

# Authentication of Bee Pollen Grains in Bright-Field Microscopy by Combining One-Class Classification Techniques and Image Processing

MANUEL CHICA<sup>1,2\*</sup>

<sup>1</sup>*Inspiralia Tecnologías Avanzadas, Estrada 10, 28034 Madrid, Spain*

<sup>2</sup>*European Centre for Soft Computing, Gonzalo Gutiérrez Quirós, 33600 Mieres (Asturias), Spain*

**KEY WORDS** microscopic imaging; pollen grains; outliers detection; classification; computer vision

**ABSTRACT** A novel method for authenticating pollen grains in bright-field microscopic images is presented in this work. The usage of this new method is clear in many application fields such as bee-keeping sector, where laboratory experts need to identify fraudulent bee pollen samples against local known pollen types. Our system is based on image processing and one-class classification to reject unknown pollen grain objects. The latter classification technique allows us to tackle the major difficulty of the problem, the existence of many possible fraudulent pollen types, and the impossibility of modeling all of them. Different one-class classification paradigms are compared to study the most suitable technique for solving the problem. In addition, feature selection algorithms are applied to reduce the complexity and increase the accuracy of the models. For each local pollen type, a one-class classifier is trained and aggregated into a multiclassifier model. This multiclassification scheme combines the output of all the one-class classifiers in a unique final response. The proposed method is validated by authenticating pollen grains belonging to different Spanish bee pollen types. The overall accuracy of the system on classifying fraudulent microscopic pollen grain objects is 92.3%. The system is able to rapidly reject pollen grains, which belong to nonlocal pollen types, reducing the laboratory work and effort. The number of possible applications of this authentication method in the microscopy research field is unlimited. *Microsc. Res. Tech.* 75:1475–1485, 2012. © 2012 Wiley Periodicals, Inc.

## INTRODUCTION

The bee-keeping sector has a notable socio-economic relevance in Europe, according to the FAO Agricultural Statistics Division. Although honey is the most important bee product, there are other well-known products that result from bee-keeping activity, such as pollen or royal jelly. Bee pollen production, for both domestic and foreign markets, is considered by many bee-keepers a means of diversification and increasing their income. Furthermore, bee pollen products are considered an important food supplement and can be used in medical treatments, although they are not scientifically recognized.

Bee-keepers, bee-keeping associations, and laboratories are interested in detecting fraud in pollen and require tools to standardize and authenticate bee pollen origin to guarantee their nutritive and health benefits. Although there are discernment methods for recognizing pollen types (Carrión et al., 2004; Chica and Campoy, 2012), the microscopic analysis of pollen grains, which form bee pollen loads, is the most precise method of identifying origin. This process requires the laboratory work of melissopalynology experts and is thus time consuming and costly.

The use of computer vision and classification techniques is not new in microscopy research and has performed well in many situations (Chen et al., 2006b; Jalba et al., 2004; Ranzato et al., 2007; Tsai

et al., 2008; Wu et al., 2008). There have been many attempts to automate pollen grain identification in microscopic images by computer algorithms, but there is no inexpensive, complete, and automated imaging process. Some systems use scanning electron microscopy images (Treloar et al., 2004). There are also systems based on laser scanning (Ronneberger et al., 2002). However, one of the requirements of our final system is to be inexpensive and easy-to-use by bee-keeping associations. Many of these associations already have conventional bright-field microscopes because of their low price and simplicity. This fact also applies in small laboratories, where expensive equipment is not available.

In literature, the first works on recognizing pollen grains by optical microscopes were presented by France et al. (2000) and Boucher et al. (2002), where some discriminative features of various pollen taxa were detected and classified. Then, Li et al. (2004) and Zhang et al. (2004) extracted more sophisticated information from pollen grains such as Gabor wavelets and moment invariants. They also implemented an artificial neural network (ANN) for classifying pollen grains.

\*Correspondence to: Manuel Chica, Inspiralia Tecnologías Avanzadas, Estrada 10, 28034 Madrid, Spain. E-mail: manuel.chica@softcomputing.es

Received 7 March 2012; accepted in revised form 18 May 2012

DOI 10.1002/jemt.22091

Published online 27 June 2012 in Wiley Online Library (wileyonlinelibrary.com).



Fig. 1. Bright-field microscopic images of pollen grains belonging to *Echium*, *Cistus*, *Rubus*, *Olea*, and *Quercus ilex*, respectively. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

Rodríguez-Damián et al. (2006) obtained an accuracy of 89%, while classifying similar species of the *Urticaceae* family using shape and texture features, ANNs, and support vector machines. An additional feasibility study on recognizing the pore and colpi structures of grass, birch, and mugwort pollen grains is done by Chen et al. (2006a). Finally, the last work on pollen grain identification is done by Landsmeer et al. (2009), where they propose a mechanism to identify pollen grains on air samples against other microscopic particles.

Nevertheless, the latter works always identify pollen grains as one of a fixed number of possible pollen types. There is no reference in literature, where pollen grains are authenticated as known or fraudulent types. Developing a system to authenticate local pollen grains in bright-field microscopic images is a highly complex task, requiring a specific solution. The classification of known local pollen grains must be made against all other world pollen types. This is an important obstacle for the designing of an automated system, as microscopic data cannot be collected from all existing pollen types.

To overcome this obstacle, we propose a novel pollen grain authentication system based on image processing and a multiclassifier formed by one-class classifiers. The use of one-class classification was introduced as a classification paradigm to detect anomalies or outliers in a data distribution (Chandola et al., 2009; Moya et al., 1993; Ritter and Gallegos, 1997; Tax, 2001). In these cases, there is a data absence or limitation in negative classes. This characteristic is ideal to deal with our authentication problem as we can model the local pollen grain types but we cannot do it with all the possible fraudulent pollen types. The application field of one-class classification is enormous, from fraud detection (Phua et al., 2004; Taniguchi et al., 1998) to image processing (Augusteijn and Folkert 2002; Pokrajac et al., 2007).

In this article, we also study the suitability of different one-class classification paradigms when solving this microscopic imaging problem. Thus, a comparison of five several models is done: a Gaussian estimator, a support vector data description (SVDD), and three variants of the  $k$ -nearest neighbors (kNN) technique. In addition to the use of one-class classification, we have applied feature selection methods (Guyon and Elisseeff, 2003; Liu and Motoda, 1998) to reduce the complexity of the initial model. A multiclassifier model is also designed to aggregate the one-class classifiers

outputs, given a unique response with a confidence measure.

The proposed methods are validated for authenticating five of the most common Spanish pollen types, *Cistus ladanifer*, *Olea*, *Rubus*, *Echium*, and *Quercus ilex*, against non-Spanish pollen types. This work is focus on Spain because it is one of the most important bee pollen producer of the European Union (CBI, 2009). Totally, a dataset of around 1,063 instances were used to train and validate the system. The classifiers were validated using receiver operating characteristic (ROC) analysis and classification accuracy indicators.

## MATERIALS AND METHODS

### Pollen Types and Microscopic Image Acquisition

The number of possible pollen types in each country is high. However, there are some types that are extremely common in bee pollen. The following types, five of the most common Spanish pollen types, have been selected for validating our proposal: *Echium*, *Cistus*, *Rubus*, *Olea*, and *Quercus*. Depending on the local region of interest, these types can be others although the proposal can still be valid. A brief description of the selected pollen types is given below. Their microscopic images are shown in Figure 1.

- *Echium*: small (10–25  $\mu\text{m}$ ), heteropolar, prolate shape, 3-colporate and perforate ornamentation.
- *Cistus*: medium-sized (26–50  $\mu\text{m}$ ), isopolar, and spheroidal. The aperture type is 3-colporate with a smooth aperture membrane and reticulate ornamentation.
- *Rubus*: medium-sized (26–50  $\mu\text{m}$ ), isopolar, spheroidal, 3-colporate and striate ornamentation.
- *Olea*: small size (10–25  $\mu\text{m}$ ), isopolar with a spheroidal shape, and 3-colporate. Its aperture membrane is ornamented being the ornamentation reticulum cristatum.
- *Quercus*: medium-sized (26–50  $\mu\text{m}$ ), isopolar, spheroidal, 3-colporate and perforate ornamentation.

A bright-field optical microscope, Nikon E200 (40 $\times$ ), is used. The USB DS-Fi1 digital camera is also used to acquire images from the microscope. This camera is charge-coupled device (CCD) capturing images at a resolution of 2560  $\times$  1920 pixels. The preparation of the pollen slides is also straight-forward and cheap. Typically, a thin flat piece of glass of 75  $\times$  25mm and about

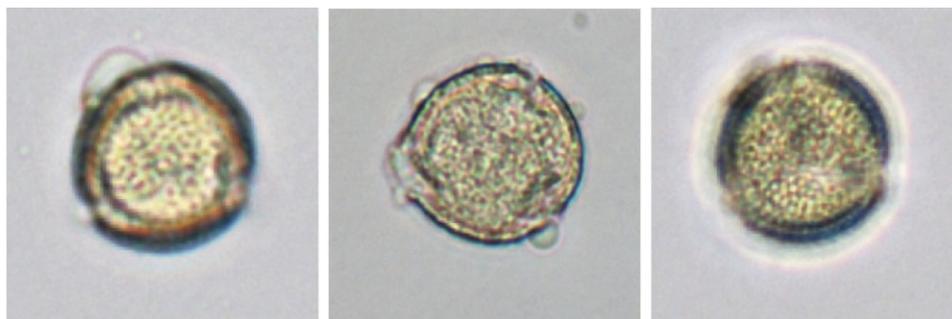


Fig. 2. Three microscopic images of *Olea* in different focal planes, A, B, and C. These images will be the input of the authentication system and are used to obtain the features required by classification. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

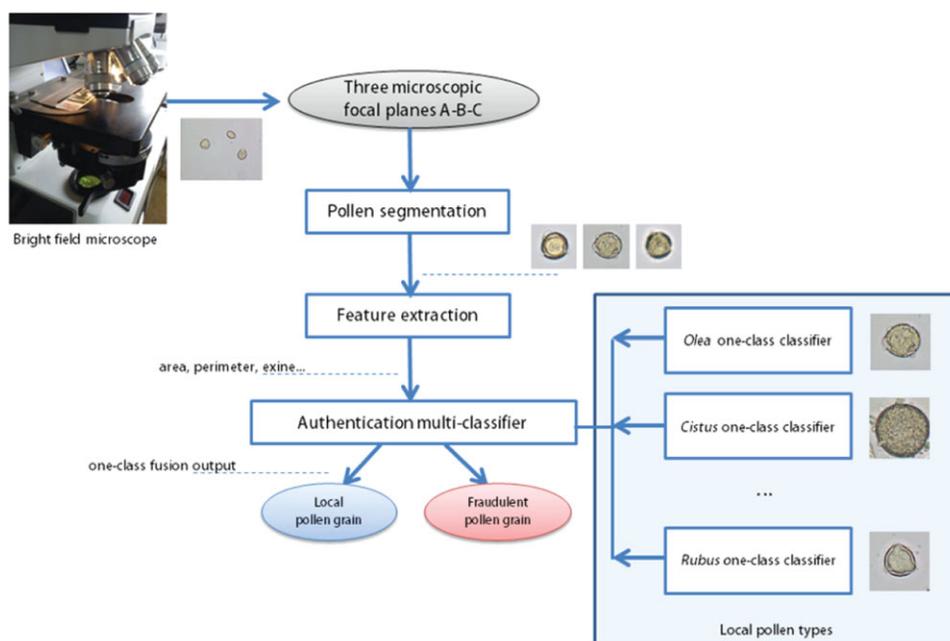


Fig. 3. Diagram of our proposed method for authenticating pollen grains. From the set of three focal images acquired by a bright-field microscope, the processing chain conducts the segmented pollen grains to a final authentication pollen output. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

1 mm thick is used together with a cover slip or cover glass over the specimen.

A motorized microscope focus stage cannot be available in our system because of its unattainable cost for small bee-keeping associations and beekeepers. Therefore, the final user must manually take three focal images by focusing on the positive sculpture of the inner part of the pollen grain (A), the exine (B), and the negative sculpture of the inner part (C). The user can use pattern examples of A, B, and C to get an approximate focus. An example of three focal planes for a pollen grain is given in Figure 2.

### Description of the Proposed Method

An overview of the proposed method can be observed in the diagram of Figure 3. First, user needs to acquire a set of three focal images for each pollen grain. Then,

pollen grain objects are segmented from background. When pollen grains are extracted, a set of discriminative features is calculated for each of them. Finally, a multiclassifier formed by one-class classifiers gives an output about the authentication of each pollen grain, classifying them as known local pollen type (*Echium*, *Rubus*, *Cistus*, *Olea*, and *Quercus*) or as nonlocal pollen type (outlier or fraudulent).

### Preprocessing and Segmentation of Pollen Grains

A contrast-limited adaptive histogram equalization is applied to enhance the contrast of the three focal gray scale images. Also, images are filtered by using a median filter to remove noise, preserving edges.

The choice of the focal plane has a considerable effect on the quality of the segmentation (Tscherepanow

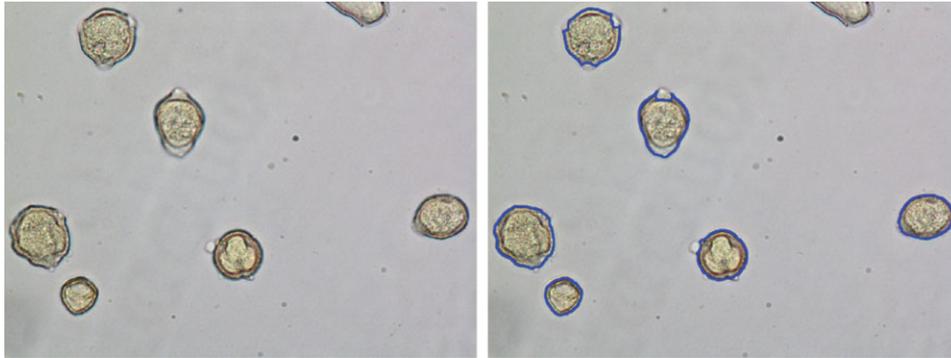


Fig. 4. On the left hand, the original microscopic image is shown. A mask showing the contours of the segmented pollen grains over the original image can be seen in the right figure. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

et al., 2008). Focal plane A is used to segment pollen grains. The reason is that the exine is more distinctive in A than in other planes as B or C.

As a bee pollen specimen does not usually contain dirt, fungal spores, and other nonpollen objects, the following steps are enough to segment isolated pollen grains:

1. Application of a thresholding method to the gray scale image. The threshold is automatically found by Otsu method (Otsu, 1979).
2. A hole filling algorithm using 4-connected background neighbors is applied to the binary thresholded image to fill the holes of the inner part of some pollen grains.
3. Opening and closing operations are also performed on the segmented image. The goal of this image processing technique is to remove small objects from the image, while preserving the shape and size of the real pollen grains.

Segmentation results are shown in Figure 4 when extracting pollen grains in a given microscopic image.

### Discriminative Pollen Grain Features

In this section, the processes to extract and select the most discriminative features of the pollen grains are, respectively, described.

**Feature Extraction.** After the segmentation process, a number of well-defined and discriminative features have to be processed for each segmented pollen grain. As already stated, the input of the extraction process is a set of three focal planes (A, B, and C). Some features such as those related to shape do not vary because of the focal plane. Then, these features are just calculated once. However, others suffer from focal variations. In these cases, we will extract three feature values, one per focal plane, not to lose important information.

The total number of features for each pollen grain is 28. We can split up them in three groups: (a) shape-related, (b) textural and color information of the inner part, and (c), exine descriptors. The description of the shape features are summarized below:

- $f_1$  (area): number of pixels representing the pollen grain area.

- $f_2$  (perimeter): number of pixels which form the pollen grain boundary.
- $f_3$  (diameter): largest distance between any two points of the pollen grain boundary.
- $f_4$  ( $d_{max}$ ): maximum distance between the center of the pollen grain and any point of its boundary.
- $f_5$  ( $d_{min}$ ): minimum distance between the center of the pollen grain and any point of its boundary.
- $f_6$  (radius dispersion): variability of the distances between the center and all the boundary points of the pollen grain.
- $f_7$  ( $d_{mean}$ ): average distance between the center of the grain and all the points of the boundary of the pollen grain.
- $f_8$  ( $d_{max}/d_{min}$ ): ratio between the maximum and minimum distance of the boundary points of the grain and the center.
- $f_9$  ( $d_{max}/d_{mean}$ ): ratio between the maximum and average distance of the boundary points of the grain and the center.
- $f_{10}$  ( $d_{min}/d_{mean}$ ): ratio between the minimum and average distance of the boundary points of the grain and the center.

The inner part of the pollen grain is calculated by extracting an internal circle of radius the half of the complete pollen grain object. Each of the following color and textural features are calculated for the inner part of the three focal planes (A, B, and C):

- $f_{11}, f_{12}, f_{13}$  ( $mean_A, mean_B, \text{ and } mean_C$ ): mean value of the gray scale histogram.
- $f_{14}, f_{15}, f_{16}$  ( $std_A, std_B, \text{ and } std_C$ ): standard deviation of the gray scale histogram.
- $f_{17}, f_{18}, f_{19}$  ( $entropy_A, entropy_B, \text{ and } entropy_C$ ): entropy value of the gray scale histogram;  $-\sum P \log_2(P)$ , being  $P$  the probability of each gray scale value.
- $f_{20}, f_{21}, f_{22}$  ( $H_A, H_B, \text{ and } H_C$ ):  $H$  component value of the HSV color space.
- $f_{23}, f_{24}, f_{25}$  ( $S_A, S_B, \text{ and } S_C$ ):  $S$  component value of the HSV color space.

Finally, three more features,  $f_{26}, f_{27}, \text{ and } f_{28}$ , are calculated to represent the information of the exine of the pollen grain. Obtaining the exact number of pores and

colpi is not useful for the problem as many pollen grains have the same number of pores and, depending on the view, few of them (or maybe none) can be observed. However, representing the morphological features of the exine is useful as it contains information about how the pores and colpi are arranged. This information is highly discriminative to authenticate the pollen types.

Just the central focal image B of the pollen grain is used to extract the exine descriptors because it is where exine is clearer. Before extracting the features, the focal image is transformed into polar coordinates being the most external pixels those involved in the feature extraction process. These three exine descriptors are detailed below:

- $f_{26}$  ( $\text{mean}_{\text{exine}}$ ): mean of the gray scale histogram of the most external 15 pixels of the focal image B in polar coordinates.
- $f_{27}$  ( $\text{std}_{\text{exine}}$ ): standard deviation of the gray scale histogram of the most external 15 pixels of the focal image B in polar coordinates.
- $f_{28}$  ( $\text{entropy}_{\text{exine}}$ ): entropy of the gray scale histogram of the most external 15 pixels of the focal image B in polar coordinates.

**Feature Selection.** Unfortunately, although the set of 28 features explained in the previous section was chosen after a complete expert and literature review study, it is usual to have irrelevant features and useless information that degrade the performance of the models both in speed, due to the high dimensionality, and accuracy, due to irrelevant information (Guyon and Elisseeff, 2003). Feature selection has the aim of choosing the smallest possible subset of features  $P$  necessary to describe a problem with an initial set of  $N$  features, being  $P \leq N$ . In other words, feature selection can be defined as a search process for removing irrelevant and/or redundant features and to obtain a simpler classification system. In some problems, feature selection ensues not only in faster performance but also in more accurate classification than using the whole set (Liu and Motoda, 1998).

The specific goal is trying to reduce the initial pool of 28 features in just the most important, without affecting the overall performance of the classification model. To achieve this objective, we have applied feature selection algorithms to rank the most important features for authenticating pollen grains. We have experimented with two evaluation measures:

- Relief (Kira and Rendell, 1992): The general idea of this method is to choose the features that can be most distinguished between classes. These are known as the relevant features. At each step of an iterative process, an instance is chosen at random from the dataset and the weight for each feature is updated according to the distance to its *nearmiss* and *nearhit*.
- Gain Ratio: It evaluates the worth of a feature by measuring the information gain ratio with respect to the class. It is a feature selection algorithm based on information theory and information gain (a variation of the MIFS algorithm of Battiti (1994)).

The results of applying the latter feature selection algorithms are described in the experiments and discussion sections.

### One-Class Classification

The most outstanding property in pollen grain authentication is the limited data to model the nonlocal pollen types (negative classes or outliers). Although it is possible to model the local pollen types, we cannot do the same with all the possible existing fraudulent pollen types from around the world. One-class classification has been selected as an appropriate paradigm to deal with this problem circumstances.

One-class classification problem is different from the conventional binary or multiclass classification problem. This distinction lies in the absence of the negative class (normally called outlier) or in the vagueness of its definition and sampling (Chandola et al., 2009; Tax, 2001). Originally, the term was given by Moya et al. (1993) and some authors refer to this problem as outlier detection (Ritter and Gallegos, 1997), novelty detection (Bishop, 1994), or concept learning (Japkowicz et al., 1995).

This absence or data limitation of the negative data makes the problem harder to solve than conventional classification problems. The goal of one-class classification is to define a classification boundary around the positive class (also called target), which maximizes the number of accepted true positive instances and minimizes the number of rejected true negative instances.

The one-class classification techniques used for anomaly or outlier detection can primarily be grouped in two categories: density-based and boundary-based classifiers (Tax, 2001). Within the first group are Gaussian models or Bayesian networks (Barbara et al., 2001; Siaterlis and Maglaris, 2004). One of the boundary-based classifiers is the well-known kNN, modified for the case of one-class classification (Byers and Raftery, 1998; Eskin et al., 2002) or SVDD (Ratsch et al., 2002; Tax and Duin, 2004).

We have used three different approaches of the latter groups: the Gaussian classifier, which could be considered as density-based classifiers. And also, SVDD and kNN which can be seen as the most representative algorithm of the boundary-based classifiers:

**Gaussian Model.** The training dataset being a set of  $p$ -dimensional instances  $x_i$ ,  $i = 1, \dots, n$ , this model simply calculates a Gaussian one-class classifier by estimating the mean  $\bar{x}$  and the covariance matrix  $S_n$  of the dataset distribution. The classifier uses the Mahalanobis distance  $(x - \mu_n)^T S^{-1}(x - \mu_n)$  to estimate the fitness of each instance to the target class. In addition, a threshold  $\theta_{p,n}$  needs to be defined during the training phase by means of a target acceptance rate, which is normally given as a parameter.

Each new testing instance  $z$  will be evaluated for acceptance or rejection as target by the following Eq. (1):

$$z \text{ is accepted as target if } (x - \mu_n)^T S^{-1}(x - \mu_n) \leq \theta_{p,n} \quad (1)$$

**SVDD.** The main idea of SVDD is to obtain a spherical-shaped boundary around the training dataset  $x$ , which can enclose as many samples as possible while

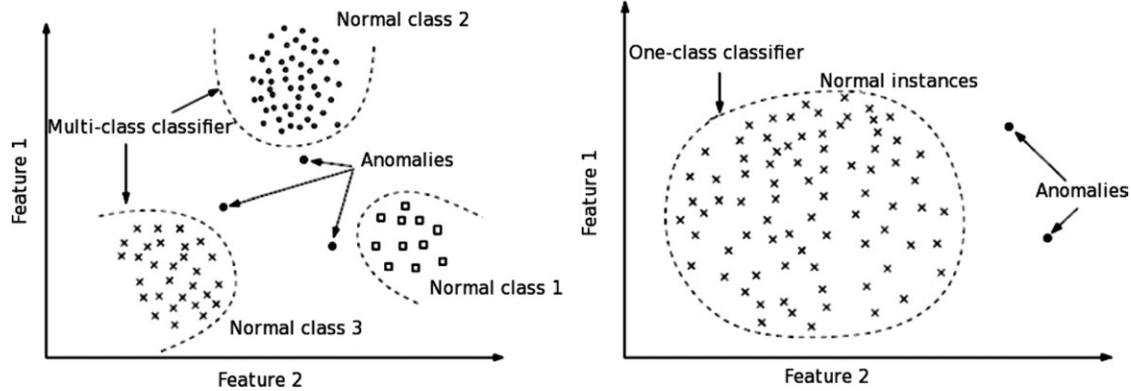


Fig. 5. Plot with the targets and outliers in a one-class classification problem on the right. On the left, a multiclass classification based anomaly detection problem (Chandola et al. (2009)).

having the minimum volume. The sphere is characterized by its center  $c$  and radius  $R > 0$ . Minimization of the sphere volume is achieved by minimizing its square radius  $R^2$ . To improve the generality of the model, some samples are located outside the sphere, but larger distances from the center should be penalized. Thus, slack variables are included in the optimization of the variables, which will determine the hypersphere. For the mathematical details, please refer to seminal paper of Tax and Duin (2004).

Given a new sample  $z$ , we compare its distance to the center of the sphere with the radius of the sphere  $R$ . If  $z$  is inside the hypersphere, it belongs to the target class. Otherwise, it is an outlier.

**One-Class  $k$ NN.** Distance-based one-class classifiers are based on the assumption that normal data instances occur in dense neighborhoods, while anomalies occur far from their closest neighbors.  $k$ NN, originally provided by Dasarathy (1991), is the best-known distance classifier. The basics of the algorithm for one-class classification is that the anomaly score of a data instance is defined as the distance with its  $k$ th nearest neighbor in a given dataset.

Nearest neighbor classifiers always require the definition of distance or similarity measures defined between two data instances. For continuous features, the Euclidean distance is the most popular choice. For our problem, we have chosen and compared three different values for  $k = 1, 3$ , and  $5$ . Each new instance  $z$  will be considered as target or outlier depending on the majority vote of its one, three, or five closest neighbors in the training data.

### Fusion of the Classifiers

Normally, as the possible local pollen types are more than just one, the authentication problem can be considered as a multiclass anomaly detection problem. An incoming pollen grain is considered anomalous if it is not classified as local by any of the one-class classifiers (see plots of Fig. 5). A confidence score with the prediction made by each particular classifier is normally provided. If none of the classifiers are confident in classifying the test instance, the instance is declared to be anomalous (Chandola et al., 2009). We have followed

this approach to aggregate the trained one-class classifiers.

Being  $C$  a set of known local bee pollen types, the training data will contain instances belonging to  $|C|$  classes. To use one-class classifiers and be able to reject unknown pollen grains, the system in  $|C|$  binary sub-problems must be decomposed. Thus,  $|C|$  one-class classifiers  $f_1, f_2, \dots, f_{|C|}$  based on densities or distances must be trained, and an ensemble scheme has to be built to fuse them in a multiclass prediction.

Therefore, for each pollen grain instance  $x$  we first map each one-class classifier output  $f_i(x)$  to a posterior probability  $P(y = c|x)$ . These probabilities are also normalized in the range  $[0, 1]$ . The posterior probability of the each classifier's target can be considered as the confidence  $CF_{oc}(y = c|x)$  that one instance  $x$  belongs to the class  $c$ .

To classify an incoming pollen grain as one of the  $|C|$  possible pollen types a multiclassifier is constructed. It compares the confidence  $CF_{oc}(y|x)$  of all the one-class classifiers and provides a global prediction from the most reliable one-class classifier. The multiclassifier prediction  $\omega$  is given by:

$$\omega = \max_{1 \leq c \leq |C|} CF_{oc}(c|x) \quad (2)$$

However, it is also necessary to estimate the confidence of the multiclassifier prediction. To do this, we first introduce two parameters as done in Goh et al. (2005):

$$T_{oc} = CF_{oc}(\omega|x) \quad (3)$$

$$T_m = T_{oc} - \max_{1 \leq c \leq |C|, c \neq \omega} CF_{oc}(c|x) \quad (4)$$

Although  $T_{oc}$  is the highest confidence factor from the  $|C|$  binary one-class classifiers and determines the multiclassifier prediction class  $\omega$ ,  $T_{oc}$  might not be sufficient to estimate the global confidence of the multiclassifier prediction. For this reason, we introduce the use of the multiclass margin  $T_m$ . Wrong predictions could have high  $T_{oc}$  but small  $T_m$  but correct predictions must have higher multiclass margin values  $T_m$ .

There is a better separation of correct from erroneous predictions if the multiclass margin variable is used (Goh et al., 2001; Schapire and Singer, 1999).

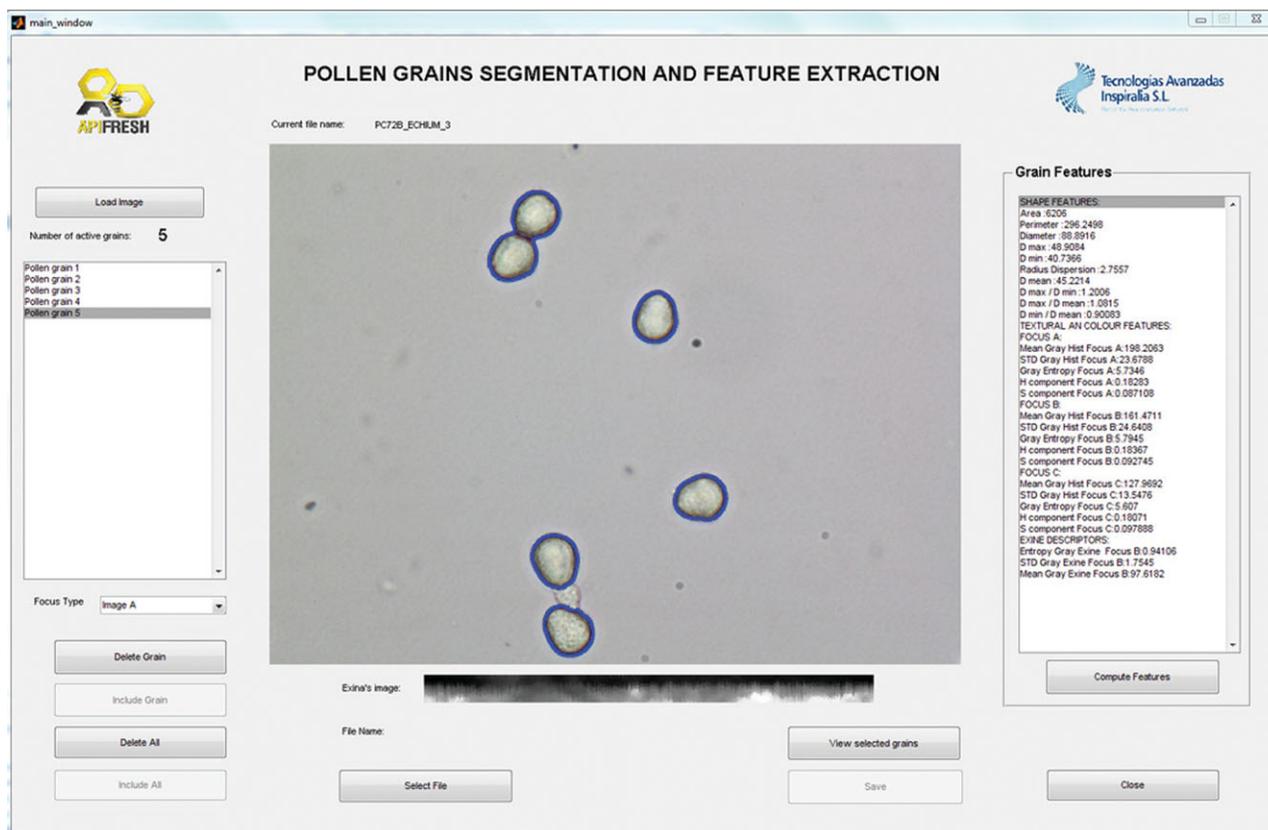


Fig. 6. Software prototype built in MATLAB to generate the datasets, test the algorithms, and validate the complete framework. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

After preliminary experimentation, we set both parameters  $T_{oc}$  and  $T_m$  to 0.5 and 0.01, respectively, for all the paper experimentation, to be used in the final decision of the multiclassifier as in Eq. (5).

$$\omega \text{ is accepted, } T_{oc} \geq 0.5 \text{ and } T_m \geq 0.01, \quad (5)$$

$$\text{outlier,} \quad \text{otherwise}$$

## RESULTS

### Validation Data and Performance Indicators

A software prototype was developed to manage data acquisition from microscopic images, process the training database, train, adjust the models, and validate the whole system (see Fig. 6). This software was programmed in MATLAB using some functions of the `dd` tools library (Tax, 2011).

Using this framework to process the bright-field microscopic images, 1,063 pollen grain samples were created. Each pollen grain sample consists of 28 input features in conjunction with its class. This class could be one of the five pollen types (i.e., *Echium*, *Rubus*, *Cistus*, *Olea*, and *Quercus*) or an outlier class (nonlocal pollen type).

To validate the algorithms in a convenient manner, we have split up the local pollen samples into training (80% of the data) and test (20% of the data). Just training positive samples are used to train the one-class

TABLE 1. Details of the training and test data partitions grouped by pollen type. These sets are created at random for each of the 10 runs of the algorithms

Pollen type	No. of samples	Training samples (80%)	Test samples (20%)
<i>Echium</i>	156	125	31
<i>Rubus</i>	113	90	23
<i>Cistus</i>	101	81	20
<i>Olea</i>	106	85	21
<i>Quercus</i>	141	113	28
Outliers	446	0	446
Total	1063	494	569

classifiers, while test instances are formed by a diverse pool of outliers (fraudulent samples) and test positive samples (local samples). These test instances were not used during the training phase, being totally classified by the trained methods in the testing stage. In addition and to avoid randomness in the data partition process, the algorithms were trained and tested 10 times by forming 10 different random data partitions. The corresponding figures of the training and test data partitions are detailed in Table 1.

The performance of the classifiers is evaluated by the classification accuracy, false negative (FN) and positive rates, and the confusion matrix. A FN occurs when the outcome of the classifier is incorrectly predicted as outlier when it is actually a target. While a false posi-

tive (FP) occurs when the outcome is incorrectly predicted as target when it is actually an outlier. The FN rate measures the number of FNs on the total number of negatives or outliers, and the FP rate calculates the fraction of FPs divided by the total number of positives or target instances (Witten and Frank, 2005).

We have also used the ROC curve analysis and the area under the curve (AUC) (Provost and Fawcett,

1997). The ROC analysis allows us to understand the performance of the classifiers without taking into account the rejection threshold of the targets considered as outliers. It represents the trade-off between the false and true positives for different values of the rejection threshold in anomaly detection and one-class classification problems (Bradley, 1997). Also, the AUC summarizes the classification performance of the classifier in the entire range [0, 1] of the FP rate and can be interpreted as the probability of authenticating pollen load outliers higher than local pollen load types. It is calculated from the ROC curve.

TABLE 2. AUC indicator values of the five one-class classifiers grouped by the five local pollentypes

	<i>Echium</i>	<i>Olea</i>	<i>Quercus</i>
Classifier 1 —Gaussian	0.99784 (0.006)	0.97698 (0.024)	0.96745 (0.018)
Classifier 2 —SVDD	0.9972 (0.006)	0.97367 (0.022)	0.97118 (0.02)
Classifier 3 —1NN	0.99763 (0.006)	0.97969 (0.022)	0.96802 (0.020)
Classifier 4 —3NN	0.99796 (0.006)	0.97538 (0.023)	0.96846 (0.019)
Classifier 5 —5NN	0.99798 (0.006)	0.9743 (0.023)	0.97118 (0.019)
	<i>Rubus</i>	<i>Cistus</i>	
Classifier 1 —Gaussian	0.98122 (0.03)	0.96111 (0.045)	
Classifier 2 —SVDD	0.98145 (0.03)	0.95426 (0.049)	
Classifier 3 —1NN	0.98156 (0.03)	0.95403 (0.049)	
Classifier 4 —3NN	0.98081 (0.03)	0.95454 (0.049)	
Classifier 5 —5NN	0.98074 (0.03)	0.95469 (0.049)	

The higher the value, the better the classifier performance. The given values are the mean ( $\bar{x}$ ) and standard deviation ( $\sigma$ ) obtained from runs on 10 independent test datasets.

### Experiments

First, the performance indicator values for the five one-class classification techniques, Gaussian classifier, SVDD, 1NN, 3NN, and 5NN, are obtained by classifying the test dataset (formed by positive and outliers samples) in 10 different runs. In Table 2, the mean and standard deviation of the AUC values generated by the one-class classifiers can be observed. Also, the ROC curves of the one-class classifiers are presented in Figure 7. These ROC curves are grouped by the five target pollen types.

The results of applying feature selection algorithms to pollen grains data are also discussed. Table 3 presents a ranking of the different features, classified in order of importance, according to the Relief and Gain Ratio measurements. These two algorithms have been used to select the most discriminative features from the initial pool of 28 features, used in the previous experimentation. Then, it is possible to compare the

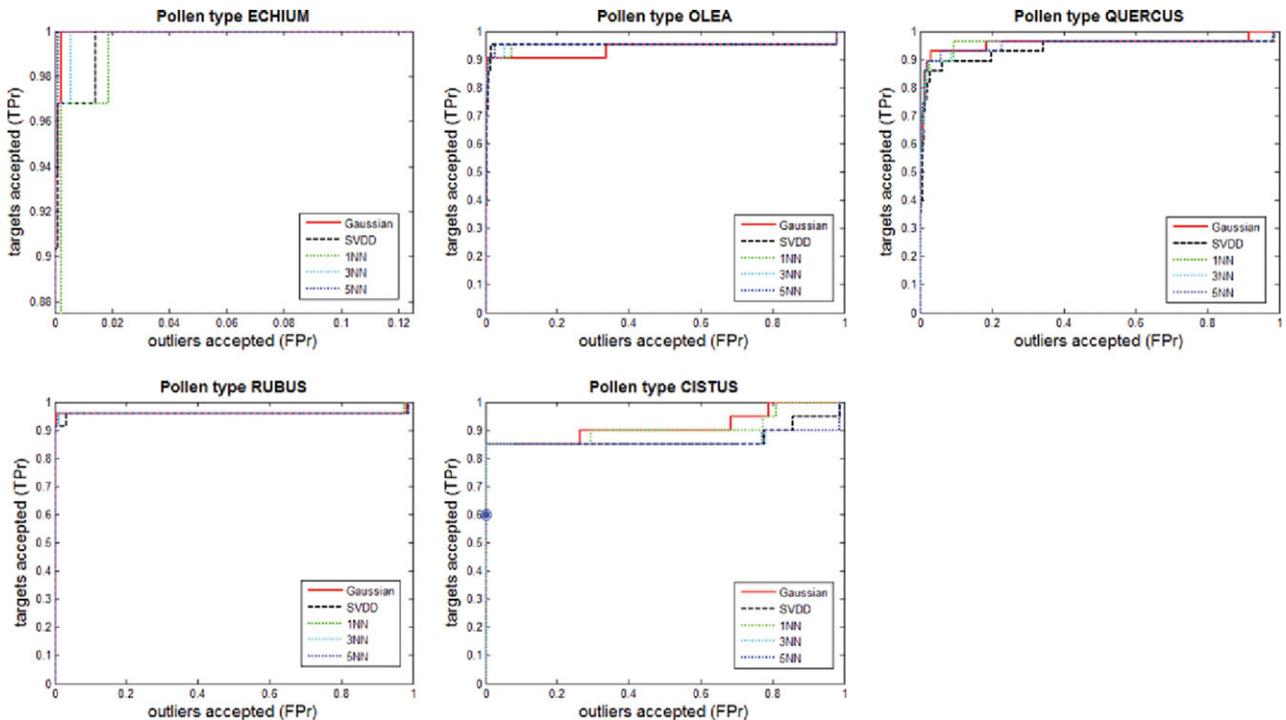


Fig. 7. ROC curves of the one-class classifiers for the five local pollen types. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 3. Ranking output of the two feature selection algorithms used in this study

Relief		Gain ratio	
Value	Feature description	Value	Feature description
0.259	$f_{15}$ std <sub>B</sub>	0.729	$f_7$ $d_{\text{mean}}$
0.248	$f_{16}$ std <sub>C</sub>	0.725	$f_2$ perimeter
0.245	$f_3$ diameter	0.724	$f_3$ diameter
0.217	$f_{14}$ std <sub>A</sub>	0.724	$f_1$ area
0.211	$f_1$ area	0.697	$f_{15}$ std <sub>B</sub>
0.184	$f_2$ perimeter	0.647	$f_{16}$ std <sub>C</sub>
0.158	$f_{23}$ $S_A$	0.642	$f_{14}$ std <sub>A</sub>
0.157	$f_7$ $d_{\text{mean}}$	0.632	$f_5$ $d_{\text{min}}$
0.144	$f_{27}$ std <sub>exine</sub>	0.553	$f_{18}$ entropy <sub>B</sub>
0.144	$f_5$ $d_{\text{min}}$	0.547	$f_{17}$ entropy <sub>A</sub>
0.143	$f_{26}$ mean <sub>exine</sub>	0.547	$f_{19}$ entropy <sub>C</sub>
0.121	$f_{24}$ $S_B$	0.532	$f_4$ $d_{\text{max}}$
0.109	$f_{25}$ $S_C$	0.508	$f_{24}$ $S_B$
0.094	$f_{19}$ entropy <sub>C</sub>	0.488	$f_{23}$ $S_A$
0.092	$f_{20}$ $H_A$	0.479	$f_{11}$ mean <sub>A</sub>
0.092	$f_{17}$ entropy <sub>A</sub>	0.454	$f_{25}$ $S_C$
0.09	$f_4$ $d_{\text{max}}$	0.407	$f_{26}$ mean <sub>exine</sub>
0.087	$f_{11}$ mean <sub>A</sub>	0.402	$f_{27}$ std <sub>exine</sub>
0.074	$f_{18}$ entropy <sub>B</sub>	0.389	$f_{20}$ $H_A$
<hr/>			
0.071	$f_{22}$ $H_C$	0.366	$f_{12}$ mean <sub>B</sub>
0.063	$f_{28}$ entropy <sub>exine</sub>	0.354	$f_6$ radiusdispersion
0.054	$f_{12}$ mean <sub>B</sub>	0.35	$f_{21}$ $H_B$
0.053	$f_{21}$ $H_B$	0.328	$f_{22}$ $H_C$
0.041	$f_{13}$ mean <sub>C</sub>	0.317	$f_{13}$ mean <sub>C</sub>
0.026	$f_9$ $d_{\text{max}}/d_{\text{mean}}$	0.282	$f_{28}$ entropy <sub>exine</sub>
0.026	$f_{10}$ $d_{\text{min}}/d_{\text{mean}}$	0.241	$f_{10}$ $d_{\text{min}}/d_{\text{mean}}$
0.015	$f_6$ radiusdispersion	0.238	$f_8$ $d_{\text{max}}/d_{\text{min}}$
0.001	$f_8$ $d_{\text{max}}/d_{\text{min}}$	0.199	$f_9$ $d_{\text{max}}/d_{\text{mean}}$

Pollen grain features are ordered by the importance given by the algorithms. The horizontal line divides the final 19 selected features from the whole pool of 28 features.

behavior of the system with and without a feature selection process.

As can be seen, the last nine features are ranked as the least discriminative by both algorithms. Consequently, we have created new one-class classification models just composed of the 19 most discriminative, selected features. The AUC indicator values obtained by these reduced one-class classifiers are shown in Table 4.

The last step of the experimentation is the selection and training of the best one-class classification paradigm for each pollen type. The best results are obtained by the 5NN classifier trained with the reduced set of 19 features. After a preliminary study of the previous ROC curves, a fraction rejection of 25% positive instances is used to train the final 5NN classifiers.

Five 5NN classifiers (one per pollen type) are fused in a single model by using the multiclassifier approach previously described. Thus, we can obtain the total accuracy and confusion matrices of the multiclassifier by classifying the test datasets in 10 different runs. The evaluation indicator values are given in Table 5, while the confusion matrix is presented in Table 6.

## DISCUSSION

In this study, we have proposed a global method to authenticate pollen grains from bright-field microscopic images. Results were promising in the Spanish case of authenticating five well-known pollen types against outliers. The final accuracy of the multiclassification system is about 92.3% with low FP and negative rates (0.123 and 0.064, respectively).

TABLE 4. AUC indicator values for five different classification paradigms authenticating the five local pollen types after the feature selection procedure

	<i>Echium</i>	<i>Olea</i>	<i>Quercus</i>
Classifier 1	0.99807 (0.006)	0.97675 (0.024)	0.97528 (0.019)
—Gaussian			
Classifier 2	0.99778 (0.004)	0.97807 (0.022)	0.96082 (0.022)
—SVDD			
Classifier 3	0.99793 (0.004)	0.97926 (0.021)	0.97632 (0.02)
—1NN			
Classifier 4	0.99877 (0.004)	0.98038 (0.022)	0.97326 (0.019)
—3NN			
Classifier 5	0.99878 (0.004)	0.98054 (0.022)	0.97371 (0.019)
—5NN			
<hr/>			
	<i>Rubus</i>	<i>Cistus</i>	
Classifier 1			
—Gaussian	0.98122 (0.029)	0.96555 (0.033)	
Classifier 2	0.98085 (0.03)	0.96559 (0.040)	
—SVDD			
Classifier 3	0.98201 (0.03)	0.97099 (0.035)	
—1NN			
Classifier 4	0.98218 (0.03)	0.95888 (0.045)	
—3NN			
Classifier 5	0.98208 (0.03)	0.95807 (0.045)	
—5NN			

The classification is made by just using 19 features. The higher the value, the better the classifier performance. The given values are the mean ( $\bar{x}$ ) and standard deviation ( $\sigma$ ) obtained from runs on 10 independent test datasets.

TABLE 5. FP and FN rate values of the final multiclassifier formed by 5NN one-class classifiers

Performance measures	Mean	Standard deviation
Accuracy (%):	92.373	1.12
FP rate	0.123	0.02
FN rate	0.064	0.02

The given values are the mean ( $\bar{x}$ ) and standard deviation ( $\sigma$ ) obtained from runs on 10 independent test datasets.

A comparison between different one-class classification paradigms was performed. The validation of the one-class classifiers showed that the most interesting pollen types are *Cistus* and *Quercus ilex* because they are the most complicated pollen types from the authentication point of view. In the case of *Quercus ilex*, kNN approaches and the Gaussian model are the best models. Even though the Gaussian model outperforms the rest of the algorithms authenticating the *Cistus* pollen type.

According to the graphs of Figure 7, *Echium* is the easiest pollen type to be authenticated, where all the classifiers perform correctly. *Rubus* and *Olea* are also well authenticated against outlier instances, although there are higher differences. kNN approaches seem to obtain better ROC curves than the rest of the approaches (see Fig. 7). In addition and by observing the AUC values of Table 2, we can also arise the following conclusions:

- kNN approaches obtain the best results in all the runs and in all the pollen types apart from *Cistus*. In this case, the Gaussian classifier outperforms the rest of the paradigms.
- Best values for kNN are  $k = 1$  and 5 according to the results.
- AUC values suggest that all the Spanish pollen types are correctly authenticated by all the one-class classifiers without being a high difference between them.

TABLE 6. Confusion matrix of the multiclassification system formed by the five one-class 5NN classifiers

		Predicted pollen type					
		Echium	Olea	Quercus	Rubus	Cistus	Outlier
Real pollen type	Echium	23.3 (2.63)	0 (0)	0 (0)	0 (0)	0 (0)	7.7 (2.63)
	Olea	0 (0)	15.7 (2.16)	0.2 (0.42)	0 (0)	0 (0)	5.1 (2.18)
	Quercus	0 (0)	0 (0)	21.7 (2.58)	0 (0)	0 (0)	6.3 (2.58)
	Rubus	0 (0)	0 (0)	0 (0)	17.3 (2.63)	0 (0)	5.7 (2.63)
	Cistus	0 (0)	0 (0)	0 (0)	0 (0)	14.8 (2.39)	5.2 (2.39)
	Outlier	0 (0)	3.9 (0.32)	9 (2.67)	0.3 (0.48)	0 (0)	432.8 (2.78)

The values represented here are the mean and standard deviation from the 10 different partitions of the training and test datasets.

An additional experimentation was performed by applying feature selection algorithms to reduce and improve the robustness and accuracy of the multiclassification system. From the results of Table 3, we can see how the applied algorithms returned similar ranking values for the initial pool of features. Then, the least discriminative nine features were removed. These features were  $f_6$ ,  $f_8$ ,  $f_9$ , and  $f_{10}$ , which are related to the  $d_{\max}$ ,  $d_{\min}$ , and  $d_{\text{mean}}$  values. Therefore, this fact arises that the information of the latter four features are given by more elemental shape features, and they do not influence the classification. Also, features  $f_{12}$ ,  $f_{13}$ ,  $f_{21}$ , and  $f_{22}$  were removed. They are features based on the processing of focal planes B and C. The entropy of the histogram of the exine ( $f_{28}$ ) was also removed as it did not favor the classification process.

The feature selection process obtained a data reduction of 32.14% with respect to the original 28 features. But, it can be seen how the reduction made by the algorithms in the feature space did not result in a loss of accuracy; it even increased the overall performance according to the AUC indicator (compare values of Tables 2 and 4). Moreover, the feature selection did not alter the ranking performance of the different classification methods. kNN methods were still the best classification paradigm, outperforming the rest of the classifiers when solving the problem. Although all the kNN methods showed similar results, the 5NN was globally the best technique.

The multiclassification system yielded a high accuracy on the validation data, 92.373%. The FP and FN rates were also low and show the good response of the authentication system. The FP rate was higher than the FN rate but still low. In the confusion matrix of Table 6, the miss-classifications can be observed. Only some nonlocal instances were classified as *Olea*, *Quercus*, and *Rubus*. The real local pollen instances classified as outliers was very low. Also, it is important to notice that there was almost no miss-classification between the local pollen types; just a mean of 0.2 instances were miss-classified as *Quercus* being *Olea*.

In conclusion, we have shown that the authentication of pollen grains by using bright-field microscopic images is possible when developing an appropriate system formed by image processing and one-class classification techniques. The main novelty of the proposed model was the classification scheme, which allows the user to detect outlier pollen grains. Although the major interest of this work is the identification of fraudulent bee pollen samples, the results should be applicable to other investigators and fields such as detecting specific pollen types in air samples or detecting allergic pollen grains in food quality testing. Future work will be

devoted to use an advanced multifocus mechanism to enrich the pollen grain information such as multifocal image fusion (Redondo et al., 2009). Also, this work may be extended by using more sophisticated segmentation methods such as level sets or deformable models (Malladi et al., 1995; Xu et al., 2000).

## ACKNOWLEDGMENTS

The work presented in this article has been performed within the scope of the APIFRESH project. APIFRESH has been cofunded by the European Commission under the R4SMEs 7th Framework Program. Author would also like to express his most sincere gratitude to Dr. Tomás Rodríguez, head of unit at Inspiralia Tecnologías Avanzadas, and Dr. Pascual Campoy, Automatics and Robotics Center of Universidad Politécnica de Madrid, for supporting this research with their experience and advice.

## REFERENCES

- Augusteijn M, Folkert B. 2002. Neural network classification and novelty detection. *Int J Remote Sens* 23:2891–2902.
- Barbara D, Couto J, Jajodia S, Wu N. 2001. Detecting novel network intrusions using Bayes estimators. In: *Proceedings of the 1st SIAM International Conference on Data Mining*, Chicago, USA.
- Battiti R. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans Neural Networks* 5:537–550.
- Bishop C. 1994. Novelty detection and neural network validation. In: *IEEE Proceedings on Vision, Image and Signal Processing. Special Issue on Applications of Neural Networks*. pp. 217–222.
- Boucher A, Hidalgo PJ, Thonnat M, Belmonte J, Galan C, Bonton P, Tomczak R. 2002. Development of a semi-automatic system for pollen recognition. *Aerobiologia* 18:195–201.
- Bradley AP. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn* 30:1145–1159.
- Byers S, Raftery A. 1998. Nearest neighbor clutter removal for estimating features in spatial point processes. *J Am Statist Assoc* 93:577–584.
- Carrión P, Cernadas E, Gálvez JF, Damián M, de Sá-Otero P. 2004. Classification of honeybee pollen using a multiscale texture filtering scheme. *Mach Vis Appl* 15:186–193.
- CBI. 2009. CBI market survey: The honey and other bee products market in the EU. CBI, Available at: <http://www.cbi.eu/>.
- Chandola V, Banerjee A, Kumar V. 2009. Anomaly detection: A survey. *ACM Comput Surv* 41:15:1–15:58.
- Chen C, Hendriks EA, Duin RP, Reiber JHC, Hiemstra PS, de Weger LA, Stoel BC. 2006a. Feasibility study on automated recognition of allergenic pollen: Grass, birch and mugwort. *Aerobiologia* 22:275–284.
- Chen X, Zhou X, Wong S. 2006b. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE Trans Biomed Eng* 53:762–766.
- Chica M, Campoy P. 2012. Discernment of bee pollen loads using computer vision and one-class classification techniques. *J Food Eng* 112:50–59.
- Dasarathy BV. 1991. Nearest neighbor (NN) norms: Nearest neighbor pattern classification techniques. Los Alamitos, CA., USA: IEEE Computational Society.

- Eskin E, Arnold A, Prerau M, Portnoy L, Stolfo S. 2002. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. D. Barabási and S. Jajodia, Eds. In: *Applications of Data Mining in Computer Security*. Kluwer Academics, Boston, USA. pp. 78–100.
- France I, Duller AWG, Duller GAT, Lamb HF. 2000. A new approach to automated pollen analysis. *Q Sci Rev* 19:537–546.
- Goh K, Chang E, Cheng K. 2001. SVM binary classifier ensembles for image classification. In: *Proceedings of the ACM Conference Information and Knowledge Management*, New York, USA. pp. 395–402.
- Goh K, Chang E, Li B. 2005. Using one-class and two-class SVMs for multiclass image annotation. *IEEE Trans Knowl Data Eng* 17:1333–1346.
- Guyon I, Elisseeff A. 2003. An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182.
- Jalba A, Wilkinson M, Roerdink J. 2004. Automatic segmentation of diatom images for classification. *Microsc Res Tech* 65:72–85.
- Japkowicz N, Myers C, Gluck M. 1995. A novelty detection approach to classification. In: *Proceedings of International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Francisco, USA. pp. 518–523.
- Kira K, Rendell LA. 1992. A practical approach to feature selection. In: *Proceedings of the 9th International Workshop on Machine Learning*, Morgan Kaufmann, San Francisco, CA, USA. pp. 249–256.
- Landsmeer SH, Hendriks EA, Weger LD, Reiber JH, Stoel BC. 2009. Detection of pollen grains in multifocal optical microscopy images of air samples. *Microsc Res Tech* 72:424–430.
- Li P, Treloar WJ, Flenley JR, Empson L. 2004. Towards automation of palynology 2: The use of texture measures and neural networks analysis for automated identification of optical images of pollen grains. *J Q Sci* 19:755–762.
- Liu H, Motoda H. 1998. *Feature selection for knowledge discovery and data mining*. Boston, USA: Kluwer Academic.
- Malladi R, Sethian J, Vemuri B. 1995. Shape modeling with front propagation: A level set approach. *IEEE Trans Pattern Anal Mach Intell* 17:158–175.
- Moya M, Koch M, Hostetler L. 1993. One-class classifier networks for target recognition applications. In: *Proceedings on World Congress on Neural Networks*, International Neural Network Society, Portland, USA. pp. 797–801.
- Otsu N. 1979. A threshold selection method from grey-level histograms. *IEEE Trans Syst Man Cybern* 9:62–66.
- Phua C, Alahakoon D, Lee V. 2004. Minority report in fraud detection: Classification of skewed data. In: *SIGKDD Explorer Newsletter*, ACM, New York, USA. pp. 50–59.
- Pokrajac D, Lazarevic A, Latecki L. 2007. Incremental local outlier detection for data streams. In: *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining*, Honolulu, Hawaii, USA. pp. 504–515.
- Provost F, Fawcett T. 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In: *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, Huntington Beach, CA, USA. pp. 43–48.
- Ranzato M, Taylor P, House J, Flagan R, LeCun Y, Perona P. 2007. Automatic recognition of biological particles in microscopic images. *Pattern Recogn Lett* 28:31–39.
- Ratsch G, Mika S, Scholkopf B, Muller K. 2002. Constructing boosting algorithms from SVMs: An application to one-class classification. *IEEE Trans Pattern Anal Mach Intell* 24:1184–1199.
- Redondo R, Sroubek F, Fischer S, Cristóbal G. 2009. Multifocus image fusion using the log-Gabor transform and a multisize windows technique. *Inform Fusion* 10:163–171.
- Ritter G, Gallegos M. 1997. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recogn Lett* 18:525–539.
- Rodríguez-Damián M, Cernadas E, Formella A, Fernández-Delgado M, de Sá-Otero P. 2006. Automatic detection and classification of grains of pollen based on shape and texture. *IEEE Trans Syst Man Cybern Part C: Appl Rev* 36:531–542.
- Ronneberger O, Schultz E, Burkhardt H. 2002. Automated pollen recognition using 3D volume images from fluorescence microscopy. *Aerobiologia* 18:107–115.
- Schapire R, Singer Y. 1999. Improved boosting algorithms using confidence-rated predictions. *Mach Learn* 37:297–336.
- Siaterlis C, Maglaris B. 2004. Towards multisensor data fusion for DoS detection. In: *Proceedings of the 2004 ACM Symposium on Applied Computing*, ACM Press, Nicosia, Cyprus. pp. 439–446.
- Taniguchi M, Haft M, Hollmn J, Tresp V. 1998. Fraud detection in communications networks using neural and probabilistic methods. In: *Proceedings of IEEE International Conference in Acoustics, Speech and Signal Processing*, IEEE Computer Society. pp. 1241–1244.
- Tax D. 2001. One-class classification; concept-learning in the absence of counterexamples, Ph.D. thesis, Delft University of Technology.
- Tax D. 2011. DDtools, the data description toolbox for Matlab. Version 1.9.0.
- Tax D, Duin R. 2004. Support vector data description. *Mach Learn* 54:45–66.
- Treloar WJ, Taylor GE, Flenley JR. 2004. Towards automation of palynology 1: Analysis of pollen shape and ornamentation using simple geometric measures, derived from scanning microscope images. *J Q Sci* 19:745–754.
- Tsai Y, Chung I, Simpson J, Lee M, Hsiung C, Chiu T, Kao L, Chiu T, Lin C, Lin W, Liang S, Lin C. 2008. Automated recognition system to classify subcellular protein localizations in images of different cell lines acquired by different imaging systems. *Microsc Res Tech* 71:305–314.
- Tscherepanow M, Zöllner F, Hillebrand M, Kummert F. 2008. Automatic segmentation of unstained living cells in bright-field microscope images. *Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry*. *Lect Notes Comput Sci* 5108:158–172.
- Witten IH, Frank E. 2005. *Data mining: Practical machine learning tools and techniques*, 2nd ed. San Francisco, USA: Morgan Kaufmann.
- Wu Q, Merchant F, Castleman K. 2008. *Microscope image processing*. Academic Press, Elsevier.
- Xu D, Pham J, Prince J. 2000. Image segmentation using deformable models. In: *Handbook of Medical Imaging*, vol. 2. Nicosia, Cyprus: SPI Press. pp. 129–174.
- Zhang Y, Fountain DW, Hodgson RM, Flenley JR, Gunetileke S. 2004. Towards automation of palynology 3: Pollen pattern recognition using Gabor transforms and digital moments. *J Q Sci* 19:763–768.