Automated Sperm Head Morphology Classification with Deep Convolutional Neural Networks

Marco Antônio Calijorne Soares, Daniel Henrique Mourão Falci, Marco Flávio Alves Farnezi, Hana Carolina Moreira Farnezi, Fernando Silva Parreiras (in memoriam), João Victor Boechat Gomide
Graduate Program in Information Systems and Knowledge Management
FUMEC University, Belo Horizonte, Brazil

***Abstract*—Background and Objective: The morphological analysis of sperm cells is considered a tool in human fertility prognosis. However, this process is manual, time-consuming and dependent on professional expertise. From a computational perspective, this is a challenging problem due to the high inter-category similarity between the objects of interest and the amount of data available. In this paper, we propose a Convolutional Neural Network model to automate morphology analysis of human sperm heads.**

Methods: We performed K-Fold cross-validation experiments over two publicly available datasets and assessed the performance of the proposed approach using Accuracy, Precision, Recall and F1-Score. We also compared the proposed model with well-known Convolutional architectures and previous approaches on the same task.

Results: Experimental evaluation showed that our approach achieved a macro-averaged F1-score of 0.95 while our best model attained an accuracy of 97.7%. The error analysis revealed a balanced classifier over different sperm head classes.

Conclusions: We proved that the proposed approach outperformed the previous state-of-the-art results on this task.

I. Introduction

Infertility, according to the World Health Organization (WHO) is "the inability of a sexually active, non-contraceptive couple to achieve pregnancy in one year" [1]. This condition affects 15% of couples, and 50% of cases are related to men [2]. Evidence shows that factors such as genetic predisposition, age, diet, and stress levels directly influence the number of abnormal sperm cells in the human species [3].The first procedure in evaluating male infertility is the health analysis of semen [1]. This investigation follows common criteria where a series of parameters are measured: volume, concentration, motility, quantification of leukocytes, and morphology. Evidently morphology is considered a good predictor of fertilization ability and yields preponderant information when deciding the course of treatment [4].

The morphology analysis examines the characteristics of its head, middle part, and tail as well and the excess residual cytoplasm. In this study we are particularly interested in sperm head morphology analysis, a challenging task that provides valuable diagnostic information. Sperm samples showing microcephaly, for instance, may indicate a high DNA fragmentation; spermatozoa with a round head and no acrosome may suggest globozoospermia, a genetic syndrome; spermatozoa with a thin head are found in cases of hyperthermia and related to cases of varicocele; samples with 100% of spermatozoa with macrocephaly indicate macrocephalic sperm syndrome [5].

Nevertheless, sperm head morphology classification has a broad spectrum of possibilities and is considered a non-trivial task [4]. The main issue is that the physical traits responsible for distinguishing cells from different classes are minimal (high inter-class similarity) and, therefore, hard to discern. This fact considerably increases the subjectivity of the analysis, and there is a significant variation in the responses obtained from different technicians on the same sample [6] even after simplifying classification rules. Therefore, this task is based on an error-prone process dependent on the professional's expertise. Computer Assisted Semen Analysis (CASA) systems analyze semen parameters, including sperm head morphology. The argument is that these systems may improve the consistency of the results while reducing the time necessary for completing the task [7]. The existing CASA systems fail to provide thorough sperm head morphology analysis, which implies losing significant diagnostic information. In recent years, there has been an increasing interest in deep supervised learning models in medical image processing. These techniques' application resulted in the new state-of-the-art in several medical image classification tasks [8]. However, to our knowledge, no prior investigation has experimented with deep learning techniques to classify sperm head morphology. Typically, these approaches require several thousands of labeled samples to achieve a good generalization capacity. This number, though, is far more than what is publicly available for the task. From a computational standpoint, the problem is amplified due to the high inter-class similarity between the objects of interest. Thus, in this context, a successful model must be able to capture slight shape variations (discreet patterns) even when exposed to small amounts of data.

This paper addresses this problem by presenting a supervised approach based on a deep Convolutional Neural Network (CNN) architecture for automated sperm head morphology classification. We validated it on two public datasets using standard performance evaluation metrics such as precision, recall, and F1 score. We also compared the proposed model, a known convolutional architecture, and previous approaches to the same task.

II. Related Work

In recent years, one may observe a growing interest in the application of deep supervised learning models in the field of medical image classification [9], [10]. Among the available techniques, we highlighted the ones based on Convolutional Neural Networks (CNN) - a technique that has been used in several domains with encouraging results [11]. Huang et. al. [12] employed this approach in identifying pancreas cancer in abdominal computed tomography. The same architecture was used by Baloni et. al [13] to identify hydrocephalus desease.

Even with the increasing applicability of CNNs in medical image classification problems, few research works have applied this state-of-the-art technology to assess the quality of human semen. Dewan et al. [14] presented a CNN model trained to distinguish between sperm and non-sperm objects. Its output enables the estimate of spermatozoa concentration directly from raw video through a frame-by-frame count. It also allows evaluating motility. In this case, whenever the binary CNN identifies a sperm cell, a second stage starts tracking its trajectory across multiple frames. The results were presented for sperm concentration, motility percentage, and tracking performance, reaching a mean miss percentage of 7.44% on 95% confidence intervals. Golomingi et. al [15] used a VGG-19 and one variation of it with 1942 images to detect sperm in glass slides in sexual assault cases reaching high levels of accuracy.

Previous studies also investigated the automated human sperm head morphology classification problem [16], [17]. The study carried out by Abbiramy and Tamilarasi [18] compared the accuracy of three distinct binary neural network models to classify sperm heads as normal or abnormal using the holdout method. The experimental results attained an accuracy of 75% using a traditional two-layer feed-forward network trained with gray-level co-occurrence matrix features. However, the authors exclusively utilized images contained in the WHO laboratory manual [1] and, therefore, there is no evidence that the proposed approach performs equally well under real usage circumstances, with images captured from microscopes with standard resolution. A recent study has proposed an approach based in pseudo-masks and unsupervised spatial prediction tasks [19] to classify the human sperm head morphology.

On the other hand, Chang et al. [16] expanded the sperm head morphology classification problem by presenting a new approach that can classify cells into five distinct groups (amorphous, normal, pyriform, small, and tapered), which is essential for the diagnostic perspective. Their two-stage system uses an ensemble of classifiers based on Support Vector Machines (cascaded SVM classifiers). The first stage aims at identifying amorphous sperm cells (filtering stage). The second (the verification stage) categorizes any positive sample from the filtering stage into the remaining classes. The goal of this schema is to minimize the average confusion matrix at the second stage. This method achieved 73% accuracy, the same accuracy that a human expert could achieve.

The same problem proposed by Chang et al. [16] was also faced by Shaker et al. [17]. The authors presented a technique based on Dictionary Learning (DL), namely ADPL, which was validated on top of datasets such as the Human Sperm Head Morphology (HuSHeM) [20] and SCIAN-MorphoSpermGS [6]. In the case of the HuSHeM dataset, four categories were considered: normal, pyriform, tapered, and amorphous. A cascade SVM was implemented as described by Chang et. al. [16] for comparison purposes and it was trained on the HuSHeM dataset using a K-fold cross-validation setup. The results showed the superiority of the DL method, reaching a mean accuracy of 92.2%, mean precision of 93.5%, mean recall of 92.3%, and mean $F1$-score of 92.9%. Another advantage of the DL method is the number of parameters compared to the ensemble of SVMs.

III. Background

CNNs are variants of multi-layer perceptron networks (MLPs). Typically arranged in a deep architecture (composed of multiple stacked layers), CNNs are focused on visual pattern recognition directly from the raw pixels of an image. This biologically inspired network architecture is based on the idea that the visual cortex recognizes objects starting from primitive shapes, iteratively elaborating more abstract concepts.

A CNN scans an image by analyzing patches of an input image (receptive field) through convolution operations of k trainable convolution kernels (local feature detectors), where each is responsible for capturing a distinct pattern on an image.

Formally, at each layer, the input image is convolved with a set of kernels, each of them generating a feature map. These feature maps are subjected to an element-wise non-linear transform, using rectifier linear unit function. The process is repeated for each convolutional layer in the architecture. A sequence of convolution layers is followed by a pooling operation where the pixel values inside a feature map are aggregated using invariant functions such as min, average or max. A sequence of convolution layers is followed by a pooling operation where the pixel values inside a feature map are aggregated using invariant functions such as min, average, or max. The pooling operation performs a subsampling that divides the feature map and, thus, reduces the overall number of parameters of the architecture which will set the boundaries for the architecture portion responsible for performing feature extraction over input images. The feature maps of the last layer are flattened to a unidimensional matrix and feed a set of fully connected layers optimized to minimize the following cost function with respect to the input images.

IV. Experiments

A. Proposed Methods

Recently, driven by the ImageNet challenge [21], highly deep and complex CNN architectures have been proposed including AlexNet, VGGNet, ZFNet, GoogLeNet, and ResNet. These models were successfully employed in applications that rely on object segmentation and classification. In this work, we compared two architectures, a tailored version of

VGGNet which we refer to as Sperm Head Morphology CNN (SHMCNN) and a bilinear CNN model (BCNN).

The VGGNet, presented by VGG group [22], has become one of the standard architectures for deep learning in computer vision reaching the top-5 accuracy on ImageNet, with 92.3%. VGGNet replaced larger convolution kernels (11×11 for instance) from Krizhevsky and Hinton (AlexNet) [23] approaches by smaller ones (3×3) arranged in sequence. VGG network preserves the ability to detect larger features while improving the ability to capture fine-level properties over the input. VGG-16 and VGG-19 are the most prominent versions of this architecture.

An off-the-shelf VGGNet architecture does not converge on small datasets with low resolution samples, such as those publicly available for the task of human sperm head morphology classification. It needs a larger number of labeled samples to prevent overfitting while requiring images with higher resolution, due to the high number of pooling operations in its structure. Therefore, we proposed an architecture that contains 4 convolutional blocks with stride one. The last convolutional layers outputs 256 feature kernels of 8x8 images. This setup was achieved after empirical evaluation and was the one that yielded the best results for this model.

The Bilinear Convolutional Neural Network (BCNN) model [24] is an architecture tailored for fine-grained object recognition. Regular CNN architectures are good at distinguishing objects with high inter-class variation such as differentiating a car from an airplane. However, when the goal is to classify highly similar objects such as birds by its species or airplane models, the accuracy decays. The BCNN architecture is composed of two parallel feature extractors for the same input image. The convolutional layer outputs of feature extractors are then combined with outer product to obtain a quadratic number of feature maps that feed a classification layer. The idea is that each output of an extractor is combined with output of the another extractor. This architecture demonstrated to be effective for fine-grained recognition tasks [25] and was relevant for the sperm head morphology classification task due the high inter-class similarity.

In our architecture we mixed two different feature detection parts as shown in Figure 1. The left one is identical to that of SHMCNN network (presented in the previous subsection). The right side follows the VGGNet architecture. The same number of layers is employed on both sides and their output dimensions are identical at the fourth convolutional block, a requirement for the outer product operation.

B. Experimental setup

The Human Sperm Head Morphology dataset (HuSHeM) [20] is a publicly available dataset comprised of 216 RGB images of sperm heads. The size of each of these images is 131×131 pixels, where the target spermatozoon, centered, may be eventually occluded by other objects. Additionally, there is no standard sperm head orientation, therefore, the cell may be geared in any direction. In this dataset, sperm heads are classified into four distinct classes,

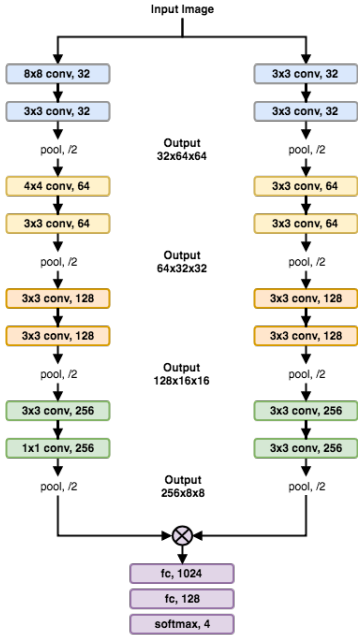

Fig. 1. The BCNN architecture tested on this study.

distributed as follows: Normal (54 images), Tapered (53 images), Pyriform (57 images), and Amorphous (52 images) (Figure 2 for samples of each class). A group of three specialists labeled all the collected images into the four proposed categories. Only those consensually labeled samples were kept in the dataset, reducing the outliers and creating a reliable gold standard for the task.

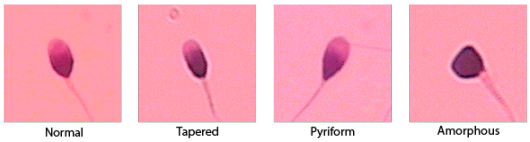

Fig. 2. A sample of each sperm head class covered at HuSHeM dataset.

To create this dataset, semen samples were collected from fifteen patients between 25 and 38 years old, following instructions and recommendations of the *WHO* manual [1]. The quick diff method was used to fix and stain the semen smears. The SCIAN-MorphoSpermGS dataset, collected by [6], consists of 1132 gray-scale images of sperm heads with no occlusion. Each image has a size of 35×35 pixels, that, differently from HuSHeM dataset, are classified into five categories: Normal (100 images), Tapered (228 images), Pyriform (76 images), Small (72 images), Amorphous (656 images). See Figure 3 for a sample image for each class.

C. Partitioning and pre-processing

In order to provide a robust approximation of the model's accuracy on unseen data, all the experiments reported in this paper were subject to a *k-fold* cross-validation method, with k=5. Hence, for each model evaluated, we randomly

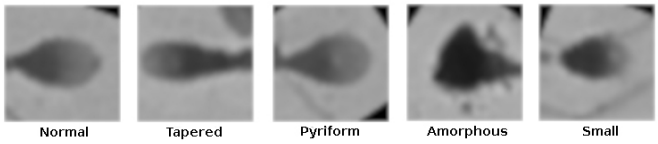

Fig. 3. Sperm head categories contained into SCIAN-MorphoSpermGS dataset.

divided the original dataset into five equal-sized folds and performed five separate training sessions, each of them using four folds for training (80% of data) and one fold for testing (the remaining 20%). At each of these training sessions, we rotated the fold selection, so that all folds were used at the test set exactly once. Additionally, we employed a minimal pre-processing stage composed of two steps. The first step was about converting original RGB images into gray-scale [1]. The second and most significant step was the data augmentation procedure.

Training a CNN model for an image classification task requires adjusting millions of parameters and, consequently, many labeled samples. As described in the Section IV, both available datasets contain only a few dozens of images representing each class. Directly using such small-sized datasets as inputs often lead to overfitting, which reduces the overall accuracy of a model. During data augmentation, one applies a series of transformations over the original images of the training set, aiming at artificially generating a more representative number of samples. Frequently used transformations involve zooming, rotating, translating, rescaling, brightness manipulation, and randomly adding noise and occlusions to the original samples. After carefully examining the results produced by different data augmentation pipelines, we chose to generate batches of new images for each training sample. We randomly zoomed on a range of 1% to 15%, distorting the image classifier. We applied standard metrics such as precision, recall, and f-measure. The measurement was made for each morphology class, yielding a local precision and recall. To compute the system's overall performance, we calculated the median results of the local precision and recall. These averaged results serve as input without de-characterizing its morphology. We also randomly flipped and rotated each image from 0 to 360 degrees.

Regarding the SCIAN-MorphoSpermGS dataset, indubitably, the number of sperm-cells forms an unbalanced distribution. For this reason we performed a down-sampling procedure where we randomly removed 556 samples from the *Amorphous* class, 138 images from *Tapered* and 10 images from *Normal*. Thus, we produced a subset with the following distribution of images amount: *Amorphous* (90 images), *Small* (72 images), *Pyriform* (76 images), *Normal* (90 images) and *Tapered* (90 images). We applied the same pipeline in this down-sampled subset (partitioning, pre-processing and data augmentation).

[1]This step was performed only for the HuSHeM dataset. SCIAN-MorphoSperm-GS samples are already at the gray-scale

D. General procedures and Evaluation

All CNNs were optimized using *Adam*, an efficient adaptive algorithm for gradient-based optimization of stochastic objective functions that is typically suited for high-dimensional parameter models [26]. The algorithm was initialized with the following settings: $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$.

We used the *Dropout* technique [27] which affects the network training procedure by randomly setting activations to zero during backpropagation at training stage. Each training session, despite the model, lasted for 200 epochs with an early stopping policy, which ensured that, if a loss reduction did occur in the last 20 training epochs, the training session would be ended[2].

To evaluate the performance of the sperm head morphology classifier, we applied standard metrics such as, precision, recall, and f-measure. The measurement was made for each morphology class, yielding a local precision and recall. To compute the overall performance of the system, we calculated the median results of the local precision and recall. These averaged results serve as input to compute the $F1$-score of the model (macro-averaged F1-score), which indicates the harmonic mean between the averaged precision and recall.

V. Results

We compared the results obtained by our proposed approaches: BCNN (Table I) and SHMCNN (Table II). The comparison revealed close performance for both methods, with a difference of 0.01 Macro-Averaged F1-score points. While the BCNN network achieved an impressive $F1$-score of 0.98 for capturing Normal sperm cells, the SHMCNN exhibited a degraded score (0.96), but on the other hand, it performed consistently across different classes and showed to be a balanced classifier. We expected that the BCNN model would be able to find better feature descriptors for sperm images, due to its natural ability to deal with inter-class similarity of objects [24]. The results suggest that the inter-class similarity, although present in the task, may be manifested through simpler shapes that SHMCNN can capture. A positive point for the SHM-CNN model is that it requires only a fraction of the parameters of a BCNN (a quadratic model) – 10M against more than 60M – which makes it faster at training and prediction, while it exhibits a small memory footprint.

TABLE I
DETAILED RESULTS FOR BCNN MODEL ON HUSHEM DATASET.

Morphology	*Precision*	*Recall*	*F1-Score*	*Support*
Amorphous	0.98	0.88	0.93	52
Normal	**1.00**	**0.96**	**0.98**	54
Pyrifom	0.85	0.98	0.91	57
Tapered	0.96	0.92	0.94	53
Macro-Averaged	0.95	0.94	0.94	216

Our best single training session of the SHMCNN model on the HuSHeM dataset achieved an accuracy of 97.7%.

[2]https://github.com/dfalci/spermheadmorphology

TABLE II
DETAILED RESULTS FOR SHMCNN MODEL ON HUSHEM DATASET.

Morphology	*Precision*	*Recall*	*F1-Score*	*Support*
Amorphous	**0.98**	0.92	0.95	52
Normal	**0.98**	0.94	**0.96**	54
Pyriform	0.92	**0.98**	0.95	57
Tapered	0.94	0.96	0.95	53
Macro-Avg	0.96	0.95	**0.95**	216

The confusion matrix is shown in Figure 4. It is the result of summing up all the confusion matrices obtained in each individual experiment of the cross-validation. Notice that the major difficulties faced by our model were at labeling Pyriform cells, with lowest Precision (0.92), with 5 samples being incorrectly classified. . The lowest recall, in turn, was at recognizing Amorphous cells (0.92), with 4 errors.

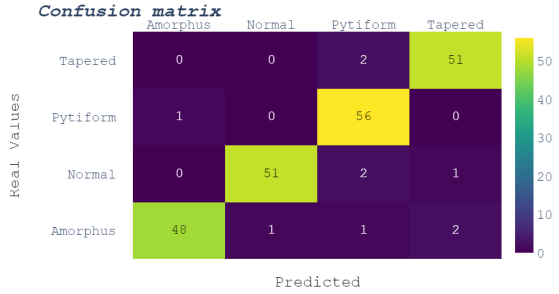

Fig. 4. Confusion matrix for the 5-fold cross-validation experiment of SHMCNN model.

Our next step was to compare the performance of SHMCNN model with that of our baseline - ADLP [17] - at HuSHeM dataset. Table III shows the results. We highlighted that the SHMCNN performance is superior in all three metrics, providing a considerable error reduction rate of 37.5% on macro-averaged F1-score. This result suggests that our approach is suitable for classifying sperm head morphology under HuSHeM dataset.

TABLE III
COMPARISON OF THE PERFORMANCE OF SHMCNN AND ADPL METHODS ON HUSHEM DATASET.

Classifier	*Precision*	*Recall*	*F1-Score*
SHMCNN	**0.96**	**0.95**	**0.95**
APDL [17]	0.93	0.92	0.92

We also applied a variant of the SHMCNN architecture to the SICAN-MorphoSpermGS dataset in the same way we did for the HuSHeM dataset. The pre-processing stage was mostly the same, despite its samples being horizontally oriented. The only change regarding the gray-scale transformation is not necessary for this dataset. Again, we used k-fold cross-validation with $k = 5$ on a down-sampled subset, following the procedures described in Section IV-C. Thus, we maintained compatibility with the setup used of previous studies. We compared our results with those of Chang et al. [16] and Shaker et al. [17]. Considering the smaller size of their images (35×35), we opted to remove the last convolutional layer of our SHMCNN architecture. The results are summarized in Table IV. Our model has achieved a higher recall rate for morphology classes. The exception lies in the amorphous category, which achieved a recall of only 13%, which is less than half of the ADPL method. We suspect this issue is due to the automated down-sampling procedure, which may have produced an under-represented sub-sample of this category. Nevertheless, the proposed method is effective for classifying sperm head morphology for the remaining classes.

VI. DISCUSSION

Despite the relevant results reached by this study, a discussion could be addressed when we analyze the applied method to different species of sperm. In fact, the basis of this study relies on the dataset and CNN techniques, hence, with a rich, correct and well-classified corpus of sperms head images. Moreover, this method will be able to accomplish similar performance. However, what could we expect if the model built in this paper was applied to classify the sperm head morphology of other species? Some studies were conducted aiming to compare a large set of attributes of different species sperm, including its morphology [28].

The attributes of 16 species of patellid limpet (marine gastropod) spermatozoa and spermatogenesis and the shape of the head have the same variances when the scientists collated them or compared to human sperm head shape [29].

An analysis was conducted among species of artiodactyls, perissodactylans, and cetaceans. As a result, the researchers found that the sperm heads of all analyzed species were dorsoventrally flattened and generally had an ovoid. In some cases, there was a variation in the relative posterior thickness of the sperm head. The study also reports variations in sperm head shapes where some species had more elongated heads while others had more rounded heads [28]. This variation could be an issue to a model trained only with human image samples. In other words, with this study we reached significant results using a dataset based on human sperm images and any kind of effort to reach these findings on another species should be followed by a previous task to create a corpus with the images of those species.

VII. CONCLUSION

The current study aimed to determine the accuracy of an automated sperm head morphology classifier based on CNN architecture. We proposed and evaluated two CNN models: SHMCNN and BCNN, tailored for this specific task. Both of them are based on state-of-the-art image recognition architectures, one used in general purpose environments and the other typically suited for fine-grained tasks. To assess their performances, we conducted a series of 5-fold cross-validation experiments using metrics such as precision, recall and f-score.

The results confirmed that SHMCNN method outperformed the previous approaches by reaching a macro averaged F1-score of 0.95 at the HuSHeM dataset. The error-analysis

TABLE IV
COMPARISON OF THE RECALL OBTAINED FROM SHMCNN AND THE ONE OBTAINED FROM ADPL METHOD FOR SCIAN-MORPHOSPERMGS DATASET.

	Amorphous	*Normal*	*Pyriform*	*Small*	*Tapered*	*Average*
SHMCNN	0.13	**0.90**	**0.71**	**0.85**	**0.91**	**0.70**
ADPL [17]	**0.35**	0.71	**0.71**	0.68	0.67	0.62

showed that we achieved a balanced classifier across different classes, which indicates that the proposed approach is effective in recognizing all four sperm head classes. The research has also demonstrated that the accuracies of the SHMCNN and BCNN models were close. However, BCNN network was heavier from a computational perspective and it required more than six times the number of parameters. To validate our SHMCNN further, the architecture was modified and applied at SICAN-MorphoSpermGS dataset. Again, our results confirmed the superiority of the proposed method over ADPL method.

The scope of this study was limited by dataset characteristics. Both SHMCNN and BCNN models with the implicit margin of error, brought segmented and cropped sperm head images. As opposed to what occurs in a real-world scenario, our evaluation was not affected by this natural error-propagation. Future studies may address this matter by focusing on end-to-end method for classifying human sperm head morphology directly from raw video stream.

ACKNOWLEDGMENT

The authors are grateful to CNPq and FAPEMIG for the financial support to this work.

REFERENCES

[1] W. H. Organization *et al.*, "Who laboratory manual for the examination and processing of human semen," Tech. Rep., 2010.

[2] W. Cui, "Mother of nothing: the agony of infertility," *Bulletin of the World Health Organization*, vol. 88, no. 12, pp. 881–882, Dec. 2010.

[3] M. D. Braekeleer, M. H. Nguyen, F. Morel, and A. Perrin, "Genetic aspects of monomorphic teratozoospermia: a review," *Journal of Assisted Reproduction and Genetics*, vol. 32, no. 4, pp. 615–623, Feb. 2015.

[4] R. Menkveld, "Sperm morphology assessment using strict (tygerberg) criteria," in *Spermatogenesis*. Springer, 2013, pp. 39–50.

[5] N. Gatimel, J. Moreau, J. Parinaud, and R. D. Léandri, "Sperm morphology: assessment, pathophysiology, clinical relevance, and state of the art in 2017," *Andrology*, vol. 5, no. 5, pp. 845–862, Jul. 2017.

[6] 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.

[7] L. Zhang, R. Y. Diao, Y. G. Duan, T. H. Yi, and Z. M. Cai, "In vitro antioxidant effect of curcumin on human sperm quality in leucocytospermia," *Andrologia*, vol. 49, no. 10, p. e12760, Jan. 2017.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[9] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, pp. 92–110, 2021.

[10] X. Wang, S. Yin, M. Shafiq, A. A. Laghari, S. Karim, O. Cheikhrouhou, W. Alhakami, and H. Hamam, "A new v-net convolutional neural network based on four-dimensional hyperchaotic system for medical image encryption," *Security and Communication Networks*, vol. 2022, 2022.

[11] M. Torres and F. Cantú, "Learning to see: Convolutional neural networks for the analysis of social science data," *Political Analysis*, vol. 30, no. 1, pp. 113–131, 2022.

[12] M.-L. Huang and Y.-Z. Wu, "Semantic segmentation of pancreatic medical images by using convolutional neural network," *Biomedical Signal Processing and Control*, vol. 73, p. 103458, 2022.

[13] D. Baloni and S. K. Verma, "Detection of hydrocephalus using deep convolutional neural network in medical science," *Multimedia Tools and Applications*, pp. 1–23, 2022.

[14] K. Dewan, T. R. Dastidar, and M. Ahmad, "Estimation of sperm concentration and total motility from microscopic videos of human semen samples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2299–2306.

[15] R. Golomingi, C. Haas, A. Dobay, S. Kottner, and L. Ebert, "Sperm hunting on optical microscope slides for forensic analysis with deep convolutional networks–a feasibility study," *Forensic Science International: Genetics*, vol. 56, p. 102602, 2022.

[16] 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.

[17] 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.

[18] V. Abbiramy and A. Tamilarasi, "A comparative study on human spermatozoa images classification with artificial neural network based on fos, glcm and morphological features," in *Advances in Digital Image Processing and Information Technology*. Springer, 2011, pp. 220–228.

[19] Y. Zhang, J. Zhang, X. Zha, Y. Zhou, Y. Cao, and D. Chen, "Improving human sperm head morphology classification with unsupervised anatomical feature distillation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[20] F. Shaker, "Human sperm head morphology dataset," 2017, available at https://data.mendeley.com/datasets/tt3yj2pf38/1.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[25] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–6.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[28] A. Downing Meisner, A. V. Klaus, and M. A. O'Leary, "Sperm head morphology in 36 species of artiodactylans, perissodactylans, and cetaceans (mammalia)," *Journal of morphology*, vol. 263, no. 2, pp. 179–202, 2005.

[29] A. Hodgson and R. Bernard, "A comparison of the structure of the spermatozoa and spermatogenesis of 16 species of patellid limpet (mollusca: Gastropoda: Archaeogastropoda)," *Journal of Morphology*, vol. 195, no. 2, pp. 205–223, 1988.