = not significant; * significant for P ≤ 0.05; ** significant for P ≤ 0.01; *** significant for P ≤ 0.001; **** significant for P ≤ 0.0001.
The heatmap shows the correlations between values of clusters 5 (C5) and 9 (C9) in the polar (P) or rectangular (R) representation obtained with the 4 different methods (M1, M2, M3a, M3b, M4) and chlorophyll and ammonia contents in rocket leaves. A different colour code is used to represent the strong of correlations; **r** is the Pearson's correlation coefficient.

Method-Cluster-Representation	Chlorophyll content (mg/100g fresh weight)		Ammonia content (μg NH_4^+/g fresh weight)	
	r	*P-value*	r	*P-value*
M1-C5-P	0.1644	ns	0.3584	ns
M1- C9-P	-0.1460	ns	0.1905	ns
M1-C5-R	-0.4051	ns	0.9015	****
M1-C9-R	0.5509	*	-0.2107	ns
M2-C5-P	-0.7004	**	0.5708	*
M2-C9-P	0.7458	***	-0.5431	*
M2-C5-R	-0.7316	***	0.6222	**
M2-C9-R	0.7838	****	-0.5991	*
M3a-C5-P	-0.7578	***	0.5489	*
M3a-C9-P	0.7578	***	-0.5489	*
M3a-C5-R	-0.7388	***	0.6061	*
M3a-C9-R	0.7912	****	-0.5713	*
M3b-C5-P	-0.7633	***	0.5300	*
M3b-C9-P	0.8648	****	-0.3701	ns
M3b-C5-R	-0.7445	***	0.6313	**
M3b-C9-R	0.3585	ns	0.3041	ns
M4-C5-P	-0.7069	**	0.5633	*
M4-C9-P	0.4480	ns	-0.0754	ns
M4-C5-R	-0.7118	**	0.7622	***
M4-C9-R	-0.1898	ns	0.6831	**



Significance: ns= not significant; * significant for P ≤ 0.05; ** significant for P ≤ 0.01; *** significant for P ≤ 0.001; **** significant for P ≤ 0.0001.

of interest. These relevant colour regions evaluate significant traits of the product at hand and are comprehensively related to chemical and physical changes induced by senescence. The simple methods built using these colour regions as features allow the tuning of computational complexity according to the requirements of the application, with minimal performance loss compared to the original Random Forest model.

The proposed approach was able (i) to identify relevant clusters of colours that are informative about the properties of the product at hand; (ii) to select the clusters more significant to estimate the desired properties; (iii) to describe shape and size of regions of the *ab*-plane in the CIELab colour representation corresponding to the clusters of interest. These results were achieved using automatic processing without the cumbersome and error-prone trial-and-error process required by manual design and selection of features.

These features were the input for objective and sound computational schemes with different execution times. Humans can choose the best trade-off between efficacy and efficiency, depending on the application constraints. All methods provided a good separation of marketable samples from non-marketable ones with computation times (from 1 ms to 22ms) lower than the Random Forest model (25ms). In particular, two of the considered methods, M3b-C9-P and M1-C5-R, provided good prediction of chlorophyll (R^2_v = 0.70) and ammonia (R^2_v = 0.72) contents, respectively. Chlorophyll and ammonia assess the state of product in an objective and robust way. The computation times of the two methods (3ms for M3b-C9-P and 1 ms for M1-C5-R) are much lower than the 25ms of the Random Forest model. The results of the experiments can be easily interpreted in terms of known processes occurring during the senescence of the product. Moreover, the identified well-grounded objective colour cues could be used to improve indications provided to human operators during their training on the quality evaluation task.

Uncited references

CRediT authorship contribution statement

Michela Palumbo : Methodology, Formal analysis, Investigation, Writing – original draft. **Maria Cefola :** Conceptualization, Writing – review & editing, Supervision. **Bernardo Pace :** Conceptualization, Writing – review & editing, Supervision. **Giancarlo Colelli :** Conceptualization, Supervision. **Giovanni Attolico :** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

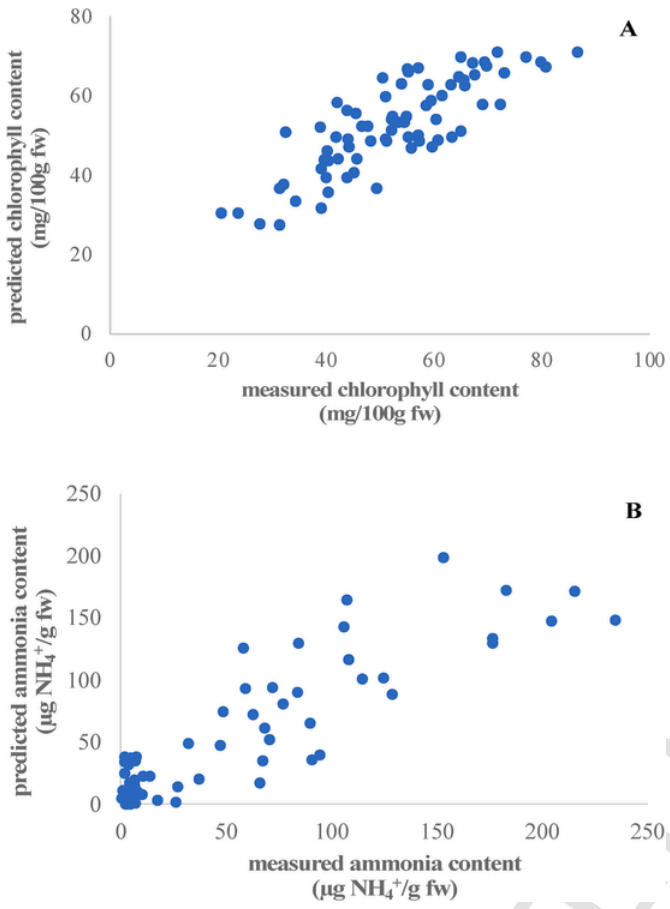


Fig. 7. Values of chlorophyll (A) and ammonia (B) content measured in the laboratory vs. values estimated by the CVS on unpackaged and packaged rocket leaves, using as variables of the model the methods M3b-C9-P and M1-C5-R, respectively.

Table 3
Root Mean Square Error (RMSE) and the coefficient of determination (R^2) in calibration ($_c$) or validation ($_v$) of the partial least square regression (PLSR) models predicting chlorophyll and ammonia contents of rocket leaves.

PLSR models	Predictors	$RMSE_c$	R^2	$RMSE_v$	R^2
Model 1 (chlorophyll content prediction)	M3b-C9-P	6.23	0.74	7.04	0.70
Model 2 (ammonia content prediction)	M1-C5-R	20.27	0.83	27.58	0.72

Data availability

No data was used for the research described in the article.

Acknowledgement

The project Prin 2017"SUS&LOW- Sustaining low-impact practices in horticulture through non-destructive approach to provide more information on fresh produce history & quality" (grant number: 201785Z5H9) from the Italian Ministry of Education University is kindly acknowledged.

References

Amodio, M.L., Colelli, G., Cantwell, M.I., 2018. Ammonia accumulation in plant tissues: a potentially useful indicator of postharvest physiological stress. Acta Hortic. 1194, 1511–1518. https://doi.org/10.17660/actahortic.2018.1194.211.

Breiman, L., 1984. Classification and Regression Trees. Wadsworth, Belmont. https://doi.org/10.1201/9781315139470.

Cavallo, D.P., Cefola, M., Pace, B., Logrieco, A.F., Attolico, G., 2017. Contactless and non-destructive chlorophyll content prediction by random forest regression: a case study on fresh-cut rocket leaves. Comput. Electron. Agric. 140, 303–310. https://doi.org/10.1016/j.compag.2017.06.012.

Cefola, M., Pace, B., 2015. Application of oxalic acid to preserve the overall quality of rocket and baby spinach leaves during storage. J. Food Process. 39 (6), 2523–2532. https://doi.org/10.1111/jfpp.12502.

Cefola, M., Amodio, M.L., Rinaldi, R., Vanadia, S., Colelli, G., 2010. Exposure to 1-methylcyclopropene (1-MCP) delays the effects of ethylene on fresh-cut broccoli raab (*Brassica rapa* L.). Postharvest Biol. Technol. 58 (1), 29–35. https://doi.org/10.1016/j.postharvbio.2010.05.001.

Chaudhry, M.M.A., Babellahi, F., Amodio, M.L., Colelli, G., Sahar, A., 2020. Image analysis. In: Khan, M.K.I. (Ed.), Advances in Noninvasive Food Analysis. CRC Press, New York, pp. 200–201.

Dange, B.J., Mishra, P.K., Metre, K.V., Gore, S., Kurkute, S.L., Khodke, H.E., Gore, S., 2023. Grape vision: a CNN-based system for yield component analysis of grape clusters. Int. J. Intell. Syst. Appl. Eng. 11 (9s), 239–244.

Fan, S., Li, J., Zhang, Y., Tian, X., Wang, Q., He, X., Zhang, C., Huang, W., 2020. On line detection of defective apples using computer vision system combined with deep learning methods. J. Food Eng. 286, 110102. https://doi.org/10.1016/j.jfoodeng.2020.110102.

Ghule, A., Deshmukh, R.R., Gaikwad, C., 2020. Distinguishing healthy and infected vegetable crops using hyperspectral leaf reflectance. CSI Journal of 42.

Gokhale, A., Chavan, A., Sonawane, S., 2023. Leveraging ML techniques for image-based freshness index prediction of fruits and vegetables. In: 2023 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, pp. 1–6.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Hierarchical clustering. In: Hastie, T., Tibshirani, R., Friedman, J. (Eds.), The Elements of Statistical Learning. second ed., Springer, New York, pp. 520–528. https://doi.org/10.1007/978-0-387-21606-5.

Hendrawan, Y., Perkasa, T.E., Prasetyo, J., Al-Riza, D.F., Damayanti, R., Hermanto, M.B., Sandra, S., 2023. Moringa leaf chlorophyll content measurement system based on optimized artificial neural network. Advances in Food Science, Sustainable Agriculture and Agroindustrial Engineering (AFSSAAE) 1–10.

Ismail, N., Malik, O.A., 2022. Real-time visual inspection system for grading fruits using computer vision and deep learning techniques. Inf. Process. Agric. 9 (1), 24–37. https://doi.org/10.1016/j.inpa.2021.01.005.

Li, B., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees (CART). Biometrics 40 (3), 358–361.

Lisboa, P.J.G., Saralajew, S., Vellido, A., Fernández-Domenech, R., Villmann, T., 2023. The coming of age of interpretable and explainable machine learning models. Neurocomputing 535, 25–39. https://doi.org/10.1016/j.neucom.2023.02.040.

Loh, W.Y., 2002. Regression trees with unbiased variable selection and interaction detection. Stat. Sin. 12, 361–386.

Loh, W.Y., Shih, Y.S., 1997. Split selection methods for classification trees. Stat. Sin. 7, 815–840.

Mastrandrea, L., Amodio, M.L., Cantwell, M.I., 2016. Modeling ammonia accumulation and colour changes of arugula (Diplotaxis tenuifolia) leaves in relation to temperature, storage time and cultivar. Acta Hortic. 1141, 275–282. https://doi.org/10.17660/ActaHortic.2016.1141.34.

Natarajan, S., Ponnusamy, V., 2023. Classification of organic and conventional vegetables using machine learning: a case study of brinjal, chili and tomato. Foods 12 (6), 1168. https://doi.org/10.3390/foods12061168.

Palumbo, M., Pace, B., Cefola, M., Montesano, F.F., Colelli, G., Attolico, G., 2022. Non-destructive and contactless estimation of chlorophyll and ammonia contents in packaged fresh-cut rocket leaves by a Computer Vision System. Postharvest Biol. Technol. 189, 111910. https://doi.org/10.1016/j.postharvbio.2022.111910.

Palumbo, M., Pace, B., Cefola, M., Montesano, F.F., Serio, F., Colelli, G., Attolico, G., 2021. Self-configuring CVS to discriminate rocket leaves according to cultivation practices and to correctly attribute visual quality level. Agronomy 11 (7), 1353.

Torkashvand, A.M., Ahmadi, A., Nikravesh, N.L., 2017. Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network (ANN) and multiple linear regressions (MLR). J. Integr. Agric. 16 (7), 1634–1644. https://doi.org/10.1016/S2095-3119(16)61546-0.

Wang, Z., Huang, C., Li, Y., Yao, X., 2023. Multi-objective feature attribution explanation for explainable machine learning. ACM Trans. on Evolutionary Learning. Association for Computing Machinery New 1–31. https://doi.org/10.1145/3617380.

Watkins, C.B., 2006. The use of 1-methylcyclopropene (1-MCP) on fruits and vegetables. Biotechnol. Adv. 24 (4), 389–409. https://doi.org/10.1016/j.biotechadv.2006.01.005.