Information Processing in Agriculture xxx (xxxx) xxx

Contents lists available at ScienceDirect

Information Processing in Agriculture

journal homepage: www.keaipublishing.com/en/journals/information-processing-in-agriculture/

A novel random forest-based approach for the non-destructive and explainable estimation of ammonia and chlorophyll in fresh-cut rocket leaves

Stefano Polimena [a,b], Gianvito Pio [a,b,*], Maria Cefola [e], Michela Palumbo [e,f], Michelangelo Ceci [a,b,d], Giovanni Attolico [c]

[a] Department of Computer Science, University of Bari Aldo Moro, Via E. Orabona 4, 70125 Bari, Italy
[b] Big Data Laboratory, National Interuniversity Consortium for Informatics (CINI), Via Ariosto 25, 00185 Rome, Italy
[c] Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing, National Research Council of Italy, via Amendola 122/D, 70126, Bari, Italy
[d] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000, Ljubljana, Slovenia
[e] Institute of Sciences of Food Production, National Research Council of Italy, c/o CS-DAT, Via Michele Protano, 71121 Foggia, Italy
[f] Department of Science of Agriculture, Food and Environment, University of Foggia, Via Napoli 25, 71122 Foggia, Italy

ARTICLE INFO

Keywords:
Fresh-cut rocket leaves
Consumer acceptability
Machine learning
Explainability

ABSTRACT

The perceived visual quality of fruits and vegetables plays a central role in the choices made by retail customers. Machine learning (ML) approaches based on image analysis have been recently proposed to overcome the poor efficiency and subjectivity of human visual evaluation as well as the expensiveness and destructiveness of physical and chemical methods that measure internal indicators. In this paper, we propose a ML method based on Random Forests for estimating the chlorophyll and ammonia contents (considered, in the literature, reliable indicators of product freshness) from images of fresh-cut rocket leaves. Our approach copes with specific issues raised by (i) the non-uniform distributions of ammonia and chlorophyll values and (ii) the need to provide insights into the features that produce a particular model outcome, aiming to enhance its trustworthiness. Our experiments, performed on real images of fresh-cut rocket leaves, proved that the proposed approach significantly outperforms 7 competitor methods, obtaining an improvement of the RSE results of 6.6% for the prediction of the ammonia and of 10.4% for the prediction of the chlorophyll over its best competitor. Moreover, a specific analysis of the explainability of the predictions showed that the learned models are based on reasonable features, empowering their acceptance in real-world applications.

1. Introduction

The research on contactless, non-destructive, rapid and accurate evaluation of the quality of fruits and vegetables has recently gained interest, to overcome the limits of traditional sensory and destructive methods. While visual inspection by humans is subjective and error prone due to intra-operator and inter-operators inconsistencies [1, 2], physical and chemical methods are generally destructive, time-consuming, expensive, polluting and not suitable for the application in an industrial line.

The product quality, together with the sustainability of the production processes, plays a central role in the choices made by customers [3]. Nowadays, emerging non-destructive methods in food technology include near infrared spectroscopy (NIR), hyperspectral imaging (HSI) and computer vision systems (CVS). Most of the research applied hyperspectral or multispectral techniques to vegetables [4–8]. The complexity of spectroscopy and hyperspectral imaging, both in terms of time and costs required for the acquisition and for the subsequent processing phases, makes their application more difficult in a pervasive way along the supply chain to enable a continuous monitoring of the parameters of interest. On the contrary, CVSs are simpler and can hopefully exploit cameras that are already available along the path from the harvest to the distribution to final consumers. They aim to mimic human visual evaluation of quality by acquiring and processing images of the whole visible surface of the products. These digital images are analyzed by extracting the most discriminative visual characteristics and by processing them through models usually learned by means of machine learning methods [1].

In the last few years, CVSs have been used to automatically evaluate several properties of different products: table grapes [9], fresh-cut

* Corresponding author at: Department of Computer Science, University of Bari Aldo Moro, Via E. Orabona 4, 70125 Bari, Italy.
E-mail addresses: stefano.polimena@uniba.it (S. Polimena), gianvito.pio@uniba.it (G. Pio), maria.cefola@ispa.cnr.it (M. Cefola), michela.palumbo@ispa.cnr.it (M. Palumbo), michelangelo.ceci@uniba.it (M. Ceci), giovanni.attolico@cnr.it (G. Attolico).

https://doi.org/10.1016/j.inpa.2024.09.002
Received 5 December 2023; Received in revised form 1 July 2024; Accepted 2 September 2024
Available online 3 September 2024

nectarines [10], apples [11], and fresh-cut rocket leaves [12]. Nowadays, baby leaf vegetables such as spinach, rocket, and lettuce are economically important products since they are basic components of ready-made salads. Among these, rocket (Eruca sativa Mill.) is a popular leafy vegetable due to its distinct taste and nutritional content. For this reason, in this paper we focus on this product, specifically on the estimation of the amount of chlorophyll and ammonia in rocket leaves through regression models.

The focus on the chlorophyll is motivated by the fact that the quality loss of fresh-cut rocket leaves during the postharvest storage is mainly due to senescence, strictly related to chlorophyll degradation that, therefore, is the most common index used to evaluate the quality and freshness of this product [13,14]. Generally, as reported by Pace et al. [15], a 30% loss of total chlorophyll content is considered the shelf life limit in rocket leaves stored for about 16 days at a temperature between 5 °C and 20 °C.

On the other hand, the focus on the ammonia comes from the fact that a low value of ammonia, which is a product of protein catabolism, generally indicates freshness [16,17], while high values of ammonia can indicate a product deterioration. Moreover, chlorophyll degradation produces yellowness during storage but also protein catabolism which contributes to ammonia accumulation: therefore its relationship with discoloration or yellowing process may be expected [18].

In the literature, we can find only a few applications regarding the adoption of CVSs for the estimation of the amount of ammonia. Pace et al. [19] applied a CVS on whole and fresh-cut lettuce for the non-destructive estimation of ammonia as a senescence indicator in leafy vegetables [20,21]. In that case, the parameters used to classify the product and to estimate the ammonia content were the amounts of green, white, and brown colors on the product. These colors were identified in the color plane by studying a proper number of samples of each color, manually extracted from images of fresh-cut lettuce. That method strongly depends on the training samples used to define the colors of interest: they must be carefully extracted each time a new product needs to be analyzed.

Palumbo et al. [22] proposed one of the first CVSs able to evaluate the visual quality level (considered as a machine learning classification task) and estimate the amount of chlorophyll and ammonia (considered as machine learning regression tasks) adopting a workflow based on Random Forests. However, for the regression task, the authors emphasized some limitations due to the skewed distribution of the target variables (i.e., the amount of ammonia and chlorophyll). Specifically, when the values are not uniformly distributed in the training set, the random sampling strategy adopted by Random Forests does not properly consider boundary values in the construction of regression trees, leading to less accurate predictions of new instances falling in such boundary areas (see Fig. 1 for an example depicting the effect of sampling performed on data following a Gamma distribution).

Moreover, these works do not include any mechanism to explain why and how the models provide specific outputs. Indeed, although ensemble methods generally lead to improve predictive accuracy, they lose the interpretability of single trees (more details will be provided in Section 2). On the other hand, it is fundamental to understand which information influenced the model output, especially if decisions of economic nature are made based on them.

In order to overcome these limitations, in this paper we design a novel approach. Specifically, we propose a different sampling procedure to be adopted during the construction of Random Forests, in order to provide the right importance to boundary values. The proposed approach exploits principles usually adopted to cope with unbalanced data in classification tasks, which provided satisfactory results in different contexts [23]. It is noteworthy that more advanced approaches to handle data unbalancing have recently been proposed in the literature, also based on Generative Adversarial Networks (GANs) [24]. However, they mainly focus on classification tasks and are not straightforwardly adaptable to the regression task that we solve.

Moreover, we analyze the learned trees to provide an explanation of the provided predictions, not only to support the model validation, but also to suggest some clues on the motivations behind the predictions which can increase the acceptability of such an automated system in the context of the efficient estimation of the quality of food and vegetables. The remaining of the paper is organized as follows: in Section 2, we describe the approach we adopt to collect and prepare data about rocket leaves, as well as our novel method based on Random Forests; in Section 3, we describe and discuss the results of our experimental evaluation; finally, in Section 5, we draw some conclusions and outline possible future work.

2. Materials and methods

2.1. Image acquisition and pre-processing

This section provides a quick summary of the procedure we followed to capture the images and extract features to learn a regression model to estimate the amount of chlorophyll and ammonia. This process was conducted according to the procedure reported in Palumbo et al. [22].

Fresh rocket leaves were selected and placed in 50 × 30 cm open polyethylene bags (Orved, Musile di Piave (VE), Italy) containing each one about 350 g of product and stored at 10 °C (as commonly occurs in the market). About 70 g of product was taken from each bag as sample at five different times (corresponding to the five commercial quality levels normally used to assess the state of products). The amount of chlorophyll and ammonia of each sample was measured in the laboratory, as described by Palumbo et al. [22]. Few images of each sample were acquired after a random shuffle of the leaves to acquire information about a larger part of its surface.

Images were acquired using a 3CCD (having a dedicated Charged Coupled Device for each color channel) digital camera (JAI CV-M9GE) having a resolution of 1024 × 768 pixels. The imaged area was about 32 × 24 cm. The 3CCD sensor was used to avoid the artifacts introduced by demosaicing, required to recover color information from a single CCD. The optical axis of the LinosMeVis 12 mm lens system was perpendicular to the black background. Eight halogen DC powered lamps were placed along two sides of the imaged area and oriented at a 45° angle with respect to the optical axis. The images were saved as uncompressed TIFF to avoid the artifacts introduced by compression algorithms.

A small X-Rite color-chart with 24 patches of known colors was placed into the scene to measure color variations due to environmental conditions and sensor characteristics. The colors in the color-chart were used to estimate the linear transformation used to correct colors.

We performed color analysis only on the part of each image belonging to the product at hand (foreground), that was separated from the background by applying a multi-threshold approach, based on the Otsu algorithm, to the Hue component of the image converted in the HSV color space as described in Cavallo et al. [14].

A linear transformation was used for color correction. The followed approach can be considered *effective*, i.e., it provides consistent color measurements, and *efficient*, since it is computationally suitable for real applications along the supply chain. In particular, given $[r_e^i g_e^i b_e^i]^T$ and $[r_m^i g_m^i b_m^i]^T$ the expected and the measured RGB values, respectively, for the ith patch, with $i = 1, \ldots, 24$. To reduce the distance between the expected and the measured values on the color chart, the following equation was adopted:

$$\begin{bmatrix} r_c \\ g_c \\ b_c \end{bmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{42} & m_{33} \end{pmatrix} \begin{bmatrix} r_m \\ g_m \\ b_m \end{bmatrix} \quad (1)$$

where $[r_c g_c b_c]^T$ are the colors corrected using the matrix. The matrix was computed by the least-square approach, and was adopted to correct all the foreground pixels of the image.

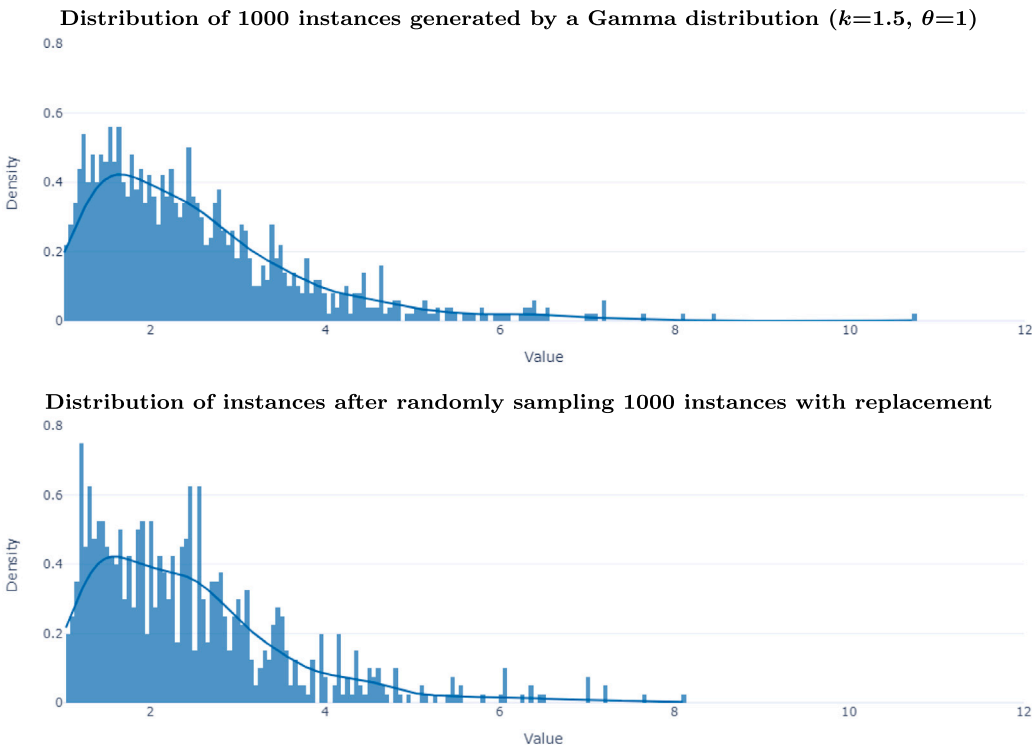


Fig. 1. On the top, the value distribution of 1000 instances synthetically generated by a Gamma distribution with k (shape) = 1.5 and θ (scale) = 1. On the bottom, the distribution of 1000 instances randomly sampled with replacement from such non-uniform data using the bootstrap approach adopted by Random Forests. As can be observed, some boundary regions (e.g., values > 8) are not represented by the sample.

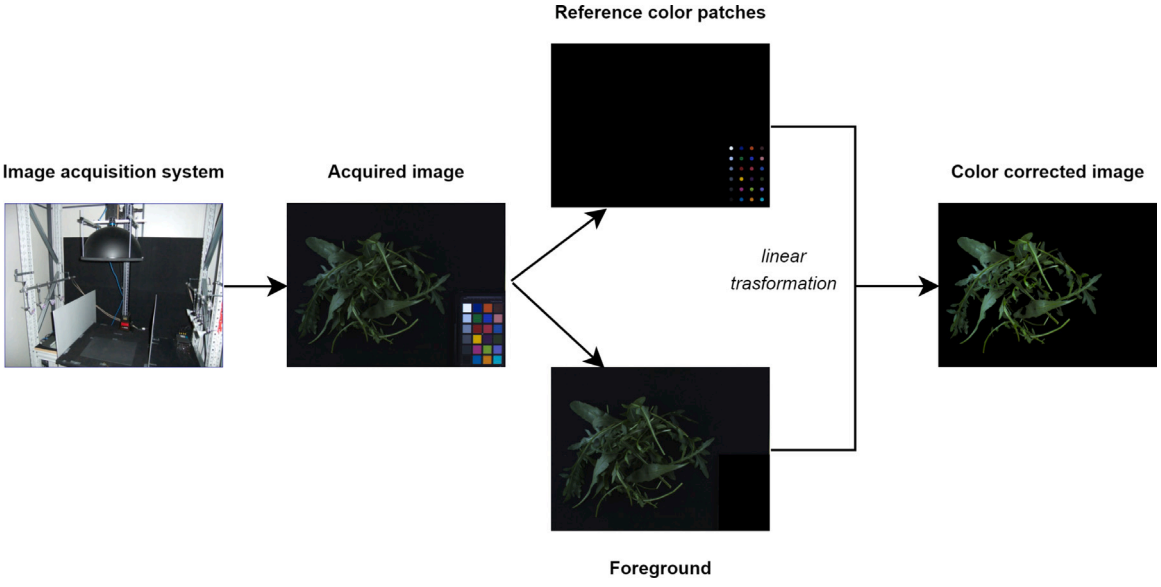


Fig. 2. Pipeline followed for image acquisition and pre-processing.

As color space to perform the analysis, we adopted the device-independent and perceptually-uniform CIE $L^*a^*b^*$. The L^* component was discarded, since it is too sensible to non-uniform illumination levels across the scene. As features to adopt to learn a regression model, we considered the color histogram of the foreground pixels. In particular, for each color in the a^*b^* plane, we counted the number of occurrences in the image. The continuous (a^*, b^*) plane was discretized into 261 integer values (in the range [−130;130]) for each axis (a^* and b^*), leading to a total of $261^2 = 68121$ colors/features.

A graphical overview of the image acquisition and pre-processing pipeline is depicted in Fig. 2.

2.2. The proposed learning approach

In this section, we describe the approach we propose to learn a model to predict the amount of chlorophyll and ammonia from the extracted features. As introduced in Section 1, our approach is based on random forests, since they already proved to be adequate for the task at hand by Palumbo et al. [22]. However, we propose a different sampling procedure in order to properly represent also boundary values, which in the classical random sampling procedure are under-represented (see Fig. 1). Before going into the details of the specific approach we propose to improve the sampling procedure, as well as our approach to explain

the provided predictions, in the following subsection, we first briefly describe the standard random forest regressor.

2.2.1. Random Forest: background

Random Forest is an ensemble approach based on learning multiple (possibly uncorrelated) regression trees from random subsets of the dataset at hand. A regression tree is a hierarchical model consisting of several decision nodes. Each decision node implements a function that splits the observations falling in such a node into disjoint subsets, according to a splitting feature/variable and a splitting point (which is a value in the case of categorical features, and a threshold, in the case of continuous features).

The learning process of regression trees is usually based on a top-down induction procedure. It starts from the root node of the tree containing all the training instances, which are recursively partitioned by identifying the best splitting variable and splitting point, among all the possible pairs of variables and splitting points. The learning process follows a greedy approach: once a decision is taken, in terms of the splitting variable and the splitting point, locally considering the observations falling into a node, it cannot be reverted even if more optimal (global) solutions could have been found considering the subsequent steps.

The identification of the best splitting feature and value/threshold relies on some heuristics that, for regression tasks, are usually based on the reduction of the variance. Precisely, given a node N, a splitting feature f and a value/threshold v, the variance reduction $\tilde{\sigma}_N(f,v)$ is computed as the difference between the variance of the target variable of the examples falling in the node N (henceforth indicated as R_N) and the sum of that of the examples falling in the left child node (henceforth indicated as $R_l(N,f,v)$) and in the right child node (henceforth indicated as $R_r(N,f,v)$). More formally:

$$\tilde{\sigma}_N(f,v) = \frac{\sum_{y_i \in R_N} (y_i - \overline{y_N})^2}{|R_N|} - \left(\frac{\sum_{y_i \in R_l(N,f,v)} (y_i - \overline{y_l})^2}{|R_l(N,f,v)|} + \frac{\sum_{y_i \in R_r(N,f,v)} (y_i - \overline{y_r})^2}{|R_r(N,f,v)|} \right) \quad (2)$$

where $\overline{y_N}$, $\overline{y_l}$ and $\overline{y_r}$ correspond to the average of the values of the sets R_N, $R_l(N,f,v)$, and $R_r(N,f,v)$, respectively.

The recursive approach stops when no partitioning is possible (or useful), namely, when the learned tree reaches a maximum pre-defined depth, or when the number of instances falling into a node falls below a given threshold, or when the variance of the target attribute in the node is below a given threshold. When the process stops, leaf nodes of the tree are associated with the actual predictions, which are computed as the average of the target attribute of the training instances falling into the leaf nodes.

One of the main strengths of trees is their interpretability: they represent understandable rules that provide clear indications of which features and which values are relevant for the prediction. However, single regression trees generally exhibit a high variance: a small change in the data can cause a large change in the structure of the tree. This phenomenon generally translates into a low generalizability of the learned model to unseen data.

Ensemble approaches, such as Random Forests, can overcome the above-mentioned issue, by learning multiple regression trees from random subsets of the initial dataset, and by averaging their predictions. In details, the process followed by Random Forests is based on two sources of randomness to achieve high diversity of each learned model. The first comes from the generation of different subspaces of features, which is usually obtained by a simple random sampling of the available features. The second involves the random selection of training instances, which is usually based on the bootstrap strategy: a sampling of n samples from the n available training instances with replacement, which generally leads to discarding about 36.8% of the available instances in each sample. Although very effective, Random Forests could introduce additional issues: the randomness of the sampling can lead to possibly discarding rare/boundary values (this issue happens in Palumbo et al. [22], see Fig. 1), and the decision taken by the ensemble is not interpretable as the rules of single regression trees.

As for the second issue, it can happen that the sampling distribution does not properly reflect the distribution of the population, which could lead to the construction of unsatisfactory trees that provide inaccurate predictions, especially for the under-represented cases. In the classification task, this issue is exacerbated when simple random sampling is adopted on datasets with unbalanced classes, that is, when the distribution of training examples across the classes is not uniform. In Fig. 3 we can see an *extreme* situation in which one of the minority classes is not represented by any training instance in the sample, which leads the learned tree to be totally unaware of such a class. Even if such an extreme situation does not occur, learning methods usually tend to provide a high amount of false negatives for minority classes on unseen data, if proper approaches for handling unbalanced data at training time are not followed. This phenomenon happens because the learning methods aim to minimize the overall prediction error, without paying specific attention to the minority classes.

For the regression task, we can observe an analogous issue when the distribution of the target variable is not uniform. In this case, some values may not be properly considered in the sampling phase, as already shown in Fig. 1. In the following subsection, we specifically focus on the approach we propose to overcome this issue for the regression task at hand, starting with a brief overview of the approaches proposed in the literature for classification tasks.

2.2.2. Dealing with a non-uniform distribution of the target variable

In the literature, several strategies have been proposed to handle the problem of class unbalancing in classification tasks. The most common approaches are based on oversampling the minority class, or on undersampling the majority class.

Simple oversampling techniques are usually based on the random duplication of examples of the minority class to obtain a balanced dataset, where each instance may be selected multiple times. More advanced techniques, such as SMOTE [23], are based on the generation of synthetic examples starting from real examples of the minority class. The SMOTE generation process is based on the construction, for each real example e of the minority class, of k synthetic examples that are "randomly in the middle", in the feature space, between e and its nearest neighbor. It is noteworthy that oversampling methods increase the size of the dataset, and therefore the computational cost.

On the other hand, undersampling techniques randomly select and remove samples from the majority class until the classes are balanced. In the literature, we can also find more sophisticated approaches, like Roughly Balanced Bagging (RBBag), that combines undersampling and bagging [25] with the goal of making the sampling probability of minority class equal to that of the majority class. However, undersampling strategies can lead to biased results due to the information loss caused by the removal of available samples: the resulting dataset might not accurately represent the original population.

The above-mentioned techniques cannot be applied to regression tasks, where the target values are continuous. In this context, we propose a method to take into account less represented values within the learning procedure of Random Forests. The idea is to discretize the target variable to trace back to a classification problem. This discretization would allow us to adopt existing oversampling techniques, increasing, as already mentioned, the computational cost. On the contrary, we aim at achieving the same result working on the sampling procedure embedded in the Random Forests algorithm, without increasing the size of the training set. We call our approach **BAL-RF**.

First, the target variable is discretized into m bins having equal width. Then, for each regression tree learned by Random Forest, instead

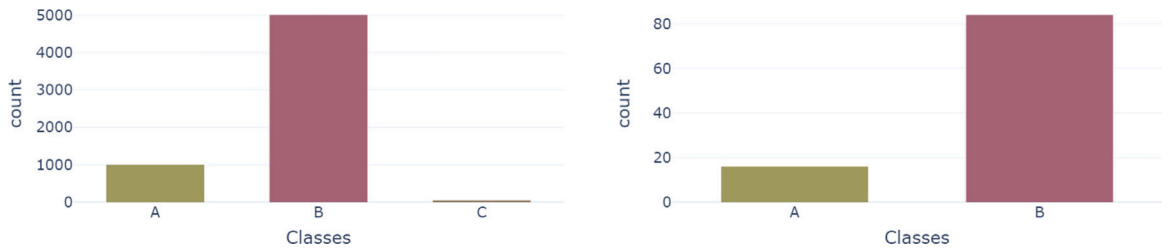


Fig. 3. An example of an unbalanced class distribution (on the left) on a dataset of 6050 instances, with 20 instances of class C, and of a sample of 100 instances obtained by simple random sampling (on the right), where the class C is not represented at all.

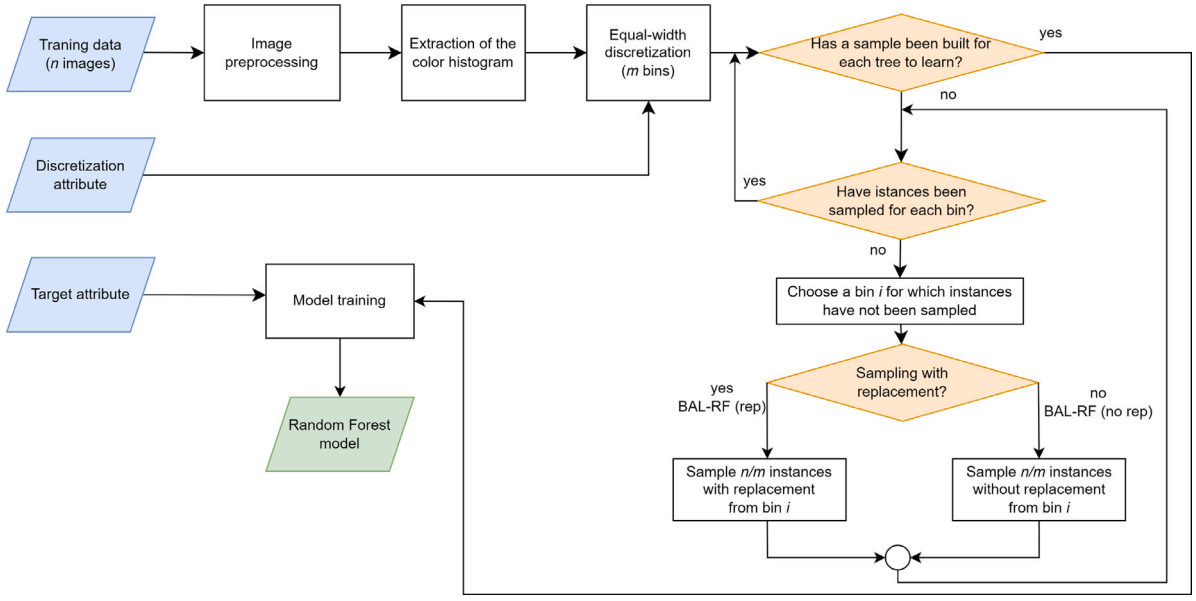


Fig. 4. Flowchart representing the proposed approach. The *Image preprocessing* step corresponds to the phases described in Section 2.1 and depicted in Fig. 2.

of sampling n examples with replacement from the n available examples, we randomly select n/m examples from each bin with replacement. In this way, each bin (that corresponds to a particular interval of the target variable) is always represented by some training examples for each learned tree.

Note that instances sampled multiple times may be discarded or kept as duplicates. In the first case, for each bin we may possibly select a number of instances which is actually less than n/m. This strategy, henceforth called **BAL-RF (no rep)**, generally leads to undersampling from bins with the majority of instances, and to guarantee that bins with a small number of instances are always represented in the training set. In the second case, each bin is always represented by n/m instances, with possible duplicates. This strategy, henceforth called **BAL-RF (rep)**, generally leads to a simultaneous undersampling from bins with a lot of instances, and oversampling from bins with few instances. Both the proposed strategies **BAL-RF (no rep)** and **BAL-RF (rep)** do not lead to increase the overall number of training instances.

It is noteworthy that the discretization may be performed on variables other than the target one: in the specific case of the prediction of the chlorophyll and ammonia, which are inherently correlated, the discretization (and the sampling) performed on one of them may also be adopted when learning the model to predict the other variable. This specific scenario, which exploits the correlation in the target space, has been explored in our experiments described in Section 3.

A flowchart of the whole proposed approach is reported in Fig. 4, while a detailed view of the proposed sampling approaches is shown in Fig. 5.

Finally, we emphasize that the proposed method can also be applied to huge amounts of data. Indeed, looking at Fig. 5, it is clear that each sample can be identified independently of the others, through the proposed sampling approach. Moreover, each tree of the forest can be learned independently of the others. These aspects make the proposed workflow scalable and easily applicable in the context of big data. However, it is out of the scope of this paper to empirically evaluate the scalability of the approach over a cluster of machines.

2.2.3. Explaining the predictions

Explaining how models make predictions is becoming of paramount importance in several fields. It not only can increase the trust in the model, but allows researchers and end-users to identify potential issues in the training data or in the adopted modeling approach. In our case, extracting an explanation in terms of the characteristics that mostly influence the prediction can help to verify that the model is considering the right signals of the images.

As mentioned in Section 2.2, single regression trees are interpretable, and allow to directly understand the rules behind a given decision, as well as the features that influence the predictions. However, Random Forests lack this characteristic since their predictions correspond to the average of the prediction of multiple trees. In this case, however, it is still possible to assess the importance of the features in the provided predictions, by resembling to their importance in each tree of the forest. In general, the rationale is that features involved in splits appearing in the top of the tree(s) are considered more important, since they have been selected earlier during the training phase, thus lead to the strongest reduction of the variance.

Accordingly, we can exploit the Gini index [26] to estimate the importance $I_i(t)$ of the feature i in the tree t as the average variance reduction of the split nodes that involve the feature i, weighted by the

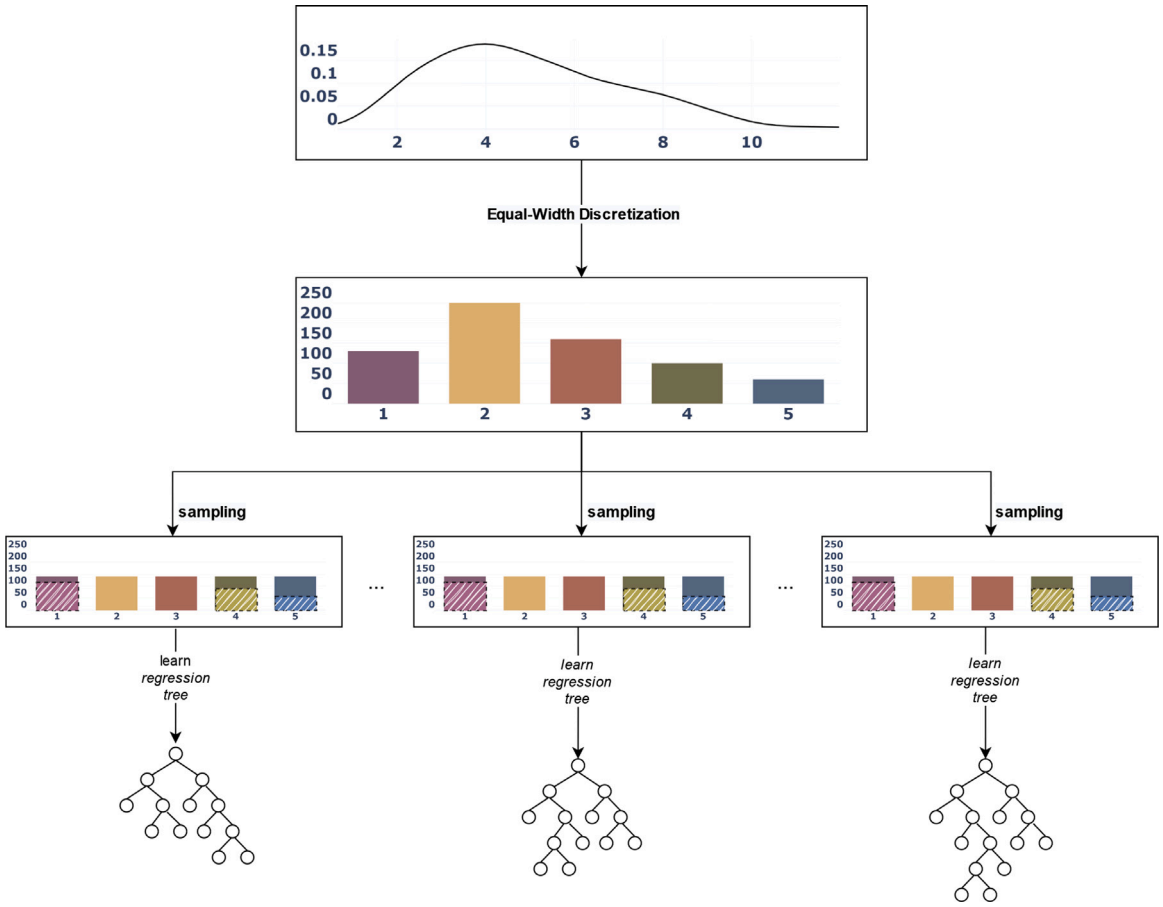


Fig. 5. A graphical overview of the proposed approach. In the depicted example, starting from $n = 700$ instances following a non-uniform distribution, the proposed method discretizes the range of values, assumed by the considered target variable in the training data, into $m = 5$ bins. Subsequently, it samples $n/m = 140$ instances from each bin to learn each regression tree of the forest. Instances sampled multiple times are kept as duplicates (solid-colored histograms in the figure) in the case of BAL-RF (rep), or discarded (dashed-colored histograms in the figure) in the case of BAL-RF (no rep).

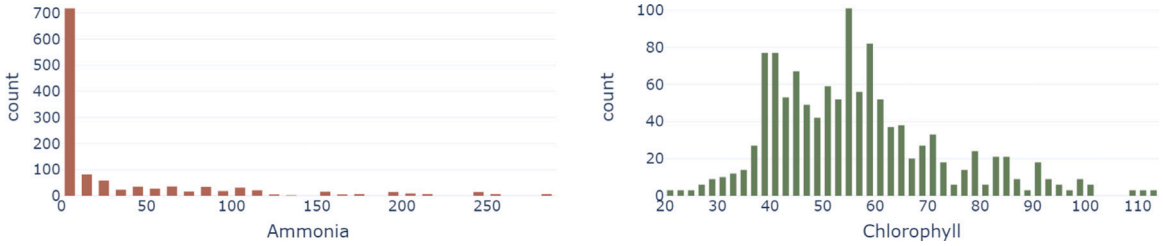


Fig. 6. Distribution of the values of ammonia (on the left) and of chlorophyll (on the right) in the considered dataset.

probability to reach such split nodes. The probability $p(N)$ to reach a given node N is estimated as the number of examples that reach the node, divided by the total number of examples n, namely: $p(N) = |N|/n$. Accordingly, $I_i(t)$ is computed as:

$$I_i(t) = \sum_{N \in t | f^*(N)=i} p(N) \cdot \tilde{\sigma}_N(f^*(N), v^*(N)) \quad (3)$$

where $f^*(N)$ and $v^*(N)$ are the best feature and value/threshold selected by the learning algorithm for the node N during the construction of the tree, and $\tilde{\sigma}_N(\cdot,\cdot)$ is the variance reduction achieved for the node N (see Eq. (2)).

Therefore, the importance of the feature i according the random forest is computed as the average of the importance of the feature i over all trees T in the forest. Formally:

$$I_i^{RF} = \frac{1}{|T|} \sum_{t \in T} I_i(t) \quad (4)$$

3. Performance evaluation

In this section, we describe our experimental evaluation, providing further details on the considered dataset and on the experimental setting. We performed our experiments on a dataset of 1191 images collected and processed following the procedure described in Section 2.1. Moreover, starting from the initial 68,121 features, representing discretized colors in the (a^*b^*) plane, we discarded those that were constantly valued with 0.0, leading to a total of 1997 features.

We considered the estimation of the amount of ammonia and of the amount of chlorophyll as two separate regression tasks. The distribution of ammonia and chlorophyll in the dataset is depicted in Fig. 6. As expected, their distribution is non-uniform. In particular, for the ammonia, we can observe that the lowest values are highly predominant.

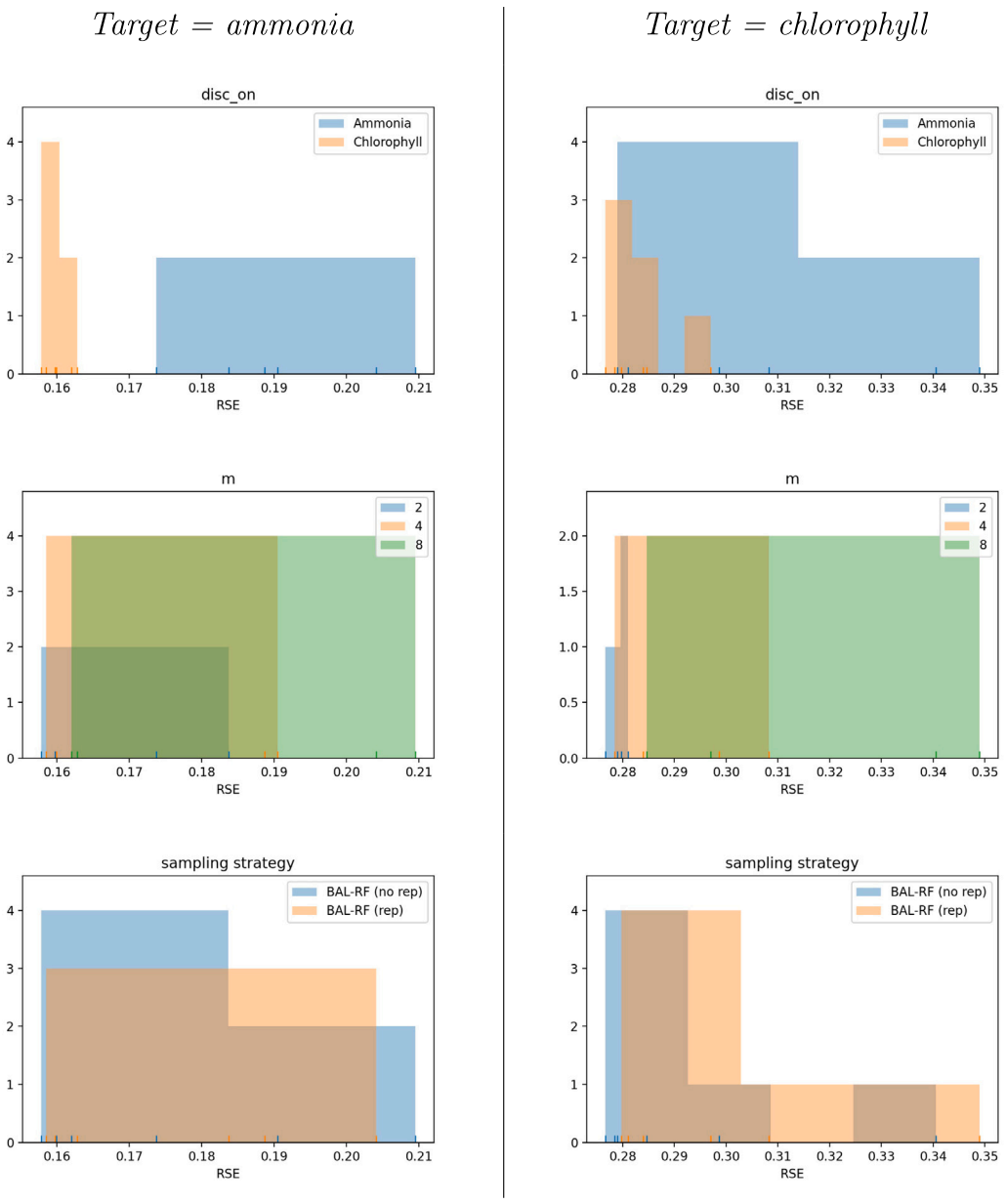


Fig. 7. Analysis of the influence of *disc_on*, *m* and *sampling_strategy* in terms of RSE on the prediction of the target variable *ammonia* (left) and *chlorophyll* (right).

As evaluation strategy, we adopted the 10-fold cross validation, where 90% of the dataset is considered as training set and the remaining 10% is considered as testing set, alternatively for 10 times.

As evaluation measures, we considered the Root Mean Squared Error (RMSE) and the Relative Squared Error (RSE). While the former provides a general idea about the actual error performed by the learned regressor, the latter estimates the accuracy of the regressor with respect to a simple baseline predictor, based on the average of the target variable in the training set. They are formally defined as:

$$RMSE = \sqrt{\frac{\sum (y_i - \widetilde{y_i})^2}{nt}} \tag{5}$$

$$RSE = \frac{\sum (y_i - \widetilde{y_i})^2}{\sum (y_i - \overline{y})^2} \tag{6}$$

where y_i and $\widetilde{y_i}$ are the true and the predicted values of the target variable, respectively, for the ith testing instance; $\overline{y}$ is the average value of the target variable in the training set; nt is the number of testing instances.

We run the experiments with both variants of our approach BAL-RF, namely **BAL-RF (rep)** and **BAL-RF (no rep)**, with different values of m, namely, $m \in \{2, 4, 8\}$. Moreover, as mentioned at the end of Section 2.2.2, we also investigated the possibility to perform the sampling guided by the discretization on a target attribute to learn a predictive model for the other target attribute. Henceforth, we indicate the attribute on which the discretization is performed with $disc\_on$, which will assume the values $disc\_on \in \{Ammonia, Chlorophyll\}$.

We compared our results with those achieved by the approach proposed by Palumbo et al. [22], that exploits Random Forests with the standard bootstrap sampling. Moreover, we also run the experiments with some other well-known regression methods, namely, Linear Regression (**LR**), Regression Tree (**RT**) and Support Vector Regressor with RBF kernel (**SVR**). The parameters of the method by Palumbo et al. [22] have been set to the values specified in the paper, while for LR, RT and SVR, we adopted their default values, as specified in the Python scikit-learn library. We also compared the results with those achieved by some approaches based on state-of-the-art neural network architectures: a classical Multi Layer Perceptron (**MLP**), Residual Network [27]

(**ResNet50**) and a Vision Transformer (ViT) [28], specifically **vit_b_16**. The MLP architecture consists of an input layer having size equal to the number of features, three hidden layers having size 512,256 and 128, and an output layer that returns the predicted value for the target attribute. As activation function, we considered ReLU. The architecture of ResNet50 and vit_b_16 were kept as they are described in their respective papers [27,28].

It is important to notice that, unlike our approach and MLP, both ResNet50 and vit_b_16 were directly trained from the original images, without extracting color histograms. This is because these architectures can also naturally capture the shape of the objects within the images, together with the pixel colors, and providing them with a pre-processed color histogram would have put them in an unfair training setting.

4. Results and discussion

All the results of our experiments are reported in Tables 1–2. Focusing on the prediction of the amount of ammonia (Table 1), we can observe that the proposed approach, in both its variants, provides advantages over the method proposed by Palumbo et al. [22] when the chlorophyll is used as discretization attribute ($disc\_on = chlorophyll$), with all the values of m. The average improvement on the prediction of the ammonia in terms of RSE, achieved by BAL-RF over the method proposed by Palumbo et al. [22], is 6.6%. This result confirms the relationship between the values of the chlorophyll and of the ammonia, and suggests that a sampling aiming to properly representing the whole range of values of the chlorophyll can lead to learn regression trees that are more accurate in the prediction of the ammonia. This may correspond to indirectly injecting some background information about the amount of chlorophyll that supports the prediction of the ammonia.

This advantage is not visible when the discretization is performed on the ammonia ($disc\_on = ammonia$), especially with high values of m. This result may be motivated by the highly skewed distribution of the ammonia (see the left part of Fig. 6), which leads to extremely undersampling from the first bin, in the case of BAL-RF (no rep), and to extremely oversampling from the other bins, in the case of BAL-RF (rep).

Comparing the results with the other competitor systems, we can observe that BAL-RF always outperforms all of them, with both considered discretization attributes and with all the values of m. Specifically, LR obtained a value of RSE higher than 1.0, which means that it performs worse than the naive predictor based on the sample mean. The results obtained overall suggest that the task at hand is not linear, and is more appropriate to adopt regressors able to capture non-linear relationships. Focusing on neural network architectures, MLP and ResNet50 obtained higher RMSE/RSE values with respect to all the configurations of BAL-RF, while vit_b_16 obtained a RMSE/RSE results comparable with those achieved by our worst configuration (BAL-RF (no-rep) - disc_on = Ammonia and m = 8). The sub-optimal results achieved by ResNet50 and vit_b_16 may be motivated by the fact that the quality of rocket leaves in terms of ammonia mainly influences the distribution of colors, rather than the shapes possibly captured by such architectures.

Switching the discussion to the prediction of the amount of chlorophyll (Table 2), we can draw similar conclusions. Also in this case, performing the sampling based on the discretization of the other target variable (i.e., ammonia, in this case) corresponds to indirectly injecting some knowledge about it in the predictor, which leads to lower prediction errors. Moreover, thanks to the lower skewness of the value distribution for chlorophyll (see the right part of Fig. 6), both variants of BAL-RF provide the expected advantages also when the discretization is performed on such a variable, which corresponds to making the learning algorithm more aware about the whole range of values assumed by the chlorophyll. Overall, the average improvement in terms of RSE, achieved by BAL-RF in the prediction of the chlorophyll over its closest competitor (i.e., the method by Palumbo et al. [22]) is 2.2%

Table 1
Results in terms of RMSE and RSE on the target variable *ammonia*. A green background indicates a better result with respect to all the competitors. The best result for each target variable is emphasized in bold.

Target = ammonia	*disc_on*	*m*	**RMSE**	**RSE**
LR	-	-	74.191	1.698
DT	-	-	32.758	0.331
SVR	-	-	39.895	0.491
MLP	-	-	28.312	0.250
resnet50	-	-	32.568	0.333
vit_b_16	-	-	26.129	0.251
Palumbo et al. [22]	-	-	23.581	0.171
BAL-RF(rep)	Ammonia	2	24.408	0.184
		4	24.735	0.189
		8	25.724	0.204
	Chlorophyll	2	22.760	0.160
		4	22.675	0.159
		8	22.977	0.163
BAL-RF(no rep)	Ammonia	2	23.734	0.174
		4	24.852	0.190
		8	26.062	0.209
	Chlorophyll	2	**22.626**	0.158
		4	22.775	0.160
		8	22.923	0.162

Table 2
Results in terms of RMSE and RSE on the target variable chlorophyll. A green background indicates a better result with respect to all the competitors. The best result for each target variable is emphasized in bold.

Target = chlorophyll	*disc_on*	*m*	**RMSE**	**RSE**
LR	-	-	31.252	3.691
DT	-	-	12.577	0.598
SVR	-	-	10.413	0.410
MLP	-	-	10.076	0.386
resnet50	-	-	10.978	0.463
vit_b_16	-	-	13.476	0.704
Palumbo et al. [22]	-	-	9.150	0.316
BAL-RF(rep)	Ammonia	2	8.624	0.281
		4	9.031	0.308
		8	9.608	0.349
	Chlorophyll	2	8.604	0.280
		4	8.668	0.284
		8	8.865	0.297
BAL-RF(no rep)	Ammonia	2	8.592	0.279
		4	8.889	0.299
		8	9.492	0.340
	Chlorophyll	2	**8.556**	0.277
		4	8.583	0.278
		8	8.678	0.285

when the discretization is performed on the ammonia and 10.4% when the discretization is performed on the chlorophyll.

The performance exhibited by the other competitors, also for the prediction of the chlorophyll, falls way behind those achieved by BAL-RF, with unsatisfactory results obtained by LR. Also for the chlorophyll, it appears that properly modeling the distribution of colors, as done by BAL-RF, MLP and the method by Palumbo et al. [22], is more beneficial than capturing differences in terms of shapes, as done by ResNet50 and vit_b_16. Moreover, the approach followed by BAL-RF for properly representing boundary values in the instance sampling procedure provides it with the advantage to outperform MLP and the method by Palumbo et al. [22].

We performed an additional analysis on the influence of the parameters m and $disc\_on$ as well as of the specific sampling strategy adopted by our method, i.e., BAL-RF (rep) and BAL-RF (no rep). The results of this analysis are depicted in Fig. 7. By observing the figure, we can confirm that the discretization performed on the chlorophyll is very beneficial for the prediction of both chlorophyll and ammonia. As regards the value of m, the best results are obtained with $m = 2$

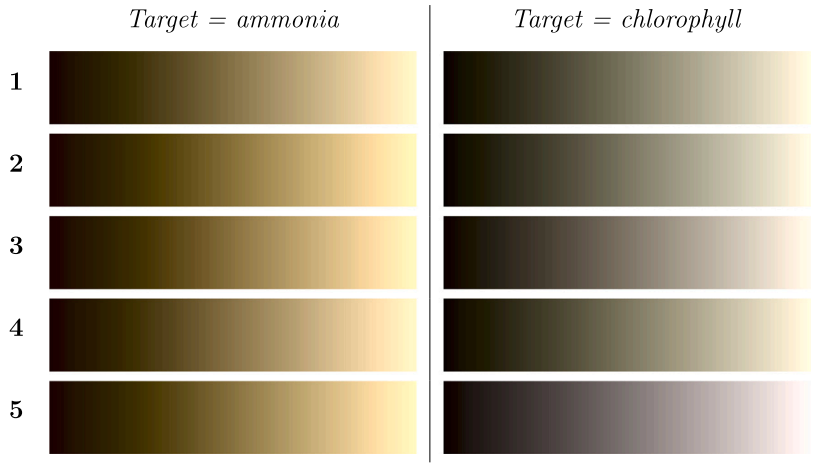


Fig. 8. Graphical representation of the 5 top-ranked features for the prediction of ammonia (left) and chlorophyll (right). Each gradient depicts the range of colors represented by a given pair of values in the a^*b^* plane, by varying the value of L in the range $[0, 100]$.

and $m = 4$ for the ammonia, and with $m = 2$ for the chlorophyll. This result indicates that the proposed sampling strategy is beneficial, provided that the range of values is not excessively fragmented. In the latter case, indeed, the effect of the combined undersampling and oversampling may significantly distort the training data distribution with respect to that observed during the prediction phase. Finally, regarding the sampling strategy, it seems that BAL-RF (no rep) and BAL-RF (rep) achieve comparable results, and that BAL-RF (no rep) provides a slight advantage. This result suggests that BAL-RF (no rep) is generally preferable, since it also feeds the learning algorithm with less training instances.

Finally, in Fig. 8 we show the 5 most important features, identified according to the strategy described in Section 2.2.3, for the configurations that led to the best results (see the bold values in Tables 1 and 2). Note that a feature corresponds to a given pair in the a^*b^* plane, while to actually draw the colors we need the triplet of values in the $L^*a^*b^*$ space. Therefore, in Fig. 8, we report, for each feature, the corresponding color gradient by varying the value of L^* in $[0, 100]$. From the figure, we can observe that brown-to-yellow colors are the most useful for the prediction of the value of ammonia, while, as expected, greens mostly support the regressors to predict the value of the chlorophyll. As introduced in Section 2.2.3, this analysis not only helps to observe and understand the features that mostly explain the predictions, but can also increase the trust in the designed approach for its application in real-world environments, since it is easy to verify that the provided output is based on reasonable characteristics of the images.

In general, the main findings of this study can be summarized as follows:

- the proposed method can be considered the new state-of-the-art for the estimation of chlorophyll and ammonia contents from images, since it outperformed 7 existing methods solving the same task;
- the obtained results confirmed the inherent relationship between chlorophyll and ammonia: using some information about the amount of chlorophyll during the discretization phase led to improvements in the estimation of the amount of ammonia;
- the considered task is clearly not linear: approaches able to model only linear relationships (e.g., LR) failed to provide accurate estimations;
- the amount of chlorophyll and ammonia in rocket leaves appears to influence the frequency of colors, rather than their location in the image: complex neural network architectures, whose features capture spatial structures within the images, obtained worse results with respect to those focusing on color histograms, such as our approach;
- exploiting the capability of our method of explaining its outputs, we observed that brown-to-yellow colors were leveraged for the estimation of ammonia, while greens mostly supported the estimation of chlorophyll. The relationships between such colors and chlorophyll/ammonia contents appear natural and reasonable for experts. This increases the belief that the estimations provided by our approach are trustworthy.

5. Conclusions

In this paper, we proposed an approach for the contactless and non-destructive evaluation of fresh-cut rocket leaves, based on the estimation of the amount of chlorophyll and ammonia. Specifically, we adopted a machine learning method that builds a regression model able to estimate such amounts from images related to rocket leaves. Methodologically, with respect to previous works, we contributed along two different aspects: (i) we tackled the issues raised by the non-uniform value distributions of the target variables in the bootstrap procedure adopted by Random Forest, by adopting a novel sampling approach; (ii) we adopted a strategy to explain the estimations provided by the model, which aims to provide clues on the colors that mostly contributed to the prediction.

Our extensive experimental evaluation proved that the proposed approach can outperform methods based on Random Forests and 6 additional competitor systems, in terms of RMSE and RSE. Specifically, the proposed sampling approach not only allowed the regressor to be more aware of the whole range of values in the dataset, but indirectly led to injecting some background information about the amount of chlorophyll to support the prediction of the ammonia, and viceversa.

As future work, we will aim to improve the learned models to make them robust to variations in the images possibly introduced by the live environment, including light/color perturbations and the presence of complex backgrounds. These issues are currently coped with by color correction and foreground extraction steps, that pose strict constraints on the environment used during the acquisition. At this aim, we will also investigate the possibility to exploit specific settings, also based on adversarial learning.

CRediT authorship contribution statement

Stefano Polimena: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gianvito Pio:** Writing – review & editing, Validation, Methodology, Formal analysis, Conceptualization. **Maria Cefola:** Resources, Investigation, Data curation. **Michela Palumbo:** Resources,

Investigation, Data curation. **Michelangelo Ceci:** Writing – review & editing, Validation, Supervision, Funding acquisition, Formal analysis, Conceptualization. **Giovanni Attolico:** Writing – review & editing, Validation, Supervision, Funding acquisition, Data curation, Conceptualization.

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.

Acknowledgments

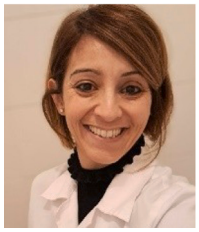
This work was partially supported by the project FAIR - Future AI Research (PE00000013), spoke 6 – Symbiotic AI, under the NRRP MUR program funded by the NextGenerationEU and by the project Prin 2017 "SUS&LOW-Sustaining low-impact practices in horticulture through non-destructive approach to provide more information on fresh produce history and quality" (grant number: 201785Z5H9) from the Italian Ministry of University and Research. This work has also been partially supported by the European Union - NextGenerationEU through the Italian Ministry of University and Research, PRIN 2022 Project "COCOWEARS" (A framework for COntinuum COmputing WEARable Systems), grant n. 2022T2XNJE, CUP: H53D23003650001. S. Polimena also acknowledges the financial support of his Ph.D. fellowship provided by the CNR STIIMA - "Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing" of the National Research Council of Italy.

References

[1] Bhargava A, Bansal A. Fruits and vegetables quality evaluation using computer vision: A review. J King Saud Univ Comput Inf Sci 2021;33:243–57.
[2] Narendra VG. An intelligent computer vision system for vegetables and fruits quality inspection using soft computing techniques. Agric Eng Int CIGR J 2019;21:171–8.
[3] Verain MC, Sijtsema SJ, Antonides G. Consumer segmentation based on food-category attribute importance: The relation with healthiness and sustainability perceptions. Food Qual Pref 2016;48:99–106.
[4] Amodio ML, Ceglie F, Chaudhry MMA, Piazzolla F, Colelli G. Potential of nir spectroscopy for predicting internal quality and discriminating among strawberry fruits from different production systems. Postharvest Biol Technol 2017;125:112–21.
[5] Chaudhry MM, Amodio ML, Amigo JM, de Chiara ML, Babellahi F, Colelli G. Feasibility study for the surface prediction and mapping of phytonutrients in minimally processed rocket leaves (diplotaxis tenuifolia) during storage by hyperspectral imaging. Comput Electron Agric 2020;175:105575.
[6] Chaudhry MM, Amodio ML, Babellahi F, de Chiara ML, Rubio JMA, Colelli G. Hyperspectral imaging and multivariate accelerated shelf life testing (maslt) approach for determining shelf life of rocket leaves. J Food Eng 2018;238:122–33.
[7] Løkke MM, Seefeldt HF, Skov T, Edelenbos M. Color and textural quality of packaged wild rocket measured by multispectral imaging. Postharvest Biol Technol 2013;75:86–95.
[8] Sánchez MT, Garrido-Varo A, Guerrero J-E, Pérez-Marín D. Nirs technology for fast authentication of green asparagus grown under organic and conventional production systems. Postharvest Biol Technol 2013;85:116–23.
[9] Cavallo DP, Cefola M, Pace B, Logrieco AF, Attolico G. Non-destructive automatic quality evaluation of fresh-cut iceberg lettuce through packaging material. J Food Eng 2018;223:46–52.
[10] Pace B, Cefola M, Renna F, Attolico G. Relationship between visual appearance and browning as evaluated by image analysis and chemical traits in fresh-cut nectarines. Postharvest Biol Technol 2011;61:178–83.
[11] Arivu CVG, Prakash G, Sarma ASS. Online image capturing and processing using vision box hardware: Apple grading. Int J Mod Eng Res 2012;2:639–43.
[12] Palumbo M, Pace B, Cefola M, Montesano FF, Serio F, Colelli G, et al. Self-configuring cvs to discriminate rocket leaves according to cultivation practices and to correctly attribute visual quality level. Agronomy 2021;11(1353).
[13] Limantara L, Dettling M, Indrawati R, Brotosudarmo THP, et al. Analysis on the chlorophyll content of commercial green leafy vegetables. Procedia Chem 2015;14:225–31.
[14] Cavallo DP, Cefola M, Pace B, Logrieco AF, Attolico G. Contactless and non-destructive chlorophyll content prediction by random forest regression: A case study on fresh-cut rocket leaves. Comput Electron Agric 2017;140:303–10.
[15] Pace B, Cavallo D, Cefola M, Burbaci S, Attolico G. Colour analysis to predict the total chlorophyll content of rocket leaves. In: VI international symposium on applications of modelling as an innovative technology in the horticultural supply chain model-IT 1311. 2019, p. 107–12.
[16] Chandra D, Matsui T, Suzuki H, Kosugi Y. Postharvest changes in some physiological traits and activities of ammonia-assimilating enzymes in lettuce during storage. Asian J Plant Sci 2006;5:378–84.
[17] Cefola M, Amodio ML, Rinaldi R, Vanadia S, Colelli G. Exposure to 1-methylcyclopropene (1-mcp) delays the effects of ethylene on fresh-cut broccoli raab (brassica rapa l.). Postharvest Biol Technol 2010;58:29–35.
[18] Amodio M, Colelli G, Cantwell M. Ammonia accumulation in plant tissues: a potentially useful indicator of postharvest physiological stress. Acta Hortic 2018;1194:1511–8.
[19] Pace B, Cefola M, Da Pelo P, Renna F, Attolico G. Non-destructive evaluation of quality and ammonia content in whole and fresh-cut lettuce by computer vision system. Food Res Int 2014;64:647–55.
[20] Tudela JA, Marín A, Garrido Y, Cantwell M, Medina-Martínez MS, Gil MI. Off-odour development in modified atmosphere packaged baby spinach is an unresolved problem. Postharvest Biol Technol 2013;75:75–85.
[21] Cefola M, Pace B. Application of oxalic acid to preserve the overall quality of rocket and baby spinach leaves during storage. J Food Process Preserv 2015;39:2523–32.
[22] Palumbo M, Pace B, Cefola M, Montesano FF, Colelli G, Attolico G. Non-destructive and contactless estimation of chlorophyll and ammonia contents in packaged fresh-cut rocket leaves by a computer vision system. Postharvest Biol Technol 2022;189:111910.
[23] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artificial Intelligence Res 2002;16:321–57.
[24] Zhu B, Pan X, Broucke Svanden, Xiao J. A gan-based hybrid sampling method for imbalanced customer classification. Inform Sci 2022;609:1397–411.
[25] Hido S, Kashima H, Takahashi Y. Roughly balanced bagging for imbalanced data. Stat Anal Data Min ASA Data Sci J 2009;2:412–26.
[26] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[27] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. CVPR, 2016.
[28] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020, arXiv preprint arXiv:2010.11929.

Stefano Polimena is a Ph.D. student in Computer Science and Mathematics at the Department of Computer Science, University of Bari Aldo Moro, Italy. His research interests include Computer Vision, Deep Learning, Data Mining and Big Data Analytics, with applications to the non-destructive and contactless quality control in the agroalimentary supply chain. During his Ph.D. studies, he spent three months at the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), University of South Brittany (UBS), focusing on innovative regression techniques, also based on contrastive learning, applied on images related to fruits and vegetables.

Gianvito Pio is an Assistant Professor (Tenure Track) at the Department of Computer Science of the University of Bari Aldo Moro. His research interests include data mining, machine learning, and big data analytics, with applications in bioinformatics, energy, social network analysis and law. On these topics, he (co-)authored more than 55 papers in conference proceedings and in international journals with a high impact factor. He participated in the organization of international workshops and conferences on the topics of machine learning and data mining. He is responsible of a research unit of the national research project COCOWEARS, funded by the Italian Ministry for Universities and Research. He is member of the editorial board of Springer Machine Learning Journal, Elsevier Expert Systems with Applications (Associate Editor), BMC Bioinformatics (Associate Editor) and Springer Medical & Biological Engineering & Computing (Associate Editor).



Maria Cefola is a researcher of Institute of Sciences of Food Production, National Research Council of Italy (CNR-ISPA). From 2006 to 2010, she attended the Ph.D. School in Innovation Management in Agri-Food Systems in the Mediterranean, researching post-harvest management of fresh and fresh-cut fruit and vegetable products of the Mediterranean region at the University of Foggia. During her doctorate, she spent six months at the Department of Plant Science of UC Davis. In 2004, she obtained a Master's Degree in Food Science and Technology at the University of Basilicata. She has had significant experience in post-harvest management of fruits and vegetables within international, national and regional research projects and has attended international courses and congress. Her research activities regard: the improvement of the quality of fresh-cut products through the application of innovative pre-treatment, packaging and/or storage conditions; innovation in logistic cold chain; the study of innovative non-destructive systems for quality evaluation.

Dr. Michela Palumbo earned her master degree in Agricultural Science and Technology at the University of Foggia in 2011. From 2012 to 2014 she won a scholarship at University of Foggia about soil and soilless cultivation systems and pre-harvest practices to produce high quality horticultural products for postharvest storage. In 2019 she had a scholarship at the Institute of Science of Food Production of Italian National Research Council (ISPA-CNR) on postharvest practices to preserve the quality and to improve the shelf-life of fresh fruit and vegetables. From 2019 to 2023 she attended a Ph.D. school in Management of Innovation in the Agricultural and Food System of the Mediterranean Region at University of Foggia, studying the development and validation of a Computer Vision System, as a non-destructive methodology to predict the quality of fresh-cut rocket leaves. In 2023 she discussed her Ph.D. thesis entitled: "Computer Vision System for non-destructive and contactless evaluation of quality traits in fresh rocket leaves (Diplotaxis tenuifolia L.)". Actually, she is a research fellow at ISPA-CNR working on innovative postharvest techniques (dipping treatments, controlled and modified atmospheres, active packaging, cold temperatures, non-destructive technologies) to monitor the quality and preserve the shelf-life of fresh and fresh-cut fruit and vegetables.

Michelangelo Ceci is a full professor at the Department of Computer Science of the University of Bari Aldo Moro. His research interests are in data mining, ML, big data analytics and ensemble learning. He coordinated UNIBA's participation in numerous national and international research projects, including FP7-FET-MAESTRA, H2020-IMPETUS and H2020-CounteR, as well as many national projects, including BA-PHERD, MAD, Vi-POC and ComESto. He is an action editor of the Data Mining and Knowledge Discovery journal, Machine Learning journal and Journal of Intelligent Information Systems. He published over 220 international publications and was PC chair of five international conferences: SEBD2007, Discovery Science 2016, ISMIS 2018, ISMIS 2022, SEBD 2024. He also was General Chair of ECML-PKDD 2017 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases in 2017.

Giovanni Attolico received the degree in Information Science from the University of Bari. He worked on Signal and Image Processing, 3D Modeling, Sensory Processing and Interpretation, Robotics, Interactive and Multimodal Systems, Haptic Interfaces, as research leader in a private company (Olivetti), as visiting scientist in the Vision and Modeling Group at the MediaLab (MIT) and as senior researcher in the Multi-spectral and Hyperspectral Imaging Laboratory of STIIMA (Institute on Intelligent Industrial Systems and Technologies for Advanced Manufacturing) of the Consiglio Nazionale delle Ricerche (CNR). His current research interests are in the application of computer vision and machine learning to quality control and evaluation of agricultural products. He has been scientific group leader in research initiatives founded by public institutions and by private companies and co-authored more than one hundred scientific papers on journals and proceedings of conferences and an international patent on a visual system for events' detection. He acts as referee for journals in the areas of artificial intelligence, computer vision and their application to cultural heritage and food science. He serves as scientific referee for the Italian Ministry of Research and University, for the Ministry of Development and for other public institutions.