2022 15th International Conference on Human System Interaction (HSI) | 978-1-6654-6822-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/HSI55341.2022.9869511

Sperm segmentation and abnormalities detection during the ICSI procedure using machine learning algorithms

1st Aleksandra Fraczek
Dept. of Biomedical Engineering
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Gdansk, Poland

2nd Gabriela Karwowska
Dept. of Biomedical Engineering
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Gdansk, Poland

3rd Mateusz Miler
Dept. of Biomedical Engineering
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Gdansk, Poland

4th Joanna Lis
Research and Development Center, INVICTA
Department of Medical Biology and Genetics, University of Gdansk
Gdansk, Poland

5th Anna Jezierska
Dept. of Biomedical Engineering
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Gdansk, Poland

6th Magdalena Mazur-Milecka
Dept. of Biomedical Engineering
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Gdansk, Poland
magmilec@pg.edu.pl

***Abstract*—(1) About 15-20% of couples struggle with the problem of infertility. 30 to 40% of these cases are caused by abnormalities in the structure and motility of sperm. Sometimes the only possibility for such people is to use the procedure of artificial insemination. CASA systems are used to increase the efficiency of this procedure by selecting the appropriate sperm cell. (2) This paper presents an approach to the sperm classification on the basis of its entire structure analysis, including flagella - often poorly visible and therefore ignored in the CASA systems element. The training of the Mask R-CNN architecture was performed on 2 publicly available and one specially created for this purpose sperm database. A 14-element feature vector was also proposed for the classification of 4 classes of typical head defects (amorphous, normal, tapered and pyriform) by the Support Vector Machine. (3) The sperm head (mAP 94.28%) and the whole flagellum (mAP 90.29%) were successfully detected. However, the flagella segmentation results were significantly lower (50.88%) than that the head segmentation (88.32%). Classification with SVM scored 82% accuracy. (4) Research has shown that segmentation and the use of a simple SVM classifier allow for quite good results in the classification of sperm defects. However, it is important to develop a larger whole sperm database, to improve the segmentation results.**

***Index Terms*—image segmentation; computer vision; deep learning**

I. Introduction

The introduction of the ICSI procedure has improved the treatment of infertility, mainly of male origin. This procedure involves the injection of a single immobilized spermatozoon directly into an oocyte, and therefore can be used not only for cases in which there are extremely low numbers of sperm, but also in the treatment of qualitative or functional sperm disorders. The sperm morphology is accepted as one of the best indicators of fertilization outcome [1]. Many studies have proven that there is a close relationship between the correct anatomical sperm morphology and male fertility. The World Health Organization [2] distinguishes 4 categories of sperm defects, i.e. head, neck and midpiece, tail defects and excess residual cytoplasm.

The visual assessment of the morphology of each part of the sperm is difficult and subjective. Often, the assessment carried out by various embryologists is not consistent. Therefore, it is extremely important to create an objective, universal tool that would help in the correct classification of a given defect. This kind of evaluation can be performed on computer-based systems also known as Computer-Aided Sperm Analysis (CASA) systems. Comparing to the visual assessment, CASA system is more reliable and objective [2].

In 2017, Chang et al. [3] were first to introduce an open and free dataset, SCIAN-MorphoSpermGS, for the analysis and evaluation of human sperm heads morphological classification. The dataset has five classes of the shapes of human sperm heads listed in the WHO laboratory manual [2]. They proposed a two-phase analysis system (cascade ensemble of support vector machine - CE-SVM) that classifies heads on the basis of shape-based descriptors [4].The system achieved an average true positive rate of 58% in classifying sperm heads in case where two or more out of three experts agreed.

Shaker et al. improved this approach by creating a dictionar-

This work has been partially supported by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdansk University of Technology. This work was supported in part through the European Regional Development Fund as part of the Project entitled: Academy of Innovative Applications of Digital Technologies under Grant The Operational Programme "Digital Poland 2014-2020" number POPC.03.02.00-00-0001/20-00.

ies trained on square patches extracted from images [5]. The results of an average true positive rate achieved 62% on the SCIAN dataset and 92.3% on the authors dataset (HuSHeM [6]) for cases where all three experts agreed.

Work of Riordon et al. [7] describes an approach that does not require pre-extraction of shape descriptors. It uses VGG16 network initially trained on the ImageNet database and then re-trained using transfer learning on labeled sperm head images from HuSHeM and SCIAN datasets. This approach produces an average true positive rate of 94.1% on the HuSHeM and 62% on the SCIAN dataset.

Despite the good results, automatic classification still does not achieve the precision required for clinical use [4]. Although the midpiece and tail are also important to consider for the understanding of the fertilization process, particular emphasis in most of the CASA systems is given to the form of the head. Precise segmentation seems to be a key step before the classification of all sperm parts.

An example of the detection and analysis of different parts of sperm is presented in the paper [8]. Javadi and Mirroshandel have designed a deep neural network architecture and trained it to detect morphological deformities in head, acrosome, and vacuole. They have also introduced a new dataset called the Modified Human Sperm Morphology Analysis (MHSMA), based on the HSMA dataset where each image was labeled for normal or abnormal acrosome, head, vacuole, and tail. 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.

The systems described above are available for quantifying the morphology of the sperm head, sometimes midpiece or part of the tail and also to perform classification. However, tail defects affecting motility can be more directly assessed by motion measurements. CASA systems are well suited for the kinematic analysis due to the greater accuracy of cell mobility measurements compared to the observer/embryologist [2].

In this study, an approach to segment the head and the tail of spermatozoon using available datasets is presented. Proper segmentation (including flagellum) is essential not only for the correct classification of sperm defects, but also for motility assessment.

Commonly used sperm morphology classifiers focus only on the head and on a few selected classes of structural defects. The use of a neural network for classification purposes is burdened with the risk of incorrect prediction for atypical images (not included in the training set), e.g. with a different lighting, zoom, background. Semen quality parameters are precisely defined by the WHO, so it is possible to build a classifier based on the specific feature vector obtained from the image, and not on the entire images. Here we proposed and tested a 14-element feature vector for the classification of 4 basic classes of sperm head morphology.

The paper is organized as follows: Section 1 presents the literature review and the purpose of this work. Section 2 describes the Materials and Methods used for sperm head and tail segmentation and detection as well as sperm head defects classification. The results are demonstrated in Section 3. Section 4 concludes the works.

II. Materials and Methods

A. Database

Five different databases were tested for the purpose of this work: HuSHeM [6], MHSMA [9], SCIAN [3], VISEM [10] and INV. However, as the efforts of this study were made to focus on the whole spermatozoon including tail in highest possible resolution, SCIAN and VISEM databases were finally rejected from further research. Images in SCIAN are cropped to show only the head, while the tails in the VISEM images are blurred and hardly visible.

HuSHeM contains 216 images annotated by embryologists, 131x131 pixels in size. Images are classified according to the head defects or their absence into: normal, tapered, pyriform and amorphous. These images are also cropped, but contain more of the tail than SCIAN images (see Fig. 1).

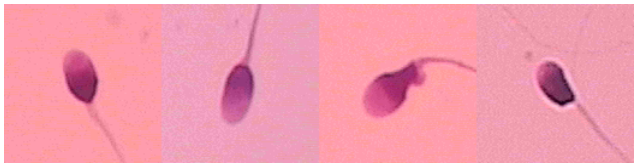


Fig. 1: Sample images from the HuSHeM database labeled as normal, tapered, pyriform and amorphous respectively

A similar range of image cropping is in the MHSMA database (Fig. 2). This database contains 1,540 expert-annotated photos in two different crop sizes: 128x128- and 64x64-pixel. Each image is labeled by experts as normal or abnormal due to the morphology of sperm acrosome, head, vacuole, and tail.

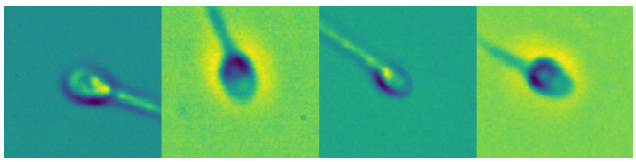
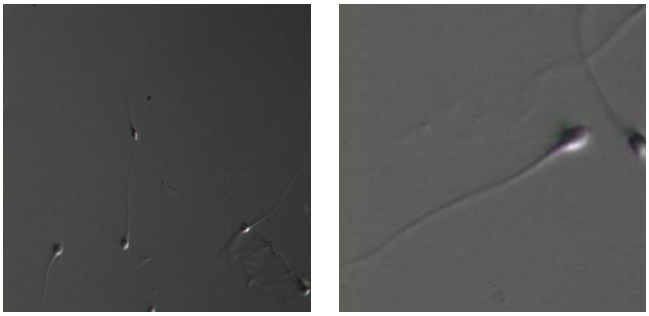


Fig. 2: Sample images from the MHSMA database

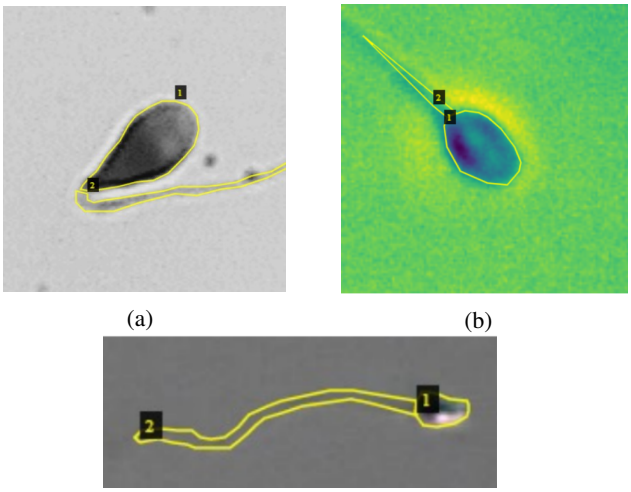
For the purposes of this study, a new database called INV, was created based on the recordings of the ICSI procedure carried out in the infertility treatment clinic. Single frames were selected from the recording of the spermatozoon choosing procedure (Fig. 3a) and then cropped so that the head and the entire sperm tail was visible (Fig. 3b).

The segmentation training database has been created from 100 HuSHEM, 100 MHSMA and 150 INV images. The sperm head and the tail area were annotated separately in each frame using VGG Image Annotator [11], [12]. Various combinations of databases were created for training the model, including combinations of: HuSHeM and INV (250 images), MHSMA and INV (250 images), HuSHeM and MHSMA (200 images) and all three databases combined together (350 images).



(a) (b)

Fig. 3: Sample images from the INV database showing a) several sperm cells and b) a cropped image of the sperm cell



(a) (b)

(c)

Fig. 4: Ground-truth images of a) HuSHEM, b) MHSMA and c) INV databases.

Examples of ground truth images are shown in Figures 4. Number 1 indicates the head and number 2 - the tail of the sperm.

B. Instance Segmentation

For the instance segmentation of sperm head and tail the pre-trained Mask R-CNN with ResNet-101-FPN backbone from the Detectron2 library [13] was used. Different parameter values were applied during transfer learning training in order to obtain the best segmentation results. There were a total of 13 combinations for: the number of iterations (1000, 3000, 4000, 5000, 6000, 8000 and 10,000) and the batch size (2 or 4). The final selected training parameters were: 6000 iterations, 2 or 4 batch size and 0.00025 learning rate. For these values, the mAP achieved the highest scores. There were also no significant decrease in the loss value for more than 6000 iterations.

C. Head classification

The classification of the head was based on 14 features, including:

- radius - the radius of the smallest surrounding circle,
- area - the area of the sperm head contour,
- ellipticity - the ratio of the semi-major (M) to the semi-minor (m) axis of the smallest surrounding ellipse,
- epsilon parameter - the maximum distance from the detected contour to its approximation [14],
- narrowness - the ratio of the width to the height of the smallest rectangle surrounding the shape,
- perimeter - a contour perimeter or a curve length,
- convexity - a boolean parameter that determines whether the curve is convex (0) or not (1),
- 7 Hu moments (h_0 - h_7) - introduced in [15].

Six parameters marked in green are illustrated in Fig. 5. All the features have been extracted using the OpenCV library [16] from binary images of segmented sperm heads.

To identify the most discriminating parameters between all, as well as between individual classes, we have used methods for calculating dependency between variables. SelectKBest and mutual information classifier [17] were implemented. Mutual information between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, higher values mean higher dependency.

In the next step, the Support Vector Machine classifier was used to check if proposed features can be used to automatically group sperm images into 4 groups in terms of common deformations. We have tested various values of hyperparameters and different SVM kernels. The classification results were presented using Accuracy as well as a weighted-averaged F_1 score defined as follows:

$$F_{1weighted} = \sum_{c} (F_{1c} \cdot SP_c)$$

where:
c - class,
F_{1c} - per-class F_1 score,
SP_c - class c support proportion.

The Principle Component Analysis was also carried out in order to achieve the best possible results in inter-class discrimination.

III. Results

A. Instance Segmentation

The results of detection and segmentation are presented in Table I and Table II respectively. Both tables show the training results (mAP) for the best-selected training (6000 iterations) for different database configurations.

The best results for the detection of the head and tail separately were achieved for the HuSHeM and MHSMA database with mAP equal to 94.28% and 93.92% respectively (see Table I). Results presented in Table II also achieved the

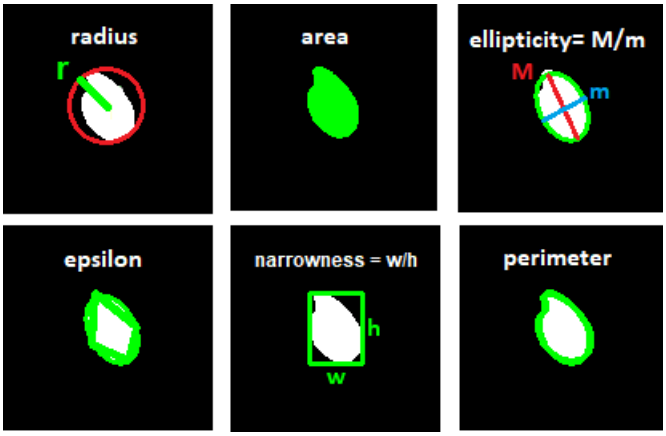


Fig. 5: Head shape parameters marked in green: radius, area, ellipticity, epsilon, narrowness and perimeter

TABLE I: Detection results for different database combinations.

Databases	Batch	mAP head	mAP tail	mAP	mAP_{50}	mAP_{75}
HuSHeM & INV	2	87.31%	85.88%	86.60%	99.41%	98.47%
	4	90.61%	89.46%	90.04%	99.79%	98.48%
HuSHeM & MHSMA	2	92.13%	89.19%	90.66%	99.98%	89.06%
	4	**94.28%**	93.92%	**94.37%**	**100.0%**	**100.0%**
MHSMA & INV	2	83.20%	84.15%	83.67%	98.93%	98.45%
	4	88.75%	**90.29%**	89.52%	99.84%	98.50%
MHSMA & INV & HuSHeM	2	82.33%	82.10%	82.22%	98.82%	98.14%
	4	90.13%	89.24%	89.68%	99.83%	97.91%

highest mAP for this database: 88.32% for head and 60.10% for the tail segmentation. It is worth noting, however, that both HuSHeM and MHSMA datasets are very similar to each other and contain images with cropped tail. This significantly improves the results of segmentation in particular due to the lack of the flagellum tip, which is often less visible. The best detection and segmentation results for bases containing images of the entire flagellum are 90.29% and 50.88% respectively. The head mAP scores for the remaining databases oscillate between 82.33% and 90.61% for detection and between 77.44% and 83.62% for segmentation. The lowest result was achieved for the tail segmentation (43.47-50.88%). The mAP of the tail detection ranged from 82.10% to 90.29%.

Fig. 6 presents testing results for HuSHeM and MHSMA images. The head and the tail are very well detected, even if the model has not been trained on one of the database. This

TABLE II: Segmentation results for different database combinations.

Databases	Batch	mAP head	mAP tail	mAP	mAP_{50}	mAP_{75}
HuSHeM & INV	2	81.46%	44.65%	63.05%	97.09%	65.44%
	4	83.07%	48.14%	65.60%	98.49%	98.48%
HuSHeM &MHSMA	2	86.87%	57.04%	71.95%	98.48%	83.52%
	4	**88.32%**	60.10%	**74.21%**	**98.87%**	**86.86%**
MHSMA &INV	2	77.89%	47.28%	62.58%	98.30%	66.50%
	4	81.76%	**50.88%**	66.13%	98.81%	72.41%
MHSMA&INV &HuSHeM	2	77.44%	43.47%	60.45%	97.14%	62.50%
	4	83.62%	49.33%	66.47%	96.22%	74.04%

is probably due to the great similarity of these two databases. In both sets, the photos are cropped, the head is under high magnification and the tail is short. However, detection errors do occur: sometimes the tail is not detected at all (Fig. 6 first column) or round and dark spots are detected as the head (Fig. 6 bottom row).

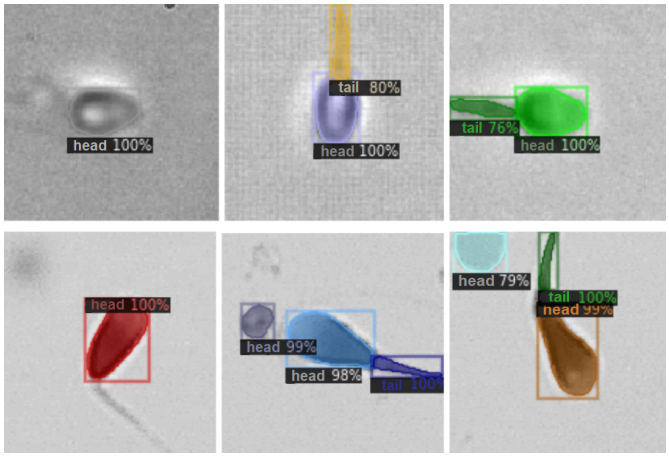


Fig. 6: Test images from HuSHeM and MHSMA dataset with segmented parts of the sperm

In the INV test images the head and the tail are properly segmented only if the algorithm has been trained on this set (Fig. 7). In the case of training the model on other databases, the prediction of the INV images segmentation achieves low results (Fig. 8). Random spots are classified as a head, while long lines or the syringe - as a sperm tail. In some cases no cell parts are detected, in others the entire sperm is labeled as a flagellum. Detection and segmentation of the head is improved by adapting the INV images to the HuSHeM/MHSMA image standard, i.e. cropping the sperm image presented in Fig. 9. This procedure, however, did not improve the tail detection.

B. Head classification

The values for the 14 proposed parameters were calculated for four classes (43 HuSHeM images in each class). Fig. 10 shows exemplary binary images for the considered classes of head defects (amorphous, tapered, normal and pyriform) for which the parameters were calculated. Minimum and maximum parameter values are presented in Table III.

Table IV summarizes the discriminant properties of the features by presenting the three most differentiable parameters together with the achieved test scores. The results were calculated between all four classes together and for each class separately versus the other classes. The score is in the range [0-1], where 1 is a high discriminant property of the parameter and 0 is none.

For all classes, the Hu moments - h_0 (0.4621), h_1 (0.4180) and ellipticity (0.3882) turned out to be the most differentiating features. The greatest differentiation in relation to the other classes was observed for the normal class followed by the tapered class (scores above 0.2), both using

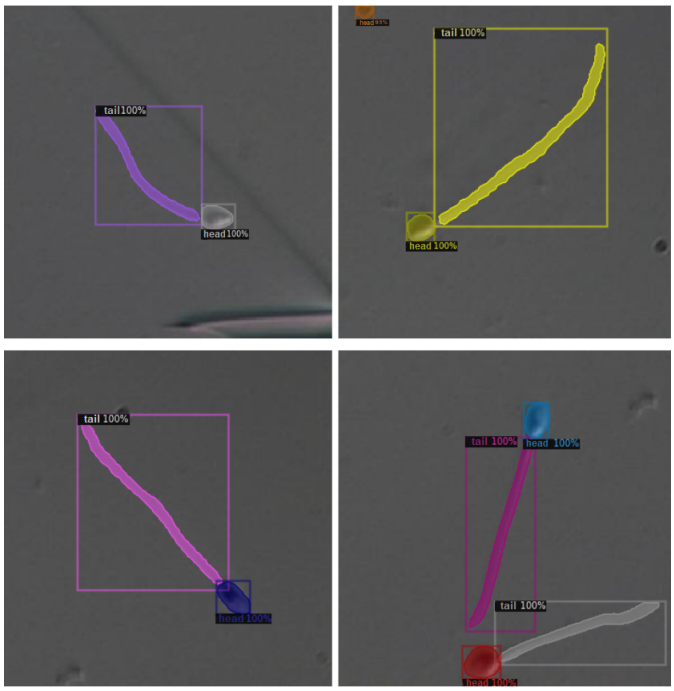


Fig. 7: Test images from INV database with segmented parts of the sperm. The model was trained on INV images

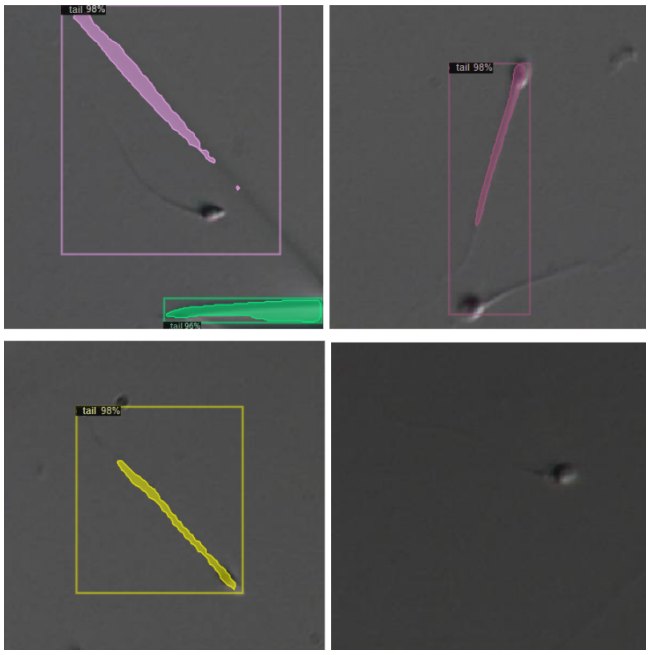


Fig. 8: Test images from INV database with segmented parts of the sperm. The model was not trained on INV images

Fig. 9: Cropped test images from INV database with segmented parts of the sperm. The model was not trained on INV images

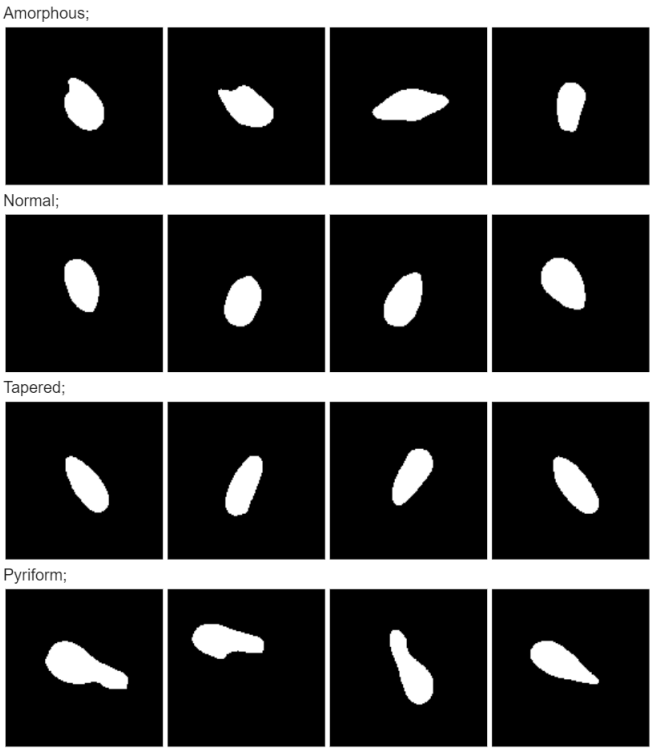


Fig. 10: Binary images from HuSHeM database labeled as (from top): amorphous, normal, tapered and pyriform.

the same parameters (h_0, h_1 and ellipticity). For the pyriform class the most differentiating was the epsilon - 0.1856, radius - 0.1785 and perimeter - 0.1774. The amorphous class was the least differentiated class. The most important feature in this class was radius with the score value equal 0.1407.

The results of the sperm head classification using the proposed parameters are presented in Table V. Four different types of SVM kernels have been tested: linear, radial basis function (rbf), sigmoid and polynomial (poly). For each kernel we have also transformed data space using PCA.

The best accuracy (0.82) and $F_{1weighted}$ (0.80) was achieved for a linear kernel without using Principle

TABLE III: Minimum and maximum parameter values used for the head abnormalities classification.

Parameter	Minimum value	Maximum value
radius	14	37
ellipticity	0.33	0.96
epsilon parameter	7.95	18.17
perimeter	79	182
area	373	1500
narrowness	0	3
convexity	0	0
h_0	0.00063	0.0011
h_1	7.21e-10	8.29e-07
h_2	7.50e-15	2.63e-10
h_3	4.03e-16	1.10e-10
h_4	-1.14e-24	1.77e-20
h_5	-3.42e-16	9.50e-14
h_6	-7.96e-22	2.47e-21

TABLE IV: The most differentiating features of sperm head abnormalities

Class of the sperm head abnormality	The three most differentiating features	Mutual information score
All classes	h_0	0.4621
	h_1	0.4180
	ellipticity	0.3882
normal vs tapered, pyriform, amorphous	h_0	0.2819
	h_1	0.2420
	ellipticity	0.2099
tapered vs normal, pyriform, amorphous	h_1	0.2577
	h_0	0.2339
	ellipticity	0.2243
pyrifrom vs normal, tapered, amorphous	epsilon	0.1856
	radius	0.1785
	perimeter	0.1774
amorphous vs normal, tapered, pyrifrom	radius	0.1407
	area	0.0689
	perimeter	0.0471

Component Analysis. In most cases, the use of PCA made the results worse or remained unchanged. Only the use of PCA improved the results for the radial basis function kernel.

IV. Conclusions

It is necessary that the systems supporting the work of embryologists achieve better results in the detection and classification of cells. For this purpose, the precise segmentation of the entire sperm, including the entire flagellum, may be helpful. Flagella segmentation requires a properly prepared training base where the entire sperm is visible. For the databases tested in the study, the algorithms detected the flagellum correctly in most cases (mAP above 80%), but they did not precisely segmented its area (mAP about 50% for the entire tail). The small tail area is also likely to have an influence on the lower mAP value. In the future, it is worth applying segmentation not to individual images, but entire video sequences.

Detailed segmentation will enable the calculation of shape parameters characteristic for each structural defect defined by the WHO. Therefore, a universal semen quality classifier will be possible to create.

TABLE V: Classification results using SVM classifier

PCA	Kernel type	Accuracy	F1 weighted
No	linear	0.82	0.80
Yes	linear	0.79	0.73
No	rbf	0.76	0.73
Yes	rbf	0.79	0.77
No	sigmoid	0.71	0.69
Yes	sigmoid	0.65	0.62
No	poly	0.76	0.66
Yes	poly	0.76	0.66

References

[1] K. Elder, B. Dale, Y. Ménézo, J. Harper, and J. Huntriss, *In-Vitro Fertilization*, 3rd ed. Cambridge University Press, 2010.

[2] *WHO laboratory manual for the examination and processing of human semen SIXTH EDITION*. World Health Organization, 06 2021.

[3] V. Chang, A. Garcia, N. Hitschfeld, and S. Härtel, "Gold-standard for computer-assisted morphological sperm analysis," *Computers in Biology and Medicine*, vol. 83, pp. 143–150, 2017.

[4] V. Chang, L. Heutte, C. Petitjean, S. Härtel, and N. Hitschfeld, "Automatic classification of human sperm head morphology," *Computers in Biology and Medicine*, vol. 84, pp. 205–216, 2017.

[5] F. Shaker, S. A. Monadjemi, J. Alirezaie, and A. R. Naghsh-Nilchi, "A dictionary learning approach for human sperm heads classification," *Computers in Biology and Medicine*, vol. 91, pp. 181–190, 2017.

[6] F. Shaker, "Human sperm head morphology dataset, hushem," 2017, https:// doi.org/10.17632/tt3yj2pf38.1.

[7] J. Riordon, C. McCallum, and D. Sinton, "Deep learning for the classification of human sperm," *Computers in Biology and Medicine*, vol. 111, p. 103342, 2019.

[8] S. Javadi and S. A. Mirroshandel, "A novel deep learning method for automatic assessment of human sperm images," *Computers in biology and medicine*, vol. 109, pp. 182–194, 2019.

[9] ——, "A novel deep learning method for automatic assessment of human sperm images," *Computers in Biology and Medicine*, vol. 109, pp. 182–194, 2019.

[10] T. B. Haugen, S. A. Hicks, J. M. Andersen, O. Witczak, H. L. Hammer, H. Borgli, P. Halvorsen, and M. A. Riegler, "Visem: A multimodal video dataset of human spermatozoa," 2019, http://doi.acm.org/10.1145/3304109.3325814.

[11] A. Dutta, A. Gupta, and A. Zissermann, "VGG image annotator (VIA)," http://www.robots.ox.ac.uk/ vgg/software/via/, 2016, version: X.Y.Z, Accessed: 21.04.2021.

[12] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: ACM, 2019. [Online]. Available: https://doi.org/10.1145/3343031.3350535

[13] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019, https://github.com/facebookresearch/detectron2.

[14] D. K. Prasad, M. K. Leung, C. Quek, and S.-Y. Cho, "A novel framework for making dominant point detection methods non-parametric," *Image and Vision Computing*, vol. 30, no. 11, pp. 843–859, 2012.

[15] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[16] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.