Automated identification and visualization of food defects using RGB imaging: Application to the detection of red skin defect of raw hams

Alessandro Ulrici, Giorgia Foca, Maria Cristina Ielo, Luisa Antonella Volpelli, Domenico Pietro Lo Fiego

PII: S1466-8564(12)00117-8
DOI: doi: 10.1016/j.ifset.2012.09.008
Reference: INNFOO 930

To appear in: *Innovative Food Science and Emerging Technologies*

Received date: 16 July 2012
Accepted date: 23 September 2012

Please cite this article as: Ulrici, A., Foca, G., Ielo, M.C., Volpelli, L.A. & Fiego, D.P.L., Automated identification and visualization of food defects using RGB imaging: Application to the detection of red skin defect of raw hams, *Innovative Food Science and Emerging Technologies* (2012), doi: 10.1016/j.ifset.2012.09.008

AUTOMATED IDENTIFICATION AND VISUALIZATION OF FOOD DEFECTS USING RGB IMAGING: APPLICATION TO THE DETECTION OF RED SKIN DEFECT OF RAW HAMS

Alessandro Ulrici [1,2*], Giorgia Foca [1,2], Maria Cristina Ielo [2], Luisa Antonella Volpelli [1,2], Domenico Pietro Lo Fiego [1,2]

[1] *Department of Life Sciences and* [2] *Interdipartimental Research Centre for Agri-Food Biological Resources Improvement and Valorisation, University of Modena and Reggio Emilia, Padiglione Besta, Via Amendola 2, 42122 Reggio Emilia, Italy*

* Corresponding author: Alessandro Ulrici, Department of Life Sciences, University of Modena and Reggio Emilia, Padiglione Besta, Via Amendola 2, 42122 Reggio Emilia, Italy. Tel: +39 0522 522043. Fax: +39 0522 522027. E-mail: alessandro.ulrici@unimore.it

Abstract

Colourgrams are signals that codify the colour-related information content of a Red-Green-Blue (RGB) image, and which can be elaborated by means of proper multivariate analysis/feature selection techniques to easily identify those image features that are more useful to solve a specific problem. The reconstruction of the selected features as segmented images allows to evaluate in a critical manner the choices made automatically by the algorithm. In the present paper colourgrams are used for the detection of the red skin defect of raw hams, in order to render more objective and transferable the evaluation usually made by expert assessors. To this aim, after a preselection of 95 raw ham samples by a panel test, the corresponding RGB images were converted into colourgrams, which in turn were used to build classification models using Partial Least Squares-Discriminant Analysis (PLS-DA) and a Wavelet Packet Transform-based feature selection/classification algorithm (WPTER). Feature selection allowed to discriminate the defective samples using only three variables, with a Classification Efficiency in prediction of an external test set equal to 97.8%. The reconstruction of the samples images using only the selected features confirmed the reliability of the obtained classification model.

Industrial Relevance: The evaluation of pig thighs is currently carried out by subjective methods, i.e. expert, long-trained personnel is needed to detect the presence or absence of defects. The method presented here would allow to uniform and drastically shorten the time needed for evaluation, and to avoid the main problems connected with human evaluation, i.e., subjectivity, possible unreliability, non-transferability and difficulty to collect historical data. Furthermore, it might represent a first step for setting up a comprehensive method of evaluation, aiming to take into account also other types of defects of raw hams destined to seasoning. More in general, thanks to its flexibility, this approach could be also successfully applied for the detection of other types of aspect-related features, even to monitor different kinds of products.

Keywords: RGB Images; Multivariate Classification; Feature Selection; Wavelet Transform; Defect Detection; Raw Ham Red Skin Defect.

1. Introduction

In the production of Protected Denomination of Origin (PDO) ham, the assessment of qualitative characteristics of fresh pig thighs is of utmost importance, in order to early define the final destination of the seasoned product. A preliminary classification of fresh thighs on the basis of the presence/absence and of the extent of defects, that are responsible – together with other factors – for the final quality and for the price of the seasoned product, is in fact highly relevant by the logistic and by the financial points of view. The red skin defect appears as a more or less extended and intense red colour of the ham rind, which turns into a dark-brown shading after seasoning (Lo Fiego et al., 2006). Even if this defect does not affect the sensory quality of the product, the dark-brown colour makes the seasoned ham unattractive to consumers, which in turn results in a lowering of its price. The origin of the red skin defect has not been clearly identified, but the incidence of this defect may be influenced by the slaughtering techniques, such as stunning voltage of animals and scalding methodologies of the carcass (Lo Fiego et al., 2009).

At present, the evaluation of the defects of raw ham destined to processing is based on the estimation, made by expert assessors at the moment of trimming. Unfortunately, the human visual evaluation is subjected to a series of drawbacks: i) the evaluation is subjective, i.e. operator dependent, and therefore it is not easily transferable among different production lines and/or industries; ii) the human eye could sometimes be inconsistent, so that a certain assessor may provide contradictory evaluations for the same sample and iii) the evaluation is dependent on the availability of specialised manpower. In this context, automated systems capable of acquiring and elaborating aspect-related data are definitely valuable tools, since they can furnish objective, reproducible and transferable information about the appearance of the analysed products (Lo Fiego et al., 2007). The need to use automated methods, i.e. not based on human visual evaluation, for estimating ham quality is clearly described in the review by Valous et al. (2010).

The use of spectrophotometers or light sensitive cells is commonly used for the quantification of food sample colour–related aspects. The instruments traditionally used for these kinds of measurements can be ascribed to two categories: spot-colorimeters and integrating spheres. Spot colorimeters analyse only restricted areas of the sample, therefore they are not appropriate for products showing an inhomogeneous aspect like raw hams. On the other hand, integrating spheres estimate the overall light reflectance from the whole sample surface, giving only a global colour evaluation and, consequently, losing information about its spatial variability. Fortunately, nowadays it is possible to analyse the sample colour both locally and globally, since the recent progress in image acquisition technology allows to use high-performance equipments, available at very low costs. Digital cameras are able to perform a detailed colour evaluation of food products with

inhomogeneous aspect, since the colour of every single portion of the analysed sample can be accounted for by one or more image pixels. These are the reasons why in the last 20 years the field of Red-Green-Blue (RGB) image analysis techniques has gained an increasing interest in industrial applications in general, and in the field of food analysis in particular (Geladi & Grahn, 1996; Panigrahi & Gunasekaran, 2001; Zheng et al., 2006; Prats-Montalbán, et al., 2011; Garrido-Novell et al., 2012).

In this context, the objective evaluation of some characteristics of fresh pig thighs, such as the presence/absence of a specific defect, can be efficiently handled by digital image analysis. Some research works on meat samples reported the use of RGB image-based systems for the detection of defects associated to peculiar chromatic characteristics. In particular, the detection of defects on chicken meat before packing has been reported (Barni et al., 1997): in this work, possibly defective areas are first extracted in chicken images by means of morphological image reconstruction, and then classified according to a predefined list of defects. More recently, Marthy-Mahe et al. (2003) developed and tested different procedures to segment the images of raw hams, that are irregular three-dimensional objects with random shape and size, in order to detect the presence of different defects, Carnier et al. (2004) used computer image analysis for measuring lean and fatty areas in cross-sectioned dry-cured hams and Faucitano et al. (2005) to measure pork marbling characteristics.

Other research works were published, where the performances of image analysis for meat quality evaluation where correlated with the results of human assessment. When referring to scores or classes defined by human assessors, the correctness of the reference values coming from human evaluation is fundamental to obtain reliable automated systems. In a recent paper (Foca et al., 2007), some of us have already highlighted the intrinsic subjectivity of human evaluation, even by expert assessors, of food-related properties. Iqbal et al. (2010) classified the quality of pork and turkey hams based on image colour and textural features and their relationships with consumer responses, using Mahalanobis distance and feature inter-correlation analyses to select the optimal descriptors, among a set of global parameters like means, standard deviations and entropies calculated using different colour spaces like RGB, HSV and L*a*b*. Sanchez et al. (2008) quantified the lean, fatty and connective tissue areas on the ham surface and determined the relationship of those areas to salt gain during the salting process, comparing the performance of the automated method with the results of a human classification. Their work was based on image segmentation using two parameters, corresponding to the differences between red and blue channels and between green and blue channels, calculated for each RGB image pixel. Tan et al. (2000) compared the ability of colour machine vision and untrained panellists to evaluate the colour of

fresh pork meat, using a neural network image classifier trained with single pixel data coming from preselected areas of interest of reference images.

In this paper, we propose the application of the colourgrams-based approach for the detection of the red skin defect of raw hams using RGB imaging. Colourgrams (Antonelli et al., 2004; Foca et al., 2011) are one dimensional signals (vectors) that codify the whole colour-related information content of a RGB image, and can be therefore considered as "fingerprints" of the corresponding images for their subsequent elaboration by means of proper multivariate analysis techniques, like data exploration, calibration or classification. When coupled with proper feature selection techniques, their use makes possible the identification of the colour-related aspects of interest that are more useful to solve a specific problem like, in the present case, the detection of the red skin defect. Moreover, the reconstruction of the selected features under the form of segmented images allows to evaluate in a critical manner the choices made automatically by the algorithm.

With respect to previous research works dealing with RGB image analysis for the automated detection of meat defects, the proposed blind analysis approach presents the advantage to render more flexible the classification model creation phase. In this case, in fact, the search of the optimal descriptors is not limited to few colour parameters based on *a priori* assumptions on the colour characteristics of interest, since this may pose limits to the classification performance. On the contrary, few useful descriptors can be automatically selected starting from a wide number of potentially useful features. Moreover, it is not necessary to perform manually a preselection of the image areas useful for classification, since the feature selection of colourgrams does it automatically and in an independent manner from probably unreliable human intervention. Once the useful features have been automatically selected, the time needed to apply the classification model to a new set of images and to visualize the areas of interest (i.e., where the defect is present) is very short, making an on-line implementation feasible.

In particular, in the present research work colourgrams were used for the detection of the red skin defect of raw hams, in order to render more objective and transferable the evaluation that is usually made by expert assessors. To this aim, digital RGB images were acquired from a set of raw ham samples and then, by means of a graphical user interface implemented *ad hoc*, they were classified into three quality categories related to the red skin defect by a panel of six expert assessors. The panel test results were then used to select the images of those samples whose class assignation was sufficiently consistent.

The RGB images were converted into colourgrams, which in turn were used to build classification models using Partial Least Squares-Discriminat Analysis (PLS-DA) and a Wavelet Transform-based feature selection / classification algorithm, WPTER (Cocchi et al., 2003). Finally, the

colourgram features selected by WPTER were used to reconstruct the RGB images, making it possible to highlight in the original image domain the areas of ham surface where the red skin defect is located.

2. Material and methods

2.1 Samples and images acquisition

The left thighs of 198 heavy pigs, slaughtered in one plant during 5 different days and destined to the PDO "Prosciutto di Parma" production, were considered for this study. After carcass slicing, the thighs were stored at 0-4°C for 24 hours and subsequently trimmed, then RGB images of the external surface of each thigh were acquired using a Nikon Coolpix 5400 digital camera with a 5.8-24 mm focal length (Nikon corp., Tokyo, Japan).

Digital images were acquired in JPEG format with a spatial resolution of 2592×1944 pixels, using white balance, with a 1/125 s shutter speed and an f/5.6 lens aperture. The choice to use a compressed image file format (i.e., JPEG file format, with average file size equal to 1.9 MB) instead of RAW images (i.e., uncompressed images with file size equal to 14.4 MB) was made on the basis of preliminary trials. In particular, images were acquired on a subset of samples both in RAW and in JPEG modes. For each file format, the corresponding dataset of colourgrams was created, then each dataset was analysed using Principal Component Analysis (PCA). The comparison of the data structure of the two datasets (i.e., colourgrams obtained by RAW images vs. colourgrams obtained by JPEG images), which was made using the score plots of the first four Principal Components, revealed similar patterns, suggesting that the loss of useful information in the colourgrams deriving from JPEG compressed images is small. This result was somehow expected, since JPEG compression affects more texture than colour, which is the main characteristic used for colourgrams-based detection of the red skin defect. In case of different classification problems, mainly related to texture and in particular concerning the detection of small details, then the effect of JPEG compression could be more marked. Based on these considerations, and focusing on the usability of the method (in terms of data storage and of computational power requirements), JPEG images were used.

In order to have constant and homogeneous lighting conditions, the camera was mounted on a white painted wooden box, containing the thigh to be photographed, equipped with 8 tungsten lamps (Philips 25 W 240 V SES Argenta Lustre), that were turned on 1 hour before starting measurements to allow the stabilisation of their emission spectra. Furthermore, the effect of

possible variations of the illumination conditions was evaluated by preliminary tests using colour standard references to correct the RGB values. However, no improvements were obtained when using the corrected images, and for this reason all the subsequent image elaborations were performed directly on the RGB images without any pretreatment.

2.2 Human visual assessment

Six assessors, indicated with letters from A to F, were asked to independently classify the images in three categories related to the extent of the red skin defect. The expert assessors involved in the panel are specialised technicians and researchers with a long experience in the evaluation of raw ham defects, therefore no specific training was required for the visual evaluation of the samples considered in this work. The assessors were invited to compare each sample with three reference images, each one representative of a specific class, where 1 = defect absent, 2 = slight defect, 3 = severe defect.

For each one of the 198 digital images two evaluations were performed by each assessor. Firstly, all the 198 images were evaluated following their original sequence, then a second evaluation was performed by sorting the same images in random order. The overall sequence of 396 images was the same for all the assessors, and the assessors were simply asked to evaluate 396 different samples.

A graphical user interface was implemented *ad hoc* for the panel, allowing the assessors to evaluate each image by comparing it with one reference image for each class, as reported in Figure 1. The software was conceived to show in sequence each one of the 396 images to be evaluated (whose progressive number was reported in the upper left corner of the image) and required that the proper class was assigned by the assessor before passing to the following one. The program did not allow a user to return to previously evaluated samples, so that the assessor was forced to consider each sample independently. Moreover, in order to avoid inconsistent evaluations due to weariness of the assessors, the software allowed to exit at any time during the evaluation and to restart it from the point where it was interrupted. The assessors performed the evaluation independently each other.

2.3 Elaboration of the panel test data

In order to analyse the results of the panel, as a starting point the “correct” class of each sample had to be defined. In fact, since the evaluation of the extent of red skin defect was based on human assessment, the definition of the correct class was subjective and questionable. Nevertheless, the

definition of "correct assignation" is fundamental for the subsequent data analysis. As a consequence, we decided to categorically assign each sample to the most frequently selected class (i.e. to the mode of the 12 class assignation values for each sample), that we called "correct".

The estimates of the assessors performance were expressed in terms of Validity and Reliability. Validity (V) was defined as the percentage of assignations to the correct class for each assessor, which is an estimate of the agreement of the assessor with the whole panel. Reliability (R) was defined as the percentage of assignations to the same class over the two repeated evaluations of each sample, and it reflects the ability of each assessor to reproduce his own results, independently from the assignation to the correct class. The sum of Validity and Reliability scores for each assessor was defined as the assessor Global Performance, which was used to evaluate the uniformity of the panel by means of a chi-squared test.

To estimate the agreement in the attribution of each sample, the percentage level of Overall Agreement on Samples Attributions (%OASA) for every single sample was defined as follows:

$$\%OASA = \frac{N_{CORR}}{N_{TOT}} \times 100 \qquad \text{(eq. 1)}$$

where N_{CORR} is the number of attributions of the considered sample to the correct class, while N_{TOT} is the total number of attributions (6 assessors × 2 replicated estimates = 12).

The %OASA results were then exploited to select those samples whose class assignation was sufficiently univocal, to be used for the development of the automated classification models. In particular, the 95 images of samples presenting %OASA values greater than 70% (i.e., at least 9 evaluations in agreement out of 12) were selected to build and validate the classification models.

2.4 Conversion of RGB images into colourgrams

The Matlab function that was developed to convert each RGB image into the corresponding colourgram goes through the following steps:

- read from the hard disk each RGB image file, which is a 3D array of size {*r*, *c*, 3}, where *r* is the number of pixel rows, *c* is the number of pixel columns, and 3 is the number of channels, i.e., the R, G and B values of each pixel. For the images analysed in the present work, the size of the 3D array is therefore equal to {1944, 2592, 3};
- unfold the 3D array to a 2D matrix with size {(*r* × *c*), 3}, which contains all the pixels in rows and the R, G and B channels in columns. For the images analysed in the present work, the size of the 2D array is therefore equal to {5038848, 3};

- expand this 2D matrix by adding a series of columns, corresponding to parameters calculated for each pixel starting from the R, G and B values. In particular: i) column 4 contains the values of Lightness (L), i.e., the R + G + B sum; ii) columns 5-7 contain the ratios between each channel (R, G, B) and L, which are defined as "relative colours": relative Red (rR), relative Green (rG) and relative Blue (rB); iii) columns 8-10 contain the Hue (H), Saturation (S) and Intensity (I) values, obtained by converting the RGB data into the HSI colour space. In this way, the 2D matrix has now size equal to $\{(r \times c), 10\}$;
- further expand the number of columns of the 2D matrix by calculating PCA models on the unfolded RGB data (X), i.e. on the first 3 columns of the 2D matrix. In particular, three PCA models are calculated: the first model (PCA_RAW) is calculated on the raw (i.e., not pretreated) data, the second one (PCA_MNCN) is calculated on the mean centered data, and the third one (PCA_AUTO) is calculated on the autoscaled data. Since the number of variables of X is equal to 3, each PCA model will have 3 Principal Components (PCs). The expansion of the number of columns of the 2D matrix is then accomplished in as follows: i) columns 11-13 contain the three score vectors of PCA_RAW; ii) columns 14-16 contain the three score vectors of PCA_MNCN; ii) columns 17-19 contain the three score vectors of PCA_AUTO. In this way, the 2D matrix has now size equal to $\{(r \times c), 19\}$;
- for each one of the 19 columns of the 2D matrix, calculate the corresponding frequency distribution vector with a length of 256 points;
- create the first part of the colourgram by joining in sequence the 19 frequency distribution vectors: a vector with length equal to $(19 \times 256) = 4864$ points is obtained;
- create the second part of the colourgram by joining in sequence the values of the loading vectors (3 values for each loading vector × 3 PCs = 9 points) and of the eigenvalues of the 3 PCs (3 points), for each one of the 3 PCA models (PCA_RAW, PCA_MNCN and PCA_AUTO). This leads to a vector with length equal to $[(9 + 3) \times 3] = 36$ points;
- create the whole colourgram by joining in sequence its first and second part, thus obtaining a vector with length equal to $(4864 + 36) = 4900$ points, which describes the colour properties of the image.

For a more detailed description of the algorithm used to build colourgrams, the reader is referred to Antonelli et al. (2004).

This data compression is particular advantageous, since from the millions of data of the original images, the colour-related information is compressed in a 4900 points long signal, which can be further significantly shortened up to few units, by proper feature selection methods. The colourgram

can be used as fingerprint of the colour content of the image, in the same way as a NIR spectrum is the fingerprint of the chemical composition of a sample. Colourgrams can then be analysed by means of suitable signal processing / multivariate analysis techniques, which enable to: i) rapidly explore the whole dataset of images, e.g. to highlight the presence of outliers and/or of clusters of similar images; ii) create calibration models, allowing to predict the value of specific properties for each sample, such as the content of a particular type of pigment; iii) create classification models, allowing to assign a sample to a specific class, based on specific colour-related characteristics. The colourgrams approach has already been applied to build calibration and classification models in several studies, concerning different food matrices and different issues related to food industry, leading to satisfactory results, also in comparison with the performance of panel tests or more traditional colorimetric measurements and chemical analyses (Antonelli et al., 2004; Lo Fiego et al., 2007; Foca et al., 2011).

In this work, all the 198 digital images were converted into the corresponding colourgrams. The 95 colourgrams deriving by the samples selected on the basis of the panel test results were used to build and validate the classification models, and the remainder 103 colourgrams were subsequently employed as a second test set to compare the performance of the results of the best classification model with the performance of the assessors.

2.5 Multivariate classification methods

The 95 colourgrams corresponding to the samples selected on the basis of the panel test results were organised in a matrix with size {95, 4900}, that was randomly split in a training set (TRN) with size {60, 4900} and in a test set (TST) with size {35, 4900}. After the elaboration of the panel test data only one sample was selected for class 3 (severe defects), therefore the classification models were built considering only 2 classes, i.e., in control samples (25 objects in TRN and 12 objects in TST) and defective samples (35 objects in TRN and 23 objects in TST), including the sample with severe defects in class 2 of TST. Classification models were then calculated both on the whole colourgram using PLS-DA and on the features selected by WPTER (Cocchi et al., 2003).

The performance of the classification models are reported in terms of Sensitivity (SENS), i.e., the percentage of objects of each class accepted by the class model, Specificity (SPEC), i.e., the percentage of objects of the other class rejected by the class model, and Classification Efficiency (EFF), i.e., the geometric mean of SENS and SPEC (Forina et al., 2009).

2.5.1 PLS-DA classification

PLS-DA is the application of PLS2 to classification issues (Wise et al., 2007). While in PLS regression the response block consists in the response variables, in PLS-DA the Y block is a matrix composed by as many columns as the number of the existing classes, where each column is a binary class vector, with ones for the objects belonging to the class and zeros otherwise. In addition, PLS-DA calculates a threshold value based on Bayesian statistics (Pérez et al., 2009), so that a sample is assigned to a class if the corresponding Y predicted value is higher than the threshold value.

PLS-DA classification models were calculated using raw (i.e., not pretreated) data, as well as on meancentered and autoscaled data. The number of Latent Variables (LVs) was chosen by minimizing the classification error estimated in cross-validation (random subsets, 6 groups, 20 iterations).

2.5.2 WPTER classification

The WPTER algorithm decomposes the signals (i.e., colourgrams) dataset into the Wavelet Packet Transform (WPT) domain, in order to find a limited number of variables (wavelet coefficients) leading to an effective separation among the samples corresponding to different classes. In fact, generally the information contained in the whole colourgram is partially redundant; a certain degree of correlation among variables could be present and other uninformative sources of variation could overwhelm the information brought by the features of the signal useful for classification. Hence, the use of a feature selection algorithm able to take also into account the signal shape could be helpful. The Wavelet Transform (WT) (Walczak, 2000; Cocchi et al., 2003). allows to represent each analysed signal in an alternative domain, where the different frequencies are separated, but maintaining at the same time the localisation in the original domain. In this manner, in addition to the single intensity values, other useful aspects like peak widths, slopes of selected portions of the signal or discontinuities are also taken into account.

In particular, WPT consists in decomposing a signal by applying iteratively a couple of wavelet filters, i.e. a low-pass filter and a high-pass filter: the first one preserves the low-frequency content of the signal into the approximation vector (A), while the second one preserves the high-frequency content into the detail vector (D). This decomposition scheme can be then applied again both to the approximation and to the detail vectors and repeated for j decomposition levels, outlining a binary tree structure. For example, at the second decomposition level (j = 2), A is further decomposed in an approximation (AA) and in a detail vector (AD), and similarly D is decomposed in DA and DD. In this manner, at each decomposition level j the signal is split into 2^j vectors (often called blocks).

The perfect reconstruction of the original signal can be achieved by properly joining each one of the many possible combinations of blocks in a way to cover horizontally the whole binary tree, without vertical overlaps. Each one of these combinations of orthonormal blocks is a complete basis. Alternatively, one can decide to select only some blocks of the complete basis, in a way to discard frequency components related to noise or to other uninformative variation (e.g., background effects). Moreover, the selection of single variables (wavelet coefficients) can be performed also within each single blocks, in a way to select single portions of the signal, and at different frequencies.

Schematically, the WPTER algorithm works as follows (for a more detailed description refer to Cocchi et al., 2001, and Ulrici et al., 2008):

- the training set of signals is decomposed in the WPT domain using a couple of wavelet filters up to a given maximum decomposition level, chosen by the user. The obtained matrix can be considered as a redundant representation of the original signal matrix, since different bases can be used to represent it;
- the best basis is selected as the one leading to the best discrimination among the signals belonging to different classes. In a preselection phase, only a fixed percentage of the wavelet coefficients (defined by the user) is retained for each block. In particular, those coefficients showing the higher discriminant capability – as evaluated by the between-class/within-class variance ratio – are retained. Then, the best basis selection is performed by using the classification ability (CA) criterion, an Euclidean distance based method which reaches optimal values when the best separation among the signals belonging to different classes and, simultaneously, the best clustering of the signals belonging to the same class is obtained. The best discriminant basis is identified as the one containing the blocks giving the optimal CA values;
- the selected wavelet coefficients of the best discriminant basis are reconstructed back into the original domain; these reconstructed signals can be viewed as the projection of the selected wavelet coefficients in the original (colourgram) domain. For comparison purposes, the mean original signal of each class is also plotted highlighting the regions corresponding to the selected features;
- the selected wavelet coefficients can be used as input variables for the calculation of discriminant models. In particular, in the present work the same method adopted to classify the whole colourgrams was used, i.e., PLS-DA with random subsets cross-validation (6

groups, 20 iterations) to define the number of LVs, calculated using raw, meancentered and autoscaled data;

- the classification model is applied to the set of test signals for validation. The test set signals are decomposed into the WPT domain, and the wavelet coefficients previously selected during the training phase are used to validate the PLS-DA classification models.

Different parameters can be varied in order to optimize the classification model, such as the wavelet filters and the percentage of wavelet coefficients to be retained in the preselection phase. In the present work, 9 different wavelet filters (db1, db2, coif1, coif5, sym4-sym8) and 5 percentages of preselected wavelet coefficients (0.1%, 0.5%, 1%, 5%, 10%) were used, setting the maximum decomposition level equal to 5. The combination of all these parameters led to a total of 45 cycles of calculation.

For the identification of the best cycle, we considered the cross-validated classification efficiency values (CV EFF) of the PLS-DA models, selecting the cycle that led to the maximum value. The possible presence of multiple optimal solutions, i.e. of models calculated on WPTER cycles showing statistically equivalent CV EFF values, was also evaluated by means of one-way ANOVA and Tukey's multiple comparison test.

2.6 Comparison between assessors and best classification model

The performance of the optimal WPTER model was also compared with the performance of the assessors, expressing the results in terms of Validity and considering all the 198 samples. To this aim, since the 60 images of the training set were used for the selection of the useful features and to build the classification model, for a fair comparison of the results we performed three separate comparisons, i.e. one for TRN (considering the results obtained in cross-validation), one for TST, which was considered separately since it contains the 35 samples whose class assignation is sufficiently univocal, and one for TST2, which includes the remainder (198-60-35 =) 103 samples.

As for the evaluations performed by the assessors, in view of the fact that the classification models were built considering only 2 classes, i.e., in control samples and defective samples, the samples assigned by the assessors to class 3, i.e. to the class of samples with severe defects, were included in class 2, therefore considering only the discrimination between presence and absence of defects also for the assessors.

Then, the evaluations of the assessors were compared with the classes predicted by the optimal WPTER model, calculating the Validity scores in the same manner as described in Section 2.3, and considering the three datasets TRN, TST and TST2 separately each other.

2.7 Image reconstruction using selected features

Notwithstanding the transformation of an image into a colourgram implies the loss of spatial information, it is still possible to represent the selected features back into the original image domain, in a way to visually evaluate the correctness of the choices made by the algorithm (Foca et al., 2011). To this aim, one randomly selected image for each class was reconstructed using only the features selected by WPTER. The procedure adopted for image reconstruction with the only selected features consisted in the following steps:

- for each colourgram portion selected by WPTER, the corresponding range of pixel values is considered. For example, if one of the selected portions corresponds to a part of the frequency distribution curve of the blue values, the corresponding range (e.g., blue values from 100 to 200) is kept;
- the sample image to be represented is then segmented according to the range of the selected values. For example, only those pixels whose values in the blue channel range from 100 to 200 are kept, while the remaining ones are set equal to 0 for all the R, G and B channels;
- for each colourgram selected region the corresponding segmented image is then displayed, allowing to localize the image areas that contain the colourgram selected features;
- alternatively, when the number of colourgram selected regions is ≤ 3, a more compact representation can be obtained by segmenting separately each one of the RGB channels accordingly with the values of each colourgram selected region. For instance, in the present work the optimal WPTER model selected three regions, corresponding to the Blue, Relative Red and Saturation parameter values; to represent in a unique image all the three regions, the red channel was segmented according to the colourgram range selected for Relative Red, the green channel according to the range selected for Saturation, and the blue channel according to the range selected for Blue.

3. Results and Discussion

3.1 Panel test

The Validity (V) and Reliability (R) values calculated for each assessor are reported in Table 1, together with the corresponding Global Performance values. The mean values of V and R (about 70%) highlight the partial subjectivity and a certain degree of unreliability of the human evaluation

of this kind of defect, even when this is performed by expert assessors, confirming that the classification of raw hams based on the presence/absence of the red skin defect is not a trivial task. It can be observed that V and R values are comparable, which means that the difference between replicated estimates made by a single assessor is similar to the difference between the evaluations made by different assessors. This observation was confirmed by a two-tailed *t*-test on paired data (P = 0.59), performed in order to compare the V values of each assessor with the corresponding R values.

In order to verify the possible presence of assessors whose performance is significantly different, in particular to evaluate whether the performance of assessor E is significantly better than the others, the uniformity of the panel was also tested. To this aim, the distribution of the Global Performance of each assessor was compared by a chi-squared test to the corresponding uniform distribution, and the results confirmed the uniformity of the performance of the six assessors (P = 0.30).

As for the percentage of assignations of each sample to the correct class, Figure 2 shows the distribution histograms of the %OASA values separately for the three classes. On the whole, 70 samples were assigned to class 1 (in control), 124 samples to class 2 (moderate defects), while only 4 samples were assigned to class 3 (severe defects). Only for 6 out of the 198 samples a perfect agreement among the 12 evaluations was reached.

The 95 images of samples presenting %OASA values greater than 70% were selected, subdivided into training (TRN) and test (TST) sets, and converted into colourgrams for the subsequent classification. The remainder 103 samples with %OASA values lower than or equal to 70% were collected into a second test set (TST2) and used to compare of the results of the best classification model with the performance of the assessors.

3.2 PLS-DA classification models

The results of the PLS-DA classification models calculated on the whole colourgrams considering the raw (not pretreated), the mean centered and the autoscaled colourgrams are reported in Table 2. The SENS and SPEC values are referred to class 1; since this is a discriminant model with only two classes, SENS(class 2) = SPEC(class 1) and SPEC(class 2) = SENS(class 1). All the three models show good classification efficiency values, both in cross-validation of the training set (TRN) and in prediction of the external test set (TST). The cross-validated classification efficiency values (CV EFF) show a slight increase in the raw-meancentered-autoscaled sequence of

pretreatments, while at the same time the model dimensionality progressively decreases following the sequence 3-2-1.

Figure 3 reports the values calculated (TRN) / predicted (TST) by the best PLS-DA model for class 1 (in control), where the class 1 objects (triangles) lying above the threshold value (horizontal dashed line) are correctly assigned to class 1, while the class 2 (defective) objects (asterisks) lying below the threshold line are correctly assigned to class 2, and where the vertical dotted line separates the TRN objects (1-60) from the TST ones (61-95). Only three objects of TRN (objects n. 11 and 22 of class 1 and object n. 41 of class 2) and one object of TST (object n. 92 of class 2) are misclassified. Three of the misclassified objects (n. 11, 22 and 92) have the lowest possible %OASA value (75 %, i.e. 9 out of 12 attributions to the correct class), which could partly justify this result; on the other hand, this does not hold for object n. 41, whose %OASA value is equal to 92% (11 correct attributions out of 12). Concerning the only selected image of a sample with severe defects included in the test set (object n. 95, with a %OASA equal to 100 % for class 3), this is correctly attributed to class 2, but with a value (-150.3) which is completely out of the range of class 2 objects. On the one hand, this could reflect the presence of strong red skin defects for this sample, but on the other hand its outlying values both of the predicted Y and of the PLS residuals suggest that the correct estimate of samples with severe defects using PLS-DA models on colourgrams needs to be confirmed by further studies, using a representative number of samples with severe defects.

3.3 WPTER classification models

The best classification model was obtained by PLS-DA on the raw (not pretreated) wavelet coefficients obtained using a sym7 wavelet filter and a percentage of wavelet coefficients retained in the preselection phase equal to 0.5 % (cycle number 27), which led to the selection of only three variables.

The performance of the best PLS-DA classification model calculated using the three selected variables is reported in Table 3, together with the performance of the other PLS-DA classification models that showed similar results as for the CV EFF value. In particular, Table 3 also reports the classification performance of:

- the two models showing the second best CV EFF value, calculated on raw coefficients of cycle number 28 (obtained with a sym7 wavelet and a preselection percentage equal to 1%) and of cycle number 43 (coif5, 1%);

- the model showing the third overall best CV EFF value, calculated on mean centered coefficients of cycle number 5 (db1, 10%);
- the model showing the best CV EFF value using autoscaling as pretreatment, calculated on coefficients of cycle number 26 (sym7, 0.1%).

In order to check the possible presence of multiple optimal solutions, i.e. of models with statistically equivalent CV EFF values, we compared the performances of all the models reported in Table 3. In particular, for each one of them, the random CV procedure was additionally repeated 10 times. Then, the 11 CV EFF results obtained for each model were transformed into the corresponding arcsin(sqrt(CV EFF)) values (Zar, 1996), which in turn were compared using one-way ANOVA followed by Tukey's multiple comparison test ($P < 0.05$). The results of ANOVA demonstrated that on the whole the analyzed models show absolutely different performances ($P = 2 \times 10^{-36}$), and the Tukey test (*see* superscript letters in the first column of Table 3) highlighted that the best WPTER model (27) leads to results that are statistically equivalent to the results of model 43, which in turn are statistically equivalent to the results of model 28. Conversely, models 5 and 26 differ significantly from each other and from the best WPTER model.

A comparison of the best PLS-DA model of Table 2 (on autoscaled colourgrams) with the best PLS-DA model of Table 3 (on raw selected variables) shows that the selection of only three variables allows to obtain optimal classification performances in cross-validation of TRN (Classification Efficiency = 100%), maintaining the same predictive performances on TST. Only one object of TST is misclassified, which is the same one of the best PLS-DA model calculated on the whole colourgram (object n. 92 of class 2, %OASA = 75). Object n. 95, corresponding to the raw ham sample with severe defects, is correctly attributed to the defective samples, and in this case its Y prediction and PLS residual values do not show an outlying behaviour. This is reasonably due to the fact that WPTER tends to select variables leading at the same time to a good clustering between objects belonging to the same class and to an efficient separation between the clusters of objects corresponding to different classes. Figure 4, which represents the samples of both TRN and TST in the space of the three selected variables, confirms that WPTER reached this goal. The groups of objects corresponding to the two classes are in fact separated each other and each group is well clustered; in particular, as one could expect, the class of defective samples is more sparse than the class of in control samples.

3.4 Best WPTER classification model vs. Assessors

Table 4 reports the Validity scores calculated on the basis of the evaluations of each assessor (from A to F) and of the predictions obtained using the best PLS-DA classification model reported in Table 3 (best WPTER). The Validity scores were calculated separately for TRN (using the results in cross-validation), TST and TST2, for the reasons described in Section 2.6. For each dataset, the best Validity value is highlighted in gray colour. The obtained results indicate that the predictions made by the classification model are similar to those of the assessors showing the best performances.

In order to evaluate whether the performances of the classification model and of the assessors differ significantly from each other, the distributions of the Validity scores obtained for each dataset were compared by a chi-squared test to the corresponding uniform distributions. The probability values calculated for each dataset, reported in the last row of Table 4, confirm the uniformity of the performance of the classification model with the performance of the six assessors. The low $P(\chi^2)$ value obtained for TST2 is due to the low performance of assessors C and D. In fact, the results obtained by repeating the chi-squared test after the elimination of the values of assessors C and D changed drastically ($P = 0.98$).

3.5 Analysis and reconstruction of the selected features

Figure 5 reports in the upper part the mean colourgrams of the two classes (class 1 in red, class 2 in blue), where the regions selected by the best WPTER model are highlighted with vertical dotted lines, and in the lower part the zoom of the frequency distribution curves where three selected regions are located, each one corresponding to one of the three selected wavelet coefficients. The three narrow regions belong to the frequency distribution curves of the Blue (B), relative Red (rR) and Saturation (S). A comparison of the mean frequency distribution curves of the two classes for the three selected features highlights that the images of defective samples have a higher number of pixels with lower B intensity values and a higher number of pixels with higher rR and S intensity values. Since the human evaluation of the red skin defect is somehow related to the visual perception of an excessive and patchy red hue of the swine rind, it is not surprising that the algorithm selected two regions of the colourgram, i.e. the rR and S ones, bringing information related to the Red channel and to Saturation, while the selection of the lower intensity values of the Blue channel is less straightforward.

Interestingly, the same three regions were also selected by the second best WPTER models (cycles 43 and 28, Table 3), which in turn differ for the additional selection of regions

corresponding to the frequency distribution curves of green and of hue (cycle 43), and of green and of PCA loadings (cycle 28). This is a further confirmation of the reliability of the feature selection obtained with the best WPTER cycle, which is also more parsimonious than cycles 43 and 28.

In order to obtain more direct information about the choices made by the algorithm, one sample was randomly selected for each class, and the corresponding images were used for the reconstruction and visualisation of the features selected by the best WPTER cycle. The positions in the selected features space of the objects corresponding to the two randomly selected images are highlighted in Figure 4, using a red filled circle and a blue filled square for the object of class 1 and the object of class 2, respectively.

In Figure 6 the two original RGB images are reported together with the corresponding images segmented according with each selected feature taken separately. First of all, it is clear that the selected pixels correspond almost exclusively to the sample (raw ham) and no parts of the background are taken, with the only exception of few pixels of the B channel. This observation suggests that the classification model selected by WPER was not based on chance, since only the informative part of the image was considered. As a further confirmation of the fact that the selected variables correspond to actually informative features for the detection of the red skin defect, from the comparison of the segmented images with the corresponding original ones it can be noticed that the selected areas are in effect those where the red skin defect is more evident. Also the areas selected for the sample of class 1 (in control) correspond to image portions where the skin is more reddish, but in this case the size of the red skin area is much smaller than for the defective sample, as it can be also noticed when comparing the two samples by looking at each channel separately. Comparing the size of the areas selected for the three channels, for both the original images it grows following the rR<B<S order, where the rR selected features highlight the parts of the image with more evident red skin defects, while B and in particular S account for much wider areas, probably also in order to take into account the overall sample colour. Finally, a high degree of superimposition of the areas selected for the three cannels can be noticed for both the samples. This fact can be better appreciated by looking at the more compact representation given in Figure 7, where a unique image of the features selected for each sample is given in the right part (images *a2* and *b2*). In these images, the red channel corresponds to the features selected for rR, the green channel to the features selected for S, and the blue channel to the features selected for B, as it has been described in Section 2.7. From the same Figure 7, a comparison between images *b1* and *b2* highlights that the central portion of the dark red area in the left-down part of image *b1*, corresponding to a region where the red skin defect has a very high intensity, is ignored in the reconstructed image *b2*. Probably, in consideration of the overall nature of the defect in the training

set images of class 2, the presence of a widely diffused red skin defect was considered more significant than the presence of narrow areas where the defect is more intense.

In the future perspective of an on-line application of an automated red skin defect detection system like the present one, image outputs like those given in Figure 7 could be very useful to help operators in deciding whether rejecting or not a suspect sample. On the one hand, by comparing the reconstructed image with the original one, an operator could evaluate the correctness of the automated classification system (e.g., a sample could be misclassified due to the presence of an extraneous object, like a bloodstain, within the image scene); on the other hand, the features highlighted in the reconstructed image could help the operator itself in making a decision based on objective and reproducible elements, minimizing the risk of subjectivity and / or unreliability.

Finally, the time required for the classification of a new image and for the visualization of the corresponding reconstructed image was also estimated, using a personal computer running with Microsoft Windows 7 - 64 bit ® and equipped with an Intel Core ® i7-2600 CPU @ 3.40 GHz processor and 4.00 GB RAM. In particular, a Matlab function was implemented which:

- reads from the hard disk the image file of the sample to evaluate;
- converts the image into the corresponding colourgram;
- decomposes the colourgram into the wavelet domain and extracts the selected wavelet coefficients;
- uses the selected wavelet coefficients to calculate the predicted probabilities for each class with the best PLS-DA model of Table 3.b;
- segments the RGB image according to the selected intervals for B, rR and S;
- plots the original image and the reconstructed one, showing an output similar to the one reported in Figure 7, where the predicted probabilities for both classes are also reported.

This Matlab function was then tested on all the images of the dataset, and the average time needed to perform all the operations for each single image (with size 2592 × 1944) resulted equal to about 4.07 seconds, which is lower than the time spent by an expert assessor to perform an evaluation (not less than 5-10 seconds for the most skilled assessors). Of course, in view of possible future applications, the time needed to process each image could be drastically shortened by compiling this Matlab function into an executable file. Moreover, it must be stressed out that we have not focused our attention on the optimization of the data dimensions of the image files, since this was beyond the aims of the present work. However, a reduction of the image size made consistently with the possibility to maintain good classification performances, could lead to a

further significant reduction of the computational times. As an example, a test performed applying the same function to the images after reducing their spatial dimensions to the 29 % of the original size (752 × 564 pixels) led to an average time equal to 0.37 seconds per image. These results suggest that this approach is already suitable for on-line applications and that, after a proper optimization of the image dimensions and compilation of the Matlab functions into a standalone application, a totally-automated classification could lead to a drastic decrease in the time needed for the evaluation.

4. Conclusions

In the present paper, a flexible approach for the development of automated classification models for the detection of product defects is presented. This blind analysis approach is based on the automated extraction from RGB images of the colour-related properties, which are the most effective descriptors for the identification of the defect, and allows a fast visualization of the sample areas where the defect is present.

In particular we reported the application of this method to the detection of the red skin defect of raw hams, which led to the construction of efficient classification models, and to the possibility of visualizing the presence of the defect by means of images reconstructed using only the pixels corresponding to the selected features.

The ability of the proposed method in detecting the red skin defect was compared with the skill of a panel of trained assessors. The obtained results indicate that the predictions made by the classification model are equivalent to those made by the panel, and that the proposed method behaves as the best assessors.

One-dimensional signals, named colourgrams, were used in place of the 3D data of the original RGB images to build the classification models. Colourgrams can be considered as an inexpensive way to obtain useful information about colour-related properties of inhomogeneous samples. The main advantages of the colourgrams approach can be summarised as follows: i) the approach can be applied to any kind of sample for any kind of analysis that involves some colour changes, since it is not based on a priori assumptions on the nature of the analysed sample; ii) the time needed for the acquisition of an RGB image of an unknown sample and for its subsequent processing is short; iii) the method is cheap, since the instrumentation needed for the image acquisition and elaboration can be easily purchased at quite low cost.

Moreover, the present work demonstrated that the use of a proper feature selection algorithm can be of great help to find those features that are more effective to solve a specific task. The selection

of a restricted number of features, in fact, allows both to shorten considerably the time needed for computation, and to critically evaluate the choices made by the algorithm, through the easy use of the images reconstructed with the only selected features.

Acknowledgements

The authors wish to thank the anonymous assessors that contributed with their experience to the execution of the visual evaluation.

The research was supported by Regione Emilia Romagna (R.L. 28/98) and Fondazione Cassa di Risparmio Pietro Manodori (Reggio Emilia) .

References

Antonelli, A., Cocchi, M., Fava, P., Foca, G., Franchini, G.C., Manzini, D. & Ulrici, A. (2004). Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm. *Analytica Chimica Acta, 515***,** 3-13.

Barni, M., Cappellini, V. & Mecocci, A. (1997). Colour-based detection of defects on chicken meat. *Image and Vision Computing, 15*, 549-556.

Carnier, P., Gallo, L., Romani, C., Sturaro, E., & Bondesan, V. (2004). Computer image analysis for measuring lean and fatty areas in cross-sectioned dry-cured hams. *Journal of Animal Science, 82, 808–815.*

Cocchi, M., Seeber, R. & Ulrici, A. (2001). WPTER: wavelet packet transform for efficient pattern recognition of signals. *Chemometrics and Intelligent Laboratory Systems, 57*, 97-119.

Faucitano, L., Huff, P., Teuscher, F., Gariepy, C., & Wegner, J. (2005). Application of computer image analysis to measure pork marbling characteristics. Meat Science, 69, 537–543.Cocchi, M., Seeber, R. & Ulrici, A. (2003). Multivariate calibration of analytical signals by WILMA (Wavelet Interface to Linear Modelling Analysis). *Journal of Chemometrics, 17*, 512-517.

Foca, G., Ulrici, A., Corbellini, M., Pagani, M.A., Lucisano, M., Franchini, G.C. & Tassi, L. (2007). Reproducibility of the Italian ISQ method for quality classification of bread wheats: An evaluation by expert assessors. *Journal of the Science of Food and Agriculture, 87(5)*, 839-846.

Foca, G., Masino, F., Antonelli, A. & Ulrici, A. (2011). Prediction of compositional and sensory characteristics using RGB digital images and multivariate calibration techniques, *Analytica Chimica Acta, 706***,** 238-245.

Forina, M., Oliveri, P., Jäger, H., Römish, U. & Smeyers-Verbeke, J. (2009). Class modeling techniques in the control of the geographical origin of wines. *Chemometrics and Intelligent Laboratory Systems, 99*, 127-137.

Garrido-Novell, C., Pérez-Marin, D., Amigo, J.M., Fernández-Novales, J., Guerrero, J.E., Garrido-Varo, A. (2012). Grading and color evolution of apples using RGB and hyperspectral imaging vision cameras. *Journal of Food Engineering, Article in Press*.

Geladi, P. & Grahn, H. (1996). *Multivariate Image Analysis*. John Wiley & Sons, Chirchester.

Iqbal, A., Valous, N.A., Mendoza, F., Sun, D.W. & Allen, P. (2010). Classification of pre-sliced pork and Turkey ham qualities based on image colour and textural features and their relationships with consumer responses. *Meat Science, 84*, 455–465.

Lo Fiego, D.P., Bertolini, D., Comellini, M., Ielo, M.C. & Righetti, R. (2006). Caratterizzazione del difetto della "cotenna rossa" del prosciutto di Parma. *Atti Soc. Ital. Sci. Vet. 60, 493-494.*

Lo Fiego, D.P., Comellini, M., Ielo, M.C., Tassone, F. & Volpelli, L.A. (2009). Effect of stunning voltage and scalding method on the incidence of the "red skin" defect of Parma ham. *Veterinary Research Communications, 33 (Suppl 1)*, S285–S288.

Lo Fiego, D.P., Comellini, M., Ielo, M.C., Ulrici, A., Volpelli, L.A., Tassone, F. & Nanni Costa, L. (2007). Preliminary investigation of the use of digital image analysis for raw ham evaluation. *Italian Journal of Animal Scienc*e, 6 (Suppl. 1), 693-695.

Marty-Mahe, P., Loisel, P. & Brossard, D. (2003). Color image segmentation to detect defects on fresh ham. *Proceedings of SPIE - The International Society for Optical Engineering, 5132*, 45-50.

Pérez, N.F., Ferré, J., Boqué, R. (2009). Calculation of the reliability of classification in discriminant partial least-squares binary classification. *Chemometrics and Intelligent Laboratory Systems, 95 (2)*, 122–128.

Panigrahi, S. & Gunasekaran, S. (2001). Computer Vision. In Gunasekaran, S. (Ed.), *Nondestructive Food Evaluation – Techniques to Analyse Properties and Quality* (pp. 39-98). Marcel Dekker, New York.

Prats-Montalbán, J.M., de Juan, A. & Ferrer, A. (2011). Multivariate image analysis: A review with applications. *Chemometrics and Intelligent Laboratory Systems, 107 (1)*, 1-23.

Sanchez, A.J., Albarracin, W., Grau, R., Ricolfe, C. & Barat, J.M. (2008). Control of ham salting by using image segmentation. *Food Control, 19*, 135–142.

Tan, F.J., Morgan, M.T., Ludast, L.I., Forrest, J.C. & Gerrard, D.E. (2000). Assessment of fresh pork color with color machine vision. *Journal of Animal Science, 78(12)*, 3078-3085.

Ulrici, A., Cocchi, M., Durante, C., Foca, G., Marchetti, A. & Tassi, L. (2008). Multivariate analysis of analytical signals to decipher relevant chemical information. In Tassi, L., Colombini, M.P. (Eds.), *New trends in analytical, environmental and cultural heritage chemistry* (Chpt. 5). Research Signpost, Trivandrum.

Valous, N.A., Mendoza, F. & Sun, D.W. (2010). Emerging noncontact imaging, spectroscopic and colorimetric technologies for quality evaluation and control of hams: a review. *Trends in Food Science and Technology, 21*, 26-43.

Walczak, B. (2000). *Wavelets in Chemistry*. (1st ed.). Amsterdam: Elsevier.

Wise, B.M., Gallagher, N.B., Bro, R., Shaver, J.M., Windig, W. & Scott Koch, R. (2007). *PLS Toolbox 4.2*, Eigenvector Research Inc., Wenatchee, WA.

Zar, J.H. (1996). *Biostatistical Analysis*. (3rd ed.). Upper Saddle River: Prentice Hall., (Chapter 13).

Zheng, C., Sun, D.W. & Zheng L. (2006). Recent developments and applications of image features for food quality evaluation and inspection – a review. *Trends in Food Science and Technology, 17*, 642-655.

CAPTION OF FIGURES

Figure 1. Graphical user interface used by the assessors for the visual evaluation of each image (left side), together with the reference images of the three quality classes (right side, with the respective class numbers specified in the upper right corner).

Figure 2. Frequency distribution histograms of % Overall Agreement on Samples Attributions (%OASA). The dashed square delimits the 95 samples retained to build the classification models.

Figure 3. Results of the best PLS-DA model for class 1, calculated on the whole colourgram. Class 1 samples are represented with triangles, class 2 samples with asterisks. The vertical dotted line separates the training set samples (on the left) from the test set ones (on the right). The threshold value is indicated with the horizontal dashed line.

Figure 4. 3D plot of the three wavelet coefficients selected in the best WPTER model. The red filled circle and a blue filled square correspond to the object of class 1 and of class 2, respectively, that were randomly selected for the image reconstruction of the selected features.

Figure 5. Regions selected by the best WPTER model highlighted on the mean colourgrams of class 1 (red) and class 2 (blue).

Figure 6. Original RGB images and areas segmented according with the B, rR and S channels, for the two randomly selected samples.

Figure 7. Original RGB images of in control (a_1) and of defective (b_1) samples, compared with the corresponding reconstructed images (a_2 and b_2, respectively).

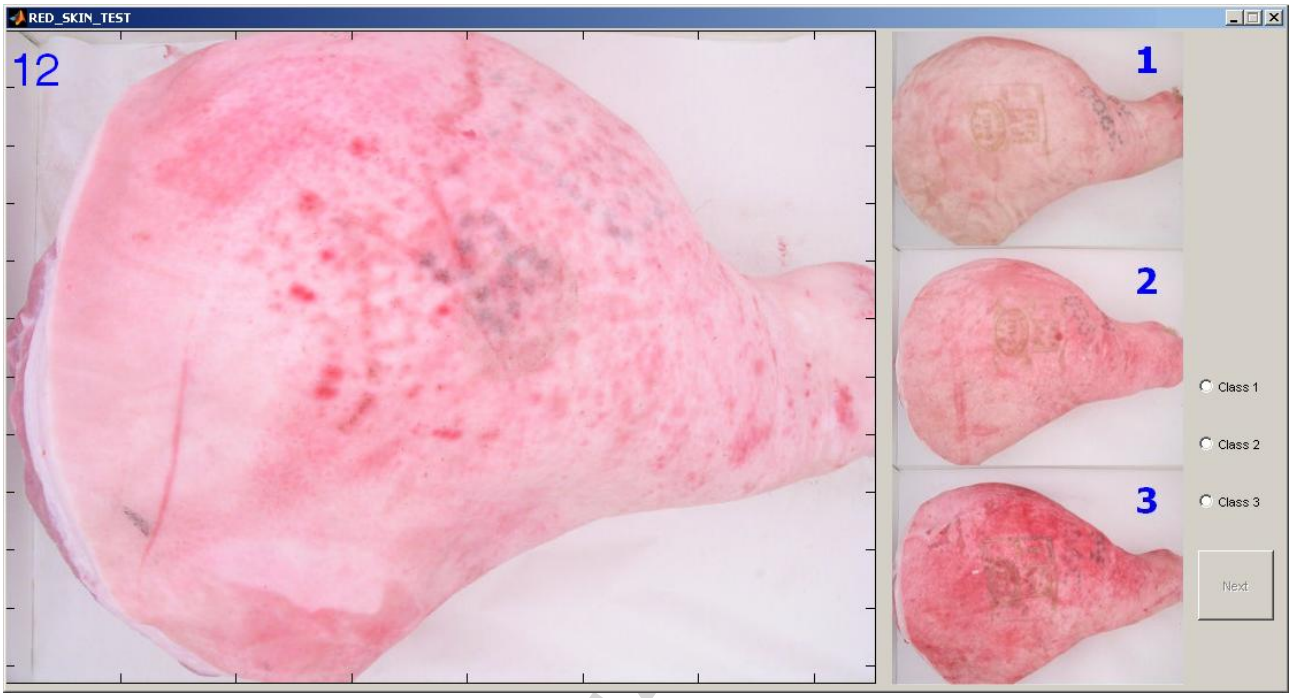

Fig. 1

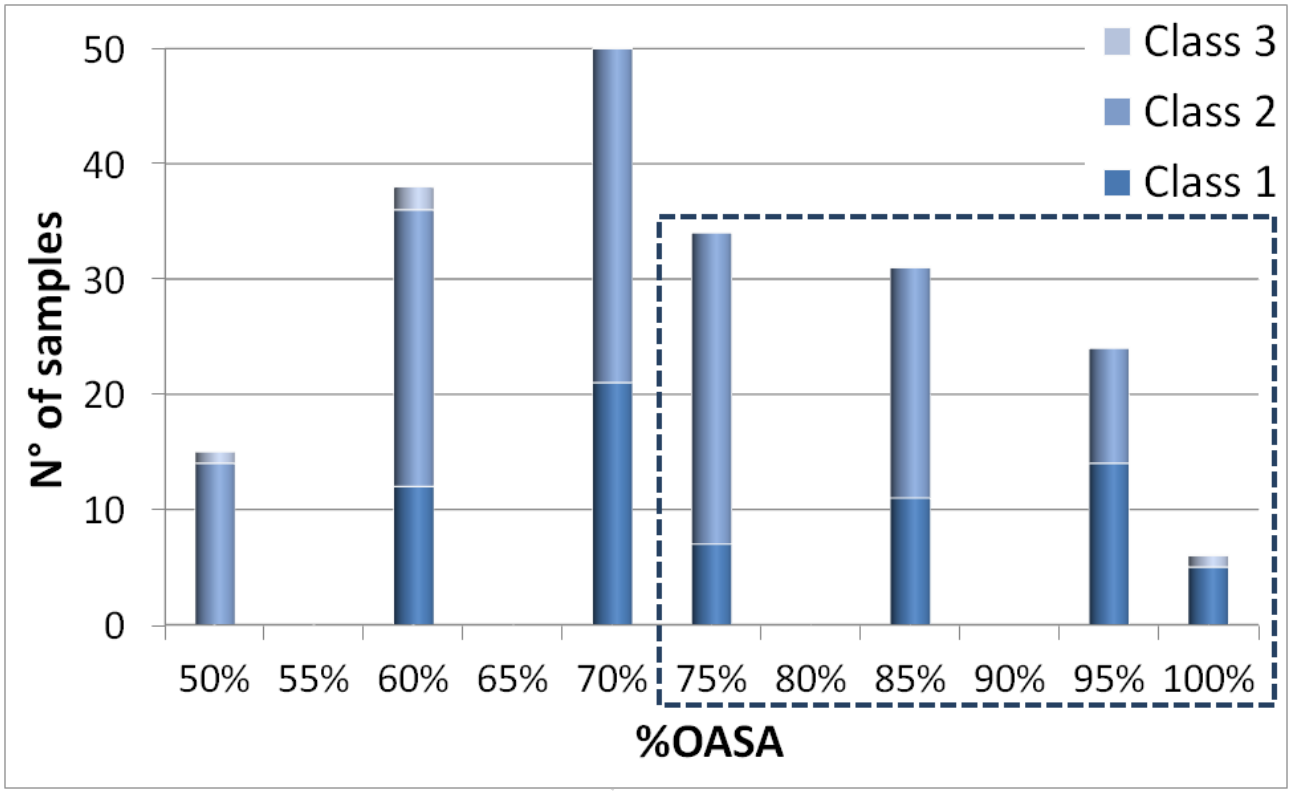

Fig. 2

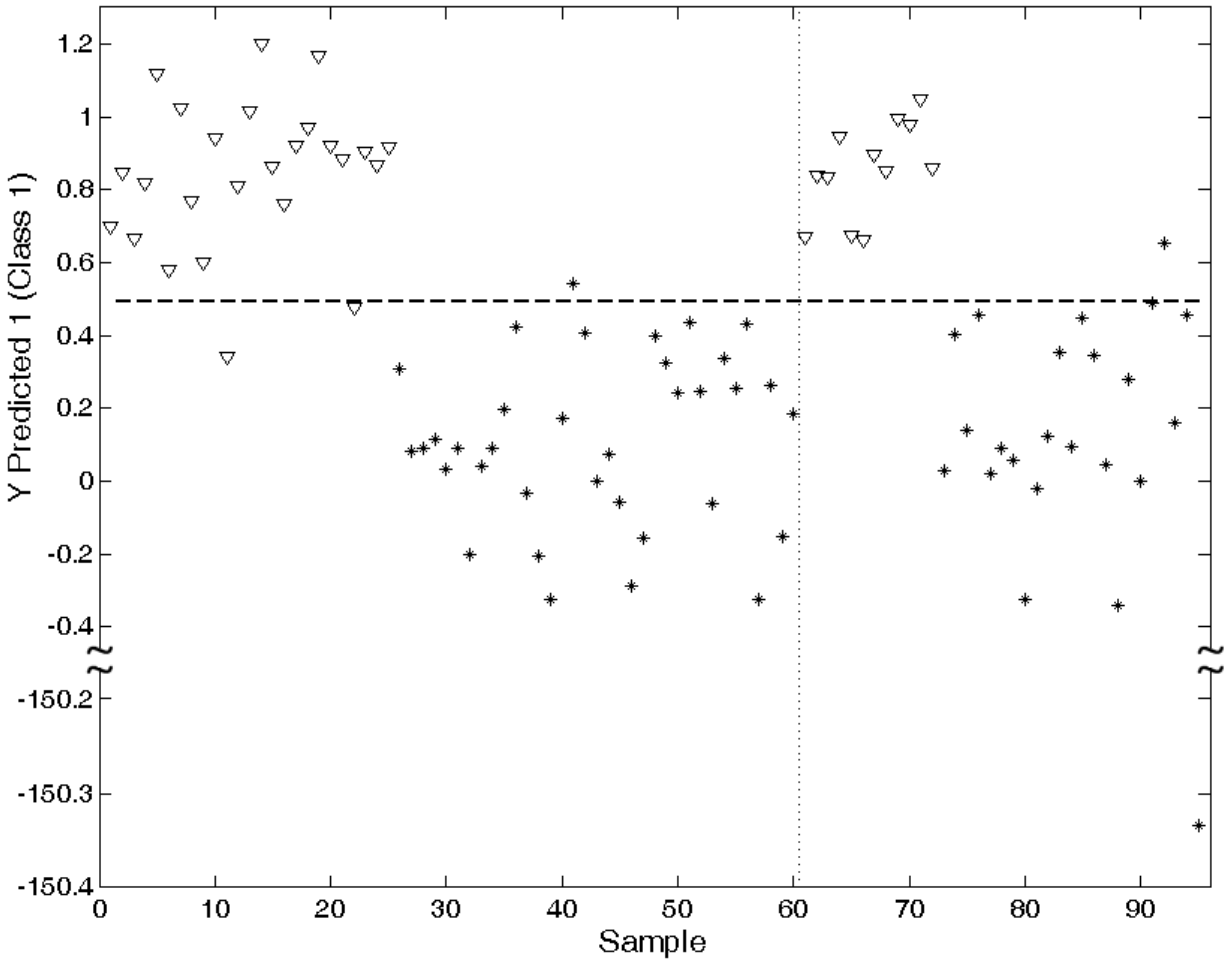

Fig. 3

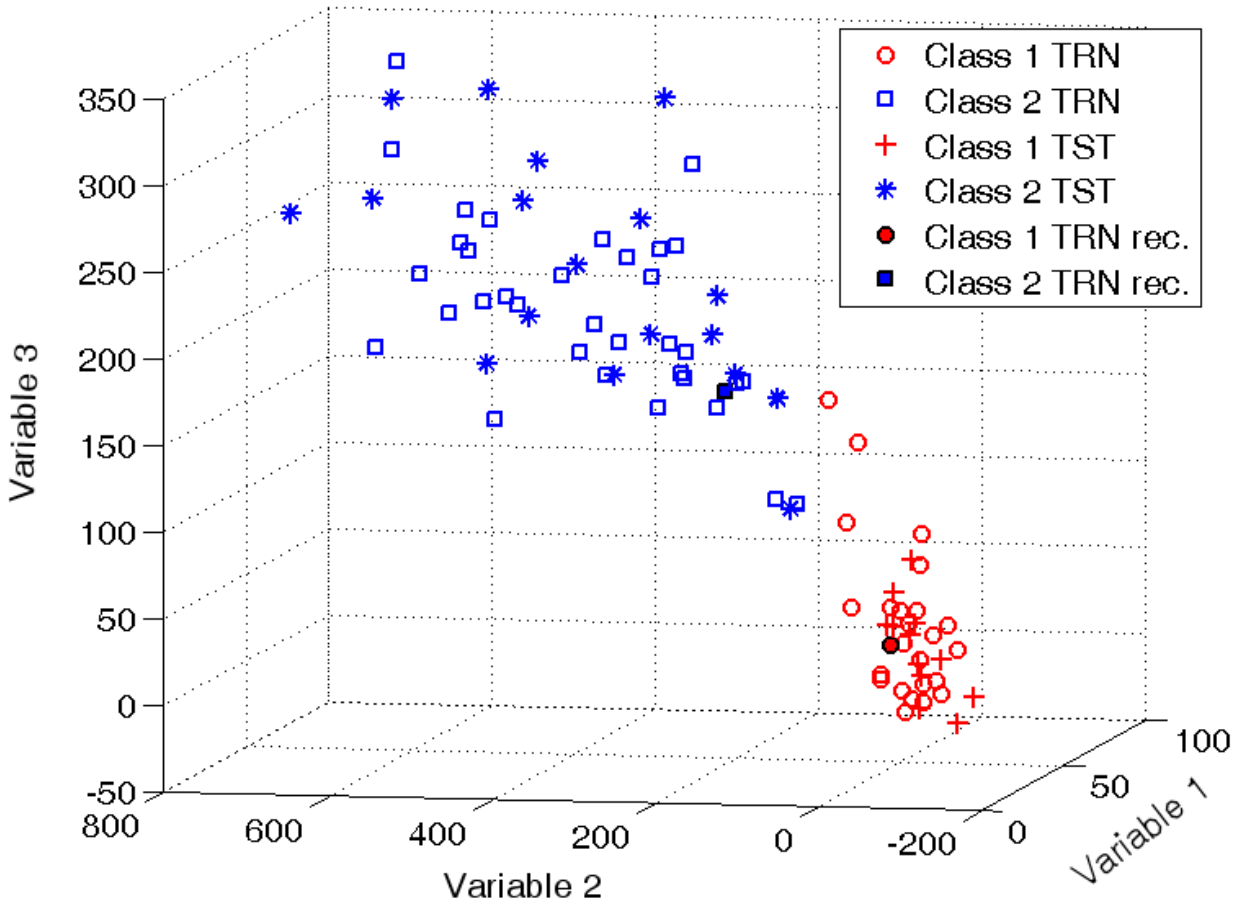

Fig. 4

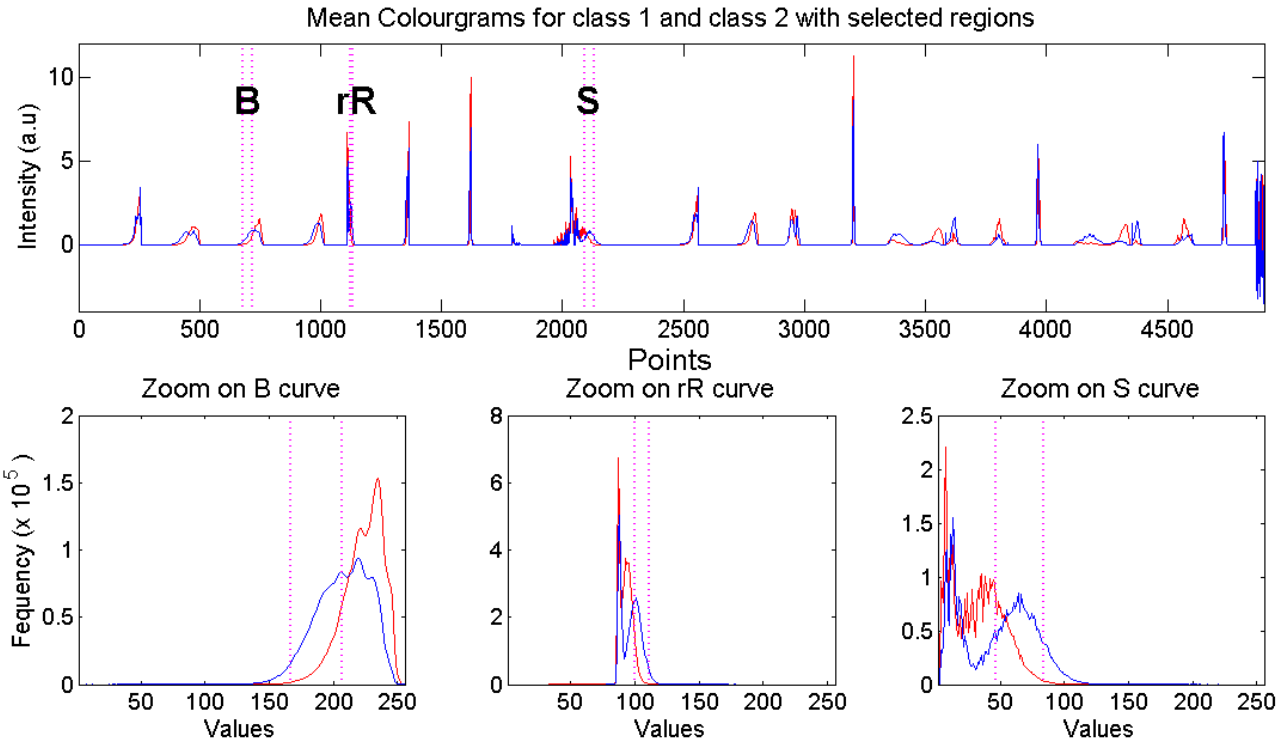

Fig. 5

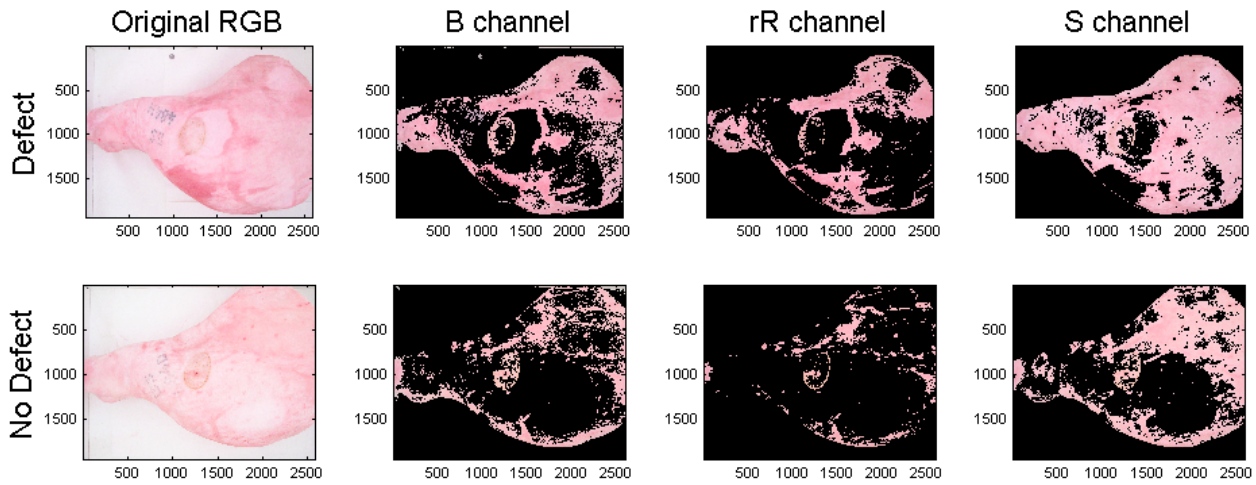

Fig. 6

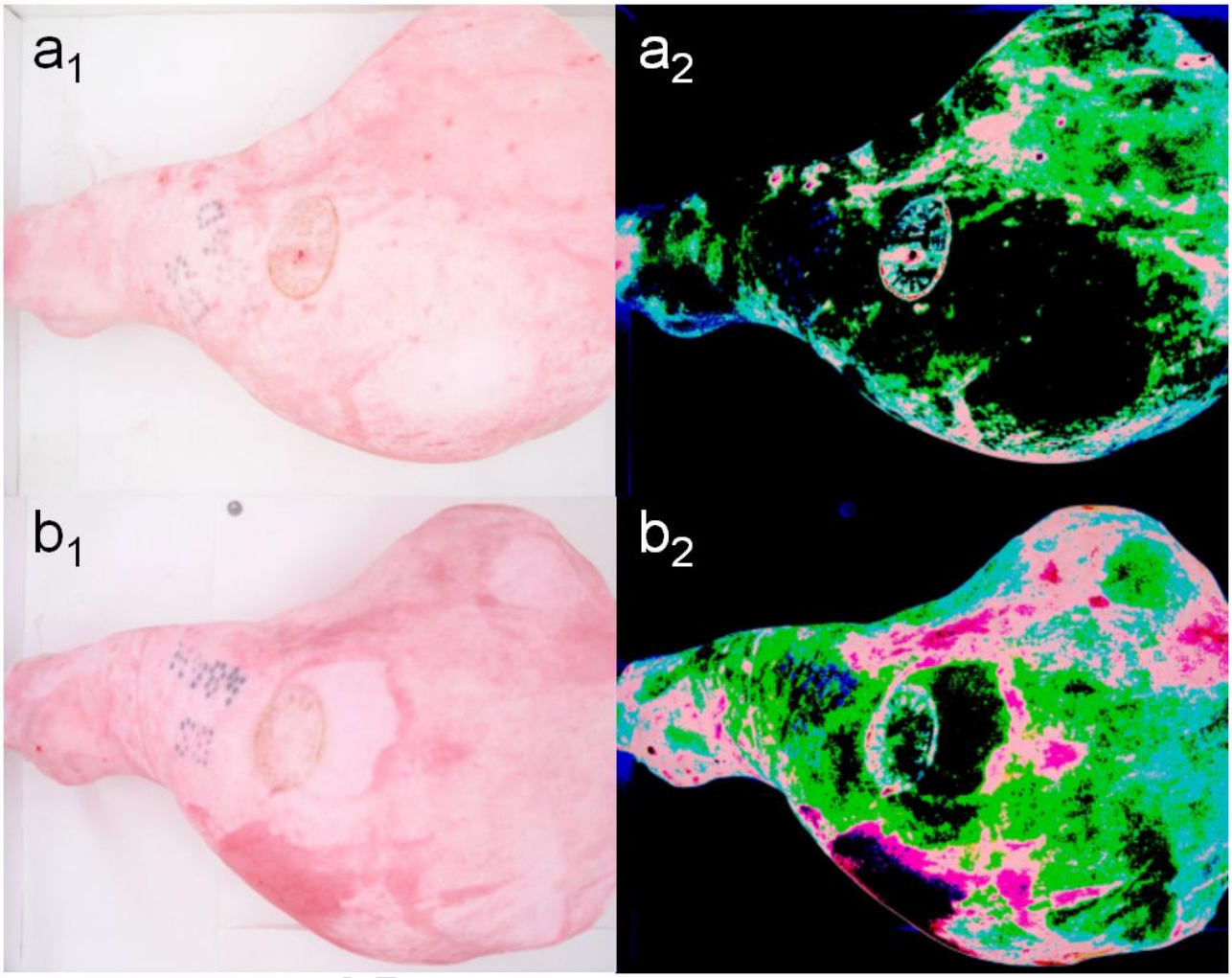

Fig. 7

Assessor	Validity	Reliability	Global Performance
A	71.7	64.6	136.3
B	75.3	62.1	137.4
C	65.9	72.2	138.1
D	64.4	75.3	139.7
E	85.9	81.8	167.7
F	68.2	62.6	130.8
mean	**71.9**	**69.8**	**141.7**
s.d.	**7.9**	**8.0**	**13.1**

Table 1. Performances of assessors involved in the visual evaluation of raw hams images.

Pretreatment	LVs	TRN (calculated)			TRN (CV)			TST		
		SENS	SPEC	EFF	SENS	SPEC	EFF	SENS	SPEC	EFF
None	3	0.920	1.000	0.959	0.914	0.926	0.920	1.000	0.913	0.956
Mean center	2	0.920	0.971	0.945	0.916	0.930	0.923	1.000	0.913	0.956
Autoscale	1	0.920	0.971	0.945	0.910	0.964	0.937	1.000	0.957	0.978

Table 2. Results of the PLS-DA classification models calculated on the whole colourgrams. The best classification model is highlighted in gray colour.

WPTER Cycle*	PLS-DA Pretreat	Vars	LVs	TRN (calculated)			TRN (CV)			TST		
				SENS	SPEC	EFF	SENS	SPEC	EFF	SENS	SPEC	EFF
27 [a]	None	3	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.957	0.978
43 [a,b]	None	6	2	1.000	1.000	1.000	1.000	0.997	0.999	1.000	0.957	0.978
28 [b]	None	6	2	1.000	1.000	1.000	1.000	0.997	0.999	1.000	0.957	0.978
5 [c]	Mean center	123	9	1.000	0.971	0.986	0.992	0.970	0.981	0.833	0.957	0.893
26 [d]	Autoscale	2	1	0.960	1.000	0.980	0.956	1.000	0.978	1.000	1.000	1.000

* Different superscript letters (a, b, ...) indicate significant differences of the CV EFF values at $P < 0.05$ by one-way ANOVA followed by Tukey test.

Table 3. WPTER cycles leading to the best PLS-DA classification models. The overall best classification model is highlighted in gray colour.

	TRN	TST	TST2
A	82.50	85.71	73.79
B	91.67	92.86	72.82
C	85.00	84.29	51.46
D	78.33	80.00	52.43
E	96.67	100.00	78.16
F	91.67	95.71	74.76
best WPTER	100.00	97.14	79.61
$P(\chi^2)$	0.67	0.71	0.06

Table 4. Validity values of the assessors (from A to F) and of the PLS-DA predictions made with the best WPTER model, estimated separately for the samples belonging to the training set (TRN), for the original test set of 35 samples (TST), and for the test set of the remainder 103 samples (TST2). The last row reports the probability values of the chi-squared test, made to test the uniformity of the Validity results.

Highlights

- Classification of raw ham images to detect the red skin defect
- Use of a blind-analysis method based on image fingerprints (colourgrams)
- Image classification made by expert assessors used as reference measurement
- Classification Efficiency of an external test set = 98% using only 3 variables
- Image-like visualization of the selected features for easy results interpretation