Computers in Biology and Medicine 109 (2019) 182–194

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# A novel deep learning method for automatic assessment of human sperm images

Soroush Javadi, Seyed Abolghasem Mirroshandel*

*Department of Computer Engineering, University of Guilan, Rasht, Iran*

ARTICLE INFO

*Keywords:*
Human Sperm Morphometry
Automatic image analysis
Sperm defects
Infertility
Deep learning

ABSTRACT

Sperm morphology analysis (SMA) is a very important factor in the diagnosis process of male infertility. This research proposes a novel deep learning algorithm for malformation detection of sperm morphology using human sperm cell images. Our proposed method detects and analyzes different parts of human sperms. First of all, we have prepared an image collection, called the MHSMA dataset, which can be used as a standard benchmark for future machine learning studies in this problem. This collection consists of 1,540 sperm images from 235 patients with male factor infertility. This unique dataset is freely available to the public. After applying data augmentation techniques, we have proposed a sampling method for fixing data imbalance. Then, we have designed a deep neural network architecture and trained it to detect morphological deformities in different parts of human sperm—head, acrosome, and vacuole. Our proposed method is one of the first algorithms that considers the acrosome. In addition, our method can work very well with non-stained and low-resolution images. Our experimental results on the proposed benchmark show the high accuracy of our deep learning algorithm for detection of morphological deformities from images. In these experiments, the proposed algorithm has achieved $F_{0.5}$ scores of 84.74%, 83.86%, and 94.65% in acrosome, head, and vacuole abnormality detection, respectively. It should be noted that our algorithm achieves a better accuracy than existing state-of-the-art methods in acrosome and vacuole abnormality detection on the proposed benchmark. Also, our method works very fast. It can classify images in real-time, even on a mainstream laptop computer. This allows an embryologist to quickly decide whether or not the analyzed sperm should be selected.

## 1. Introduction

Infertility is the problem of almost 15% of couples. The lack of pregnancy after 12 months of intercourse without protection is defined as infertility. About 30 to 40 percent of infertile couples suffer from male factor abnormalities [20,43]. One of the problems in male factor infertility is spermatozoa morphology abnormalities, which may show teratozoospermia or oligoasthenoteratozoospermia.

The quality of spermatozoa is one of the most important parameters for oocyte fertilization and embryo quality. It is shown that abnormalities in sperm correlate with cleaving embryo morphology at later stages [7]. In other words, the shape of sperm is reflected by sperm development during spermatogenesis. As a result, problems in sperm maturation causes abnormalities in sperm morphology and the functionality of egg fertilization [2]. By assessment of sperm parameters and seminal plasma characteristics such as semen pH and sperm morphology, viscosity, concentration, and motility, male factor infertility can be determined [6].

The first successful live birth of a child using intra-cytoplasmic sperm injection (ICSI) method happened in 1992 [33]. These days, the ICSI method is widely used for the treatment of various couples who need assisted reproductive technologies. These couples can have normal, mildly, or severely abnormal semen parameters [5]. In several studies, the positive correlation between high ICSI outcomes and normal sperm morphology has been proven. In other words, serious abnormalities of the sperm head cause low fertilization, implantation, and pregnancy rates [30].

The normal sperm morphology is defined in previous studies by Menkveld et al. [28]. The length of a normal sperm head is between 3 and 5 micrometers. This range for the width of the sperm head is be-

* Corresponding author. Department of Computer Engineering, Faculty of Engineering, University of Guilan, P.O. Box: 1841, Rasht, Iran.
*E-mail addresses:* soroush.javadi@gmail.com (S. Javadi), mirroshandel@guilan.ac.ir (S.A. Mirroshandel).

https://doi.org/10.1016/j.compbiomed.2019.04.030
Received 19 December 2018; Received in revised form 22 April 2019; Accepted 22 April 2019

tween 2 and 3 micrometers (two-thirds or three-fifths of head length). The length of midpiece is one and a half of the head length and it is axially presented $1\mu m$ in width. The normal tail should be also visible, uniform, uncoiled and thinner than the midpiece. Its length must be about $45\mu m$.

In the intracytoplasmic morphologically selected sperm injection (IMSI) procedure, sperm selection is performed at high magnification (usually 6000×) [30]. However, existing microscopes of laboratories commonly have a low magnification (400 × and 600×). These magnification levels are routinely utilized for sperm selection in the ICSI procedure. The visual assessment of sperms are also commonly performed manually and it is only based on the judgment of embryologists. This method is inexact, subjective, non-repeatable, and hard to teach. Another solution for the assessment of men fertility is computer-aided sperm analysis (CASA) using different staining procedures [6]. Due to the flaws of manual solutions, automatic techniques are essential for analyzing human sperm morphology. As a result, designing efficient and accurate algorithms for analyzing and classification of sperms and selecting the best one before ICSI is a challenging and trending task [38]. In other words, automatic methods for selection of the best sperm during the ICSI process and without staining will be more desirable for embryologists. This will lead to higher fertilization and pregnancy rates.

In this paper, we propose a novel deep learning algorithm for automatic extraction of sperm morphological features. This problem is a challenging task, due to the following reasons: 1) the number of sperm images is not enough for the training phase; 2) the normal and abnormal sperm classes are highly imbalanced, thus making the problem harder; 3) the pictures are taken using a low-magnification microscope and the details of these images are not clear; 4) the pictures are very noisy; 5) the sperms should not be stained; and 6) the analysis should be done in real-time to be useful for treatment purposes.

As the first step, we introduce a new dataset for our deep learning purpose. This dataset is based on the Human Sperm Morphology Analysis dataset (HSMA-DS) [15]. It is a unique dataset based on the number of images and its characteristics. For solving problems of training sample scarcity and class imbalance, we apply data augmentation and sampling techniques, respectively. Then, we propose a deep neural network architecture that can be trained to classify the normality of sperm acrosome, head, and vacuole. The proposed model has 24 convolution, two max pooling, one average pooling, and two fully-connected layers. Our proposed model, unlike a lot of existing methods, is able to work well with non-stained pictures. In other words, our method is applicable for treatment purposes. In addition, the time required for checking each sperm is very short (under 25 milliseconds) and our method can work in real-time. Our experimental results also show the effectiveness of our method, which can be regarded as the state-of-the-art for this dataset.

The rest of the paper is organized as follows: Some previous works are reviewed in Section 2. The prepared dataset is presented in Section 3. Our proposed algorithm, including the designed deep learning model, is depicted in Section 4. Sections 5 and 6 contain a comprehensive description of experimental setup, comparisons, and discussion. Finally, conclusions and future work are summarized in Section 7.

## 2. Previous work

There are lots of research on automatic selection of sperms. In one of these studies, the fraction of boar spermatozoa heads was calculated and a pattern of intracellular density distribution was recognized as normal [37,38]. In this method, a deviation model was defined and computed for each sperm's head. Then, for each sperm classification, an optimal value was considered. Afterward, by applying morphological closing, sperms tails were removed and the holes in the contours on the heads were filled. Finally, using Otsu's method [32], sperm's head were separated from background.

Ramos et al. [34] combined DNA-specific stain (Feulgen) and computerized karyometric image analysis (CKIA) system in order to evaluate ICSI-selected epididymal sperms using a high magnification (1000 ×).

In another method, sperms were classified into normal and abnormal classes during four stages [2]: 1) image preprocessing: RGB images were converted to grayscale images and noises were removed by applying median filter; 2) detection and extraction of each sperm: this step was done using Sobel edge detection algorithm; 3) segmentation: each sperm was segmented into different parts (i.e., head, tail, and midpiece); and 4) statistical measurement: a classification was done to detect normal and abnormal sperms.

The sperm nuclear morphometric subpopulations for different species including sheep, goat, cattle, and pig were compared by Vicente-Fiel et al. [44]. ImageJ software [3] was used for processing of the sperm images and the results were used for clustering. In this method, computer-assisted sperm morphology analysis-fluorescence (CASMA-F) technology and multivariate cluster analyses were combined.

In another study, a fully automatic identification and discrimination method was proposed that was able to extract acrosome, nucleus, midpiece, and tail of each sperm on stained human semen smear [6]. In their method, the Bayesian classifier was utilized in order to segment different parts of sperm. This segmentation was done using the entropy-based expectation-maximization (EM) algorithm and Markov random field (MRF) model. The images were captured on stained sperms with a high-magnification (1000 ×) microscope.

The influence of different staining methods on human sperm head was reported by the research of Maree et al. [26]. In their study, stained sperms were compared with fresh ones.

Yániz et al. [46] proposed an automatic assessment of ram sperm morphology. In this work, fluorescence microscope was used to capture images from stained sperms. The analysis of sperm morphometry was done using the ImageJ program.

Combining digital image processing and learning vector quantization (LVQ) was used in another study in order to automatically classify the boar sperm acrosome [4]. Using a phase-contrast microscope, the images of sperm heads were obtained and the acrosome status was evaluated based on its stain color. Their experimental results have shown 6.8% of classification error on sperm heads.

Chang et al. [11] proposed a framework to detect and segment acrosome and nucleus of human sperm head. The head segmentation was applied using histogram statistical analysis and clustering techniques. They also tried the combination of different color spaces. Above 98% of accuracy achieved in sperm head detection. In this work, all algorithms were performed on stained sperm images.

In other work, principal component analysis (PCA) was used to extract features from sperm images. K-nearest neighbors (KNN) was also utilized to diagnose normal sperm. By applying these methods, an accuracy of 87.53% was achieved. However, this work was performed on a small dataset [25].

One of the most successful algorithms in sperm selection is the work of Ghasemian et al. [15]. This method (i.e., the SMA method) is able to work with fresh human sperm in real-time with a low-magnification microscope (400 × and 600×) and it can be used in the ICSI process. It has achieved over 90% of accuracy in sperm morphological features classification. One of the other advantages of this method is that sperm

defects can be detected with a low computation cost, i.e., in less than 9 s on a personal computer. However, this method cannot extract the size and shape of the acrosome. Our proposed deep learning model is more precise than the SMA method. In addition, it can extract the characteristics of the sperm acrosome accurately and it is also faster.

Zhang [47] has proposed a novel method for animal sperm morphology analysis. In this method, by applying image processing techniques, parameters such as elongation of the head, ellipticity, percentage of acrosome, and mid-piece angle have been calculated. For achieving this goal, different algorithms such as K-means, thinning algorithm, active contour model, and image moment have been utilized. Consequently, based on the extracted parameters, the algorithm can decide about the morphological quality of each sperm. There are also some researches that are not focused on sperm morphological analysis, but they are related to the task of automatic sperm analysis. One of these works is the method of Medina-Rodríguez et al. [27] that combines the Lambertian model based on surface reflectance with mathematical morphology for sperm cells segmentation. There are also some algorithms that work on microscopic videos in order to segment sperms and calculate their motilities [8,18,19].

## 3. Dataset

We introduce a new dataset called the Modified Human Sperm Morphology Analysis dataset (MHSMA), which is based on the Human Sperm Morphology Analysis dataset (HSMA-DS) [15]. The version of HSMA-DS we used contains 1,540 RGB images with a size of 1280 × 1024 pixels. These images were taken at a magnification of either 400 × (706 images) or 600 × (834 images) using a microscope (IX70, Olympus, Japan) equipped with a CCD camera (DP71, Olympus, Japan) with chromatic infinity objective lenses [15]. Each image captures a single sperm. In the rare case where more than one sperm is present in the image, the one closer to the center of the image would be considered the sperm in question. Each image is labeled by experts for normal (positive) or abnormal (negative) sperm *acrosome*, *head*, *vacuole*, and *tail and neck*. Positive and negative classes for *vacuole* are regarded as absence and presence of vacuole, respectively. The distribution of samples is shown in Table 1.

Since all the three channels in HSMA-DS images contain very similar data, we decided to convert them to grayscale images. To do this, we took the Y channel of the YCbCr color space. This conversion is done using equation (1).

$$Y = 0.299 \times R + 0.587 \times G + 0.114 \times B \quad (1)$$

We manually marked the position of head of sperms in each image. We used these points as centers for cropping the images into smaller ones. We cropped the 400 × images into 128 × 128-pixel ones. For the 600 × images, first, we cropped them into 192 × 192-pixel ones, and then resized them to 128 × 128 pixels—this mimics a 400 × magnification. These modified images make up the MHSMA dataset. Fig. 1 shows some sample images from the two datasets.

The MHSMA dataset contains 1,540 grayscale sperm images with a size of 128 × 128 pixels. Each image is centered on the head of the sperm it captures. The actual sperm head in every image fits into a centered 64 × 64-pixel box, therefore, images can also be cropped to this smaller size. However, parts of the sperm tail will be lost in both crop sizes. MHSMA is freely available at http://nlp.guilan.ac.ir/datasets in two flavours: 128 × 128- and 64 × 64-pixel size. We have randomly selected 240 images (15.58% of the total) as the *validation set* and 300 images (19.48% of the total) as the *test set*. The remaining 1000 images (64.94% of the total) form the *training set*. Also, the dataset is shuffled and samples occur in random order.

**Table 1**
Distribution of samples in MHSMA.

| Set | Label | # Positive | # Negative | % Positive |
|---|---|---|---|---|
| Whole dataset | Acrosome | 1,086 | 454 | 70.52 |
| | Head | 1,122 | 418 | 72.86 |
| | Vacuole | 1,301 | 239 | 84.48 |
| | Tail and neck | 1,471 | 69 | 95.52 |
| Training set | Acrosome | 699 | 301 | 69.90 |
| | Head | 727 | 273 | 72.70 |
| | Vacuole | 830 | 170 | 83.00 |
| | Tail and neck | 954 | 46 | 95.40 |
| Validation set | Acrosome | 174 | 66 | 72.50 |
| | Head | 176 | 64 | 73.33 |
| | Vacuole | 209 | 31 | 87.08 |
| | Tail and neck | 233 | 7 | 97.08 |
| Test set | Acrosome | 213 | 87 | 71.00 |
| | Head | 219 | 81 | 73.00 |
| | Vacuole | 262 | 38 | 87.33 |
| | Tail and neck | 284 | 16 | 94.67 |

## 4. Proposed method

We propose a machine learning-based approach for detecting abnormalities in sperm morphology. We train a convolutional neural network to classify sperm images into normal (positive) and abnormal (negative) classes for different morphological features.

As noted in Table 1, there are only 69 samples for abnormal tail and neck. Also, while detecting abnormalities in sperm head is complicated, detection of abnormal tail and neck is usually fairly easy for human experts. As a result, we will focus on the classification of sperm acrosome, head, and vacuole. This is also why we cropped away parts of the sperm tail and neck in our dataset. In other words, three classification problems will be addressed in this research. These are difficult problems, due to the low-quality, noisy images which are taken by a low-magnification microscope.

The whole process of our proposed method is described in Algorithm 1. We train a deep convolutional neural network on mini-batches generated from the training set, and save the checkpoint with minimum loss value on the validation set. The size of mini-batches in our algorithm is 64, which is a common value for this issue. A mini-batch is a small batch of training samples used to calculate model error and update model parameters. There are three variants of gradient descent, which differ in the amount of data used to compute the gradient of the objective function and update model parameters. In batch gradient descent, the gradient of the loss function is computed for the entire training dataset. As it requires to calculate the gradients for the whole dataset in order to perform a single update, batch gradient descent can be very slow to converge. In contrast, stochastic gradient descent performs a parameter update for each training example. These frequent, high-variance updates can cause the objective function to fluctuate heavily and lead to an unstable convergence. Finally, mini-batch gradient descent performs an update for every mini-batch of $n$ training examples. We choose mini-batch gradient descent for training the model because it offers a balance between the two other variants—its convergence tends to be faster than batch gradient descent and more stable than stochastic gradient descent.

Also, to counter the issues of class imbalance and training sample scarcity, we employ oversampling and data augmentation techniques, as described in algorithms 3 and 3. The rest of this section will explain the proposed method in more details.

**Algorithm 1.** The whole process of our proposed method.

**Data:** $S_0$ and $S_1$ list of all positive and negative training samples, respectively
**Result:** $P$ learned model parameters
$l_{min} \leftarrow +\infty$ `// minimum loss`
**repeat** *10000* **times**
  $B \leftarrow GenerateMiniBatch(S_0, S_1, 64)$ `// see algorithm 2`
  train the model on the mini-batch $B$
  $l_{cur} \leftarrow$ value of the loss function for the current model parameters on the validation set
  **if** $l_{cur} < l_{min}$ **then**
    $l_{min} \leftarrow l_{cur}$
    $P \leftarrow$ current model parameters

**Algorithm 2.** Our oversampling method.

**Input:** $S_0$ and $S_1$ list of all positive and negative training samples, respectively; $n$ size of mini-batch
**Output:** a balanced mini-batch
$i_0 \leftarrow 0$ `// index of the last used sample from` $S_0$
$i_1 \leftarrow 0$ `// index of the last used sample from` $S_1$
**function** $GenerateMiniBatch(S_0, S_1, n)$
  $B \leftarrow$ an empty list `// the mini-batch`
  **repeat** $n$ **times**
    draw $c$ randomly from $\{0, 1\}$ `// a randomly selected class`
    **if** $i_c = 0$ **then**
      shuffle $S_c$ `// the passed list itself is shuffled`
    $X \leftarrow ModifyImageRandomly(S_c[i_c])$ `// see algorithm 3`
    append $X$ to $B$
    $i_c \leftarrow (i_c + 1) \bmod size(S_c)$
  **return** $B$

**Algorithm 3.** Our data augmentation method.

**Input:** $X$ a 128×128-pixel image
**Output:** a 64×64-pixel crop of $X$ with various random modifications
**function** $ModifyImageRandomly(X)$
  consider a 64×64-pixel crop area centered on $X$
  `// rotate`
  draw $\theta$ randomly from the uniform distribution of $[0, 360)$
  rotate the crop area by $\theta$ degrees
  `// shift`
  draw $t_x$ and $t_y$ randomly from $\{x \in \mathbb{Z} \mid -5 \leq x \leq 5\}$
  shift the crop area horizontally by $t_x$ pixels
  shift the crop area vertically by $t_y$ pixels
  `// flip`
  draw $h$ and $v$ randomly from $\{0, 1\}$
  **if** $h = 1$ **then**
    flip $X$ horizontally
  **if** $v = 1$ **then**
    flip $X$ vertically
  `// change brightness`
  draw $\alpha$ randomly from the uniform distribution of $[-\log 1.25, \log 1.25]$
  multiply $X$ by $e^{\alpha}$
  **return** the crop area of $X$

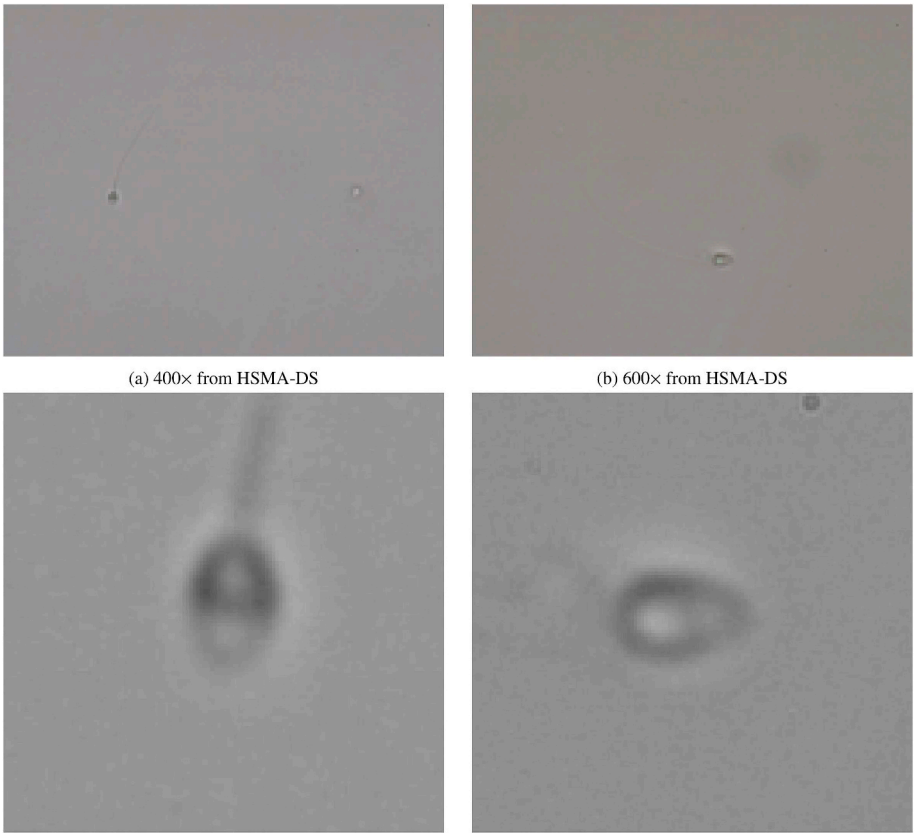
### 4.1. Sampling

As shown in Table 1, the classes are highly imbalanced in our dataset. For such datasets, most classifiers are biased towards the majority class (i.e., *normal* in our case) and show poor accuracy on the minority class. It is also possible for a classifier to predict every sample as normal and ignore the minority class. This is especially problematic in this task, since misclassification of an abnormal sperm as normal is very costly in a real-world scenario.

To combat the issue of class imbalance, we train the model using balanced mini-batches. For creating a balanced mini-batch, we employ an oversampling method described as follows. Positive and negative training samples are kept in two separate lists. To add one sample to a mini-batch, one of the lists is chosen randomly with a fair chance, and the sample at the top of that list is picked. After all samples in a list are used, the list will be shuffled. This process is described in Algorithm 2. Using this oversampling method, classes in each mini-batch will be balanced, regardless of the actual distribution of samples.

### 4.2. Data augmentation

In the training phase of deep learning algorithms, parameters of the network are tuned in such a way that the model maps an input (i.e., a



(a) 400× from HSMA-DS (b) 600× from HSMA-DS

(c) a in MHSMA-128 (d) b in MHSMA-128

**Fig. 1.** Sample images from dataset.

sperm image) to a label (i.e., *normal* or *abnormal*). State-of-the-art neural networks commonly have parameters in the order of millions. As the number of parameters increases (i.e., when the problem becomes more complex), the neural network needs more training examples to properly tune the parameters. On the other hand, collecting human sperm images is a very hard and costly task. To the best of our knowledge, it can be said that MHSMA is the largest available dataset in this field of study.

To counter the problem of training example scarcity, we employ a data augmentation technique in order to virtually enlarge the dataset and prevent overfitting. Before feeding the images to the model, we extract a $64 \times 64$-pixel crop from the processed $128 \times 128$-pixel images, with various random modifications. This process is described in Algorithm 3. Also, Fig. 2 shows a few augmentation samples. The random modifications are as follows:

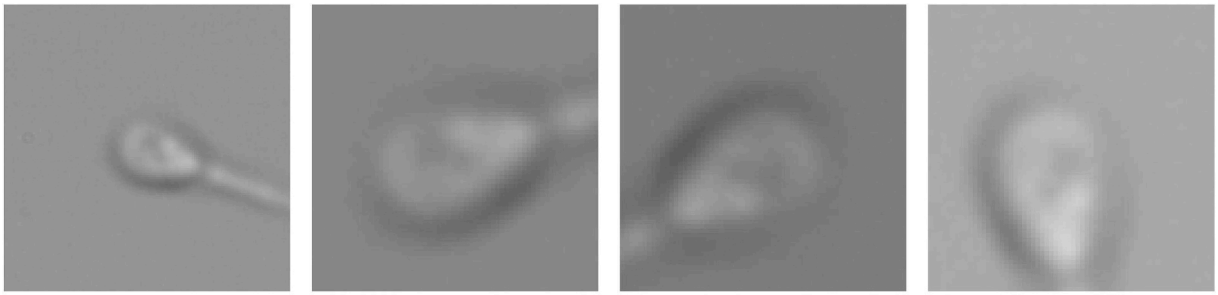
*Rotating.* We rotate the crop area by $\theta$ degrees, where $\theta$ is randomly picked from the uniform distribution of [0,360) (see Fig. 3).

*Shifting.* We move the crop area towards the horizontal and vertical axis by $t_x$ and $t_y$ pixels, which are randomly drawn from $\{x \in \mathbb{Z} \mid -5 \leq x \leq 5\}$ (see Fig. 4).

*Flipping.* We flip the image horizontally (mirror) and vertically (reflect), each with a chance of 0.5 (see Fig. 5).

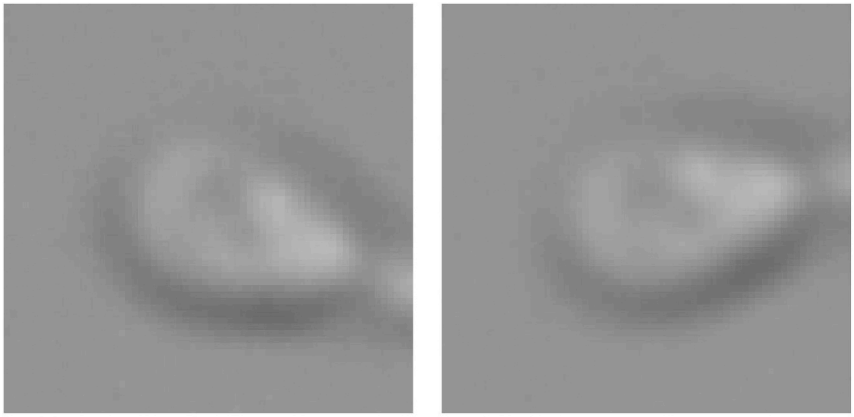
*Changing brightness.* We multiply the pixel values by $e^{\alpha}$, where $\alpha$ is randomly drawn from the uniform distribution of $[-\beta, \beta]$ and $\beta = \log(1.25)$ (see Fig. 6).

All of these steps are applied to every training example. Finally, we normalize the image by subtracting its mean and dividing by 255, as shown in equation (2), where $x$ is a single image.



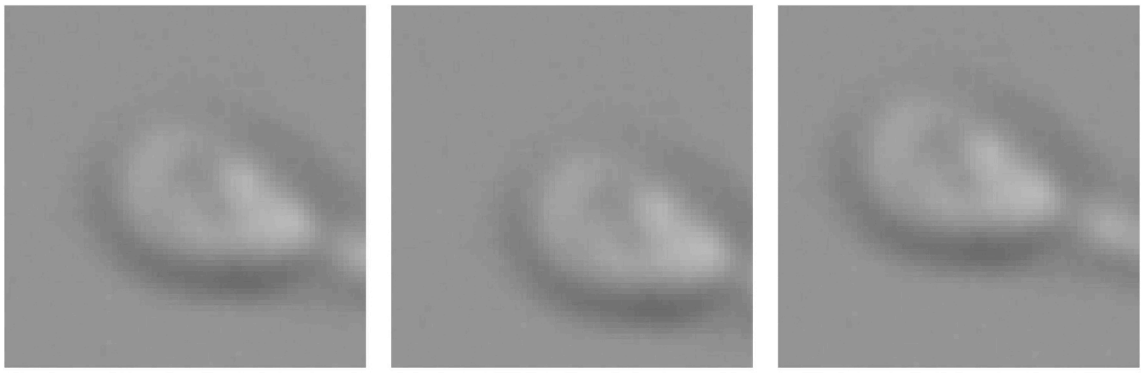
(a) Original (b) Augmented (c) Augmented (d) Augmented

**Fig. 2.** Data augmentation samples.



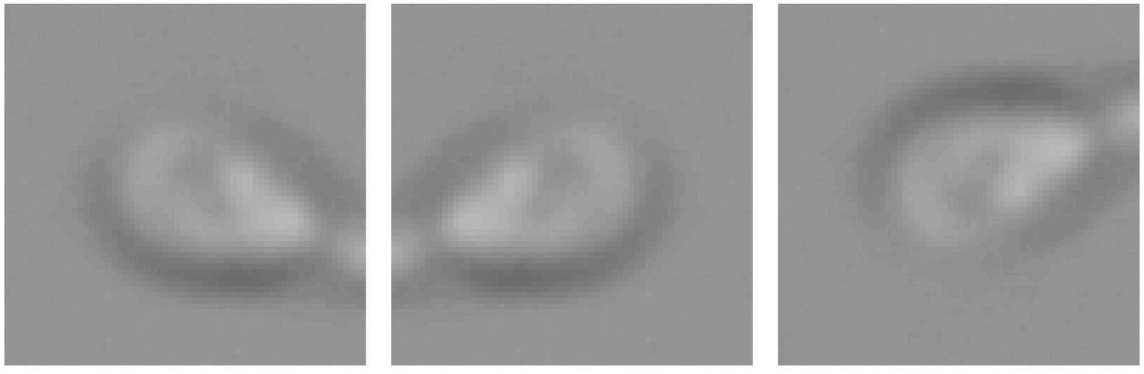
(a) No modifications (b) Rotated 30 degrees

**Fig. 3.** Data augmentation: rotating.



(a) No modifications (b) Shifted (−5, −5) pixels (c) Shifted (5, 5) pixels

**Fig. 4.** Data augmentation: shifting.



(a) No modifications (b) Flipped horizontally (c) Flipped vertically

**Fig. 5.** Data augmentation: flipping.

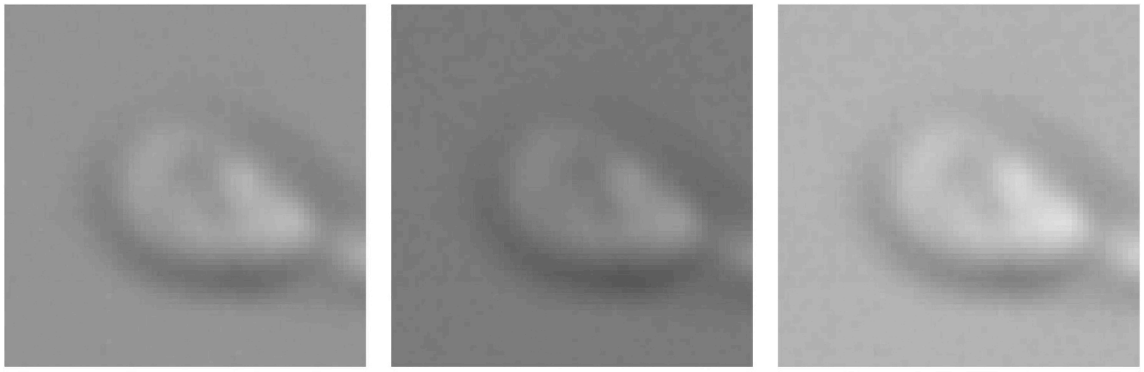
$$normal(x) = \frac{x - mean(x)}{255} \quad (2)$$

It should be noted that data augmentation is done only on the training set—validation and test sets are not augmented. Also, in our data augmentation method, we intentionally avoid skewing and scaling the images in order to maintain the morphological features of the sperms.

### 4.3. Model architecture

The architecture of our convolutional neural network is inspired by the VGG network [41]. Details of the proposed architecture is shown in Fig. 7 and Table 2. This network consists of 24 convolution, three pooling, and two fully-connected layers. The model has a total of 5,637,649 trainable parameters.

We initialize the weights using the LeCun normal initializer [22,24] and biases with zeros. The LeCun normal initializer draws samples from a truncated normal distribution centered on zero with a standard



(a) No modifications (b) Decreased brightness (×0.80) (c) Increased brightness (×1.25)

**Fig. 6.** Data augmentation: changing brightness.

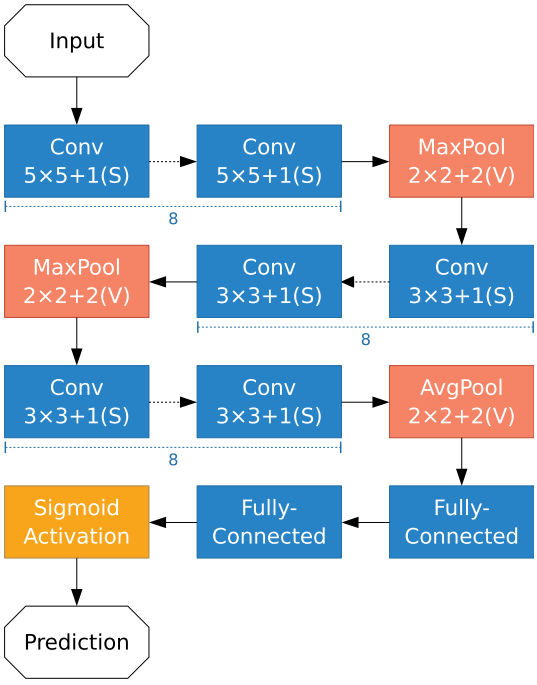


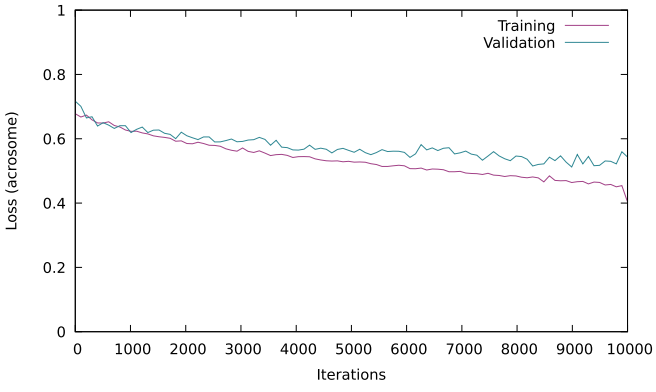

**Fig. 7.** Our proposed architecture (dotted arrows represent repeated layers).

deviation of $\sigma = \sqrt{\frac{1}{fanin}}$, where *fanin* is the number of input units in the weight tensor [12].

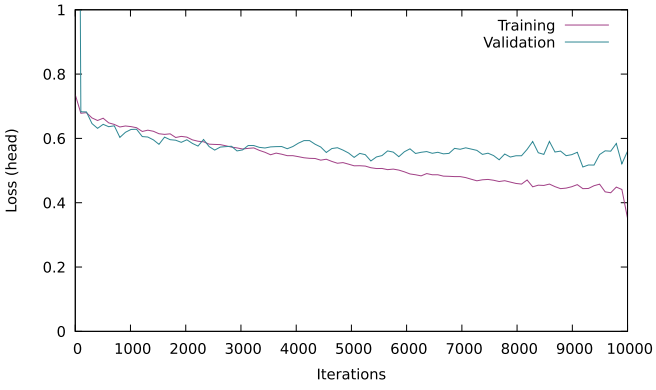
We use Scaled Exponential Linear Units (SELUs) [22] as the activation function for the convolutional and fully-connected layers, while applying the sigmoid activation function to the output layer. The SELU

**Table 2**
Our proposed architecture.

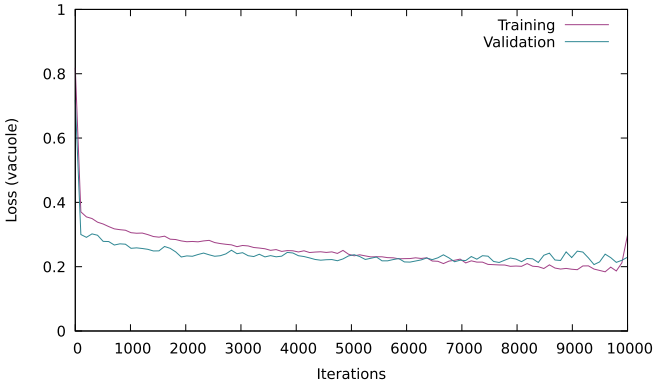
| Layer | Details | Count | Params |
|---|---|---|---|
| Convolution | 16 filters, 5 × 5 kernel, stride 1 same padding, selu activation | 8 | 45,328 |
| Max Pooling | 2 × 2 pool, stride 2, valid padding | 1 | 0 |
| Convolution | 32 filters, 3 × 3 kernel, stride 1 same padding, selu activation | 8 | 69,376 |
| Max Pooling | 2 × 2 pool, stride 2, valid padding | 1 | 0 |
| Convolution | 64 filters, 3 × 3 kernel, stride 1 same padding, selu activation | 8 | 276,992 |
| Average Pooling | 2 × 2 pool, stride 2, valid padding | 1 | 0 |
| Fully-connected | 1024 units, selu activation | 2 | 5,244,928 |
| Fully-connected | 1 unit, sigmoid activation | 1 | 1,025 |




(a) Acrosome




(b) Head

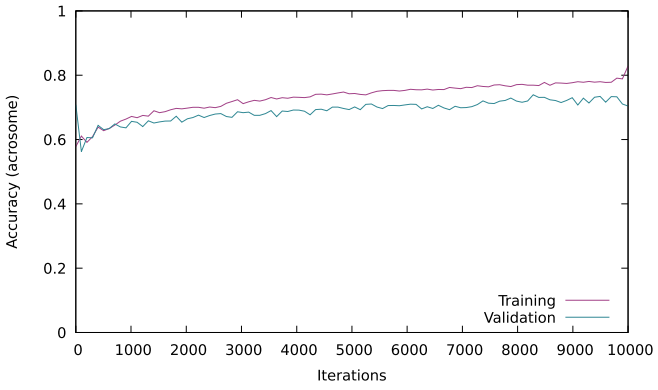



(c) Vacuole

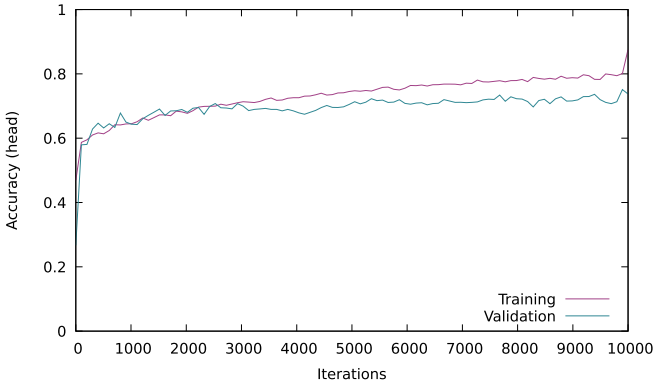
**Fig. 8.** Training and validation loss over iterations of training.

activation function is defined in equation (3), where *scale* and $\alpha$ are predefined constants, chosen so that the mean and variance of the inputs are preserved between two consecutive layers [12]. SELU is closely related to Exponential Linear Unit (ELU) [13], which is defined in equation (4). Also, the sigmoid activation function is defined in equation (5).

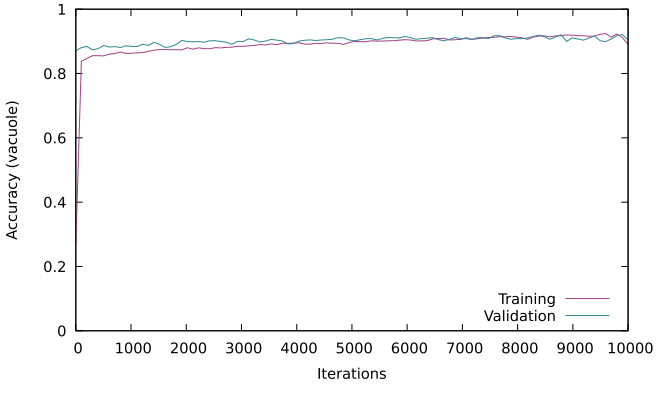
$$selu(x) = scale \times elu(x, \alpha) \tag{3}$$




(a) Acrosome




(b) Head




(c) Vacuole

**Fig. 9.** Training and validation accuracy over iterations of training.

$$elu(x, \alpha) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha(e^x - 1), & \text{otherwise} \end{cases} \tag{4}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

We utilize the Adam optimization method for training. Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. This method is computationally efficient, has little memory requirements, and is well suited for problems that, like ours, are large in terms of parameters [21]. We use a constant learning rate of 0.0001, while exponential decay rates for the moment estimates are set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Also, we use binary cross-entropy as the loss function. This function is defined in equation (6), where $\theta$ denotes the model parameters, $n$ is the number of samples in a mini-batch, and $y_i$ and $p_i$ denote respectively the labels and predicted probabilities for the samples.

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{6}$$

The implementation is done using TensorFlow [1] and Keras [12]. Training the model took approximately sixteen hours on a single Intel Xeon CPU, with no GPU.

## 5. Results

We trained the model independently for classification of sperm acrosome, head, and vacuole. For each one, the model is trained for 10,000 iterations on the training set (1000 samples). After each iteration, loss value is calculated on the validation set (240 samples) and the checkpoint with the lowest validation loss is saved. Once training is done, we evaluate the saved checkpoint on the held-out test set (300 samples).

Fig. 8 shows training and validation loss for each label over iterations of training. Similarly, Fig. 9 shows accuracy. In these figures, curves are smoothed for better visualization.

Results of the evaluations on the test set are shown in Table 3. In this table, ROC AUC is the area under the receiver operating characteristic curve and MCC is the Matthews correlation coefficient. The evaluation metrics are defined in equations (7)–(13), where $TP$, $TN$, $FP$, and $FN$ denote respectively true positive (i.e., normal sperms classified correctly), true negative (i.e., abnormal sperms classified correctly), false positive (i.e., abnormal sperms classified incorrectly), and false negative (i.e., normal sperms classified incorrectly). In equation (12), $X_1$ is the score for a positive instance and $X_0$ is the score for a negative instance. It should be noted that for this task, the $F_{0.5}$ score is a very

**Table 4**
Confusion matrix for evaluation on test set.

| Label | Actual class | | Predicted class |
|---|---|---|---|
| | Normal | Abnormal | |
| Acrosome | 171 true positives | 28 false positives | Normal |
| | 42 false negatives | 59 true negatives | Abnormal |
| Head | 187 true positives | 37 false positives | Normal |
| | 32 false negatives | 44 true negatives | Abnormal |
| Vacuole | 251 true positives | 15 false positives | Normal |
| | 11 false negatives | 23 true negatives | Abnormal |

**Table 3**
Results of evaluation on test set (all values except MCC are in percent).

| Label | Accuracy | Precision | Recall | $F_{0.5}$ score | G-mean | ROC AUC | MCC |
|---|---|---|---|---|---|---|---|
| Acrosome | 76.67 | 85.93 | 80.28 | 84.74 | 83.06 | 83.89 | +0.4618 |
| Head | 77.00 | 83.48 | 85.39 | 83.86 | 84.43 | 77.80 | +0.4053 |
| Vacuole | 91.33 | 94.36 | 95.80 | 94.65 | 95.08 | 88.08 | +0.5910 |

**Table 5**
Comparison of our results with those achieved by Ghasemian et al. [15] (all values except MCC are in percent).

| Label | Method | Accuracy | Precision | Recall | $F_{0.5}$ score | G-mean | ROC AUC | MCC |
|---|---|---|---|---|---|---|---|---|
| | **Proposed** | **77.00** | **83.48** | **85.39** | **83.86** | **84.43** | **77.80** | **+0.4053** |
| Head | G. et al. | 61.00 | 76.71 | 71.79 | 75.68 | 74.21 | 47.26 | −0.0511 |
| | **Proposed** | **91.33** | **94.36** | **95.80** | **94.65** | **95.08** | **88.08** | **+0.5910** |
| Vacuole | G. et al. | 80.33 | 83.21 | 93.56 | 85.09 | 88.23 | 63.95 | +0.3492 |

**Table 6**
Statistical tests for comparison of our results with those achieved by Ghasemian et al [15].

| Label | Test | Statistic | *p*-value |
|---|---|---|---|
| | Student's paired *t*-test | 4.56 | 0.000007 |
| Head | McNemar's test | 19.53 | 0.000010 |
| | Student's paired *t*-test | 4.28 | 0.000026 |
| Vacuole | McNemar's test | 17.29 | 0.000032 |

**Table 7**
Contingency table for comparison of our results with those achieved by Ghasemian et al [15].

| | Ghasemian et al. | | |
|---|---|---|---|
| Label | Correct classification | Incorrect classification | Proposed method |
| | 148 | 83 | Correct classification |
| Head | 35 | 34 | Incorrect classification |
| | 226 | 48 | Correct classification |
| Vacuole | 15 | 11 | Incorrect classification |

important metric, because, in the real world, the cost of a false positive (an abnormal sperm misclassified as normal) is much more than the cost of a false negative. In other words, precision is much more important than recall. Table 4 shows the confusion matrix for each label.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$F_\beta \text{ score} = (1 + \beta^2) \times \frac{\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision+Recall}} \tag{10}$$

$$\text{G}- \text{ mean} = \sqrt{\text{Precision} \times \text{Recall}} \tag{11}$$

$$\text{ROC AUC} = P(X_1 > X_0) \tag{12}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{13}$$

For all three labels, training is done with the same settings. The only exception is that, for vacuole, we did not use the oversampling method.

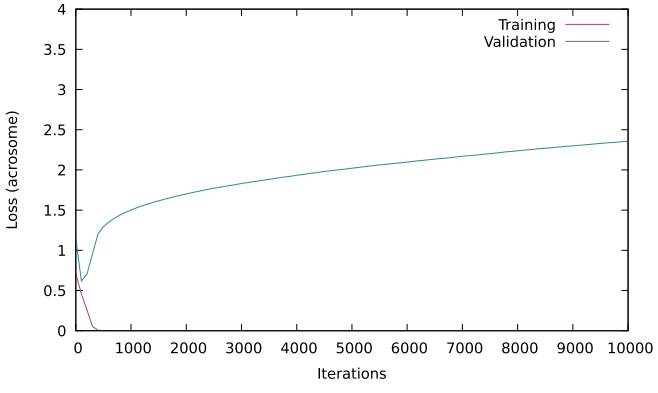
Inference time is calculated with a batch size of 32 and averaged over 100 executions. Average inference time per sample is 24 milliseconds on an Intel Core i5-6200U ("Ultra-low power") CPU, and 4 milliseconds on an Nvidia GeForce 940MX GPU. These processing units can be found on mainstream laptops. This means that using our method, abnormalities can be detected in real-time, even on a mainstream laptop computer.

**Table 8**
Five-fold cross-validation results for each fold plus mean and standard deviation (SD) for three labels.

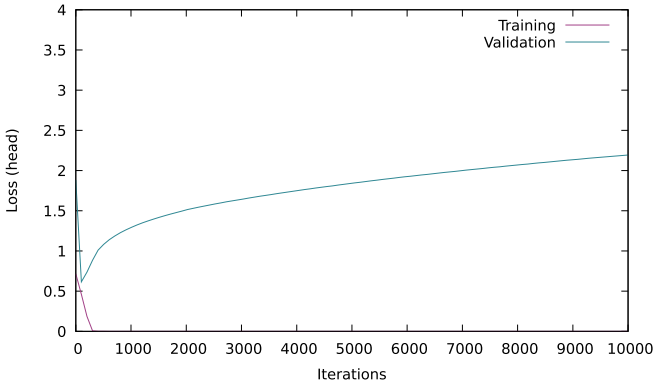
| Label | Fold | Accuracy | Precision | Recall | $F_{0.5}$ score | G-mean | ROC AUC | MCC |
|---|---|---|---|---|---|---|---|---|
| Acrosome | 1 | 78.25 | 82.30 | 85.15 | 82.85 | 83.71 | 82.53 | +0.5111 |
| | 2 | 77.27 | 82.06 | 85.92 | 82.81 | 83.97 | 81.80 | +0.4527 |
| | 3 | 82.14 | 90.45 | 85.41 | 89.40 | 87.89 | 85.24 | +0.5454 |
| | 4 | 74.35 | 81.17 | 83.03 | 81.53 | 82.09 | 77.48 | +0.3699 |
| | 5 | 78.90 | 88.18 | 81.36 | 86.72 | 84.70 | 83.82 | +0.5155 |
| | **Mean** | **78.18** | **84.83** | **84.17** | **84.66** | **84.47** | **82.17** | **+0.4789** |
| | **SD** | **0.0282** | **0.0419** | **0.0192** | **0.0329** | **0.0214** | **0.0293** | **0.0696** |
| Head | 1 | 75.00 | 76.34 | 93.02 | 79.18 | 84.27 | 81.23 | +0.3395 |
| | 2 | 77.27 | 86.57 | 82.02 | 85.62 | 84.27 | 83.16 | +0.4385 |
| | 3 | 78.90 | 84.35 | 87.00 | 84.86 | 85.66 | 82.44 | +0.4589 |
| | 4 | 79.22 | 87.89 | 84.12 | 87.11 | 85.99 | 84.52 | +0.4620 |
| | 5 | 74.03 | 78.95 | 87.44 | 80.51 | 83.09 | 76.37 | +0.2946 |
| | **Mean** | **76.88** | **82.82** | **86.72** | **83.46** | **84.65** | **81.54** | **+0.3987** |
| | **SD** | **0.0231** | **0.0498** | **0.0416** | **0.0343** | **0.0118** | **0.0313** | **0.0767** |
| Vacuole | 1 | 89.94 | 94.27 | 93.92 | 94.20 | 94.10 | 93.98 | +0.6004 |
| | 2 | 83.77 | 89.92 | 89.92 | 89.92 | 89.92 | 87.35 | +0.4825 |
| | 3 | 89.94 | 94.49 | 93.39 | 94.27 | 93.94 | 92.91 | +0.6445 |
| | 4 | 90.58 | 96.11 | 92.86 | 95.44 | 94.47 | 94.07 | +0.6375 |
| | 5 | 87.34 | 96.34 | 88.76 | 94.72 | 92.48 | 90.45 | +0.5660 |
| | **Mean** | **88.31** | **94.23** | **91.77** | **93.71** | **92.98** | **91.75** | **+0.5862** |
| | **SD** | **0.0283** | **0.0258** | **0.0228** | **0.0218** | **0.0187** | **0.0286** | **0.0659** |

Ghasemian et al. [15] have proposed an image processing-based approach to this problem. For a better comparison of the performance of their method with ours, we ran their own implementation of their proposed algorithm on the same test set we used for evaluation. It should be noted that HSMA-DS images were used with their method, since the algorithm is designed to work with this dataset. As presented in Table 5, our method outperforms theirs for sperm head and vacuole on all the metrics discussed earlier. Moreover, they have not proposed a method for detecting abnormalities in acrosome, which is a very complex problem in this dataset. To compare the results statistically, Student's paired *t*-test and McNemar's test were performed. As shown in Table 6, in both tests and for both labels, $p-\text{value} < 0.0001$. This shows that the difference in performance is statistically significant. Also, Table 7 shows the contingency table for the two classifiers.

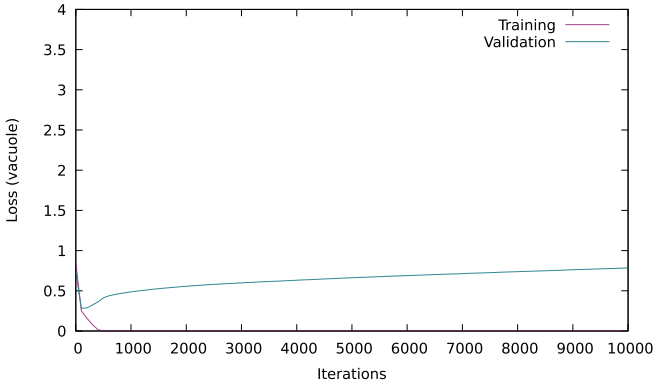
In order to further evaluate the performance of the proposed method, we performed five-fold cross-validation. In five-fold cross-validation, the size of the test set is 308 (a single fold). Of the remaining 1,232 samples (four folds), 1,000 are randomly selected as the training data, and the remaining 232 samples compose the validation set. Training and evaluation are done in the same way as described earlier in this section. The results of five-fold cross-validation are shown in Table 8. This table shows the results of our proposed algorithm for each fold. We also reported the mean and standard deviation results for five folds. As it is shown in this table, the achieved results are close to the results from Table 3.




(a) Acrosome




(b) Head




(c) Vacuole

**Fig. 10.** Training and validation loss with no data augmentation.

In our experiments, data augmentation proves to be a very effective technique for preventing the model from overfitting. As shown in Fig. 10, without data augmentation, the model quickly overfits to training data and fails to generalize.

Also, we experimented with the dropout technique, but it did not improve our results. Our experiments includes applying either regular dropout [42] or *alpha dropout* [22] to the fully-connected and/or pooling layers.

For sperm acrosome and head, our oversampling method improves accuracy, precision, and the $F_{0.5}$ score, while harming recall. As mentioned earlier, in this task, precision is much more important than recall, therefore, this is an acceptable trade-off. For vacuole, however, oversampling led to slightly worse accuracy and $F_{0.5}$ score. As a result, we did not use oversampling during the training of the final model for vacuole. Table 9 shows the effect of our oversampling method on the results.

As shown in Table 9, for all three labels, oversampling improves precision while harming recall. The reason behind this phenomenon is that, by creating mini-batches that on average contain an equal number of positive and negative samples, our oversampling method prevents the model from getting biased towards the majority class (i.e., the positive class). This leads to fewer false positives, which in turn results in improved precision. On the other hand, this also increases the number of false negatives and consequently decreases the recall rate. Table 10 shows the effect of our oversampling method on the confusion matrix.

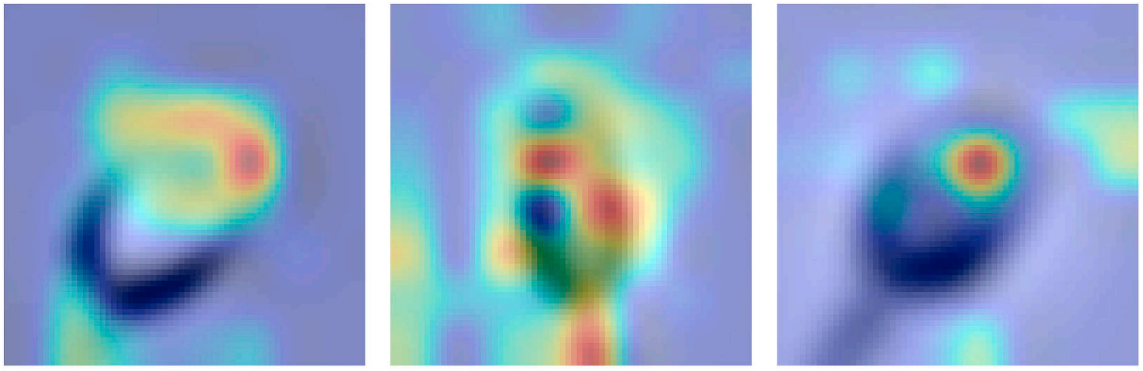
The oversampling method has two opposite effects on accuracy. On one hand, by preventing the model from getting biased towards the majority class, this method can decrease the accuracy. The more imbalanced a dataset, the greater this effect will be. On the other hand, by feeding more samples from the minority class, oversampling can enable the model to learn better features for this class and thus increase the accuracy. This effect can be less significant on an extremely imbalanced dataset, due to the small number of samples from the minority class

**Table 9**
Effect of oversampling on results (all values are in percent).

| Label | Oversampling | Accuracy | Precision | Recall | $F_{0.5}$ score |
|---|---|---|---|---|---|
| Acrosome | No | 74.33 | 77.64 | **89.67** | 79.78 |
| | **Yes** | **76.67** | **85.93** | 80.28 | **84.74** |
| Head | No | 75.67 | 78.29 | **92.24** | 80.74 |
| | **Yes** | **77.00** | **83.48** | 85.39 | **83.86** |
| Vacuole | **No** | **91.33** | 94.36 | **95.80** | **94.65** |
| | Yes | 90.00 | **94.62** | 93.89 | 94.47 |

**Table 10**
Effect of oversampling on the confusion matrix (numbers in parentheses indicate results achieved without oversampling).

| Label | Actual class | | Predicted class |
|---|---|---|---|
| | Normal | Abnormal | |
| Acrosome | 171 (191) true positives | 28 (55) false positives | Normal |
| | 42 (22) false negatives | 59 (32) true negatives | Abnormal |
| Head | 187 (202) true positives | 37 (56) false positives | Normal |
| | 32 (17) false negatives | 44 (25) true negatives | Abnormal |
| Vacuole | 246 (251) true positives | 14 (15) false positives | Normal |
| | 16 (11) false negatives | 24 (23) true negatives | Abnormal |



(a) Acrosome (b) Head (c) Vacuole

**Fig. 11.** Grad-CAM visual explanations for different labels (warmer colors indicate more attention).

available for training. The results show that our oversampling method improves the accuracy of the model for acrosome and head while leading to a decrease in accuracy for vacuole. This can be linked to the extreme class imbalance for vacuole, i.e., 84.48% positive in the whole dataset. This figure is substantially lower for the other two labels, i.e., 70.52% and 72.86% for acrosome and head, respectively.

### 5.1. Visual explanation

We employ Gradient-weighted Class Activation Mapping (Grad-CAM) [23,39] to generate *visual explanations*—visualization of regions of input images that are "important" for predictions. This helps us better understand our image classification models. Fig. 11 shows visual explanations for some samples that were classified correctly. These visual explanations show that our models have indeed learned to pay attention to the image regions that are actually important for the classification task, i.e., regions of acrosome, head, and vacuole of sperms, as shown in Fig. 12.

## 6. Discussion

Ability to work with non-stained images from low-magnification microscopes is one of the main features of our proposed deep learning method. Almost all existing algorithms work on stained images from high-magnification microscopes. Our method is also capable of processing each sperm image in real-time (i.e., less than a second). These characteristics are desirable features that are important for treatments applications.

As mentioned earlier, we marked the position of the sperm head in each image manually as the first step in data preparation. In order to have a fully automatic system for assessment of human sperm cell images, this step should be automated too. While our proposed method does not address this issue, the dataset we introduced provides an opportunity to develop such a system. This is an interesting topic for future work. Recently, deep learning-based approaches to object detection such as the work of Girshick et al. [17]; Girshick [16]; Ren et al. [36]; and Redmon et al. [35] have been very successful. These approaches can be employed to develop a method for automatic detection of sperm head regions in an image.

Extracting midpiece and tail morphological features is another important task that we have not considered in this paper. It should be noted that comparing to head, vacuole, and acrosome analysis, these two problems are easier tasks. In addition, our method cannot measure the motility of each sperm, because the goal of the proposed method is sperms' morphological analysis from images, not videos. However, motility is an important parameter for measuring the quality of each sperm. In the future, we will work on these issues, in order to have a fully automatic sperm quality software.

To the best of our knowledge, there do not exist many algorithms that can be fairly compared with the proposed method. Most of the current algorithms work on high-resolution and/or stained images, which is not the same as our problem. In other words, our problem is far more difficult. On the other hand, there are also some methods that work on sperm detection, not on morphological analysis.

One of the most similar works to our algorithm is the SMA method [15,29]. This method is also able to work with non-stained human sperm in real-time with a low-magnification microscope (400× and 600×) and it can be used in the ICSI process. As it is shown in Table 5, our method works better in vacuole and head abnormality detection than theirs. In addition, their method cannot extract the size and shape of the acrosome. SMA method also processes sperm images in real-time, however, the proposed method is about seven times faster than this method. SMA method is sensitive to noise and microscope light, and for achieving best results, they may need to change some threshold values. In other words, by changing dataset, some thresholds and algorithms need to be changed. On the other hand, our proposed deep learning algorithm is totally automatic. It just needs a dataset, and all following process will be done automatically. The proposed method does not need an explicit noise removal step and threshold setting. Our method only needs images as input and all other processes will be done without any human intervention.

There are very limited deep learning based algorithms in sperm detection and morphology analysis. One of these algorithms uses a set of convolutional neural network (CNN) architectures in order to segment and detect sperm cells in semen sample images [31]. One of the strengths of their algorithm is the ability to work on non-stained sperms

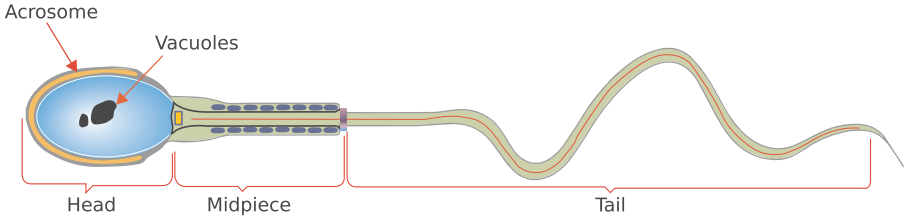



**Fig. 12.** Diagram of a human sperm cell.

for full sperm quality analysis. They have also constructed a dataset of 765 images. These images are grayscale and collected from 35 independent sperm samples. Their best network has achieved 93.87% precision and 91.89% recall on their predefined test set. However, compared to the proposed method, no morphological analyses were done in this method and they have only focused on sperm cell counting and segmentation.

In another study, a dataset of sperm head images with expert annotation, named SCIAN-MorphoSpermGS, has been constructed [10]. There are five different labels in this dataset: normal, tapered, pyriform, small, and amorphous. This dataset describes a gold-standard for evaluating and comparing known techniques for morphological analysis of human sperm heads. The authors of this paper have not proposed new algorithms for sperms' morphological analysis. However, they applied some existing baselines for this analysis. They utilized three shape-based descriptors (Hu moments, Zernike moments, and Fourier descriptors) for building models using four supervised learning methods (Nearest Neighbor, naive Bayes, decision trees, and Support Vector Machine). Their best classification accuracy has been achieved using the Fourier descriptor and SVM: only 49%.

In another research, a Dictionary Learning (DL) technique has been utilized for construction of a dictionary for sperm head shapes [40]. This dictionary has been used for classifying the sperm heads into four different classes. From sperm head images, square patches have been extracted. These patches from each class of sperm have been used for learning class-specific dictionaries. They have evaluated their method on two datasets and have achieved an average accuracy of 92.2% on the HuSHeM dataset and an average recall of 62% on the SCIAN-MorphoSpermGS dataset. Comparing the proposed method, this DL method is not able to detect abnormalities regarding vacuole and acrosome.

Dai et al. [14] have presented an automated technique for non-invasive measurement of motility and morphological parameters of individual sperms. Their method also processes each sperm in nearly real-time. They have proposed an adapted joint probabilistic data association filter for multi-sperm tracking. This filter has shown acceptable results in identifying sperms that have small spatial distances or intersect. For morphological analysis, they have integrated total variation norm into the quadratic cost function method in order to tackle the inhomogeneous image intensity problem. Based on their evaluation, an accuracy about 90% has been achieved for morphology measurement problem.

In another study, the problem has been divided into two stages: 1) detection and extraction of individual spermatozoon and 2) feature extraction and classification [9]. For the first stage, a novel method, called nth-fusion has been introduced. The accuracy of this method on 250 sperm images was 95.78%. For the second stage, several features were extracted from images and classification is done using the WHO [45] criteria. It should be noted that this method works on stained images.

## 7. Conclusion

In this paper, we have proposed a deep learning method for selecting the best sperms in the ICSI procedure. First of all, we have prepared a unique dataset that contains 1,540 images from different types of sperms. This dataset is freely available to the public. Our proposed deep learning model is a compact and accurate model for classification of sperms. This model extracts features of acrosome, head shape, and vacuole from sperm images. In almost all previous studies, sperms were fixed, stained, and photographed. As a result, these sperms are not useful for treatment purposes in the ICSI procedure. However, our proposed method selects the best fresh sperms, which can be used for injection. Another feature of our model is its ability to work with low-magnifications (400× and 600×) and noisy images. Measuring the level of acrosome in the sperm head is another important feature of our algorithm, since the size and shape of the acrosome are particularly important for sperm binding to the oocyte. The experimental results show the high accuracy of the proposed deep learning model. It can be said that this method is the state-of-the-art for selecting the best sperms. In addition, the time required for processing each fresh sperm image is less than a second and our model can be regarded as a real-time method.

Currently, we are working on finding ways to improve the accuracy of our algorithm. Transfer learning and designing more complex deep learning models are our further directions of research for solving this problem.

## Conflict of interest statement

None declared.

## Acknowledgement

The authors thank Dr. Fatemeh Ghasemian for her valuable efforts for dataset annotation and her scientific advises.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.04.030.

## References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a System for Large-Scale Machine Learning, OSDI, 2016, pp. 265–283.
[2] V. Abbiramy, V. Shanthi, Spermatozoa segmentation and morphological parameter analysis based detection of teratozoospermia, Int. J. Comput. Appl. 3 (2010) 19–23.
[3] M.D. Abràmoff, P.J. Magalhães, S.J. Ram, Image processing with ImageJ, Biophot. Int. 11 (2004) 36–42.
[4] E. Alegre, M. Biehl, N. Petkov, L. Sánchez, Automatic classification of the acrosome status of boar spermatozoa using digital image processing and LVQ, Comput. Biol. Med. 38 (2008) 461–468.
[5] S.N. Babayev, C.W. Park, O. Bukulmez, Intracytoplasmic sperm injection indications: how rigorous? Seminars in Reproductive Medicine, Thieme Medical Publishers, 2014, pp. 283–290.
[6] A. Bijar, A.P. Benavent, M. Mikaeili, et al., Fully automatic identification and discrimination of sperm's parts in microscopic images of stained human semen smear, J. Biomed. Sci. Eng. 5 (2012) 384.
[7] E. Blahová, J. Máchal, L. Máchal, I. Milaković, Š. Hanuláková, Eliminating the effect of pathomorphologically formed sperm on resulting gravidity using the intracytoplasmic sperm injection method, Exp. Ther. Med. 7 (2014) 1000–1004.
[8] K. Boumaza, A. Loukil, K. Aarizou, Automatic human sperm concentration in microscopic videos, Med. Technol. J. 2 (2018) 301–307.
[9] H. Carrillo, J. Villarreal, M. Sotaquira, A. Goelkel, R. Gutierrez, A computer aided tool for the assessment of human sperm morphology, 2007 IEEE 7th International Symposium on BioInformatics and BioEngineering, IEEE, 2007, pp. 1152–1157.
[10] V. Chang, A. Garcia, N. Hitschfeld, S. Härtel, Gold-standard for computer-assisted morphological sperm analysis, Comput. Biol. Med. 83 (2017) 143–150.
[11] V. Chang, J.M. Saavedra, V. Castañeda, L. Sarabia, N. Hitschfeld, S. Härtel, Gold-standard and improved framework for sperm head segmentation, Comput. Methods Progr. Biomed. 117 (2014) 225–237.
[12] F. Chollet, et al., Keras, https://keras.io, (2015).
[13] D.A. Clevert, T. Unterthiner, S. Hochreiter, Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), (2015) arXiv preprint arXiv:1511.07289.
[14] C. Dai, Z. Zhang, J. Huang, X. Wang, C. Ru, H. Pu, S. Xie, J. Zhang, S. Moskovtsev, C. Librach, et al., Automated non-invasive measurement of single sperms motility and morphology, IEEE Trans. Med. Imaging 37 (2018) 2257–2265.
[15] 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 Progr. Biomed. 122 (2015) 409–420.
[16] R. Girshick, Fast R-CNN, Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[17] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[18] T.B. Haugen, J.M. Andersen, O. Witczak, H.L. Hammer, S.A. Hicks, R.J. Borgli, P. Halvorsen, M.A. Riegler, Visem: A Multimodal Video Dataset of Human Spermatozoa, (2019).
[19] H.O. Ilhan, N. Aydin, A novel data acquisition and analyzing approach to spermiogram tests, Biomed. Signal Process. Control 41 (2018) 129–139.
[20] A. Isidori, M. Latini, F. Romanelli, Treatment of male infertility, Contraception 72

(2005) 314–318.
[21] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, Proc. 3rd Int. Conf. Learn. Representations, 2014.
[22] G. Klambauer, T. Unterthiner, A. Mayr, S. Hochreiter, Self-normalizing neural networks, Advances in Neural Information Processing Systems, 2017, pp. 971–980.
[23] R. Kotikalapudi, et al., keras-vis, https://github.com/raghakot/keras-vis, (2017).
[24] Y.A. LeCun, L. Bottou, G.B. Orr, K.R. Müller, Efficient backprop, Neural Networks: Tricks of the Trade, Springer, 2012, pp. 9–48.
[25] J. Li, K.K. Tseng, H. Dong, Y. Li, M. Zhao, M. Ding, Human sperm health diagnosis with principal component analysis and k-nearest neighbor algorithm, Medical Biometrics, 2014 International Conference on, IEEE, 2014, pp. 108–113.
[26] L. Maree, S. Du Plessis, R. Menkveld, G. Van der Horst, Morphometric dimensions of the human sperm head depend on the staining method used, Hum. Reprod. 25 (2010) 1369–1382.
[27] R. Medina-Rodríguez, L. Guzmán-Masías, H. Alatrista-Salas, C. Beltrán-Castañón, Sperm cells segmentation in micrographic images through lambertian reflectance model, International Conference on Computer Analysis of Images and Patterns, Springer, 2015, pp. 664–674.
[28] R. Menkveld, C.A. Holleboom, J.P. Rhemrev, Measurement and significance of sperm morphology, Asian J. Androl. 13 (2011) 59.
[29] S.A. Mirroshandel, F. Ghasemian, Automated morphology detection from human sperm images, Intracytoplasmic Sperm Injection, Springer, 2018, pp. 99–122.
[30] G.L. Monte, F. Murisier, I. Piva, M. Germond, R. Marci, Focus on intracytoplasmic morphologically selected sperm injection (IMSI): a mini-review, Asian J. Androl. 15 (2013) 608.
[31] M.S. Nissen, O. Krause, K. Almstrup, S. Kjærulff, T.T. Nielsen, M. Nielsen, Convolutional neural networks for segmentation and object detection of human semen, Scandinavian Conference on Image Analysis, Springer, 2017, pp. 397–406.
[32] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1979) 62–66.
[33] G. Palermo, H. Joris, P. Devroey, A.C. Van Steirteghem, Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte, The Lancet 340 (1992) 17–18.
[34] L. Ramos, P. de Boer, E.J. Meuleman, D.D. Braat, A.M. Wetzels, Evaluation of ICSI-selected epididymal sperm samples of obstructive azoospermic males by the CKIA system, J. Androl. 25 (2004) 406–411.
[35] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[36] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[37] L. Sánchez, N. Petkov, E. Alegre, Statistical approach to boar semen head classification based on intracellular intensity distribution, International Conference on Computer Analysis of Images and Patterns, Springer, 2005, pp. 88–95.
[38] L. Sánchez, N. Petkov, E. Alegre, Statistical approach to boar semen evaluation using intracellular intensity distribution of head images, Cell. Mol. Biol. 52 (2006) 38–43.
[39] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-Cam: Why Did You Say that? (2016) arXiv preprint arXiv:1611.07450.
[40] F. Shaker, S.A. Monadjemi, J. Alirezaie, A.R. Naghsh-Nilchi, A dictionary learning approach for human sperm heads classification, Comput. Biol. Med. 91 (2017) 181–190.
[41] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, (2014) arXiv preprint arXiv:1409.1556.
[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[43] K. Stouffs, H. Tournaye, J. Van der Elst, I. Liebaers, W. Lissens, Is there a role for the nuclear export factor 2 gene in male infertility? Fertil. Steril. 90 (2008) 1787–1791.
[44] S. Vicente-Fiel, I. Palacin, P. Santolaria, J. Yániz, A comparative study of sperm morphometric subpopulations in cattle, goat, sheep and pigs using a computer-assisted fluorescence method (CASMA-F), Anim. Reprod. Sci. 139 (2013) 182–189.
[45] World Health Organisation, WHO Laboratory Manual for the Examination of Human Semen and Sperm-Cervical Mucus Interaction, Cambridge University Press, 1999.
[46] J. Yániz, S. Vicente-Fiel, S. Capistrós, I. Palacín, P. Santolaria, Automatic evaluation of ram sperm morphometry, Theriogenology 77 (2012) 1343–1350.
[47] Y. Zhang, Animal sperm morphology analysis system based on computer vision, 2017 Eighth International Conference on Intelligent Control and Information Processing (ICICIP), IEEE, 2017, pp. 338–341.