Discrimination of Wheat Grain Varieties Using X-Ray Images

Małgorzata Charytanowicz, Jerzy Niewczas, Piotr Kulczycki, Piotr A. Kowalski and Szymon Łukasik

Abstract A study was conducted so as to develop a methodology for wheat variety discrimination and identification by way of image analysis techniques. The main purpose of this work was to determine a crucial set of parameters with respect to wheat grain morphology which best differentiate wheat varieties. To achieve better performance, the study was done by means of multivariate discriminant analysis. This utilized both forward and backward stepwise procedures based on various sets of geometric features. These parameters were extracted from the digitized X-ray images of wheat kernels obtained for three wheat varieties: Canadian, Kama, and Rosa. In our study, we revealed that selected combinations of geometric features permitted discriminant analysis to achieve a recognition rate of 89–96 %. We then compared the correctness of classification with results obtained by way of employing the nonparametric approach. The discriminant analysis proved effective in differentiating wheat varieties.

Keywords Wheat grain · Morphological features · X-ray imaging · Image processing · Image analysis · Discrimination · Classification · Nonparametric density estimation

M. Charytanowicz (✉) · P. Kulczycki · P.A. Kowalski · S. Łukasik
Systems Research Institute, Centre of Information Technology
for Data Analysis Methods, Polish Academy of Sciences, Newelska 6, Warsaw, Poland
e-mail: mchmat@ibspan.waw.pl

M. Charytanowicz · J. Niewczas
Institute of Mathematics and Computer Science, The John Paul II Catholic
University of Lublin, Konstantynów 1H, Lublin, Poland

P. Kulczycki · P.A. Kowalski · S. Łukasik
Faculty of Physics and Applied Computer Science, Division for Information
Technology and Biometrics, AGH University of Science and Technology,
Mickiewicza 30, Cracow, Poland

E. Piętka et al. (eds.), *Information Technologies in Medicine*,
Advances in Intelligent Systems and Computing 471,
DOI 10.1007/978-3-319-39796-2_4

1 Introduction

Recent research in computer-aided data analysis based on image processing, has brought about the high development of accurate and automated systems for the discrimination and classification between wheat grain categories [3, 9, 11]. One of the more well-known and widely used imaging techniques is soft X-ray photography. This is an objective and precise method which provides high quality visualization of the internal kernel structure. Together with photo-scanning procedures, this technique provides sufficient resolution for reflecting distinct features important for the accurate characterization of kernels. Moreover, this is non-destructive and considerably cheaper than other more sophisticated techniques such as magnetic resonance imaging, scanning microscopy or laser technology. It has to be stressed, however, that sole visualization of the kernels, irrespective to the techniques used, does not provide quantitative evaluation of the overall quality of the kernel, e.g., measures of geometric parameters of internal grain features, their correlation and distribution within the structure, etc. In order to carry out an accurate grain quality assessment, specialized image processing procedures need to be employed for the measurement, detection and interpretation of kernels in X-ray images. Various grading systems using different morphological features for the classification of different cereal grains and varieties have been reported in literature [12, 15, 17]. As well as geometric features, the authors used grain surface texture and color to develop a statistical model that effectively employs these variables to classify wheat grain.

The main objective of this work is to investigate whether it is possible to recognize wheat varieties by way of multivariate discriminant analysis as applied to certain basic, critical geometric parameters of kernels that have been extracted from X-ray images. The accomplished study consists of two phases: the construction of a model based on the training set that has been created for cases with known belonging to different groups, and, subsequently, to extend the use of this model for classifying new cases. The classification results given by the classical approach were then compared with nonparametric classification results.

2 Mathematical Preliminaries

Classical methods of data analysis assume that the data are drawn from one of a known parametric family of distributions, determined by their parameters [2, 5]. The density underlying the data could then be estimated by finding from the data estimates of unknown parameters, using optimization criteria. Such attitude requires performing goodness-of-fit tests on the data, in which the null hypothesis states that our data follows a specific distribution. This rigidity can be overcome by nonparametric estimation methods that assume no pre-specified functional form for a density function. In this section, both approaches to discriminant analysis are shortly described.

2.1 Nonparametric Density Estimation

Suppose that there is the n-dimensional random variable X, with a distribution characterized by the density f. Its kernel estimator $\hat{f} : R^n \to [0, \infty)$, calculated using experimentally obtained values for the m-element random sample

$$x_1, x_2, \ldots, x_m, \tag{1}$$

in its basic form is defined as

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right), \tag{2}$$

where $m \in N\backslash\{0\}$, the coefficient $h > 0$ is called a smoothing parameter, while the measurable function $K : R^n \to [0, \infty)$ of unit integral $\int_{R^n} K(x)\mathrm{d}x = 1$, symmetrical with respect to zero and having a weak global maximum in this place, takes the name of a kernel.

The choice of the form of the kernel and the calculation of the smoothing parameter is made most often by way of established optimization criterions. The choice of the kernel has no practical meaning, and, due to this situation, it is possible to take into account primarily the properties of the obtained estimator. In practice, for the one-dimensional case, the function is assumed usually to be the density of a common probability distribution. Most often the standard normal kernel given by the formula

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{3}$$

is used. In the multidimensional case, two natural generalizations of the above concept are employed: radial and product kernels, but from an applicational point of view, the difference is insignificant. The fixing of the smoothing parameter has importance in establishing the quality of estimation. For the one-dimensional case, the effective plug-in method is especially recommended. For the multidimensional case, two natural attitudes are usually used: the plug-in method for product kernels and the cross-validation method for radial kernels.

Broader discussion of this task, as well as additional procedures improving the quality of the estimator obtained, such as modification of the smoothing parameter and linear transformation, as well as general aspects of the theory of statistical kernel estimators are found in [6, 7, 13].

2.2 Discriminant Analysis

Classificatory discriminant analysis is used to allocate new observations into one of the two or more known groups or clusters on the basis of the measured characteristics.

The basic problem of discrimination can be stated as follows. Consider K independent simple random samples containing $m_1, m_2, \ldots, m_K$ elements, respectively drawn from K different populations (classes):

$$\begin{array}{c} x_{11}, x_{12}, \ldots, x_{1m_1} \\ x_{21}, x_{22}, \ldots, x_{2m_2} \\ \ldots \\ x_{K1}, x_{K2}, \ldots, x_{Km_K} \end{array} \tag{4}$$

where $x_{kj} \in R^n$ constitutes the j-multivariate observation from k-th class. These samples form the training set. Given a new observation $z \in R^n$, the main task is to assign that observation to one of K classes.

Suppose, first of all, that observations are drawn from populations having probability density functions $f_1, f_2, \ldots, f_K$. The classical Bayesian approach would then allocate the tested element to the class for which the value

$$c_1 f_1(z), c_2 f_2(z), \ldots, c_K f_K(z) \tag{5}$$

is the largest [4, 8]. The constants $c_k, k = 1, 2, \ldots, K$, are chosen by reference to the probabilities of misclassification, or by considering the prior probability that z comes from the population k and other utilities of correct classification. In practice, the densities $f_k, k = 1, 2, \ldots, K$, cannot be assumed to be known, and so the discriminate rule must be estimated from the training set. One natural approach is to suppose that the unknown densities come from some parametric family and then to estimate the parameters. When the within-class covariance matrices are assumed to be equal, the parametric approach generates a linear discriminant function, otherwise it can be extended to a quadratic form.

When the distribution within each group is not assumed to have any specific distribution or is assumed to have a distribution different from the multivariate normal distribution, nonparametric methods can be used to estimate the densities $f_1, f_2, \ldots, f_K$ and derive classification criteria. A natural description of these distributions allow the specifying of the kernel estimators $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_K$, constructed under given random samples (4), and consequently, the choosing of the population for which the value $c_k \hat{f}_k(z)$, $k = 1, 2, \ldots, K$, is the greatest. The main concept of the kernel density methodology is described in Sect. 2.1.

3 Methodology

The proposed methodology for wheat variety discrimination and identification by way of image analysis techniques, is summarized as follows:

- selecting wheat grains,
- capturing the wheat grain photograms,

- producing bitmap graphics files for reflecting geometric features of grains,
- rescaling images to standardize the unit of measurement,
- measuring geometric features,
- selecting wheat kernel geometric parameters,
- performing the discriminant analysis and classification process.

3.1 Image Processing and Feature Extraction

The study was undertaken at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin, and conducted using the combine harvested wheat grain of three different varieties: Canadian, Kama, and Rosa, originating from experimental fields. Randomly selected grain samples of these varieties contained 108, 72, and 108 kernels, respectively. A high quality visualization of the internal kernel structure (Fig. 1) was then generated using a soft X-ray technique.

The images were obtained in the form of photograms of size 13 $\times$ 18 cm, at the scale of 5 : 1. For each X-ray exposure, 10–12 grain kernels were evenly positioned groove down. The photograms were then scanned by way of an Epson Perfection V700 table photo-scanner that was equipped with a transparency adapter, at 600 dpi resolution and 8 bit gray scale levels. This produced bitmap graphics files with a sufficient resolution for reflecting distinct features important for the accurate characterization of objects. Before taking the measurements of certain geometric features, all images were rescaled to standardize the unit of measurement. Figure 2 presents the X-ray images of these kernels.

Image processing allowed a determination of size measures of wheat grain, as well as a possibility to compute shape coefficients [14, 16]. The complete analysis procedure of the obtained bitmap graphics files was based on the computer software package `Grains`, specially developed for the X-ray diagnostic of wheat kernels [10]. The most important feature of the package `Grains` is that it provides functions for ascertaining the particular characteristics of any selected kernel. In our work,

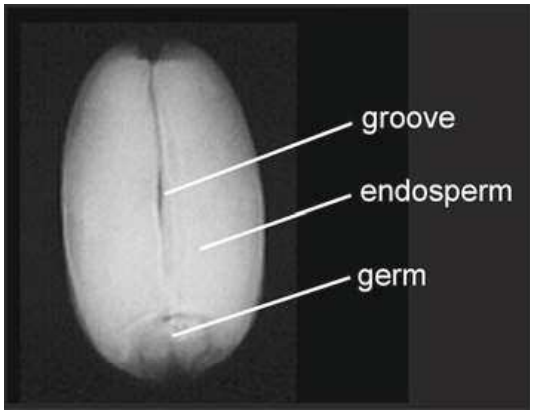

Fig. 1 X-ray image of an individual kernel

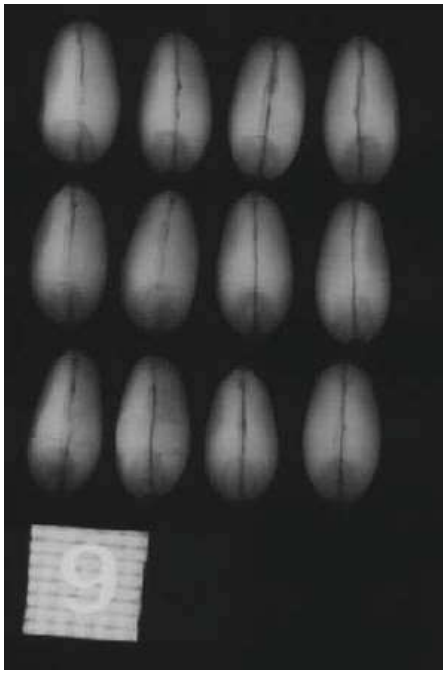

Fig. 2 X-ray photogram (13 × 18 cm) of kernels

using the commands available in the `Grains` menu analysis, automatic boundary detection, and, subsequently, calculation of main geometric and statistical parameters was accomplished for each individual grain kernel. Figure 3 provides an illustration of the contour indicating boundary of a kernel, as well as the obtainable measurements of the detected features.

To construct the data, the following descriptors were determined for each kernel:

- area A, understood as an area of a grain projection,
- perimeter P, understood as a perimeter of a grain projection,
- compactness,[1] given as $C = 4\pi A/P^2$,
- length of a kernel,
- width of a kernel,
- length of a kernel groove,
- asymmetry coefficient

where the asymmetry coefficient, given as a percentage, is a ratio between two quantities: the absolute value of the difference between areas of the left and right part of a kernel, and the total area of a kernel. Moreover, two parameters were taken by direct measurement using the package `Grains`. These being:

- area of a germ,
- length of a germ.

In addition to these numerical characteristics, the following parameters were obtained indirectly:

[1] The maximum value of the compactness is equal to one and is taken for a circle.

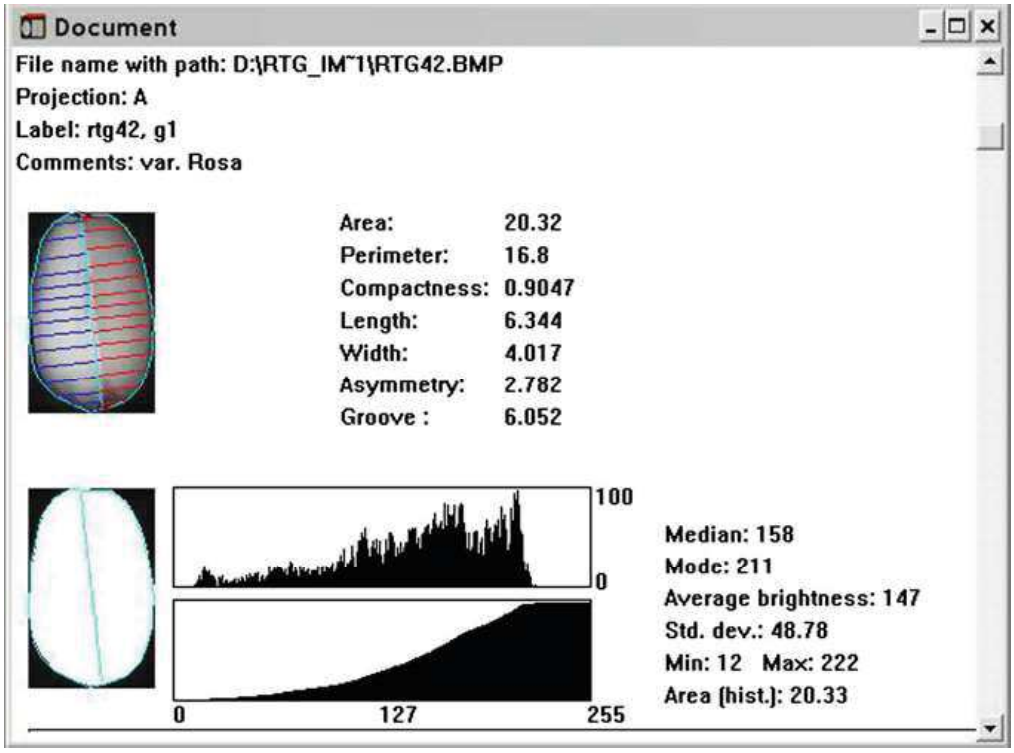

Fig. 3 Document window with geometric parameters of a kernel and statistical parameters of its image (as a unit of measure, millimeters were utilized)

- the ratio between the length of a germ and the length of a kernel,
- the ratio between the area of a germ and the area of a kernel,
- the ratio between the width of a kernel and the length of a kernel,
- the Feret coefficient.[2]

Computer image analysis of whole kernels allowed size and shape descriptors. Finally, the data contained the vectors of real-valued parameters determined for 288 samples belonging to three different varieties of wheat: Canadian, Kama, and Rosa. Randomly selected sets containing 100, 65, and 100 grains respectively, were then used to derive a discriminant procedure. The remaining observations of the test samples, as batches of 8, 7, and 8 grains, respectively, were considered as unknown categories, and, hence, comprised the test data set that was employed in the classification procedure. The level of procedure performance was evaluated by rate of correct classification.

[2]The ratio between the maximum diameter of the kernel in the vertical direction and the maximum diameter of the kernel in the horizontal direction, the measure of elongation.

3.2 Feature Selection

All measurements were made automatically from a total of 288 samples. Grains of the Canadian (the average length of kernels being equal to 5.250 ± 0.203, the average width of kernels being equal to 2.892 ± 0.201) and Kama (the average length of kernels being equal to 5.504 ± 0.230, the average width of kernels being equal to 3.236 ± 0.182) varieties are seen to be less in size than the Rosa (the average length of kernels being equal to 6.246 ± 0.390, the average width of kernels being equal to 3.749 ± 0.245).

In order to ascertain the usability of the determined geometric features in wheat grain discrimination, the distribution of the obtained histograms for each variable was analyzed, and, subsequently, correlation analysis was performed. A histogram distribution for all individual variables was symmetric and unimodal. Due to the significant relationships between some of these variables, these were reduced to the inclusion of the two best six-element sets that minimized correlations (Table 1).

Once the variables were selected, as an input, multidimensional analysis was performed for both data sets of feature vector components. This was done so as to extract the data set that better discriminates the investigated wheat varieties. The analysis was carried out through the separate use of forward and backward stepwise discriminant analysis. Firstly, the assumptions concerned with the discriminant analysis were assessed. Subsequently, multidimensional analysis was undertaken in order to discriminate the varieties. Finally, the classification process employing the discriminant functions, was performed. In so doing, no significant improvement was noticed when forward and backward stepwise discrimination was analyzed. Thus, the results for forward stepwise discriminant analysis will be shown alone.

Table 1 Feature vector components for discriminant analysis

Data set I	Data set II
1. Perimeter of a kernel	1. Area of a kernel
2. Compactness	2. Length of a kernel
3. Asymmetry coefficient	3. Asymmetry coefficient
4. The ratio between the length of a germ and the length of a kernel	4. Length of a germ
5. The ratio between the area of a germ and the area of a kernel	5. Area of a germ
6. The ratio between the width of a kernel and the length of a kernel	6. The Feret coefficient

3.3 Results and Discussion

To compute discriminant functions for the three-group classification, a training set containing 265 observations and two different feature vector components were used. Firstly, for both data sets of geometric features, discriminant analysis allowed the determination of which variables had significant contribution to the discrimination. In the data set I, the compactness was found to be insignificant and was removed from the model. In the data set II, all variables demonstrated significant contribution to the model.

Next, Wilks' Lambda criterion was used to assess feature discriminatory power. In the data set I, the kernel's perimeter, and subsequently, the ratio between the length of a germ and the length of a kernel, and the ratio between the area of a germ and the area of a kernel, were established as being most important in discrimination, while the asymmetry coefficient and the ratio between the width of a kernel and the length of a kernel were seen to be the least important. In the data set II, the length of a germ, and, subsequently, the Feret coefficient, the asymmetry coefficient, the area of a germ and the area of a kernel had the strongest discriminatory power, whereas the length of a kernel had the weakest power.

For both models, two discriminant functions were statistically significant. Thus, the first function discriminated mostly between Rosa, and Canadian and Kama combined. The level of discrimination was equal to 89 % for the data set I, and 88 % for the data set II. The second function was found to discriminate Kama variety, but the level of discrimination was lower and equaled 11 % and 12 %, respectively (Figs. 4 and 5).

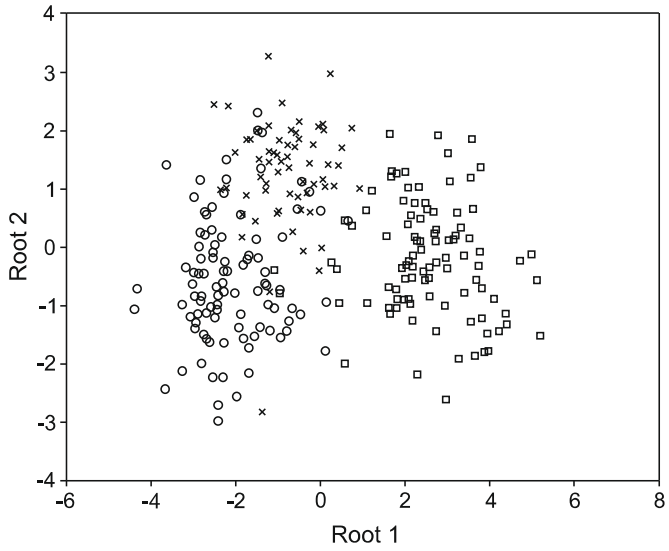

Fig. 4 Results of discrimination by scatterplots of canonical scores for data set I: (○) the Canadian wheat variety, (×) the Kama wheat variety, (□) the Rosa wheat variety

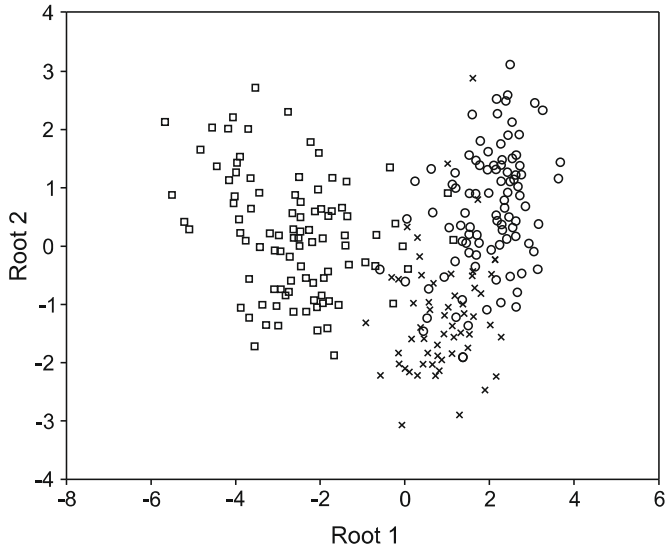

Fig. 5 Results of discrimination by scatterplots of canonical scores for data set II: (○) the Canadian wheat variety, (×) the Kama wheat variety, (□) the Rosa wheat variety

Table 2 The classification correctness for discriminant analysis

Geometric feature	Canadian (%)	Kama (%)	Rosa (%)	Total (%)
Data set I	92	89	96	92
Data set II	89	88	93	90

The classification correctnesses employing both sets of geometric features as an input for computing discriminant functions were summarized in Table 2. Use of the discriminant analysis for data set I resulted in higher proper classification, which, accordingly, equaled 92, 89, and 96 % for Canadian, Kama, and Rosa varieties. For the Rosa variety, 96 of 100 kernels were classified properly, and no kernels of the other two varieties were classified as being of the Rosa variety. For the Canadian variety, 92 of 100 kernels were correctly classified, while 2 kernels of the Rosa variety and 7 kernels of the Kama variety were mistakenly classified, as the Canadian variety. For the Kama variety 58 of 65 kernels were correctly classified, however, 2 kernels of the Rosa variety and 8 kernels of the Canadian variety were mistakenly classified as the Kama variety.

The classification rate was a bit worse when data set II was used. The correct classification results dropped to 89 % for Canadian, 88 % for Kama, and 93 % for Rosa varieties. For the Rosa variety, 93 of 100 kernels were classified properly and none of the other two varieties were misclassified as the Rosa variety. For the Canadian variety, 89 of 100 kernels were correctly classified, while 3 kernels of the Rosa variety and 8 kernels of the Kama variety were mistakenly classified as the Canadian variety. For the Kama variety, 57 of 65 kernels were correctly classified,

however 4 kernels of the Rosa variety and 11 kernels of the Canadian variety were mistakenly classified as the Kama variety.

To test the accuracy of the discrimination, a set of new observations utilizing 23 grains, was subjected to a classification process employing the evaluated discriminant functions. The results confirmed earlier conclusions on the discrimination: proper classification, accordingly, equaled six of 8, five of 7, and eight of 8 kernels for Canadian, Kama, and Rosa varieties. For both geometric feature sets all Rosa kernels were correctly classified, while Kama and Canadian varieties appeared to be misclassified, albeit to a insignificant degree. Herein, two Canadian kernels were classified as being Kama, one Kama kernel was classified as Canadian, and one Kama kernel was misclassified as Rosa. In this assessment a comparable correctness of classification was obtained to that when the nonparametric kernel discrimination was used. The classification rate, of the latter, though, was a bit better: two Kama kernels and only one Canadian kernel, were incorrectly classified. Hence, the nonparametric character of this method provided better fitting to the real data structure.

It is worth noting that the results of the classification confirmed the results obtained by the clustering analysis, which was conducted in the earlier stage of the research study [1].

4 Summary

Recent advances in digital image processing technique provide non-destructive tools for improving insight into seed morphology in terms of image acquisition and automatic feature detection. The accomplished study demonstrated that the image analysis commonly employed in discriminant methods gives reliable results in classifying wheat grain. In such work, selected combinations of geometric features permitted discriminant analysis to achieve a recognition rate of 89–96 %. The Rosa variety is, however, better discriminated, whilst Kama and Canadian varieties are harder to differentiate. Such a summation of results was also confirmed by the results of the subsequent classification of new observations, using both parametric and nonparametric methods. The conducted study confirmed the practical usefulness and effectiveness of the evolved methods in classification practices. Thus, discriminant analysis should be considered as being very effective in separating out wheat varieties.

References

1. Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Łukasik, S., Żak, S.: Complete gradient clustering algorithm for features analysis of X-ray images. In: Pietka, E., Kawa, J. (eds.) Information Technologies in Biomedicine, pp. 15–24. Springer, Berlin (2010)
2. Draper, N.R., Smith, H.: Applied Regression Analysis. Wiley, New York (1981)
3. Guevara-Hernandez, F., Gomez-Gil, J.: A machine vision system for classification of wheat and barley grain kernels. Span. J. Agric. Res. **9**, 672–680 (2011)

4. Kowalski, P.A., Kulczycki, P.: Interval probabilistic neural network. Neural Comput. Appl. 1–19 (2015). doi:10.1007/s00521-015-2109-3
5. Krzyśko, M., Wołyński, W., Górecki, T., Skorzybut, M.: Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości (Learning systems. Pattern recognition, cluster analysis and dimensionality reduction). WNT, Warszawa (2008) (in Polish)
6. Kulczycki, P.: Estymatory jadrowe w analizie systemowej (Kernel Estimators in Systems Analysis). WNT, Warszawa (2005). (in Polish)
7. Kulczycki, P.: Kernel estimators in industrial applications. In: Prasad, B. (ed.) Soft Computing Applications in Industry. Springer, Berlin (2008)
8. Kulczycki, P., Kowalski, P.A.: Bayes Classification for nonstationary patterns. Int. J. Comput. Methods **12**(2), 155008–19 (2015)
9. Majumdar, S., Jayas, D.S.: Classification of cereal grains using machine vision: I. Morphology models. Am. Soc. Agric. Eng. **43**(6), 1669–1675 (2000)
10. Strumiłł o, A., Niewczas, J., Szczypiński, P., Makowski, P., Woźniak, W.: Computer system for analysis of X-ray images of wheat grains. Int. Agrophys. **13**, 133–140 (1999)
11. Utku, H.: Application of the feature selection method to discriminate digitized wheat varieties. J. Food Eng. **46**, 211–216 (2000)
12. Utku, H., Koksel, H., Kayhan, S.: Classification of wheat grains by digital image analysis using statistical filters. Euphytica **100**, 171–178 (1998)
13. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1994)
14. Wiwart, M., Suchowilska, E., Lajszner, W., Graban, Ł.: Identification of hybrids of spelt and wheat and their parental forms using shape and color descriptors. Comput. Electron. Agric. **83**, 68–76 (2012)
15. Zapotoczny, P.: Discrimination of wheat grain varieties using image analysis and neural networks. Part I. Single kernel texture. J. Cereal Sci. **54**, 60–68 (2011)
16. Zarychta, P.: Features extraction in anterior and posterior cruciate ligaments analysis. Comput. Med. Imaging Graph. **46**, 108–120 (2015)
17. Zhu, M., Hastie, T.J.: Feature extraction for nonparametric discriminant analysis. j. comput. graph. stat. **12**(1), 101–120 (2003)