Computers in Biology and Medicine 83 (2017) 143–150

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Gold-standard for computer-assisted morphological sperm analysis

Violeta Chang[a,b,*], Alejandra Garcia[b], Nancy Hitschfeld[a], Steffen Härtel[b]

[a] Department of Computer Science, University of Chile, Beauchef 851, 3rd Floor, Santiago, RM, Chile
[b] Laboratory for Scientific Image Analysis, SCIAN-Lab, Centro de Espermiograma Digital Asistido por Internet (CEDAI SpA), Centro de Informatica Medica y Telemedicina (CIMT), Centro Nacional en Sistemas de Informacion en Salud (CENS), Biomedical Neuroscience Institute (BNI), Instituo de Ciencias Biomedicas (ICBM), Faculty of Medicine, University of Chile, Av. Independencia 1027, Independencia, RM, Chile

ARTICLE INFO

Keywords:
Infertility
Gold-standard
Morphological sperm analysis
Sperm head classification
Sperm classification base-line

ABSTRACT

Background and Objective: Published algorithms for classification of human sperm heads are based on relatively small image databases that are not open to the public, and thus no direct comparison is available for competing methods. We describe a gold-standard for morphological sperm analysis (SCIAN-MorphoSpermGS), a dataset of sperm head images with expert-classification labels in one of the following classes: normal, tapered, pyriform, small or amorphous. This gold-standard is for evaluating and comparing known techniques and future improvements to present approaches for classification of human sperm heads for semen analysis. Although this paper does not provide a computational tool for morphological sperm analysis, we present a set of experiments for comparing sperm head description and classification common techniques. This classification base-line is aimed to be used as a reference for future improvements to present approaches for human sperm head classification.

Methods: The gold-standard provides a label for each sperm head, which is achieved by majority voting among experts. The classification base-line compares four supervised learning methods (1 − Nearest Neighbor, naive Bayes, decision trees and Support Vector Machine (SVM)) and three shape-based descriptors (Hu moments, Zernike moments and Fourier descriptors), reporting the accuracy and the true positive rate for each experiment. We used Fleiss' Kappa Coefficient to evaluate the inter-expert agreement and Fisher's exact test for inter-expert variability and statistical significant differences between descriptors and learning techniques.

Results: Our results confirm the high degree of inter-expert variability in the morphological sperm analysis. Regarding the classification base line, we show that none of the standard descriptors or classification approaches is best suitable for tackling the problem of sperm head classification. We discovered that the correct classification rate was highly variable when trying to discriminate among non-normal sperm heads. By using the Fourier descriptor and SVM, we achieved the best mean correct classification: only 49%.

Conclusions: We conclude that the SCIAN-MorphoSpermGS will provide a standard tool for evaluation of characterization and classification approaches for human sperm heads. Indeed, there is a clear need for a specific shape-based descriptor for human sperm heads and a specific classification approach to tackle the problem of high variability within subcategories of abnormal sperm cells.

1. Introduction

Up to 15% of couples worldwide are affected by infertility [1]. In the evaluation of the male factor, the first step consists of a semen analysis according to standard criteria [2] that sets the basis for possible medical treatment of the couple [3]. The morphology of the sperm cells is useful to illustrate the potential fertility of a sample [3] and to make a decision about infertility treatment [4].

Sperm morphology reflects different kinds of anomalies in human semen samples. Depending on the anomalies, abnormal sperm cells generally have a lower fertilizing potential and may also have abnormal DNA [2]. The categories of defects include head, neck and mid-piece, tail defects, as well as excess residual cytoplasm (see Fig. 1).

As a result of morphological semen analysis, all the sperm cells in the semen sample are classified as *normal* or *abnormal* [5]. Many studies have demonstrated the close relationship between fertility and morphologically normal sperm [6–10]. The morphology is considered a clinical tool dedicated to the fertility prognosis and serves as a way of making decisions regarding the options of assisted reproduction technologies [4]. In addition to a rigorous application of existing

* Corresponding author at: Laboratory for Scientific Image Analysis, SCIAN-Lab, University of Chile, Av. Independencia 1027, Independencia, RM, Chile.
E-mail addresses: vchang@dcc.uchile.cl (V. Chang), al_garcia@med.uchile.cl (A. Garcia), nancy@dcc.uchile.cl (N. Hitschfeld), shartel@med.uchile.cl (S. Härtel).

http://dx.doi.org/10.1016/j.compbiomed.2017.03.004
Received 10 November 2016; Received in revised form 31 January 2017; Accepted 1 March 2017

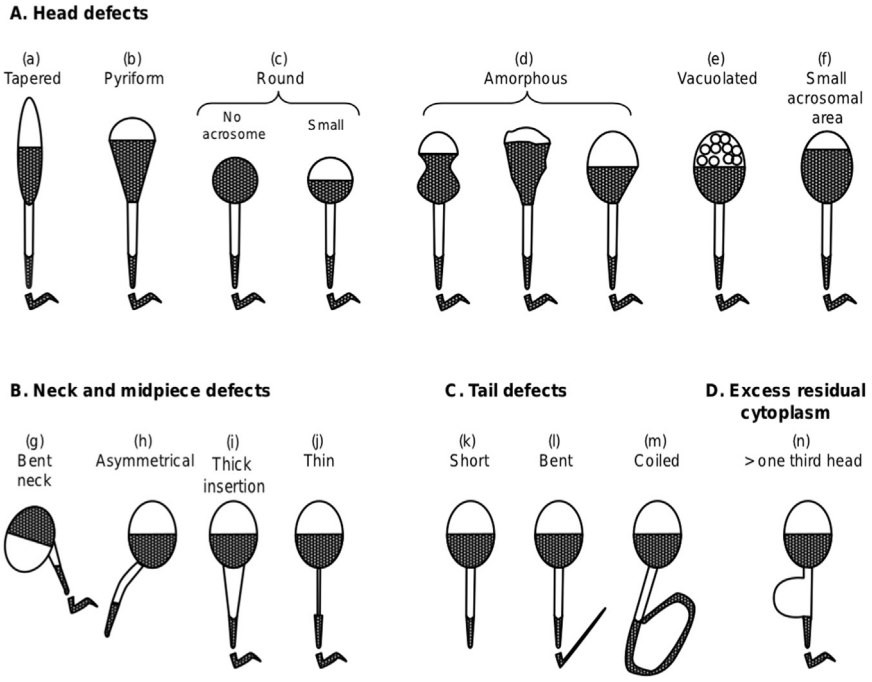

Fig. 1. Human sperm abnormalities. Image reproduced exactly as appears in [2], showing schematic drawings of some abnormal forms of human sperm.

guidelines and respect to high laboratory standards [11], emphasis on identifying the categories of abnormal sperm heads, and morphological patterns of sperm heads may have significant clinical utility when deciding on an infertility treatment. For example, tapered heads may be due to stress caused by a male urogenital tract infection [12]. There is also clinical significance regarding the shape of sperm heads, and each class is associated with different genetic and environmental factors that impact clinical decisions pertaining to an infertility treatment [12]. A complete analysis of normal and abnormal sperm cells therefore turns out to be critical.

The inherent lack of objectivity in the evaluation of human sperm morphology, the difficulty in standardizing, implementing and controlling manual methods, and the high degree of variation within and between laboratories and technicians have fueled the computer-assisted sperm morphology assessment [13]. Despite decades of research on computer-assisted morphological sperm analysis [14,15,3,16,17], there are still no standard ways of comparing the results achieved with different methods. The results of sperm morphology assessment methods are usually evaluated according to how well they correlate with expert-generated classifications, though it seems that each research group has its own dataset of images, staining protocol, and evaluation metrics. Until today, no publications existed that present results of test comparing approaches using the same data.

A ground-truth represents the absolute truth for a certain application. For instance, in cancer detection from medical images, a suspicious region is malignant or benign. The absolute truth (whether there is cancer or not) can be obtained from biopsies and an appropriate staining. These biopsy results constitute the ground-truth for those medical images [18]. Unfortunately, for morphological analysis of sperm cells, it is impossible to count with a ground-truth because of the subjectivity of the task [19]. A valid alternative consists of asking many experts in the field for their opinion about specific cases to generate a gold-standard [20].

In this paper, we introduce and describe the SCIAN Gold-standard for Morphological Sperm Analysis (SCIAN-MorphoSpermGS), a dataset of sperm head images with expert-classification labels. The dataset contains 1854 sperm head images obtained from six semen smears and classified by three Chilean referent domain experts according to World Health Organization (WHO) criteria [2], in one of the following classes: normal, tapered, pyriform, small and amorphous. This gold-standard is aimed for use in evaluating and comparing not only known techniques, but also future improvements to present approaches for classification of human sperm heads for semen analysis. This is a very significant contribution to the scientific community, because at present there is no public gold-standard for human sperm head classification, so the few existing methods cannot be properly evaluated and compared. To show the usability of the proposed gold-standard, we conducted experiments to define a five-class classification base-line.

This paper is organized as follows. In Section 2 we review the research work in the area, justifying the need for a gold-standard for classification of human sperm heads. Section 3 is devoted to describing in detail the staining method, the features of the equipment we used to capture the images and specific details about image sources, as well as the description and analysis of our proposed gold-standard. In Section 4 we briefly discuss the common shape-based descriptors and supervised classification techniques used for the construction of the classification base-line, and we present the results of applying those descriptors and classifiers to the SCIAN-MorphoSpermGS. The summary and conclusions can be found in Section 5.

2. Related work

The importance of having an image database containing ground-truth labelings has been well-demonstrated in many applications of computer vision: hand-writing recognition [21], face recognition [22], indoor/outdoor scene classification [23] and mammal classification [24]. As said before, a ground-truth represents the absolute truth for a certain application that is not always available or costly. Unfortunately, for many applications, especially in biomedicine, it is impossible to have a ground-truth and a valid alternative consists of asking experts in the field for their opinion about specific cases, in order to generate a gold-standard [20]. The need for a gold-standard in biomedical applications has been demostrated in PAP-smear classification [25], human sperm segmentation [26], and sub-celullar structures classification [27,28], among others.

No gold-standards are available for morphological sperm analysis. Instead, several research groups have independently gathered sperm smear images and run different sets of tests, with different performance

Table 1
Summary of previous databases for morphological sperm analysis.

Publication	Number of classes	Number of images or samples	Image size	Source
Yi et al. 1998 [29]	4	300–360	640×480	Andrology Clinic, Seoul National University Hospital
Ramos et al. 2002 [30]	2	590	not reported	Division of Reproductive Medicine, University Medical Center St. Radboud Nijmegen, The Netherlands
Soler et al. 2003 [13]	*	5 samples	262×144	Institute of Reproductive Medicine of the University, Münster, Germany
Abbiramy et al. 2011 [31]	2	91	86×100	World Health Organization, 5th Laboratory Manual for the Examination and Processing of Human Semen
Aksoy et al. 2012 [32]	*	67 samples	not reported	Assisted Reproductive Techniques Unit, Faculty of Meram Medicine, University of Selcuk, Turkey
Lammers et al. 2014 [33]	*	250 samples	not reported	Andrology Laboratory, Service de Médecine et Biologie de la Reproduction, University Hospital of Nantes, France
Ghasemian et al. 2015 [34]	2	1457 samples	576×764	Infertility Therapy Center, Alzahra Educational and Remedial Center, Guilan, Iran

* Oriented to assessment of morphology parameters, not sperm classification.

measures. In Table 1, we list several sperm image datasets currently used in publications on morphological sperm analysis.

3. Gold-standard

3.1. Sample preparation

Sperm samples were stained with a modified Hematoxylin/Eosin procedure, in order to distinguish different parts of the sperm cell. First, the sperm smear was fixed with ethanol 70% and immersed in Harris' Hematoxylin for ten seconds for nuclear staining. To remove residual staining, slides were washed with tap water for ten minutes. Then, slides were immersed in 1% Eosin for two minutes to stain the acrosome in a pink-orange color, mid-piece and tail. Finally, the sample was washed with distilled water for one minute and air-dried. This staining procedure allows samples to be used for more than one year and is the most commonly used staining protocol in clinical laboratories to enhance morphologic characteristics of sperm heads, suggested by WHO [2].

3.2. Image acquisition

We use optical, bright field microscopy (Axiostar Plus, Carl Zeiss Inc, Wetzlar, Germany), a 63x objective (oil, NA 1.4) with an adapter of 0.63x and a digital camera (scA780-54gc, Basler AG, Ahrensburg, Germany) to acquire digital images.

Bright field microscopy was used, because it represents the most common method to acquire images of sperm cells in sufficient detail. In comparison to alternative techniques, bright field microscopy is cheap, easy to use, and offers reproducible conditions for the observation of sperm head morphology. The generation of the gold-standard under standard conditions sets the basis for the direct comparison of different algorithms that can improve the analysis of sperm characteristics in clinical practice.

Besides bright field microscopy, there are alternative techniques that enhance spermatozoa head structures, as Memmolo et al discussed in [35]. Since unstained spermatozoa are essentially transparent under a bright field microscope, alternative acquisition techniques such as phase contrast microscopy [36], Differential Interference Contrast (DIC) microscopy, also called Nomarski Interference Contrast (NIC), or Digital Holographic (DH) microscopy can be used. Phase contrast microscopy, DIC, and DH bear the advantage to enhance contrast of internal head structures without altering the sample through staining, labeling, or physical stress, and can therefore be used as in vivo techniques for sperm analysis. [35] propose DH microscopy for computer-assisted sperm head morphometry and compare two different techniques to identify and measure the region of spermatozoon heads. In Merola et al. [37], the authors suggest high-throughput analysis of label-free microfluidic based cytofluorimeters for biovolume estimation of bovine sperm morphology (length, width and height) and prognostic examination.

3.3. Source of sperm smears

We obtained semen smears from volunteers between 28 and 35 years old at the Laboratory of Spermiogram, Program of Anatomy and Developmental Biology (ICBM), Faculty of Medicine, University of Chile, Santiago, Chile.

3.4. Gold-standard description

We collected semen smears containing 1872 sperm head images that could be classified according to 11 head defects as WHO defines [2]. We decided to build the gold-standard with 1854 observable and evaluable sperm cells whose class was one of the following: Normal, Tapered, Pyriform, Small, or Amorphous. Fig. 2 shows representative sperm cells from each class. The manual classification process was performed independently, per patient/smear, by three referent Chilean experts with vast experience in morphological sperm analysis.

3.5. Analysis and discussion

A very important aspect in the analysis of the gold-standard is the discussion of the inter-expert agreement distribution. As this gold-standard was built with the cooperation of three experts, there are three different agreement scenarios: one (basis set), two experts (partial agreement - PA), or three experts agree on the same label for a given sperm head (total agreement - TA). The first set contains 1854 sperm head labels, but a sperm head can be classified into three different classes by the three different experts. The second set contains 1132 sperm heads, meaning that there are 1132 sperm heads with partial agreement and without overlapping. The third set contains only 384 sperm heads, with total agreement between the three expert technicians.

Table 2 shows the number of sperm cells per class for each agreement scenario. Considering the manual classification agreement by at least one, two, or three experts, the class Amorphous was the largest class in all cases, concentrating over 68% in the total agreement scenario (see Fig. 3). The class Tapered was the second largest class, slightly decreasing assignment percentage as agreement among experts increases and concentrating around a fifth of the samples. The Pyriform and Small classes decreased their assignment percentage while increasing the agreement among experts (from 188 to 7 and from 152 to 11 sperm cells, respectively). It is important to note that in the case of class Pyriform, less than 2% of the samples had total agreement of the experts (only 7 sperm cells). The only class that maintained its assignment percentage, without statistical differences between agreement scenarios, was the class Normal consisting of less than 10% of the

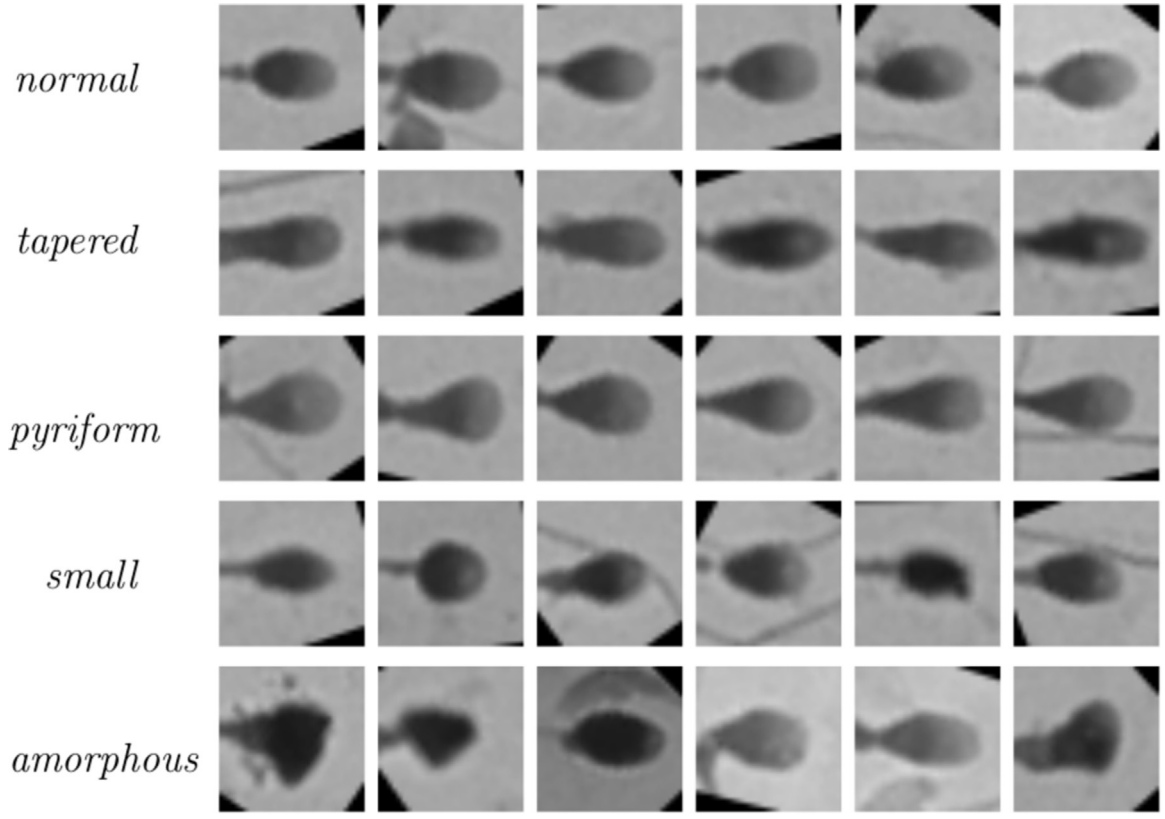

Fig. 2. Classification gold-standard. Representative images of normal, tapered, pyriform, small and amorphous sperm cells that showed total agreement (TA) among experts (Image size: 35×35 pixels ≈7×7 μm).

Table 2
Inter-expert agreement.

Agreement among experts	Normal	Tapered	Pyriform	Small	Amorphous	Total
At least one (Basis set)	175	420	188	152	919	1854
Partial agreement (PA)	100	228	76	72	656	1132
Total agreement (TA)	35	69	7	11	262	384

samples. Statistical differences between agreement scenarios in each class were evaluated by Z-test and considered significant at $p < 0.05$.

The underlying complexity of the sperm head classification task can be studied by evaluating the degree of agreement between different experts. Fig. 4 shows the inter-expert agreement per class. Considering the number of heads that had been categorized as a certain class by at least one expert as the size of the basis set, Fig. 4 shows the percentage of partial agreement and total agreement. There are some classes in which it was really difficult to reach an agreement, for example, class Pyriform. Only 40% of sperm heads that were classified as Pyriform by at least one expert reached partial agreement and less than 4% that reach total agreement. It seemed to be the most difficult class from which to find agreement among experts. While the most morphologically ambiguous class in theory, Amorphous, turned out to be the class

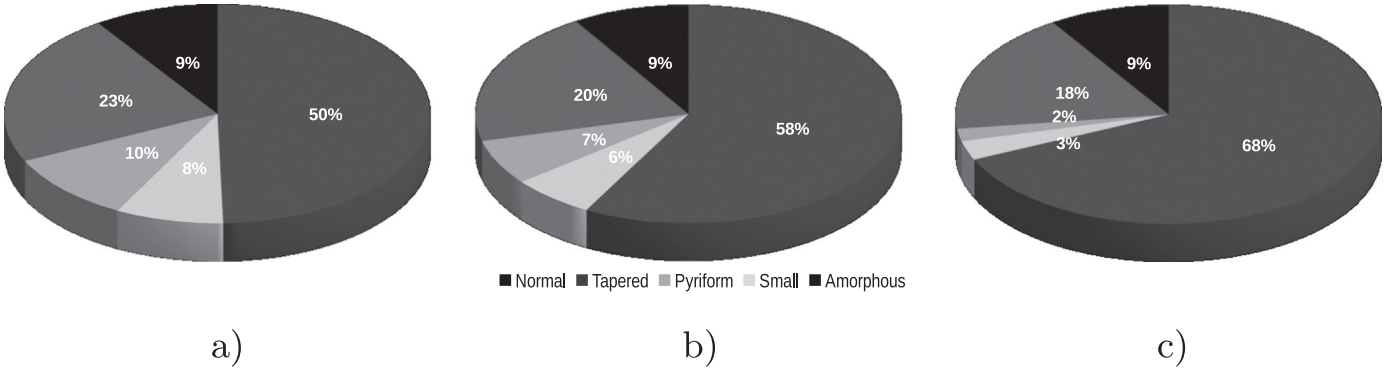

Fig. 3. Inter-expert agreement. (a) Manual classification by at least one expert assigning a class label Amorphous amounts to 50% with similar presence of classes Normal, Pyriform and Small (around 10%). (b) For partial agreement, the class Amorphous is the biggest class (almost 60%), while classes Tapered, Small and Pyriform sightly decrease. (c) For total agreement, the class Amorphous amounts almost 70%, while Pyriform covers almost 2%. The only class that maintains its assignment percentage is the class Normal, regardless if one considers the label agreement of at least one, two, or three experts.

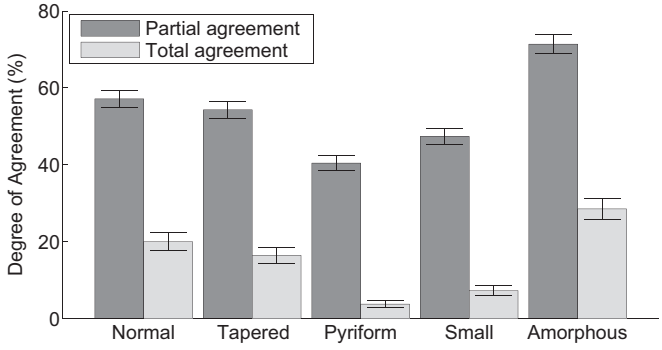

Fig. 4. Partial and total inter-expert agreement. For each class, we show the percentage ± SE of partial and total agreement among experts normalized by the size of the basis set.

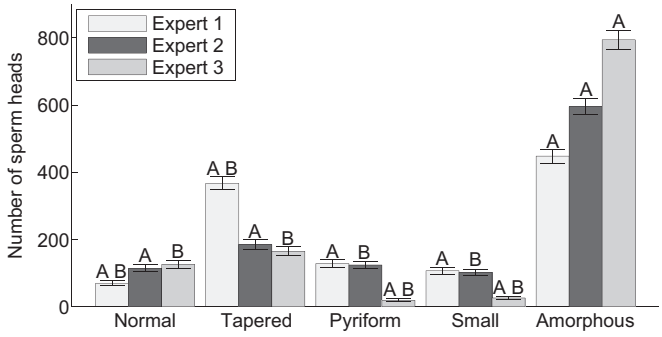

Fig. 5. Inter-expert variability in five-class classification. For research purposes, the expert classifies sperm heads in a number of classes (five in our case: Normal, Tapered, Pyriform, Small and Amorphous). We show the number of sperm heads ± SE that belong to each class according to each of the three experts. Statistical differences between experts in each class were evaluated by Fisher's exact test and considered significant at $p < 0.05$ and indicated as A and B. For instance, for class Pyriform, the A in Expert1 and Expert3 means that there is a significant difference between those experts. The expert manual classification shows a fair agreement among experts with Fleiss' Kappa coefficient of 0.36($\alpha = 0.05$).

that had the greatest agreement among experts in both agreement scenarios, partial and total agreement.

To demonstrate the subjectivity of morphological analysis and dependence of the specialist who performs it, Fig. 5 shows inter-expert variability per class. Pyriform and Small were the most defined classes according to their morphological features, and both showed a high degree of agreement between two of the three experts, while the discrepancy with the third expert was really significant. In the case of classes Normal and Tapered, a high degree of agreement was reached between two technicians, whereas the discrepancy with the remaining expert was very high in the case of class Tapered. Class Amorphous showed a high degree of variability among all experts. In general, the inter-expert variability analysis showed 60% of pairwise expert agreement. We calculated the Fleiss' Kappa coefficient [38] as a way of measuring the inter-expert agreement, and it showed a fair degree of agreement, with a coefficient of 0.36($\alpha = 0.05$). Furthermore, assuming that a semen analysis would provide the percentage of sperm heads in each of the five selected classes, Table 3 shows how the seminogram would be quantified by each expert considered in this study, working with the partial agreement data set.

4. Classification base-line

4.1. Feature extraction

Feature extraction is one of the basic steps in a classification process and consists of quantifying properties of the objects derived from the segmented regions of interest (ROI). For instance, an object defined by a ROI can be described in terms of its shape, texture, and color, among other features [39]. To describe the content of an image semantically, shape-based descriptions have been proven to be much more effective than other descriptions, such as those based on texture or color [40]. However, when invariance with respect to the number of possible transformations such as scaling, shifting, and rotation is required, the construction of shape-based descriptors is more complicated [41].

Shape-based descriptors are categorized as contour-based shape descriptors and region-based shape descriptors. The contour-based ones exploit information just at the boundary points focusing on contour features, which is critical for human perception of shapes [42]. Region-based shape descriptors are useful for describing non-connected and disjointed ROIs because they combine information across an entire object, so they can capture the interior content of an object defined by a ROI [42]. The Fourier descriptor is an example of a contour-based descriptor, while image moments are an example of a region-based descriptor [43]. In this work, we conducted experiments with three different shape-based descriptors: Hu moments [44], Zernike moments [45] and Fourier descriptors [46].

4.2. Classification methods

Classification is the process that assigns objects to a set of classes. There are many approaches used for classification purposes and are categorized as supervised and unsupervised methods. Supervised classification techniques involve the participation of an expert who is responsible for teaching the classifier with examples. After training, the classifier is expected to classify similar objects, that are previously unseen, to the correct classes. A classification paradigm uses a set of training examples of the form $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ for the projection of the function $f(x)$. The values x are usually vectors of real or discrete values of the form $<x_{i1}, x_{i2}, \ldots, x_{in}>$. The values y are the expected outputs for given x values, and usually obtained from a discrete set of classes. Consequently, the task of a learning paradigm involves the approximation of a function $f(x)$ to produce a classifier. In this work, we used four supervised classification techniques that have demonstrated their suitability while coping with classification problems in different domains: $K-$ Nearest Neighbors [47], Naive-Bayes [48], Decision Trees [49] and Support Vector Machines [50].

4.3. Performance measures

By taking advantage of having a classification gold-standard, it is possible to evaluate the performance of a classification method against

Table 3
Inter-expert variability (including standard error).

	Normal%	Tapered%	Pyriform%	Small%	Amorphous%	Other%
Expert1	6.2 ± 0.7	32.4 ± 1.7	11.3 ± 1.0	9.5 ± 0.9	39.6 ± 1.9	1.1 ± 0.3
Expert2	10.1 ± 1.0	16.3 ± 1.2	11.0 ± 1.0	9.0 ± 0.9	52.6 ± 2.1	1.1 ± 0.3
Expert3	11.1 ± 1.0	14.6 ± 1.1	1.7 ± 0.4	2.3 ± 0.4	70.1 ± 2.5	0.2 ± 0.1

Table 4
Partial agreement dataset partition. $PA - Dataset = PA - DS1 \cup PA - DS2 \cup PA - DS3$.

	PA-Dataset	PA-DS1	PA-DS2	PA-DS3
Number of Normal sperm heads	100	60	20	20
Number of Tapered sperm heads	228	137	46	45
Number of Pyriform sperm heads	76	45	15	16
Number of Small sperm heads	72	44	14	14
Number of Amorphous sperm heads	656	394	131	131
Total number of sperm heads	1132	680	226	226

that gold-standard. To this end, it is important to realize that even though there are a number of performance measures proposed in the literature, the evaluation of a classification method can be measured by using simple metrics such as accuracy and True Positive Rate (TPR) - both of them based on the information provided by the confusion matrix [51].

4.4. Dataset partition

For the experimental results that we show in this section, we used the classification gold-standard introduced in Section 3.4. We conducted experiments using the partial agreement dataset (PA-Dataset) with 1132 sperm heads without overlapping and distributed in five classes. Even though this is a very challenging problem, this paper does not aim to provide a computational tool for morphological sperm analysis but provides a public gold-standard to evaluate and compare algorithms for classification of sperm heads. On that basis, we decided to perform experiments using the PA-Dataset because of the reduced size of data in the total agreement dataset.

The dataset was partitioned in three subsets, named *Dataset 1 (DS1), Dataset 2 (DS2) and Dataset 3 (DS3)*, aiming to comprise a training (60% of the whole dataset), validating (20%) and testing (20%) dataset, respectively. In Table 4, the size and distribution of classes in each partition are presented.

4.5. Experimental results

In our experiments, we measured the accuracy of sperm head classification in five classes: Normal (N), Tapered (T), Pyriform (P), Small (S) and Amorphous (A). We computed three different and independent feature vectors using: 1) Hu moments, 2) Zernike moments, and 3) Fourier descriptors. With respect to the classification method, we used four common supervised learning techniques: 1) 1 – NN, 2) naive Bayes, 3) decision trees, and 4) SVM. We used DS1 as the training dataset and DS3 as the testing dataset. For training purposes, we balanced training data by randomly taking the same number of samples in each class. We did 100 runs for each feature extraction-classification combination. Tables 5–7, show the accuracy per class by using four supervised learning techniques and different shape-based descriptors using the PA-Dataset.

As shown in Tables 5–7, the best accuracy in the five-class classification scenario was achieved when using Fourier descriptors with SVM. In this case, 49% was the achieved correct classification regarding all classes, but achieving only 15% of correctly classified amorphous sperm heads: this demonstrates the difficulty of this task. Analyzing the performance of the different descriptors evaluated in this paper, we can conclude that neither descriptor nor classifier performs as the best for all classes. The Hu moments descriptor achieved the lowest correct classification which ranged from 33% to 46%, while Zernike moments descriptor ranged from 40% to 46%, slightly outperformed by Fourier descriptors, which achieved from 40% to 49% correct classification. From Table 5, we see an obvious inverse behavior of classes Normal and Small. While using Hu moments and SVM, normal sperm heads could be correctly classified in 100% of cases, small sperm heads only achieved 4% of correct classification. On the other hand, while using Hu moments and the naive Bayes classifier, 92% of correct classification of small sperm heads was opposed by the 1% of correct classification of normal sperm heads. With respect to the difficulty in each class, experimental results in the previous Tables confirmed that the amorphous class was the most difficult to classify: the classification accuracy ranged from 7% to 25%, without regarding the descriptor or classifier used. In Fig. 6 we show a graphical representation of the comparison of performance results of classifiers.

Table 5
Five-class base-line classification using Hu moments. *tpr* stands for True Positive Rate, *acc* stands for accuracy understood as the mean of tpr of classes Normal, Tapered, Pyriform, Small and Amorphous.

Classifier	tpr (N)	tpr (T)	tpr (P)	tpr (S)	tpr (A)	acc
1 – NN	0.20	0.50	0.49	0.54	0.21	0.39
Bayes	0.01	0.29	0.32	0.92	0.15	0.33
Decision trees	0.33	0.52	0.53	0.35	0.22	0.39
SVM	1.00	0.70	0.47	0.04	0.10	0.46

Table 6
Five-class base-line classification using Zernike moments. *tpr* stands for True Positive Rate, *acc* stands for accuracy understood as the mean of tpr of classes Normal, Tapered, Pyriform, Small and Amorphous.

Classifier	tpr (N)	tpr (T)	tpr (P)	tpr (S)	tpr (A)	acc
1 – NN	0.34	0.47	0.36	0.62	0.20	0.40
Bayes	0.92	0.37	0.15	0.73	0.07	0.45
Decision trees	0.40	0.44	0.37	0.61	0.25	0.41
SVM	0.44	0.62	0.33	0.70	0.23	0.46

Table 7
Five-class base-line classification using Fourier descriptors. *tpr* stands for True Positive Rate, *acc* stands for accuracy understood as the mean of tpr of classes Normal, Tapered, Pyriform, Small and Amorphous.

Classifier	tpr (N)	tpr (T)	tpr (P)	tpr (S)	tpr (A)	acc
1 – NN	0.50	0.49	0.36	0.44	0.20	0.40
Bayes	0.31	0.45	0.58	0.78	0.10	0.44
Decision trees	0.44	0.51	0.35	0.45	0.25	0.40
SVM	0.57	0.68	0.53	0.54	0.15	0.49

5. Summary and conclusions

To tackle the problem of lacking a gold-standard for evaluating sperm head classification methods, we have introduced the SCIAN-MorphoSpermGS. We built our gold-standard with images from Chilean laboratories following a staining protocol with Hematoxylin/Eosin. It was built with the cooperation of three experts and consisted of 1854 sperm head images. Each sperm head was manually classified by each expert in one of the following classes: Normal, Tapered, Pyriform, Small or Amorphous. This gold-standard is for evaluating and comparing not only known techniques, but also for future improvements to present approaches for classification of human sperm heads. The second contribution of this paper is a classification base-line for comparison of classification results regarding four supervised learning methods (1 – NN, naive Bayes, decision trees and SVM) and three different shape-based descriptors (Hu moments, Zernike moments and Fourier descriptors). This classification base-line is for use as a reference for future improvements to present approaches for human sperm head classification. This base-line demonstrates the utility of the gold-standard by exploring the difference in classification results between the most promising shape-based descriptors and learning techniques, and shows that there is great room for improvement in specific classification approaches for human sperm heads.

We conducted experiments that allow us to conclude that neither

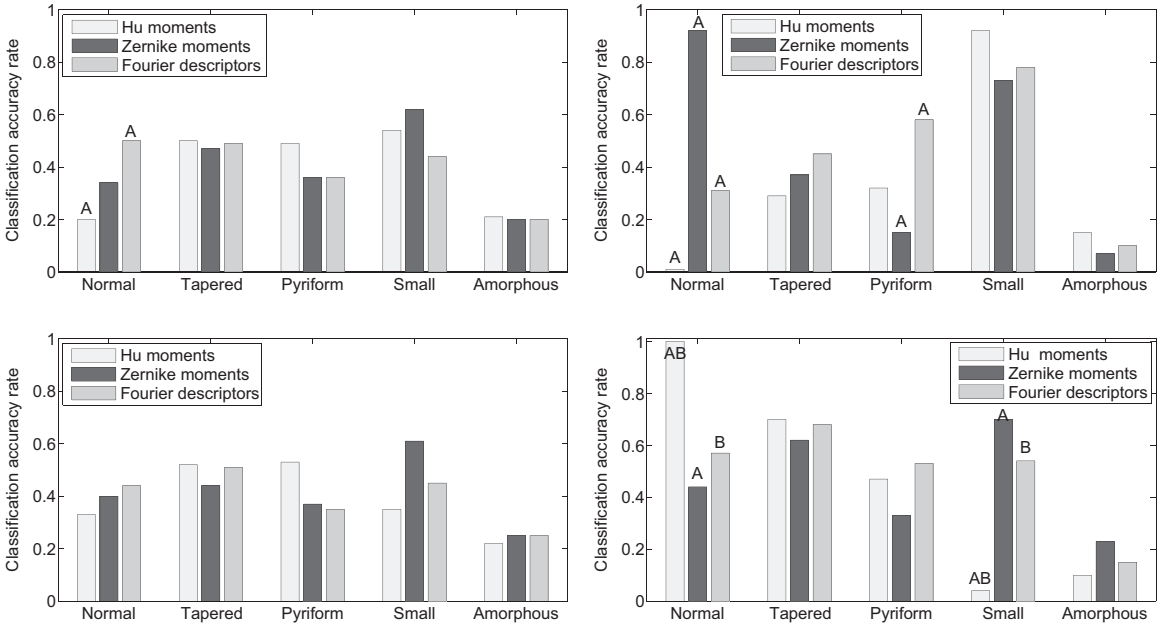

Fig. 6. Five-class base-line classification. We compare the results obtained using four different classification approaches: 1 − NN (top left), naive Bayes (top right), decision trees (bottom left) and SVM (bottom right). Statistical differences between descriptors in each class were considered significant at $p < 0.05$ and indicated as A and B. For instance, in fourth row, for class Normal, the A in Hu moments and Zernike moments means that there is a significant difference between those descriptors, while the B in Hu moments and Fourier descriptors means that there is a significant difference between those descriptors.

standard descriptor nor classification approach is best suited to tackle the problem of sperm head classification. In particular, we discovered that the correct classification rate is highly variable when trying to discriminate among abnormal sperm head. By using the Fourier descriptor and SVM, we achieved the best mean rate: 49% of correct classification. We conclude that there is a need for a specific shape-based descriptor for human sperm heads and a specific classification approach to tackle the problem of high variability within subcategories of abnormal sperm cells. In this sense, it is only possible to understand the differences between shape-based descriptors and classification approaches by looking at the results of several experiments aimed to test specific properties, as was one of the goals of this paper, along with the introduction of the gold-standard. The SCIAN-MorphoSpermGS provides a standard tool for this type of experimentation.

This paper suggests several directions for future research. First, the gold-standard should be extended to consider a larger number of domain experts and to take into account the expertise level of each domain expert. We plan to use multi-label classification approaches for consensus labeling from experts rather than simple majority voting. Second, we plan to design a particular shape-based descriptor with this application in mind, given that with three of the most promising shape-based descriptors there was great difficulty in characterizing human sperm heads towards an accurate morphological classification. Finally, as four general supervised learning approaches showed the need for an *ad-hoc* classification approach, we plan to explore the design and development of a classification scheme for human sperm heads that takes advantage of a combination of classifiers. We believe that the gold-standard described in this paper provides a solid infrastructure for continued research in these directions.

Conflicts of interest

We certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript titled Gold-standard for computer-assisted morphological sperm analysis.

Acknowledgements

The authors thank L. Sarabia, A. Acosta and F. Horta for manual classification labels of the gold-standard. The authors thank V. Castañeda for support with the web-based tool for expert labeling. Violeta Chang was partially funded by CONICYT (NAC-DoctoradoLatin 57090057/NAC-ApoyoTesis 24100118) and FONDECYT (Postdoctorado 3160559). Alejandra Garcia was partially funded by CORFO (SSAF 27061-7). Research in SCIAN-Lab (S. Härtel) is funded by FONDECYT (1151029), EQM140119, CONICYT (PIA ACT 1402), CORFO (16CTTS-66390) and the Biomedical Neuroscience Institute (BNI, ICM P09-015-F). SCIAN-Lab is a selected member of the German-Chilean Center of Excellence Initiative (DAAD 57220037 and 57168868).

References

[1] WHO, Mother or nothing: the agony of infertility, World Health Organization Bulletin 88 (12) (2010) 881–882.
[2] WHO, World Health Organization - Laboratory Manual for the Examination and Processing of Human Semen, 5th Edition, World Health Organization, 2010.
[3] D. Katz, J. Overstreet, S. Samuels, P. Niswander, T. Bloom, E. Lewis, Morphometric analysis of spermatozoa in the assessment of human male fertility, J. Androl. 7 (4) (1986) 203–210.
[4] R.K.-K. Lee, J.-W. Hou, H.-Y. Ho, Y.-M. Hwu, M.-H. Lin, Y.-C. Tsai, J.-T. Su, Sperm morphology analysis using strict criteria as a prognostic factor in intrauterine insemination, Int. J. Androl. 25 (5) (2002) 277–280.
[5] J. Auger, Assessing human sperm morphology: top models, underdogs or bio-metrics?, Assian J. Androl. 12 (1) (2010) 36–46.
[6] G. Moench, H. Holt, Sperm morphology in relation to fertility, Am. J. Obstet. Gynecol. 22 (1) (1931) 199–210.
[7] J. MacLeod, R. Gold, The male factor in fertility and infertility - sperm morphology in fertile and infertile marriage, Fertil. Steril. 2 (1) (1951) 394–414.
[8] T. Kruger, R. Menkveld, F. Stander, C. Lombard, J. van der Merwe, J. van Zyl, K. Smith, Sperm morphologic features as a prognostic factor in in-vitro fertilization, Fertil. Steril. 46 (6) (1986) 1118–1123.
[9] M. Enginsu, J. Dumoulin, M. Pieters, M. Bras, J. Evers, J. Geraedts, Evaluation of human sperm morphology using strict criteria after diff-quik staining: correlation of morphology with fertilization in vitro, Hum. Reprod. 6 (6) (1991) 854–858.
[10] T. Kobayashi, M. Jinno, K. Sugimura, S. Nozawa, T. Sugiyama, E. Iida, Sperm morphological assessment based on strict criteria and in-vitro fertilization out-come, Hum. Reprod. 6 (7) (1991) 983–986.
[11] R. Menkveld, Sperm morphology assessment using strict (tygerberg) criteria, in: Spermatogenesis, , vol. 927 of Methods in Molecular Biology, 2013, pp. 39–50.
[12] R. Menkveld, C. Holleboom, J. Rhemrev, Measurement and significance of sperm morphology, Asian J. Androl. 13 (1) (2011) 59–68.

[13] C. Soler, J. de Monserrat, R. Gutiérrez, J. Nuñez, M. Nuñez, M. Sancho, F. Pérez-Sánchez, T. Cooper, Use of the sperm-class analyzer for objective assessment of human sperm morphology, Int. J. Androl. 26 (5) (2003) 262–270.
[14] A. Schmassmann, G. Mikuz, G. Bartsch, H. Rohr, Spermiometrics: objective and reproducible methods for evaluating sperm morphology, Eur. Urol. 8 (5) (1982) 274–279.
[15] J. Jagoe, N. Washbrook, E. Hudson, Morphometry of spermatozoa using semi-automatic image analysis, J. Clin. Pathol. 39 (12) (1986) 1347–1352.
[16] J. Moruzzi, A. Wyrobek, B. Mayall, B. Gledhill, Quantification and classification of human sperm morphology by computer-assisted image analysis, Fertil. Steril. 50 (1) (1988) 142–152.
[17] F. Pérez-Sánchez, J. de Monserrat, C. Soler, Morphometric analysis of human sperm morphology, Int. J. Androl. 17 (5) (1994) 248–255.
[18] V.C. Raykar, S. Yu, L.H. Zhao, A. Jerebko, C. Florin, G.H. Valadez, L. Bogoni, L. Moy, Supervised learning from multiple experts: Whom to trust when everyone lies a bit, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, 2009, pp. 889–896.
[19] Y. Yan, R. Rosales, G. Fung, M.W. Schmidt, G.H. Valadez, L. Bogoni, L. Moy, J.G. Dy, Modeling annotator expertise: Learning when everybody knows a bit of something, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10), , vol. 9, 2010, pp. 932–939.
[20] T. Fuchs, J. Buhmann, Computational pathology: challenges and promises for tissue analysis, Comput. Med. Imaging Graph. 35 (7–8) (2011) 515–530.
[21] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[22] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1615–1618.
[23] A. Payne, S. Singh, A benchmark for indoor/outdoor scene classification, in: Pattern Recognition and Image Analysis, Vol. 3687 of Lecture Notes in Computer Science, 2005, pp. 711–718.
[24] M. Fink, S. Ullman, From aardvark to zorro: a benchmark for mammal image classification, Int. J. Comput. Vis. 77 (1–3) (2008) 143–156.
[25] J. Jantzen, J. Norup, G. Dounias, B. Bjerregaard, PAP-smear benchmark data for pattern classification, in: Proceedings of Nature inspired Smart Information Systems (NiSIS 2005), 2005, pp. 1–9.
[26] V. Chang, J. Saavedra, V. Castañeda, L. Sarabia, N. Hitschfeld, S. Härtel, Gold-standard and improved framework for sperm head segmentation, Comput. Methods Prog. Biomed. 117 (2) (2014) 225–237.
[27] M. Boland, M. Markey, R. Murphy, Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images, Cytometry 33 (3) (1998) 366–375.
[28] M. Boland, R. Murphy, A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells, Bioinformatics 17 (12) (2001) 1213–1223.
[29] W. Yi, K. Park, J. Paick, Parameterized characterization of elliptic sperm heads using Fourier representation and wavelet transform, in: Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, , vol. 2, 1998, pp. 974–977.
[30] L. Ramos, J. Hendriks, P. Peelen, D. Braat, A. Wetzels, Use of computerized karyometric image analysis for evaluation of human spermatozoa, J. Androl. 23 (6) (2002) 882–888.
[31] V. Abbiramy, A. Tamilarasi, A comparative study on human spermatozoa images classification with artificial neural network based on FOS, GLCM and morphological features, in: Advances in Digital Image Processing and Information. Technology, Vol. 205 of Communications in Computer and Information Science, Springer Berlin Heidelberg, 2011, pp. 220–228.
[32] E. Aksoy, T. Aktan, S. Duman, G. Cuce, Assessment of spermatozoa morphology under light microscopy with different histologic stains and comparison of morphometric measurements, Int. J. Morphol. 30 (4) (2012) 1544–1550.
[33] J. Lammers, C. Splingart, P. Barrière, M. Jean, T. Fréour, Double-blind prospective study comparing two automated sperm analyzers versus manual semen assessment, J. Assist. Reprod. Genet. 31 (1) (2014) 35–43.
[34] F. Ghasemian, S.A. Mirroshandel, S. Monji-Azad, M. Azarnia, Z. Zahiri, An efficient method for automatic morphological abnormality detection from human sperm images, Comput. Methods Prog. Biomed. 122 (3) (2015) 409–420.
[35] P. Memmolo, G.D. Caprio, C. Distante, M. Paturzo, R. Puglisi, Identification of bovine sperm head for morphometry analysis in quantitative phase-contrast holographic microscopy, Opt. Express 23 (2011) 23215–23226.
[36] F. Zernike, How i discovered phase contrast, Science 121 (3141) (1955) 345–349.
[37] M.F, M.L, M.P, D.C.G, G.A, P.R, B.D, C.G, N.P, P. Ferraroa, Digital holography as a method for 3d imaging and estimating the biovolume of motile cells, Lab on a Chip 23 (13) (2013) 45124516.
[38] J. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (5) (1971) 378–382.
[39] L. Costa, R. Cesar, Shape Classification and Analysis: Theory and Practice, 2nd edition, CRC Press, 2009.
[40] E. Persoon, K.-S. Fu, Shape discrimination using fourier descriptors, IEEE Trans. Syst., Man Cybern. 7 (3) (1977) 170–179.
[41] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[42] D. Zhang, G. Lu, Review of shape representation and description techniques, Pattern Recognit. 37 (2004) 1–19.
[43] A. Amanatiadis, V. Kaburlasos, A. Gasteratos, S. Papadakis, Evaluation of shape descriptors for shape-based image retrieval, IET Image Process. 5 (5) (2011) 493–499.
[44] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Inf. Theory 8 (2) (1962) 179–187.
[45] M. Teague, Image analysis via the general theory of moments, J. Opt. Soc. Am. 70 (8) (1980) 920–930.
[46] C. Zahn, R. Roskies, Fourier descriptors for plane closed curves, IEEE Trans. Comput. C.- 21 (3) (1972) 269–281.
[47] E. Fix, J. Hodges, Discriminatory analysis. nonparametric discrimination: Consistency properties, Tech. Rep. 4, USAF School of Aviation Medicine, Randolph Field, Texas, USA (1951).
[48] I. Kononenko, Comparison of inductive and naive bayesian learning approaches to automatic knowledge acquisition, in: Current Trends in Knowledge Adquisition, Vol. 8 of Frontiers in Artificial Intelligence and Applications, 1990, pp. 190–197.
[49] J. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[50] B. Boser, I. Guyon, V. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992, pp. 144–152.
[51] R. Kohavi, F. Provost, Glossary of terms, Mach. Learn. 30 (2–3) (1998) 271–274.