Computers in Biology and Medicine 91 (2017) 181–190

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A dictionary learning approach for human sperm heads classification

Fariba Shaker [a], S. Amirhassan Monadjemi [a,*], Javad Alirezaie [b], Ahmad Reza Naghsh-Nilchi [a]

[a] Department of AI, Faculty of Computer Engineering, University of Isfahan, Isfahan, 81746, Iran
[b] Department of Electrical and Computer Engineering, Ryerson University, Toronto, M5B 2K3, Canada

ARTICLE INFO

Keywords:
Sperm head classification
Sperm abnormality
Sperm morphology
Dictionary learning
Sparse representation
Infertility

ABSTRACT

Background and objective: To diagnose infertility in men, semen analysis is conducted in which sperm morphology is one of the factors that are evaluated. Since manual assessment of sperm morphology is time-consuming and subjective, automatic classification methods are being developed. Automatic classification of sperm heads is a complicated task due to the intra-class differences and inter-class similarities of class objects. In this research, a Dictionary Learning (DL) technique is utilized to construct a dictionary of sperm head shapes. This dictionary is used to classify the sperm heads into four different classes.

Methods: Square patches are extracted from the sperm head images. Columnized patches from each class of sperm are used to learn class-specific dictionaries. The patches from a test image are reconstructed using each class-specific dictionary and the overall reconstruction error for each class is used to select the best matching class. Average accuracy, precision, recall, and F-score are used to evaluate the classification method. The method is evaluated using two publicly available datasets of human sperm head shapes.

Results: The proposed DL based method achieved an average accuracy of 92.2% on the HuSHeM dataset, and an average recall of 62% on the SCIAN-MorphoSpermGS dataset. The results show a significant improvement compared to a previously published shape-feature-based method. We have achieved high-performance results. In addition, our proposed approach offers a more balanced classifier in which all four classes are recognized with high precision and recall.

Conclusions: In this paper, we use a Dictionary Learning approach in classifying human sperm heads. It is shown that the Dictionary Learning method is far more effective in classifying human sperm heads than classifiers using shape-based features. Also, a dataset of human sperm head shapes is introduced to facilitate future research.

1. Introduction

Infertility affects almost ten percent of the population and at least 30–50% of the cases are related to men [1]. To diagnose infertility in men, semen examination is conducted in which one of the steps is the assessment of sperm morphology which is referred to the specific size and shape of the sperm. Morphology assessment of sperms involves finding the percentages of morphologically normal and abnormal sperms and their type of abnormality. Morphology data are of high predictive value in diagnosing men fertility potential [2]. When infertility is diagnosed, usually some form of Assisted Reproductive Technique (ART) is utilized. Originally, sperm morphology analysis was mostly restricted to assessments of the proportion of normal spermatozoa. However, recent publications by prominent specialists such as Menkveld, suggest that in order to decide which method of treatment is appropriate, it is important to get an in-depth report on the types of existing abnormalities [2].

Studies show some association between genetic, environment, and lifestyle factors and abnormal sperm morphology [3–5]. However, the exact origins of many sperm abnormalities are not clear yet. Also, the effect of some abnormalities on sperm fertility potential is still unknown. As a consequence, the precise analysis of the number of sperm abnormalities would be a useful approach for research purposes [2,6].

However, the visual assessment of sperm abnormalities is believed to be subjective, time consuming, hard to teach, and highly dependent on the technician experience [7]. Therefore, inter and intra observer variability is common [8,9]. It has been shown that using computers for sperm morphology analysis can reduce the inter-laboratory variations [9]. Computer Assisted Semen Analysis (CASA) systems are widely used in laboratories to analyze semen parameters, including the morphology of sperms. However, none of these systems provide in-depth analysis of sperm abnormalities. Considering the need for comprehensive analysis of sperm abnormalities, it is very much important to provide automatic

* Corresponding author.
E-mail addresses: f.shaker@eng.ui.ac.ir (F. Shaker), monadjemi@eng.ui.ac.ir (S.A. Monadjemi), javad@ryerson.ca (J. Alirezaie), nilchi@eng.ui.ac.ir (A.R. Naghsh-Nilchi).

https://doi.org/10.1016/j.compbiomed.2017.10.009
Received 12 May 2017; Received in revised form 8 October 2017; Accepted 9 October 2017

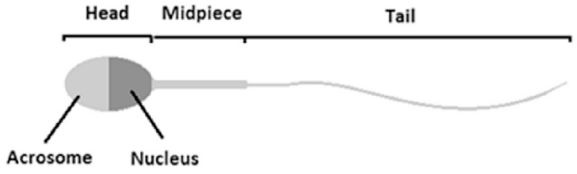

Fig. 1. Different parts of a sperm [11].

methods capable of performing such analysis. Automatic detection of all sperm shape categories will be helpful in diagnosing male infertility. It will also facilitate the research on the causes of male infertility in general, which can aid in developing methods for its treatment and even its prevention in future [10].

A sperm consists of three main parts, namely head, midpiece and tail (Fig. 1). Head itself is further divided into acrosome and nucleus. Each of these three parts should have certain characteristics to be considered morphologically normal and these characteristics are species-dependent. For human sperms, the head should be smooth, regularly contoured and generally oval in shape. Also, there should be a well-defined acrosomal region comprising 40–70% of the head area. There should not be any large vacuole nor more than two small vacuoles on the acrosomal region and there should not be any vacuoles on the post-acrosomal region (nucleus) [6]. Sperm head abnormalities are the result of one or more of these characteristics being disrupted. Sperm abnormalities are divided into three main categories: head defects, midpiece defects, and tail defects among which head defects are considered more important [6]. In this research, we focus on head defects.

According to WHO[1] 2010, there are eleven classes of abnormalities related to the head of the sperms which include small, large, amorphous, tapered, pyriform, large acrosome, small acrosome, round, vacuolated, vacuoles in the post-acrosomal region, and double heads. These classes are characterized by the specific shape, size, and texture of the head and its constituent parts.

Among these classes, the four classes known as normal, pyriform, tapered, and amorphous are differentiated by their specific shapes in addition to their size while the other classes of head abnormalities are mainly characterized by the size of the head or the acrosome or the existence of vacuoles. These four classes of abnormalities present a vast and at the same time continuous difference in size, shape, and texture which make their classification a complicated task. One of the complications in this regard is the existence of the class amorphous which makes the classification of these classes a challenging task. The reason is the infinite number of shapes and textures that are classified under this class of abnormality. Therefore it seems that finding representative features to discriminate amorphous sperms is a farfetched goal. Fig. 2 shows some examples of amorphous sperms. The other complication is that in addition to intra-class variability, there are inter-class similarities as well. For example, there are amorphous sperms that are elongated like tapered sperms or tapered sperms that are narrowed towards the tail like pyriform sperms. In this research, we focus on the classification of these four classes, since they are the most difficult to discriminate. Other classes of sperm shapes are easily distinguishable after a precise segmentation [11] followed by measuring the physical attributes of the head and its constituent parts.

Most of the works in the area of sperm classification are in the veterinary field and the works related to human sperm is scarce. Although there are similarities between these two areas but the types of abnormalities are different. Also, the motivation for assessing different classes of abnormalities in these two fields is not the same. Most of the works in the veterinary field are focused on classifying sperms in two classes of normal/abnormal or dead/alive based on the state of the sperms' acrosome.

[1] World Health Organization.

Works on morphology classification of human sperm can be divided into two categories; (1) classification using images of live, moving sperms, (2) classification using the images of fixed and stained sperms. The first category aims at quickly deciding which sperm is healthy without damaging the sperm. The resulting normal sperms will be used in ART procedures. In this case, the images have a lower quality and the type of abnormality is not important. Therefore sperms are only classified as normal and abnormal. The work presented by Ghasemian et al. [12] is the state of the art in this category. The classification method targets all three parts of a sperm, i.e. head, midpiece and tail and classifies the overall sperm as normal or malformed. Although this method is fast and highly accurate, it is not designed to distinguish specific types of the head abnormalities. In the second category, the aim is to find out the types of existing abnormalities for diagnosis or research purposes. The images have higher quality and the staining highlights the details of the spermatozoa. Most of the works in automatic morphology assessment of sperms including the current work fall under this category since this category is the recommendation of WHO. In the following, we review the works that are placed under the second category.

One of the earliest works in the field of human sperm classification is the work of Perez-Sanchez et al. [13]. They studied ten shape features to find out their relevance to classification of human sperm heads into 14 classes of shapes. These features comprise two sets of head measurements: the first set called basic features (length, width, area, perimeter, and mass) and the second set called derived features (ratio, the difference in length and width, ellipticity, form and total mass). Their method is manual and they used statistical tools to show these features are relevant in discriminating some of the shape classes.

Shaker et al. [14] used the shape features proposed by Perez-Sanchez in an automatic framework to classify sperm heads into four classes of normal, tapered, pyriform and amorphous. They also proposed new shape features called elliptic features and demonstrated that by adding these attributes to Perez-Sanchez features, the classification accuracy could be improved.

In another study by Yi et al. [15], the sperm heads are classified into four classes of normal, small, elongated, and megalo. They used the first ten coefficients of Fourier transform of the head contour points to reconstruct the head contour. Then they applied the wavelet transform to the enclosed area of the head. The root mean square error in transform space between the image and each class of abnormalities was used to classify the sperm head. The four classes used in this study are somehow trivial since these classes are easily separable by just the size criterion.

In a recent study, Jiaqian et al. [16] used Principal Component Analysis (PCA) combined with K-nearest neighbors algorithm to classify human sperm heads into two classes of normal and Abnormal. Their results show that there is a bias towards the normal class which makes the usefulness of the method questionable.

In the most recent study [17], Chang et al. showed that multiclass classification of sperm heads is a complicated task and using a single classifier will not be discriminative enough. They proposed to use a combination of classifiers, specifically a Cascade Ensemble of SVM classifiers (CE-SVM), and a combination of six shape-based feature families, i.e. Morphological descriptors, Fourier descriptors, Geometric moments, Zernike moments, Convexity measures and Ellipticity measures, to classify sperm heads into five classes of normal, tapered, pyriform, amorphous, and small. Their method tries to select the best features and the best combination of classifiers to achieve the best classification accuracy. In CE-SVM the classification is performed in two stages; in the first stage, amorphous sperms are separated from the rest of the sperms and in the second stage the rest of the sperms are classified as normal, pyriform, tapered, and small. The aim of the first stage is to separate as many amorphous sperms as possible. They showed that to distinguish amorphous sperms from other classes, the most discriminating features were Morphological, Fourier, and Zernike descriptors. So these descriptors were used in the first stage. In the second stage, there were four SVM classifiers each trained to distinguish a specific class (other than

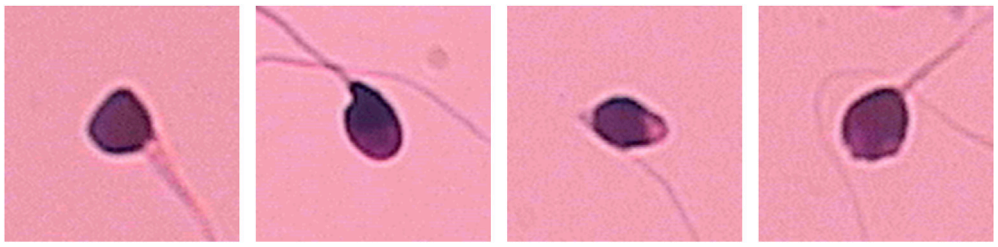

Fig. 2. Examples of amorphous sperm heads.

amorphous). They selected the best combination of descriptors for each SVM classifier experimentally. The average true positive rate achieved with this method was 58% and 74% for partial agreement and total agreement datasets respectively. The experimental results showed a very low accuracy rate for the amorphous class due to the vast differences existing within this class of abnormality.

Sparse coding has been applied successfully to a variety of problems in the fields of computer vision and medical image processing for tasks ranging from classification [18] to segmentation [19] and denoising [20, 21]. The idea of sparse coding is drawn from the fact that our observations can be described by a sparse subset of atoms taken from a redundant dictionary [22]. Although dictionaries could be constructed from off-the-shelf bases such as wavelets, it is shown that learning the dictionary adaptively from the observations can dramatically improve the signal reconstruction [20]. In this research, we propose an Adaptive Patch-based Dictionary Learning (APDL) method to classify sperm heads into four classes of normal, tapered, pyriform, and amorphous. Although all ten classes of sperms need to be reported, however, the classes of abnormalities other than these four classes are characterized by the size of the sperm heads or their constituent parts rather than their shapes. On the other hand, these four classes of sperms are differentiated by their specific shapes and therefore they are harder to differentiate. Accordingly, the focus of this research is to find a method to discriminate the four classes of sperm heads listed above. We also apply CE-SVM classification method using six groups of shape-based feature families as presented in Ref. [17] to our dataset and compare its performance with that of APDL method. We show that the results obtained using the APDL method outperform the results obtained from the conventional feature-based approach with a large margin. We also present a data set for sperm head classification denoted as Human Sperm Head Morphology dataset (HuSHeM) which is freely available for research purposes [23]. To validate our proposed method further, we used APDL method on SCIAN-MorphoSpermGS dataset as well and the results were compared with the ones reported by Chang et al. [17]. The results show

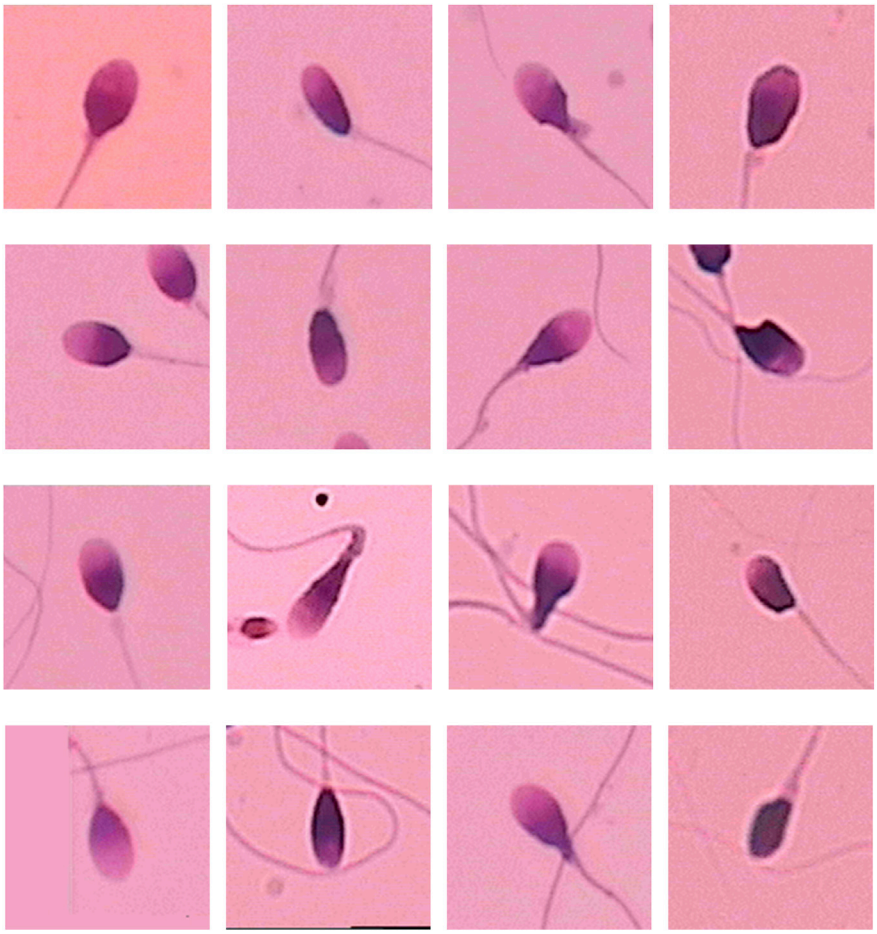

Fig. 3. Samples of the images from HuSHeM dataset. Columns from the left: normal, tapered, pyriform, and amorphous.

improvements in classification compared to the CE-SVM method using shape-based features.

2. Materials and methods

2.1. Datasets

One of the difficulties in this field of study is the lack of publicly available datasets. In this research, we attempted to provide a dataset of sperm head shape images in order to facilitate the future development of algorithms and their comparison. The purpose of collecting this dataset was to provide a reference and a common comparison tool for developing and improving computerized algorithms for sperm morphology classification. Therefore, we were mostly interested in the image processing aspects of data collection. The medical aspects of the images were documented only when it affected the image quality and its format. To build the dataset, the instructions and recommendations given by WHO 2010 [6] were followed. At Isfahan Fertility and Infertility Center (IFIC), semen samples were collected from patients between 25 and 38 years old. The semen smears were fixed and stained using Diff-Quik method. To this end, first, the air dried semen smears were fixed by immersing the slides in triarylmethane fixative for 15 s. After draining the excess solution, the fixed semen smears were sequentially immersed in eosinophilic xanthene for 10 s and in basophilic thiazine for 5 s. At last, to remove the excess stain, the slides were put under running tap water for 10–15 s. To obtain more clear images, immersion oil was applied to the dried slides. Images were taken using a Sony color camera (Model No SSC-DC58AP) attached to the third eyepiece of an Olympus CX21 microscope with a × 100 brightfield objective and a × 10 eyepiece. The resolution of each image was 576 × 720 pixels in RGB color space. The number of images taken per sample varied depending on the quality of the samples. For example, some samples had too many sperms per slide and therefore a lot of sperms were overlapped. Or some samples' staining quality was low and the images looked blurry. From these images, the sperm heads were cropped and classified into four classes of normal, tapered, pyriform, and amorphous, by three specialists at IFIC. The final dataset consists of the images of these four classes of sperm heads. In order to reduce the outliers in the dataset, only the samples with a collective consensus about their classes were kept in the dataset. The resulting dataset of sperm heads denoted as Human Sperm Head Morphology dataset (HuSHeM) includes 216 sperm heads (54 normal, 53 tapered, 57 pyriform, and 52 amorphous). The dataset contains four folders, one for each class. The folder names indicate the class of the images. The images of sperm heads are in RGB file format with the size of 131 × 131 pixels. Fig. 3 shows samples of the sperm heads from the dataset. This dataset is freely available for research purposes [23].

In this research, we also use the SCIAN-MorphoSpermGS dataset [24] in order to validate the proposed method further. This dataset consists of five classes of sperm shapes, i.e. normal, tapered, pyriform, small and amorphous. The images in each class are the results of majority agreement among three experts with a vast experience in morphological sperm analysis. So a sperm head is included in class c if two out of three experts label that sperm as class c. There are 100 normal, 228 tapered, 76 pyriform, 73 small and 656 amorphous sperms in this dataset. Since the images of the SCIAN-MorphoSpermGS dataset are only partially agreed upon by experts, the shapes are much more varied in each class compared to the HuSHeM dataset. It should be noted that images in SCIAN-MorphoSpermGS have been taken with lower magnification compared to the images in HuSHeM dataset; therefore, the size of the images is less for SCIAN-MorphoSpermGS compared to HuSHeM. More information about SCIAN-MorphoSpermGS dataset can be found in Ref. [24].

2.2. Dictionary learning

Suppose there are N training data denoted by $x_i \in R^m$ ($i = 1, \ldots N$). Dictionary learning (DL) for sparse representation learns dictionary D from training data by minimizing the following cost function over dictionary D and coefficient matrix $A = \{a_1, \ldots, a_N\} \in R^{k \times N}$

$$f_N(\boldsymbol{D}) \triangleq \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{x_i} - \boldsymbol{D}a_i\|_2^2 + \lambda \|a_i\|_p, \quad s.t. \ \ \|d_k\|_2^2 = 1, \quad \forall k = 1, .., K \tag{1}$$

where $D \subset R^{m \times k}$ is the dictionary in which each column (d_k) represents a basis vector (or atom) and $p \in \{0,1\}$. The original DL methods were designed in order to learn a dictionary which can represent signals faithfully. To adapt the DL method to classification tasks, a number of approaches have been developed in recent years. Most of the DL-based classification methods learn the dictionary either by forcing the dictionary discriminative or making the sparse coefficients discriminative. SRC [25] is one of the original DL methods used for face recognition in which each class-specific dictionary D_k is constructed using the actual training images of that class as its columns. Matrix D is constructed by concatenating the class dictionaries ($D = [D_1, D_2, \ldots, D_K]$).

The sparse representation of test sample y is then obtained by solving the l_1 minimization problem:

$$\widehat{A} = \text{argmin}_A \|A\|_1 \quad subject \ \ to \ \ \|DA - y\|_2 \leq \varepsilon \tag{2}$$

Coefficient vector $\widehat{A}$ could be thought of as the concatenation of coefficient vectors $\widehat{a}_1, \widehat{a}_2, \ldots, \widehat{a}_K$ each corresponding to one class-specific dictionary. Signal y is then assigned to class i which minimizes the residual between y and its approximation:

$$\min_i r_i(y) = \|y - D_i \widehat{a}_i\|_2 \tag{3}$$

Yang et al. [26] improved the SRC method by learning the class-specific dictionaries instead of using the original training samples as dictionary atoms. In this way, they reached a smaller sized dictionary which was also better representing the class signals.

Inspired by these works, we adapt Yang's method of face recognition to learn class-specific dictionaries for human sperm heads. However, instead of using the whole image of a sperm head as one sample, we use patches from the sample image (size m × n) by sliding a window of size $p \times q$ over the image, resulting in P = (m-p+1) × (n-q+1) patches as is discussed in the next subsections. To this end, the images of sperm heads are first turned so that the main axis of sperm heads are horizontal and the acrosome points to the right side and then each image is cropped to the size of 50 × 76 pixels so that most of the background is removed. In this research, the images from HuSHeM were turned manually. However, there are automatic algorithms for turning the sperm heads which can be used in practice [27]. The images from SCIAN-MorphoSpermGS were horizontal so we only corrected their directions. At last, the images are transformed from RGB color space to gray-level. Fig. 4 shows a sample image from HuSHeM data set and the resulting turned and cropped image.

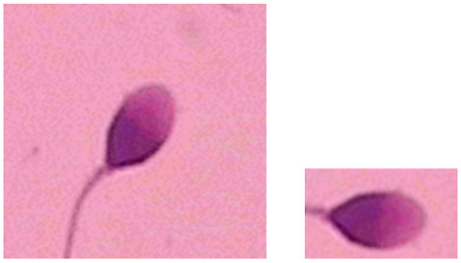

Fig. 4. A sample dataset image and the resulting turned and cropped image.

2.3. Dimensionality reduction

The size of the images of sperm heads is $50 \times 76 = 3800$ pixels which is relatively high. In some applications such as face recognition, down-sampling is used in order to reduce the dimensionality of the data. But in sperm classification application downsampling is not appropriate. The reason is that the shape and size of the sperms from different classes are very much close and down sampling removes the subtle differences that differentiate between the two classes. On the other hand, since manual classification of sperm heads has inherent difficulties, the number of training samples is limited. For these two reasons, we decided to use patches from the sperm head images as the training, and also testing, data. In this way, the number of training samples is increased and the microscopic features are preserved. Also, there are some common features among the samples of a single class such as the narrowing of the nucleus towards the tail which is extracted in this way. The other point is that using the whole image as one training vector may cause a slight misalignment between the training and the testing image to result in poor classification. Patch training eliminates this problem as well. In the next section, we will explain our method of dictionary learning using the patches, instead of the whole image or its downsampled counterpart, as the training and testing data.

2.4. The proposed adaptive patch-based dictionary learning method (APDL)

We denote $I_c = \{i_1, i_2, \ldots, i_{Nc}\}$ as the training samples of class c. If the size of each training image i_i is $m \times n$, P patches of size $p \times q$ are extracted from each training image by sliding a window of size $p \times q$ over the image ($P = (m\text{-}p+1) \times (n\text{-}q+1)$). Each patch is then transformed into a unit column vector x_j (Fig. 5). The training samples of class c are therefore denoted by $X_c = [x_1, x_2, \ldots, x_{P.Nc}]$. Our aim is to learn a dictionary of sperm heads $D_c = [d_1, d_2, \ldots, d_n]$ from X_c where $n < P.N_c$, The objective function for the dictionary learning is

$$J_{D_c,A} = \underset{D_c,A}{\operatorname{argmin}} \left\{ \|X_c - D_c A\|_F^2 + \lambda \|A\|_1 \right\} \; s.t.\; d_j^T d_j = 1 \tag{4}$$

In which A is the sparse coefficients matrix and $||\ ||_F$ is the Frobenius norm. d_j is the jth atom of the dictionary D_c. In order to classify the test sample y, first of all, patches of size $p \times q$ are extracted from y in the same way we did for training samples. From the columnized patches of y, matrix Y is constructed. Then the sparse representation of Y is obtained by l_1 minimization problem with the dictionaries of each class c as follows

$$\widehat{A}_c = \operatorname{argmin}_{A_c} \|A_c\|_1 \quad subject\ \ to\ \ \|D_c A_c - Y\|_F \le \varepsilon \tag{5}$$

Signal y is assigned to class c which minimizes the residual between Y and its approximation:

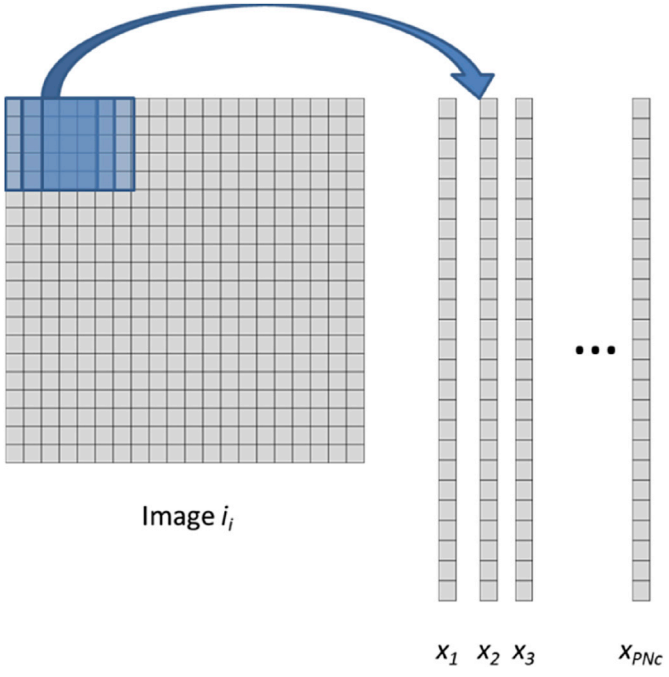

Fig. 5. Sliding window and the columnized patches.

Table 1
Evaluation metrics.

Measure	Formula	Evaluation
tp_i	M_{ii}	True positive per class
fp_i	$\sum_{j \neq i} M_{ji}$	False positive per class
fn_i	$\sum_{j \neq i} M_{ij}$	False negative per class
$precision_i$	$\frac{tp_i}{tp_i + fp_i}$	Per class precision (Specificity)
$recall_i$	$\frac{tp_i}{tp_i + fn_i}$	Per class recall (Sensitivity)
$accuracy$	$\frac{\sum_i M_{ii}}{\sum_{ij} M_{ij}}$	Average accuracy
$Average\ precision$	$\frac{\sum_i Precision_i}{l}$	The average per class precision (Macro averaging)
$Average\ recall$	$\frac{\sum_i Recall_i}{l}$	The average per class recall (Macro averaging)
$F\text{-}score$	$2 \times \frac{Precision.Recall}{Precision+Recall}$	F-score for each class

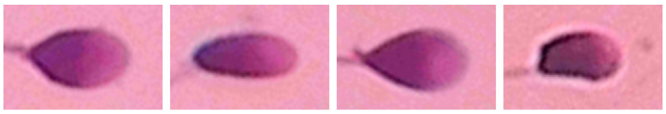

Fig. 6. Horizontal images of sperm heads.

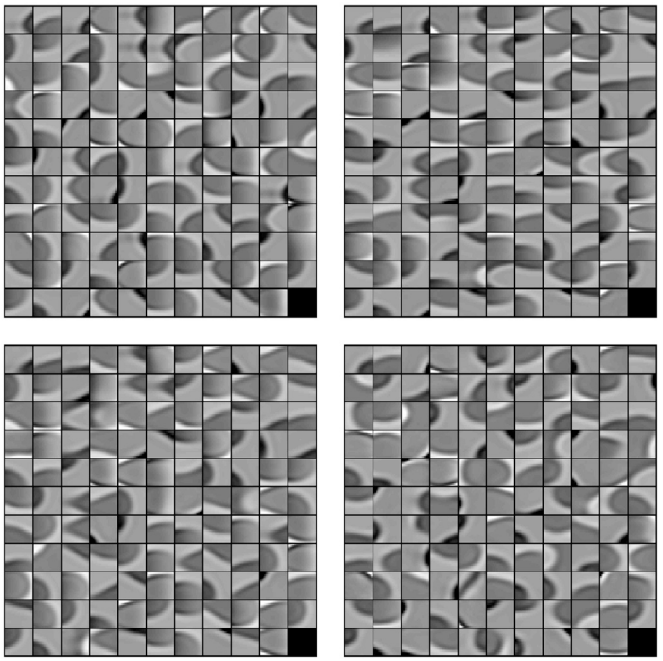

Fig. 7. One example of dictionaries learned from the training images of HuSHeM dataset. These images are dictionary atoms from four classes of normal (top left), tapered (top right), pyriform (bottom left) and amorphous (bottom right).

$$\min_c r_c(y) = \left\| Y - D_c \widehat{A}_c \right\|_F \quad (6)$$

2.5. Evaluation criteria

In order to evaluate the multiclass classification results, there are a number of metrics to consider. For an individual class C_i, the metrics are tp_i (the number of true positives for class C_i), fp_i (the number of false positives for class C_i), fn_i (the number of false negatives for class C_i), *accuracy* (average accuracy), $precision_i$ (precision for class C_i) and $recall_i$ (recall for class C_i). The formulas for calculating these metrics are summarized in Table 1 [28]. In this table M is the confusion matrix, index i is the class number, and l is the number of classes.

3. Results and discussion

As discussed in section 2, the images of sperm heads are turned horizontal and cropped into size 50 × 76 pixels (Fig. 6) and then transformed from RGB to grayscale. For learning the dictionary, k-fold cross validation with k = 5 is used which means that the images from each class are divided into 5 equal sets and each time one of these sets is used as the validation set and the rest are used to learn the dictionaries. The experiment is repeated five times covering all combinations. In each trial, the size of each class-specific dictionary is changed from 40 to 680 atoms with the increments of 40 and from 1000 to 2000 atoms with the increments of 200 and the size of the patches is changed from 15 × 15

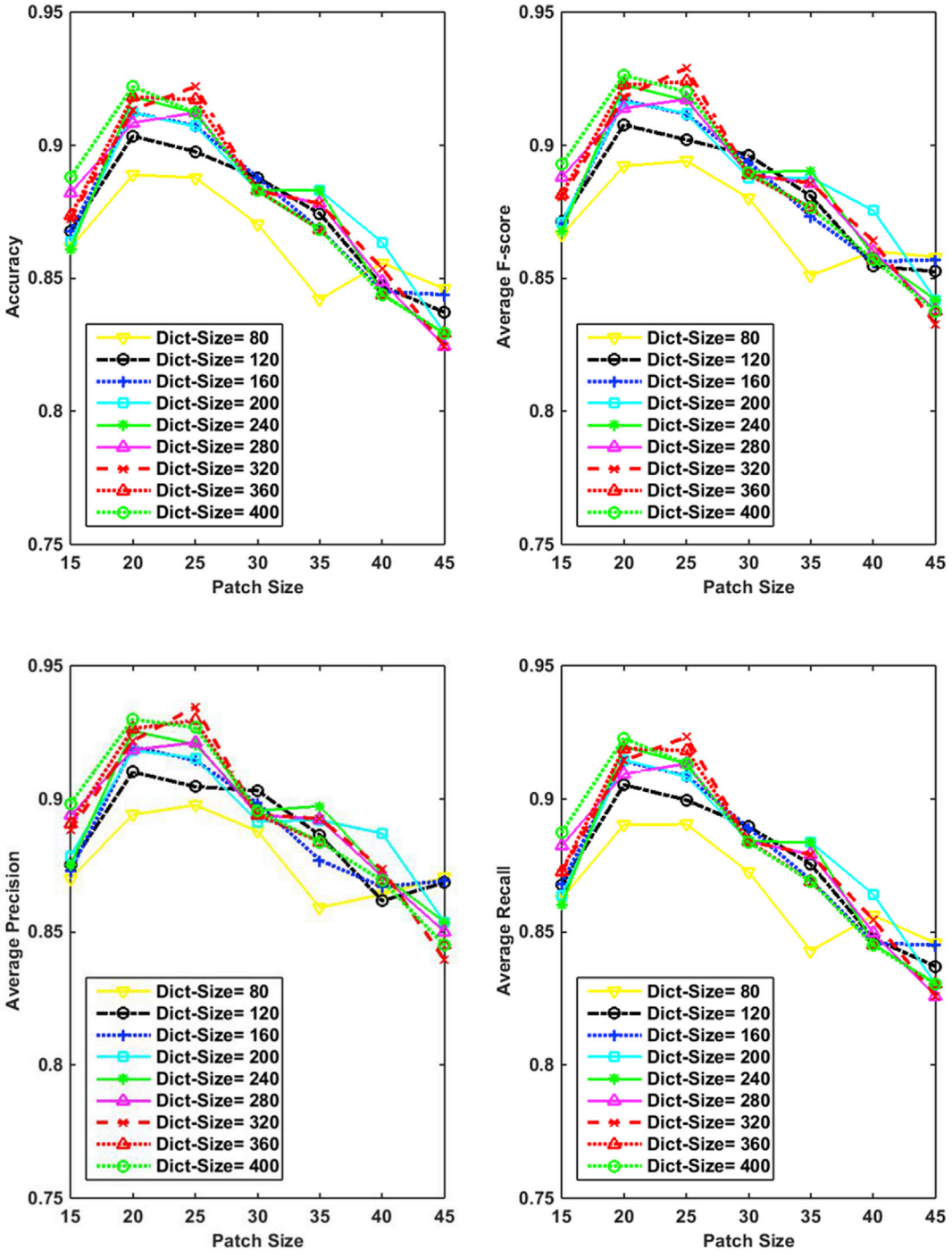

Fig. 8. Plots showing the performance of APDL based classifier.

pixels to 45 × 45 pixels with the increments of 5 × 5 pixels. It should be noted that increasing the size of the patch will reduce the number of samples per class and the dictionary size cannot be larger than the number of samples. An example of a learned dictionary, corresponding to the patch size 20 × 20 pixels and per class dictionary size of 120 atoms, for all four classes is shown in Fig. 7.

Fig. 8 shows the accuracy, average F-score, average precision and average recall for various dictionary and patch sizes. In each plot, the horizontal axis corresponds to the patch size and each curve corresponds to a certain dictionary size. At first glance, it is obvious that all four performance measures have almost the same behavior with changing the dictionary and patch sizes. As the curves show, all of the performance measures have their best values at patch size 25 × 25 pixels. For patch sizes less than 25 × 25 pixels, the performance measures increase with the size of the patch and for patch sizes larger than 25 × 25 pixels, the performance measures decrease with the size of the patch. The highest performance corresponds to the combination of dictionary size of 320 atoms and the patch size of 25 × 25 pixels. Fig. 9 shows the changes of performance measures with respect to dictionary sizes. The figure only shows accuracy and F-score. The other measures were omitted since they follow the same behavior. As demonstrated, the performance improves as the dictionary size increases, until there is no more improvement. The best size of the dictionary depends on the patch size. When the patch size is small, larger dictionary sizes result in a better performance; however, for large patch sizes, increasing the dictionary size does not improve the performance and sometimes even worsens the performance. The reason is that with larger patch sizes there are less distinct patches available from each image. Since the size of our dataset images is 50 × 76 pixels, for patch size 45 × 45 pixels, there are 192 patches from each image. Since there is only one-pixel difference between consecutive patches, the number of distinct patches is far less than 192. Therefore, in this case, increasing the number of dictionary atoms will result in repetitive atoms

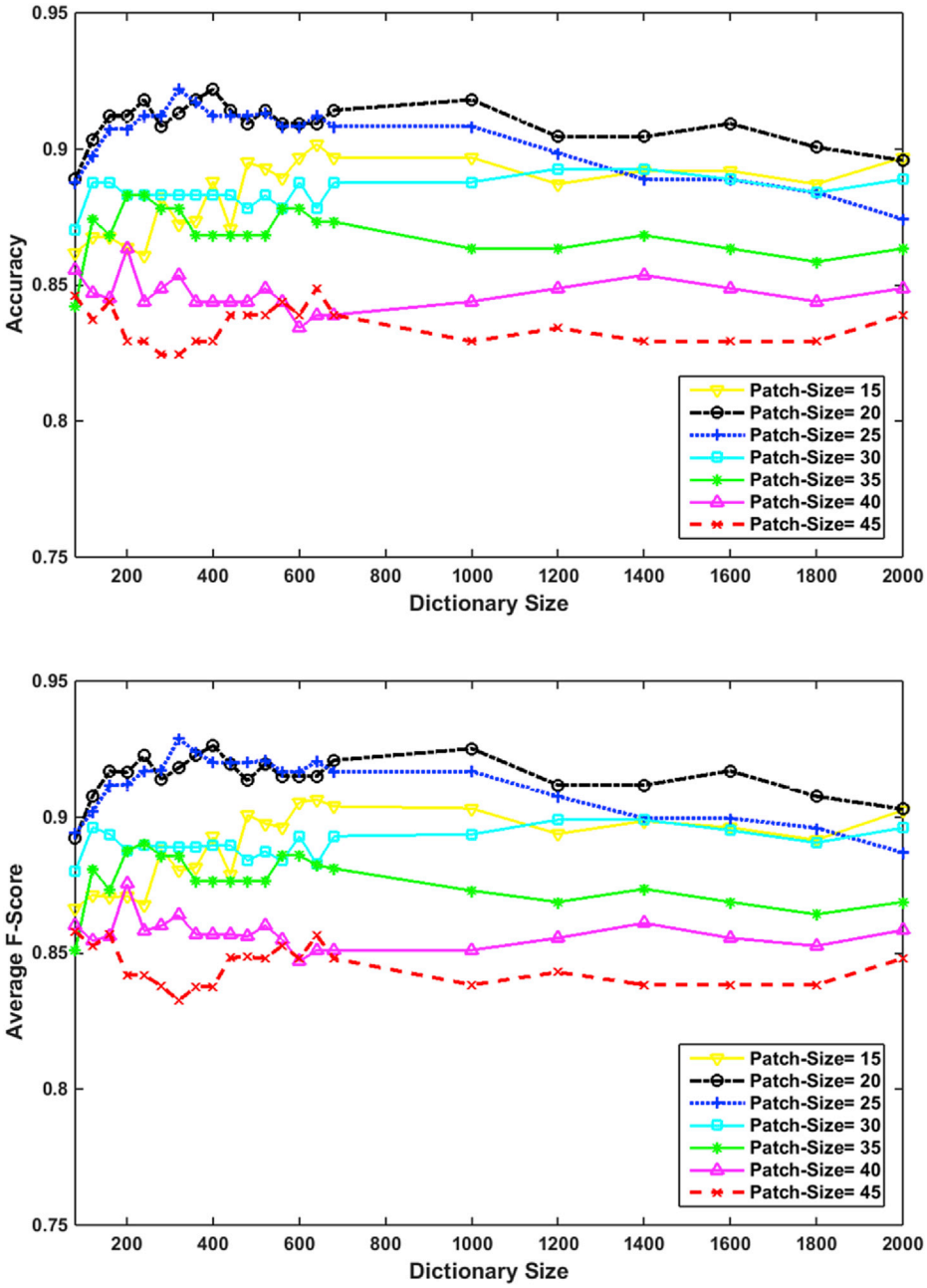

Fig. 9. Changes of performance measures for different patch sizes with respect to dictionary sizes in APDL method.

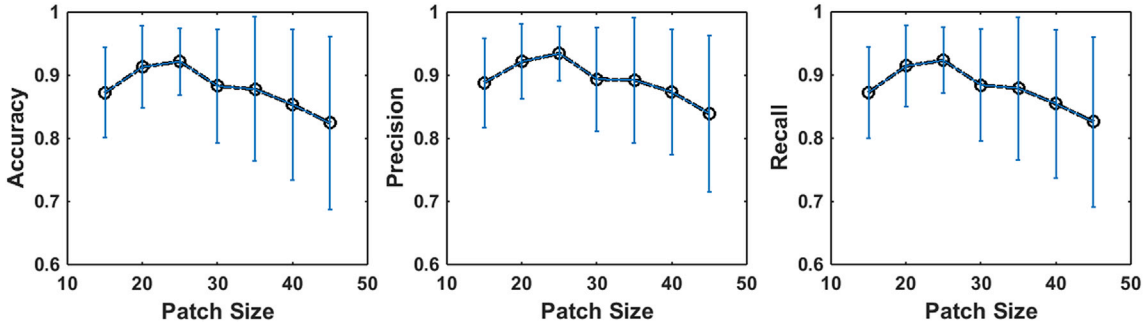

Fig. 10. Performance measures for dictionary size 320.

in each dictionary. On the other hand, the number of distinct patches is much higher when a small patch size is used and therefore a small dictionary size is not representative enough. For example, when the patch size is 15 × 15 pixels, the number of resulting patches is 2232 for each image. Obviously, with such a high number of patches, a dictionary of 80 atoms will not be able to capture all distinct structures presented in the training images and as a result, the performance drops for small dictionary sizes. Fig. 9 shows that by increasing the number of dictionary atoms for patch size 15 × 15 pixels, the performance measures improve which justifies this argument. However increasing the dictionary size beyond 1000 does not improve the performance any more. Fig. 9 shows that although the best performance happens for patch size 25 × 25 pixels, for most of the dictionary sizes, the patch size 20 × 20 pixels results in the best performance.

Fig. 10 shows the performance measures for dictionary size 320 as a function of patch size. In this figure, the black curve shows the average performance and the vertical lines show the standard deviation. As shown, for patch size 25 × 25 pixels, the average performance is the highest and at the same time, there is less variation in the performance for different runs of the algorithm. As the patch size increases, the performance drops and the standard deviation increases. As the curves show, although the best performance is for patch size 25 × 25 pixels, for other patch sizes smaller than 40 × 40 pixels, the performance remains close to the best performance. For example, the average accuracy in the worst case is 87.4% for patch size 15 × 15 pixels and in the best case, it is 92.2% for patch size 25 × 25 pixels. The same is true for average precision and average recall which shows the robustness of this method. Comparing the precision and recall shows that both performance measures increase or decrease together. This is an interesting result showing that by changing the dictionary size and the patch size both measures can improve together.

We also implemented the CE-SVM method using shape-based descriptors (SBD) presented in Ref. [17] and compared the results of this method with the results taken from APDL method. A short description of CE-SVM method is presented in the introduction. To extract the sperm head features, the contour of the sperm heads were used. We used the method presented in Ref. [11] for the segmentation of sperm head which automatically detects and segments the sperm head with a good performance. The resulting contour of the segmented sperm head was used to extract the shape features such as Morphological descriptors, Fourier descriptors, Geometric moments, Zernike moments, Convexity measures and Ellipticity measures [17]. The reader is referred to [17] for a detailed description of the features and the method of classification that is used. To evaluate CE-SVM method we used k-fold cross validation with the exact same data partitioning for training and testing that was used for APDL method.

Table 2 shows the best results for all four performance measures from the CE-SVM method and compares them with the ones obtained from APDL. As the table shows, for all performance measures, the APDL method resulted in a significant improvement (between 16% and 17% improvement in all performance measures) over the SBD-based classifier. Tables 3 and 4 show the average confusion matrices for CE-SVM and APDL methods respectively. Comparing these two matrices proves again the superiority of dictionary learning method over the feature-based method. The precision and recall for all four classes are much higher in APDL than the corresponding classes in the CE-SVM method. Also, the performance measures demonstrate more homogeneous results for all classes in APDL method while in the CE-SVM method, the performances for different classes are imbalanced. As Table 3 shows the precision for the normal class in the CE-SVM method is 1.00 while this measure for the amorphous class is 0.66. Another advantage of APDL method is the balance between precision and recall as shown in Table 4. The last column of this table presents the F-score obtained for each class. The F-score is the weighted average of precision and recall which shows how effective the classification method is. When precision and recall are equal, the F-score is equal to their average and if they are different, it is less than their average. Therefore, F-score favors a balanced classifier over an imbalanced one. Comparing the F-scores from APDL with the ones obtained from CE-SVM method indicates again the superiority of APDL method. In addition to a better performance achieved using APDL method, this method is much more straightforward than CE-SVM. In CE-SVM there are six families of descriptors and two levels of classifiers. For each classifier, a specific combination of the descriptors is used and the combination of descriptors is very much dependent on the class of abnormality. Therefore extending the CE-SVM method to new classes is not straight forward. However, APDL method could easily be extended to include more classes of sperm shapes.

We also applied the APDL method to SCIAN-MorphoSpermGS dataset in the same way we did for HuSHeM dataset. Since the images in this dataset are horizontal and cropped into size 35 × 35 pixels and their format is grayscale, the preprocessing step is omitted. Again we used k-fold cross validation with k = 5. This way the size of the validation dataset is the same as the one used in Ref. [17]. Since the images in SCIAN-MorphoSpermGS dataset are smaller, the patch sizes were changed from 5 × 5 pixels to 30 × 30 pixels and the dictionary sizes were changed from 40 to 280 per class. We compared our results with the results reported by Chang et al.in Ref. [17] by presenting the true positive rate (recall) and its average as reported by the authors. In our experiments, the dictionary that resulted in the best F-score was chosen as the best dictionary. The best results were achieved with the dictionary size 180 and the patch size 20 × 20 pixels. The results are summarized in Table 5. As demonstrated, APDL method has achieved a higher

Table 2
Comparison of the results obtained from the CE-SVM method and the ones obtained from APDL method for HuSHeM dataset. The bold font shows the improved results.

	Accuracy	Average precision	Average recall	Average F-score
CE-SVM	78.5%	80.5%	78.5%	78.9%
APDL	**92.2%**	**93.5%**	**92.3%**	**92.9%**

Table 3
Average confusion matrix for the CE-SVM method on HuSHeM dataset.

Precision	1.00	0.82	0.74	0.66		
Normal (actual)	8.2	0.2	0.8	1.6	0.76	0.86
Tapered (actual)	0	8.2	1.2	1.2	0.77	0.79
Pyriform (actual)	0	0.4	9.8	1.2	0.86	0.80
Amorphous (actual)	0	1.2	1.4	7.8	0.75	0.70
	Normal (predicted)	**Tapered (predicted)**	**Pyriform (predicted)**	**Amorphous (predicted)**	**recall**	**F-score**

Table 4
Average confusion matrix for APDL method with Dictionary size = 320 and patch size = 25 × 25 pixels on HuSHeM dataset.

	precision	0.98	0.93	0.91	0.89		
actual	**Normal**	10.2	0	0.6	0	0.94	0.96
	Tapered	0	10	0.2	0.4	0.94	0.94
	Pyriform	0.2	0.4	10	0.8	0.88	0.89
	Amorphous	0	0.4	0.2	9.8	0.94	0.91
		Normal	**Tapered**	**Pyriform**	**Amorphous**	**recall**	**F-score**
				predicted			

Table 5
Comparison of the recall (true positive rate) obtained from the CE-SVM method and the one obtained from APDL method for SCIAN-MorphoSpermGS dataset. The bold font shows the improved results.

Class	Normal	Tapered	Pyriform	Small	Amorphous	Average
CE-SVM	62%	64%	50%	82%	30%	58%
APDL	**71**%	**67**%	**71**%	68%	**35**%	**62**%

performance for all classes of sperm shapes except for "small" class. The average recall shows improvement compared to CE-SVM using SBDs. These results confirm that the proposed dictionary learning method is much more effective for discriminating sperm head shapes than the shape-feature-based classifiers.

4. Conclusion

In this paper, we presented an adaptive patch-based dictionary learning method (APDL) for classification of human sperm head images into four classes of normal, tapered, pyriform, and amorphous. We also implemented the CE-SVM method proposed in Ref. [17] which uses a cascade of SVM classifiers with a combination of six feature families to classify human sperm heads. To evaluate APDL and compare it to CE-SVM, the HuSHeM dataset was used. For both methods, k-fold cross validation was used to train and test the learners. The results show that dictionary learning method is far more effective in the classification of sperm heads than the shape-feature-based method. APDL achieved an average accuracy of 92.2%, average precision of 93.5%, average recall of 92.3%, and average F-score of 92.9%. The performance measures for each separate class show that this method is effective in recognizing all four classes of abnormalities. In addition to achieving higher accuracy, precision, and recall, APDL resulted in a more balanced classifier where the precision and recall for each class are close. The other advantage of APDL is the minimum number of parameters involved in the algorithm, which is the patch size and dictionary size, and its robustness against these parameters.

To validate APDL method further, it was applied to SCIAN-MorphoSpermGS dataset as well and the results were compared to the ones reported in Ref. [17]. Again, the results confirm the superiority of APDL method over the shape-feature-based classifiers.

One of the contributions of our work is the introduction of human sperm head morphology (HuSHeM) dataset. This dataset was created with the help of the field experts and was used in this research to evaluate the proposed method and compare it to the existing shape-feature-based method. This dataset can facilitate the development and comparison of future methods.

Acknowledgment

The authors would like to thank Isfahan Fertility and Infertility Centre for providing the samples and microscopic images; and their specialized staff for labeling the data. We would also wish to thank Ms. Mahboubeh Mehraban for providing us with valuable advice in addition to her assistance in creating the dataset.

References

[1] M.R. Maduro, D.J. Lamb, Understanding new genetics of male infertility, J. Urol. 168 (5) (2002) 2197–2205.
[2] R. Menkveld, Sperm morphology assessment using strict (tygerberg) criteria, Methods Mol. Biol. 927 (2013).
[3] M. De Braekeleer, M.H. Nguyen, F. Morel, A. Perrin, Genetic aspects of monomorphic teratozoospermia: a review, J. Assist. Reprod. Genet. 32 (4) (Apr. 2015) 615–623.
[4] J. Rubes, S.G. Selevan, D.P. Evenson, D. Zudova, M. Vozdova, Z. Zudova, W.A. Robbins, S.D. Perreault, Episodic air pollution is associated with increased DNA fragmentation in human sperm without other changes in semen quality, Hum. Reprod. 20 (10) (2005) 2776–2783.
[5] J. Auger, F. Eustache, A.G. Andersen, D.S. Irvine, N. Jørgensen, N.E. Skakkebæk, J. Suominen, J. Toppari, M. Vierula, P. Jouannet, Sperm morphological defects related to environment, lifestyle and medical history of 1001 male partners of pregnant women from four European cities, Hum. Reprod. 16 (12) (Dec. 2001) 2710–2717.
[6] WHO Laboratory Manual for the Examination and Processing of Human Semen, fifth ed., 2010.
[7] M. Freund, Standards for the rating of human sperm morphology, Int. J. Fertil. 11 (1) (1966) 97–180.
[8] G. Barroso, R. Mercan, K. Ozgur, M. Morshedi, P. Kolm, K. Coetzee, T. Kruger, S. Oehninger, Intra- and inter-laboratory variability in the assessment of sperm morphology by strict criteria: impact of semen preparation, staining techniques and manual versus computerized analysis, Hum. Reprod. 14 (8) (1999) 2036–2040.
[9] K. Coetzee, T.F. Kruger, C.J. Lombard, D. Shaughnessy, S. Oehninger, K. Özgür, K.O. Pomeroy, C. Muller, Assessment of interlaboratory and intralaboratory sperm morphology readings with the use of a Hamilton Thorne Research integrated visual optical system semen analyzer, Fertil. Steril. 71 (1) (1999) 80–84.
[10] R.J. Aitken, Whither must spermatozoa wander? The future of laboratory seminology, Asian J. Androl. 12 (1) (2010) 99–103.
[11] F. Shaker, S.A. Monadjemi, A.R. Naghsh-Nilchi, Automatic detection and segmentation of sperm head, acrosome and nucleus in microscopic images of human semen smears, Comput. Methods Prog. Biomed. 132 (2016) 11–20.
[12] 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 Programs Biomed. 122 (3) (2015) 409–420.
[13] F. Pérez-Sánchez, J.J. de Monserrat, C. Soler, Morphometric analysis of human sperm morphology, Int. J. Androl. 17 (5) (1994) 248–255.
[14] F. Shaker, S.A. Monadjemi, J. Alirezaie, Classification of human sperm heads using elliptic features and LDA, in: 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), 2017, pp. 151–155.
[15] W.J. Yi, K.S. Park, J.S. Paick, Parameterized characterization of elliptic sperm heads using fourier representation and wavelet transform 20 (2) (1998) 974–977.
[16] L. Jiaqian, T. Kuo-Kun, D. Haiting, L. Yifan, Z. Ming, D. Mingyue, Human sperm Health diagnosis with principal component analysis and K-nearest neighbor

algorithm, in: Medical Biometrics, 2014 International Conference, 2014, pp. 108–113.
[17] V. Chang, L. Heutte, C. Petitjean, S. Härtel, N. Hitschfeld, Automatic classification of human sperm head morphology, Comput. Biol. Med. 84 (December 2016) (2017) 205–216.
[18] T.H. Vu, H.S. Mousavi, V. Monga, G. Rao, U.K.A. Rao, Histopathological image classification using discriminative feature-oriented dictionary learning, IEEE Trans. Med. Imaging 35 (3) (2016) 738–751.
[19] S. Roy, Q. He, E. Sweeney, A. Carass, D.S. Reich, J.L. Prince, D.L. Pham, Subject-specific sparse dictionary learning for Atlas-based brain MRI segmentation, IEEE J. Biomed. Heal. Inf. 19 (5) (Sep. 2015) 1598–1609.
[20] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process 15 (12) (2006) 3736–3745.
[21] K. Abhari, M. Marsousi, J. Alirezaie, S. Member, P. Babyn, Computed tomography image denoising utilizing an efficient sparce coding algorithm, in: The 11th International Conference on Information Sciences, Signal Processing and Their Applications, 2012, pp. 259–263.
[22] I. Tosic, P. Frossard, Dictionary Learning, what is the right representation for my signal? IEEE Signal Process. Mag. 28 (2 March) (2011) 27–38.
[23] F. Shaker, Human Sperm Head Morphology Dataset (HuSHeM), Mendeley Data, 2017. Available at: https://data.mendeley.com/datasets/tt3yj2pf38/1.
[24] V. Chang, A. Garcia, N. Hitschfeld, S. Härtel, Gold-standard for computer-assisted morphological sperm analysis, Comput. Biol. Med. 83 (March) (2017) 143–150.
[25] J. Wright, a. Y. Yang, a. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[26] M.Y.M. Yang, L.Z.L. Zhang, J.Y.J. Yang, D. Zhang, Metaface learning for sparse representation based face recognition, in: Image Process. (ICIP), 2010 17th IEEE Int. Conf., 2010, pp. 2–5.
[27] V. Chang, J.M. 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.
[28] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45 (4) (2009) 427–437.

Fariba Shaker has received her B.S. degree in electronic engineering from the Isfahan University of Technology on 1989 and her M.Eng. degree in computer engineering from the University of Toronto on 1998. She is currently a PH.D. Candidate at the department of Artificial Intelligence, Faculty of Computer Engineering, University of Isfahan. Prior to that, she has been a lecturer at Islamic Azad University Isfahan Branch. Her research interests are in the areas of medical image processing, pattern recognition, and machine learning. She has been a visiting researcher at the department of Electrical and Computer Engineering, Ryerson University, from June to Nov 2016.

S. Amirhassan Monadjemi was born 1968, in Isfahan, Iran. He received his B.S. and M.Sc. degrees in computer engineering from Isfahan University of Technology in 1990 and Shiraz University in 1994, respectively. He received his Ph.D. in computer engineering, pattern recognition, and image processing, from University of Bristol, Bristol, England, in 2004. He is now working as an associate professor at the Department of AI, Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran. His research interests include AI, image processing, artificial neural networks, and physical detection and elimination of viruses.

Javad Alirezaie received his B.Sc. degree in Electronic Engineering from Tehran University in 1988 and his M.A.Sc. and Ph.D. degrees in Systems Design Engineering from the University of Waterloo in 1992 and 1996, respectively. He joined Ryerson in 2001 and he is currently a full Professor in the Department of Electrical and Computer Engineering. Dr. Alirezaie is the author/coauthor of over 130 research papers in refereed journals and conference proceedings to date and is involved in a variety of funded research projects. Dr. Alirezaie research interests include biomedical signals and image processing, computer-aided diagnosis, neural networks, pattern recognition, computer vision, and modeling.

Ahmad Reza Naghsh-Nilchi is a professor at the University of Isfahan, Iran. He received his B.S., M.Sc., and Ph.D., all in electrical engineering from the University of Utah. His research interests include medical image and signal processing as well as intensive computing. He has been an author or coauthor of several journal articles and conference papers and a couple of book sections. He is the editor-in-chief of the Journal of Computing and Security. He has served as the chairman of the Computer Engineering department for three terms. He has served as a research scholar at the National University of Ireland (summer 2011), and the University of California, Irvine (2012).