Computers in Biology and Medicine 136 (2021) 104687

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Impact of transfer learning for human sperm segmentation using deep learning

Ruth Marín, Violeta Chang *

Departamento de Ingeniería Informática, Universidad de Santiago de Chile (USACH), Chile

ARTICLE INFO

Keywords:
Infertility
Segmentation
U-net
Transfer learning

ABSTRACT

Background and objective: Infertility affects approximately one in ten couples, and almost half of the infertility cases are due to the malefactor. To diagnose infertility and determine future treatment, a semen analysis is performed. Evaluation of sperm morphology is one of several steps in semen analysis, in which the shape and size of sperm parts are examined. The laboratories dedicated to this use traditional methods susceptible to errors. An alternative to replace the poor visual ability to assess sperm size and shape is to analyze sperm morphology with a computer's help. However, since the automatic sperm classification rates do not show an acceptable precision rate for use in the clinical setting, it is considered an exciting approach to focus efforts on improving the precision in sperm segmentation to extract the contour sperm before classification. This work aims to assess the utility of two image segmentation deep learning models for segmenting human sperm heads, acrosome, and nucleus.

Methods: In this work, we evaluate the use of two well-known deep learning architectures (U-Net and Mask-RCNN) to segment parts of human sperm cells using data augmentation, cross-validation, hyperparameter tuning, and transfer learning. The experimental results are carried out using SCIAN-SpermSegGS, a public dataset with more than two hundred manually segmented sperm cells and widely used to validate segmentation methods of human sperm parts.

Results: Experimental evaluation shows that U-net with transfer learning achieves up to 95% overlapping against hand-segmented masks for sperm head (0.96), acrosome (0.94), and nucleus (0.95), using Dice coefficient as the evaluation metric. These results outperform state-of-the-art sperm parts segmentation methods.

Conclusions: The impact of transfer learning is substantial, significantly improving the results of state-of-the-art methods with a higher Dice coefficient, less dispersion, and fewer cases where the model failed to segment sperm parts. These results represent a promising advance in the ultimate goal of performing computer-assisted morphological sperm analysis.

1. Introduction

Human infertility is the inability of human beings to reproduce due to problems associated with various factors. Recent studies indicate that there are 15% of infertile men worldwide, and 30% of them is due to problems related to semen quality [1]. Male infertility is the inability to reproduce but is responsible for several associated problems, as has been investigated for several decades [2], Male infertility is associated with psychological disorders, namely stress, depression, low self-esteem, among others [3]. A semen analysis, according to standard criteria, is the first step in the evaluation of the malefactor and sets the basis for all posterior steps for the medical treatment of the couple [4]. A typical spermiogram considers concentration, motility, vitality, and the fragmentation of the spermatic DNA. Traditional methods for evaluating a patient's semen quality consist mainly of taking samples for subsequent biological and visual analysis of the sperm cells.

Accurate semen sample analysis is critical to fertility treatment decisions. The morphology of the sperm cell is considered an important tool to clarify the possible infertility of a man [5]. Therefore, morphology becomes the most significant technical challenge for andrology laboratories since there is much variability between technicians who perform it [6]. Laboratories dedicated to the morphological analysis of human sperm often use traditional methods susceptible to errors. Usually, this analysis is carried out manually, making the process difficult to replicate or teach. C. Brazil states *"Standardization of semen analysis is very difficult for many reasons, including the use of subjective*

* Corresponding author.
E-mail addresses: ruth.marin@usach.cl (R. Marín), violeta.chang@usach.cl, vnchangc@gmail.com (V. Chang).

https://doi.org/10.1016/j.compbiomed.2021.104687
Received 19 April 2021; Received in revised form 18 July 2021; Accepted 23 July 2021
Available online 29 July 2021

techniques without benchmarks, poor training of technicians, and resistance to changing techniques" [6]. Thus, visual analysis of sperm morphology still presents a challenge in reproducibility and objectivity. Having technological support that automatically and precisely makes this type of measurement and offers objective criteria would be very useful.

Computational advances make it possible to generate new solutions to the adversities mentioned above, especially in matters related to artificial intelligence and imaging applications, which currently work as an assistant to the specialist, both in the scientific area (investigating causes, analyzing results) and in the medical area (when supporting diagnoses and treatments). In this sense, one of the significant advances in deep learning, which, together with computer vision, allows finding characteristics or patterns that can provide clues or conclusions about an image.

However, given that automatic sperm classification rates according to various morphological defects do not show acceptable precision for clinical use [7,8], An alternative would be to improve the quality of segmentation before classifying. Thus, it would be essential to concentrate efforts on developing methods to enhance the segmentation of sperm parts as a crucial stage before classification.

New deep learning architectures for image segmentation have proven effective and accurate in microscopic image segmentation problems [9,10]. The great potential of these techniques for the analysis of biomedical images lies in their speed and efficiency. Its use can be applied to tasks as diverse as detecting, and segmentation of medical images [11,12], among others. However, not all these advances have been taken to advantages, such as in sperm parts segmentation. There is no comparison between traditional methods and the latest deep learning methods in the specialized area.

This research seeks to contribute to the medical area since it aims to improve human sperm parts segmentation. We assess the use of two deep learning networks to segment sperm cells to improve the precision in segmentation and thus take the first step for automation in the classification of sperm morphology. This can reduce resources in the diagnosis and subsequent treatment of patients with infertility or other associated pathology.

This paper is organized as follows. First, in Section 2 we review the research work in the area. Next, used methods are described in detail in Section 3. Then, in Section 4 we describe the dataset, the experimental protocol, and the experimental results. Finally, the summary, conclusion, and future works are drawn in Section 5.

2. Related work

Many research works have been published in the last years about various methods tackling automatic human sperm segmentation, mainly based on thresholding, clustering, combining multiple color spaces, active contour methods, and convolutional neural networks (CNN).

Chang et al. [13] introduced a framework for detection and segmentation of human sperm head, acrosome, and nucleus, combining three different color spaces and clustering methods. The detection of sperm heads is achieved using K-means, while the final segmentation is reached using mathematical morphology and histogram statistical analysis techniques. The experimental results showed 98% in the sperm head detection and segmentation of head, acrosome, and nucleus achieved up to 88%, 83% and 82% of overlapping against hand-segmented gold-standard, respectively. The main contribution iteChang2014 is the introduction of SCIAN-SpermSegGS, a gold-standard and public dataset used to compare and evaluate methods for detecting and segmenting sperm cell parts.

Shaker et al. [14] presented a framework for detection and segmentation of human sperm head, acrosome, and nucleus, combining different color spaces (as the previously reviewed piece of work) and active contours. Also, the authors proposed a tail point detection method to refine the head segmentation by locating and removing the midpiece. The method was evaluated using SCIAN-SpermSegGS, achieving segmentation accuracy of 92% for sperm head, 84% for acrosome, and 87% for nucleus.

Movahed et al. [15] presented a framework for segmenting sperm parts based on deep and classical learning methods. The main idea is to segment the external and internal parts of the sperms through the concatenation of learning approaches. In this sense, the external parts of the sperms are segmented using two CNN models, which produce the probability maps of the head and the axial filament regions. Inside the head region, a k-means clustering approach is applied to segment the acrosome and nucleus. Inside the axial filament, a Support Vector Machine (SVM) classifier is used to segment the tail and mid-piece. The proposed method was evaluated using SCIAN-SpermSegGS and achieved 90%, 77%, 78% as segmentation accuracy for the head, acrosome, and nucleus, respectively.

A summary of segmentation accuracy rates achieved using the public dataset SCIAN-SpermSegGS is presented in Table 1. In all cases, the segmentation accuracy was calculated using the Dice coefficient).

3. Material and methods

3.1. Data augmentation

If properly training with a large dataset of annotated samples, deep learning techniques have greatly improved segmentation accuracy. However, collecting such a massive dataset of annotated biomedical image cases is typically challenging because performing annotating new images is tedious and expensive [16]. Therefore, the most commonly adopted method to increase the size of the training dataset is data augmentation which generates various slightly different versions of images from each image in the original training set using the random application of a group of different transformations (such as flipping, rotation, and scaling, among others).

The data augmentation strategy can be implemented online or offline. The augmented images are pre-generated for offline augmentation and combined with the original images into a more extensive training set. For online augmentation, the augmentation transformations are implemented as a part of the training process. Even though data augmentation introduces variations to the original data, reducing the risk of overfitting to a small training set and improving the generalization of the data [17,18], It is not equivalent to having a set of independent training samples of comparable size [19].

In this research work, to increase the dataset and preserve the morphological features of the sperm cells, we applied different geometric transformations while keeping the shape and size of sperm cells and obtaining a homologous image with different orientations. In this sense, 54 combinations of geometric transformations were applied to the original dataset. First, we rotated the original image in a range from 0° to 180° every 5° counterclockwise. Second, the flipping technique was incorporated, reflecting the vertical or horizontal axis to each previous step's rotated images. Then, the reflected images were rotated from 0° to 45° in 5° intervals. A combination of horizontal and vertical flipping was also applied. Finally, the original dataset was offline augmented in 2500%. The details of the original dataset can be found in Section 4.1.

Table 1
Segmentation accuracy achieved on SCIAN-SpermSegGS and calculated as Dice Coefficient.

Publication	Head segmentation	Acrosome segmentation	Nucleus segmentation
Chang et al. [13]	88%	83%	82%
Shaker et al. [14]	92%	84%	87%
Movahed et al. [15]	90%	77%	78%

3.2. Transfer learning

Transfer learning is a common approach that takes advantage of a model's capability to recognize and use the knowledge learned in a previous source domain for a novel task [17]. Therefore, transfer learning is done by fine-tuning the model pre-trained on images from a different source domain. It has been demonstrated that transfer learning has better performance when the source and target model tasks are more similar [19]. However, transfer learning of very different source and target tasks has been proven to outperform randomly initialized weights for the target task [20].

In general, transfer learning can be done in two different ways. Still, the underlying idea is to initialize the weights using a pre-trained model (source model) instead of random initialization. Then, one can update the weights during the training of the target model (full adaption) or freeze the weights for the first few layers and update the last layers during the target model's training (partial adaption). The level of transfer learning will depend on the size of the target training dataset. Therefore, if the target training dataset is small and the number of parameters is large (deeper models), full adaption may result in overfitting, and partial adaption is a better choice. On the other hand, if the size of the target training dataset is relatively bigger, full adaption is a better choice because overfitting would not appear [16].

For the experimental evaluation presented in section 4, a public dataset was used to pre-training the original models that we adapted for this work. We used the Data Science Bowl 2018 [21], containing 670 images of segmented cell nuclei provided by a non-profit biomedical and genomics research institute. Because our source and target models had very similar tasks, and the target dataset is small (19 images), we decided to train some layers and leave the others frozen, as in partial adaption.

3.3. Cross-validation

One of the primary sources of variability in learning methods is the difference between the observed samples of the dataset and the actual distribution of the dataset [22]. The experimental procedure often consists of splitting the dataset into three independent subsets: training, validation, and testing to infer a segmentation with a learning model. The proportions of the different subsets are chosen concerning the sample size and can significantly affect the expected degree of generalization [22]. It is suitable to use a cross-validation strategy to avoid bias in the former data splitting and selection. A cross-validation method divides the dataset into several folds and then assigns those folds to the training, validation, and test sets. At the end of the model's training and evaluation processes, the folds are reassigned for new estimates [23]. The most popular methods are leave-one (LOO) and k-fold-out (KFO) [23].

As we show in Section 4, it was necessary to take the maximum advantage of the whole dataset that we count on to obtain robust experimental results. In this sense, we applied the LOO strategy for a 5-fold cross-validation process. We used a 70%–30% random split from the original dataset for training and evaluation for each fold. The data augmentation strategy was applied only to the training data, randomly divided in 80%–20%, leaving 20% for validation purposes. The validation set was used to partially evaluate the model while adjusting some models' parameters after using the training data. In contrast, the evaluation set was only used once the model was fully trained.

3.4. U-net

Ronneberger et al. proposed one of the most well-known architectures for biomedical image segmentation called U-net [9]. This model is built upon a Fully Convolutional Network (FCN) architecture with an increased depth of 19 layers and a design of skip connections between different stages [16]. In short, U-net is a semantic segmentation model which has a contracting and an expansive path (Fig. 1). In this way, the contracting path extracts feature maps while spatial information is decreased, while the expansive pathway combines the feature and spatial information [24].

In this way, one of U-net's main features is overcoming the trade-off between localization and context with the two paths in the architecture. Thus, there is prior knowledge that small-sized patches can only observe the small context of input. In contrast, large-sized patches require more pooling layers and consequently will reduce the localization accuracy [16].

This model has been proved to be capable of fast and precise biomedical image segmentation using different images. Fan et al. [25] showed successful results at segmenting bone and lumbosacral nerve using U-net and achieving 94.5% and 90.5% of accuracy in bone and nerve segmentation. Wong et al. [26] proposed a U-net-based segmentation method for wrist bone, achieving up to 89% of accurate segmentation rate. Ling et al. [12] used a U-net model to segment single stem cell nucleus within colonies in phase-contrast images reaching a true positive rate above 88%.

As the original U-net architecture works with 512 × 512 size images, the input images were resized to fit the model for the experimental evaluation, whose results are presented in Section 4. For U-net from scratch, the original architecture was used, considering 64 channels as usual. For U-net using transfer learning, some of the initial layers (in the contracting stage) were frozen, and only the remaining layers were trained. In this sense, the U-net architecture was trained from scratch with the pre-training dataset. Then, the obtained weights were loaded into the contracting part instead of the randomly initialized weights to perform transfer learning. Finally, the last layer of this pre-trained model was concatenated to the consecutive convolution blocks' outputs in the resulting model.

3.5. Mask-RCNN

The Mask-RCNN model [27] was presented in 2017 as an extension of the Faster-RCNN model [28] for semantic image segmentation by adding a branch for predicting segmentation masks on each *Region of Interest* (RoI), in parallel with the existing branch for classification and bounding box regression from Faster-RCNN (Fig. 2).

Mask-RCNN includes three branches for predicting class, bounding-

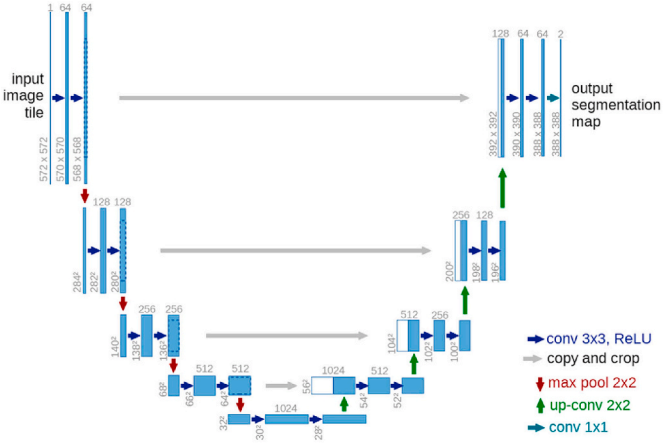

Fig. 1. U-net architecture. Contracting and expansive stages of the original U-net structure [9] work as follows: Each step in the contracting path consists of two consecutive 3 × 3 convolutions followed by ReLU nonlinearity and max-pooling using 2 × 2 window. Each step in the expansive path consists of an up-sampling of feature map followed by a 2 × 2 up-convolution. The reduced (by a 2 factor) feature map is concatenated with the corresponding cropped feature map from the contracting path. Finally, two consecutive 3 × 3 convolutions are applied followed by ReLU nonlinearity.

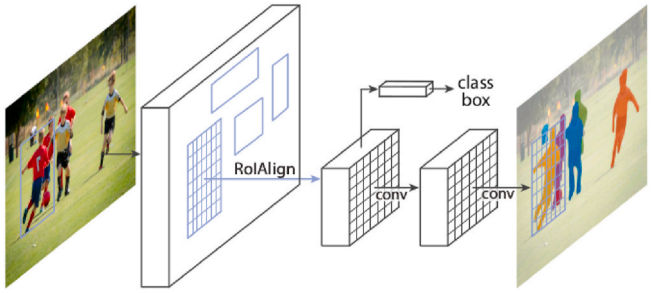

Fig. 2. Mask-RCNN architecture. The Mask-RCNN [27] model for semantic segmentation is based on a bounding box approach. It extends the Faster-RCNN model with an RoI-Align layer that preserves exact spatial locations. Thus, Mask-RCNN has three outputs for each candidate object: a class label, a bounding-box offset, and an object mask.

box, and segmentation mask for instances within an RoI. In that sense, Mask-RCNN adopts the same two-stage procedure of Faster-RCNN. The first stage is identical to Faster-RCNN and uses a Region Proposal Network (RPN) to predict object bounds and objectness scores at each location. Then, RoI-Align is used to preserve the spatial location instead of RoI-Pool as in Faster-RCNN [27]. Simultaneously predicting the class label and a bounding box offset in the second stage, MaskRCNN also outputs a binary mask for each RoI. The main idea is to apply bounding-box classification and regression in parallel, and the mask network head predicts the mask independently from the network head predicting the class. However, it was not a multi-class prediction [24].

Even the Mask-RCNN model was presented for semantic image segmentation and natural object detection; it has been demonstrated that this architecture is suitable for automatic segmentation of microscopic and medical images. For instance, Johnson [10] showed that Mask-RCNN could be used to perform highly effective and efficient automatic segmentation of a wide range of microscopy images of cell nucleus for different kinds of cells acquired under a variety of conditions, achieving up to 71% of segmentation accuracy. In the same line, Fujita et al. [11] adopted and improved the original Mask-RCNN model to simultaneously detect and segment cell nuclei in microscopy images achieving up to 89% as segmentation accuracy rate.

For this research, and aiming to have a fair comparison, we used 512 × 512 resized images from the dataset over the original architecture of the Mask-RCCN. When using transfer learning, the pre-trained weights were loaded, excluding the last layers of the model. At the same time, stochastic gradient descent (SGD) optimization was used to optimize the model leading to faster convergence from scratch and using transfer learning. An initial learning rate was reduced by 10% if the validation loss does not decrease for the following 50 epochs, where an epoch denotes an iteration over all training examples. A reduced learning rate helped to improve the model by fine-tuning to reach its local minimum.

3.6. Evaluation metric

There are various methods to evaluate the quality of segmentation given a hand-made segmentation. In general, the idea is to measure the difference between the automatic segmentation S against the manually segmented image G by computing some evaluation metrics. These metrics can be based on spatial overlap measures (e.g., Dice coefficient [29]) and on distance measures (e.g., Hausdorff distance [30]).

Our evaluation metric is similar to the one used in previous works related to sperm segmentation2, that is, the Dice coefficient, to compare our results to the ones of the state-of-the-art methods.

The Dice coefficient value is calculated as

$$D = \frac{2|S \cap G|}{|S| + |G|}$$

where S represents an automatic segmentation method, and G represents the hand-segmented masks (gold-standard). Therefore, the Dice coefficient values vary in the range [0, 1], where 0 indicates no spatial overlap between S and G, while 1 indicates complete overlap.

4. Experimental results and analysis

4.1. Dataset

For the experimental results that we show in this section, we have used the human sperm segmentation gold-standard SCIAN-SpermSegGS introduced in Ref. [13]. Semen samples from volunteers, with an age range of 28–35 years old, were obtained at the Laboratory of Spermiogram, Program of Anatomy and Developmental Biology (ICBM), Faculty of Medicine, University of Chile, Santiago, Chile. After collection, the semen samples were stained with a modified Hematoxylin/Eosin procedure (for details, review [13]). Digital images were captured using optical, bright field microscopy (Axiostar Plus, Carl Zeiss Inc, Wetzlar, Germany), a 63x objective (oil, NA 1.4) with an adapter of 0.63x and a digital camera (scA780-54gc, Basler AG, Ahrensburg, Germany). The SCIAN-SpermSegGS consists of 19 images with 264 sperm cells, where 210 are valid sperm cells (not at the image's border, without noise on it, etc.). Each image has 780 × 580 pixels. For each of these images, hand-made segmentation masks have been designed under the supervision of a referent expert in the field. Fig. 3 shows a representative image and its hand-made segmentation masks.

4.2. Experimental protocol

To face this challenging task of segmenting sperm cells in such a precise way towards an accurate morphological sperm analysis, we performed two different experiments evaluating two architectures: U-net and Mask-RCNN. The first experiment regards the original models' architecture without pre-trained weights and using the data augmentation strategy as explained in section 3.1 while evaluating the adapted models' performance using a specific and limited dataset. The second experiment aims to assess the impact of transfer learning in each of the chosen models.

For both experiments, we used a 5-fold cross-validation scheme. In this sense, we randomly split the original dataset for training and testing data. Then, we applied the data augmentation strategy only for the training dataset and reserved 20% of it for validation purposes. In each iteration, we train and perform hyper-parameter tuning, from which we choose the best model for each scenario.

In the training phase of both models, we applied hyperparameter optimization utilizing a grid search to find the best set of hyperparameters such as optimization method, initial learning rate, number of epochs, and batch size. In Table. 2 we present the ranges in which varied each one of the hyperparameters.

We implemented both models using TensorFlow and Keras framework on an NVIDIA TITAN RTX with 24 GB memory. Looking for reliability and robustness of the results, the whole training/validation/evaluation procedure was repeated 10 times, and the average results are reported. In the following subsections, we summarize the experimental results achieved using the SCIAN-SpermSegGs dataset.

4.3. Experiment 1: from scratch with data augmentation

Our goal in this first experiment is to evaluate the impact of each of the deep learning architectures revised in Sections 3.4 and 3.5 on the task of segmenting parts of human sperm cells. We adapted the original architecture for both models and used the data augmentation strategy as explained in Section 3.1.

We train both models on each choice of the training set and tune the hyperparameters on the validation set. Fig. 4 shows typical loss

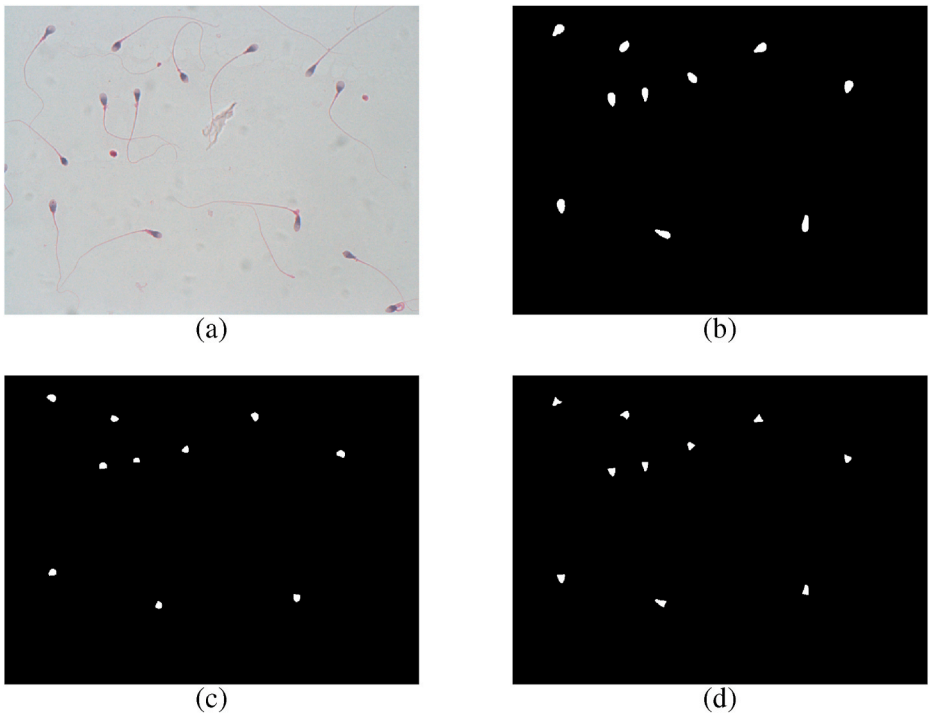

Fig. 3. Representative image from SCIAN-SpermSegGS. (a) Original image in RGB color space containing a number of human sperm cells. Image size: 780 × 580 pixels ∼ 164 × 122 μm. **(b)–(d)** Hand-made segmentation mask for head, acrosome and nucleus of valid sperm cells in (a), respectively.

Table 2
Hyperparameter grid search ranges.

Parameter	Min value	Max value
Batch size	2	16
Initial learning rate	0.00001	0.01
Number of epochs	10	200
Optimizer	{Adam, SDG with momentum}	

estimation during training on a specific choice of training and validation sets using Dice coefficient as cost function over U-net while evaluating different values for hyperparameters. It is seen that the training process converged within 150 epochs for U-net (Fig. 4a), using Adam's optimization algorithm (Fig. 4d) with an initial learning rate of 0.0001 (Fig. 4c) and batch size of 2 (Fig. 4b). For Mask-RCNN, Fig. 5 shows typical loss estimation during training on a specific choice of training and validation sets using Dice coefficient as cost function over Mask-RCNN, while evaluating different values for hyperparameters. It is seen that the training process converged when weights were optimized by stochastic gradient descent with momentum (Fig. 5d) with a 0.9 boost using an initial learning rate of 0.001 (Fig. 5c), batch size of 4 (Figs. 5b) and 180 epochs (Fig. 5a).

Fig. 6 shows the mean Dice coefficient for segmentation results of both models (U-net and Mask-RCNN) using SCIAN-SpermSegGS. The Dice coefficient assesses the quality of sperm head, acrosome, and nucleus segmentation by measuring the overlap with hand-segmented masks (gold-standard). As it is observed, U-net outperforms Mask-RCNN in segmenting heads, nucleus, and acrosomes of sperm cells using the mean Dice coefficient from scratch. Table 3 summarizes the mean Dice coefficient for the head, nucleus, and acrosome and standard deviation in each case for U-net and Mask-RCNN. As the table shows, the best results from scratch were achieved using U-net getting up to 89% of overlapping against hand segmented masks, with average Dice coefficients of 0.93, 0.88, and 0.88 for the head, nucleus, and acrosome segmentation, respectively. In contrast, the worst scenario resulted from using Mask-RCNN to achieve up to 84% of overlapping against hand segmented masks, with average Dice coefficients of 0.87, 0.85, and 0.80 for the head, nucleus, and acrosome segmentation, respectively. Furthermore, we used Wilcoxon rank-sum test with $\alpha = 0.05$ significance level to show that U-net performance is statistically significantly better compared to the performance of Mask-RCNN without transfer learning and using the limited dataset SCIAN-SpermSegGS following the data augmentation strategy presented previously.

4.4. Experiment 2: original model with transfer learning

This second experiment aims to assess the impact of transfer learning in the segmentation of human sperm parts using U-net and Mask-RCNN. Similarly to Experiment 1, we train the deep learning architectures on each choice of the training set of the dataset mentioned in the next paragraph. However, we no longer tune the hyperparameters since they have been adjusted in the scratch setting's previous case.

The dataset from the 2018 Data Science Bowl [21] was used for pre-training U-net and Mask-RCNN. This dataset contains a large number of segmented nucleus images. The images were acquired under various conditions and vary in cell type, magnification, and imaging modality (bright-field and fluorescence). We used the first stage training set of 670 images with more than 37, 000 segmented cell nucleus.

For the transfer learning process, considering that the 2018 Data Science Bowl dataset contains images of different sizes, the first task was to resize all the images to match the standard size of 512 × 512 to pre-train both models. We used the same data partition and cross-validation strategies as explained in Section 3.3.

Fig. 7 shows a comparison of mean Dice coefficient for segmentation results of U-net trained and tested with SCIAN-SpermSegGS (as in Experiment 1) against a U-net model pre-trained with the 2018 Data Science Bowl dataset. As the figure shows, pre-trained U-net outperforms U-net without transfer learning significantly in segmenting heads, nucleus, and acrosomes of sperm cells using the mean Dice coefficient as an evaluation metric. There is an improvement of up to 3.7%, 7.1%, and 5.8% overlapping against segmented masks of heads, nucleus, and acrosomes, respectively, with an average improvement of 5.5%. It is

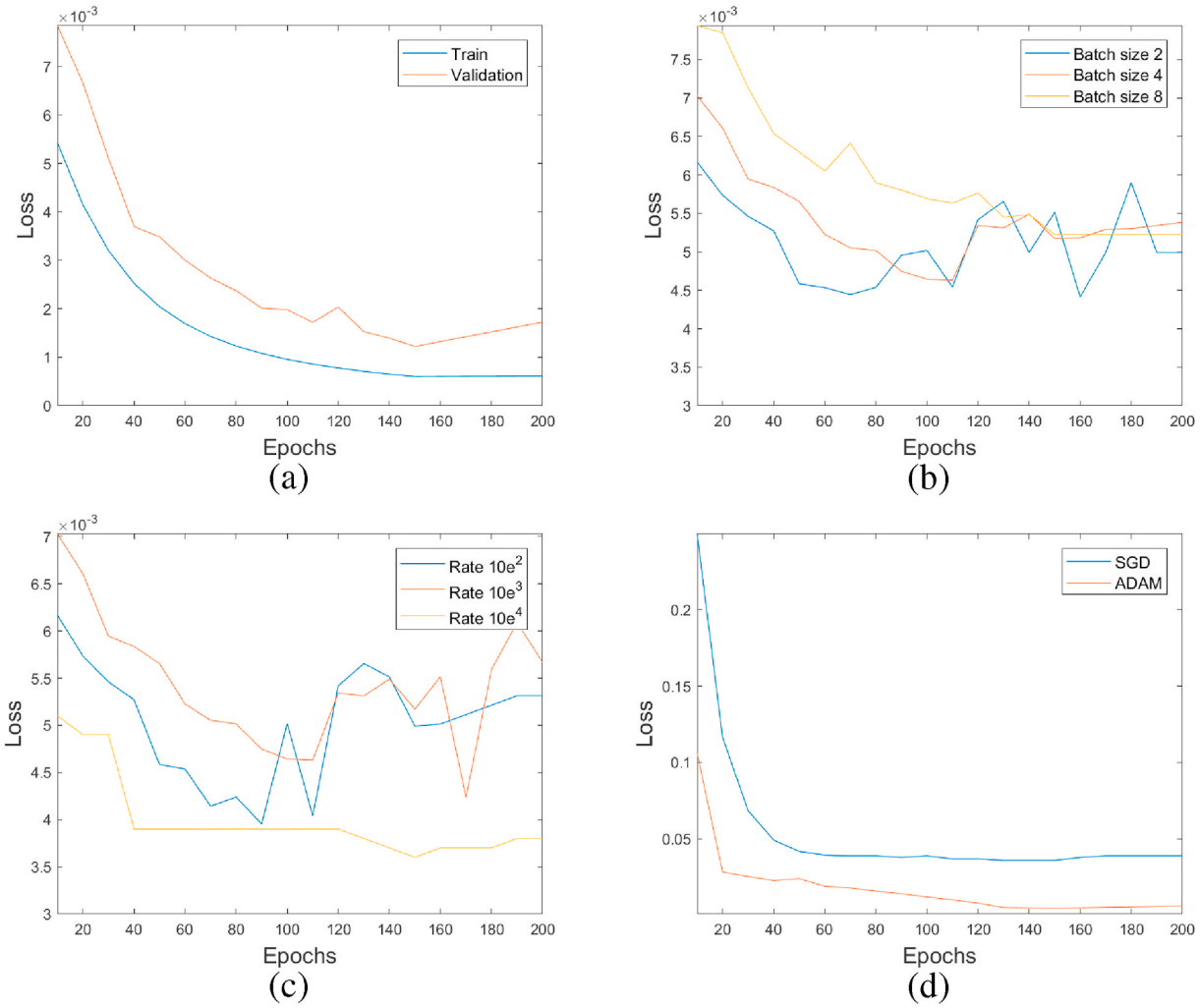

Fig. 4. Typical loss estimation during training of U-net for human sperm segmentation Loss line plots in validation and training datasets during the training stage showing how the model is learning. **a)** U-net training stopped at the point where the validation loss began to stabilize at epoch 150. **b)** Although the batch size has an irregular impact over U-net training, in most epochs batch size 2 is better. **c)** Learning rate affects the performance of Mask-RCNN, and the best performance is obtained with 0.0001 as the initial learning rate. **d)** Mask-RCNN training shows similar performance when using Adam's algorithm and SDG for optimization of weights.

imperative to notice the boxes' height for pre-trained U-net that shows that most of the segmentation results are very close to the average performance, particularly for head and nucleus. Another remarkable essential fact is the low dispersion of results while using transfer learning with U-net architecture for segmenting human sperm cell parts. As for the previous experiment, we used Wilcoxon rank-sum test with $\alpha = 0.05$ significance level to show that pre-trained U-net performance is highly statistically significant compared to the performance of U-net without transfer learning.

Fig. 8 shows a comparison of mean Dice coefficient for segmentation results of Mask-RCNN trained and tested with the SCIAN-SpermSegGS (as in Experiment 1) against a Mask-RCNN model pre-trained with the 2018 Data Science Bowl dataset. As the figure shows, pre-trained Mask-RCNN outperforms the same model results without transfer learning in segmenting heads, nucleus, and acrosomes of sperm cells. Although there is an improvement compared to Experiment 1, there was a lower impact of transfer learning using Mask-RCNN architecture to segmentation human sperm heads, acrosomes, and nucleus against U-net improvement. In particular, the improvement in nucleus segmentation is up to 0.9% only. In comparison, there is an improvement of 5.3% and 4.1% overlapping against segmented masks of heads and acrosomes, respectively, with an average improvement of 3.4%. As in previous cases, we used Wilcoxon rank-sum test with $\alpha = 0.05$ significance level to show that pre-trained Mask-RCNN is statistically significantly better compared to the performance of Mask-RCNN without transfer learning.

Table 4 summarizes the mean Dice coefficient for the head, nucleus, and acrosome segmentation and standard deviation in each case for U-net and Mask-RCNN with transfer learning. As it is observed, the best results using transfer learning were obtained using U-net achieving up to 95% of overlapping against hand segmented masks, with average Dice coefficients of 0.96, 0.95, and 0.94 for the head, nucleus, and acrosome segmentation, respectively. In comparison, the worst scenario resulted from using Mask-RCNN to achieve up to 88% of overlapping against hand-segmented masks, with average Dice coefficients of 0.93, 0.86, and 0.85 for the head, nucleus, and acrosome segmentation, respectively. We used Wilcoxon rank-sum test with $\alpha = 0.05$ significance level to show that U-net performance is statistically significantly better compared to the performance of Mask-RCNN using transfer learning.

4.5. Discussion

This paper presents a robust experimental framework for sperm cell segmentation using well-known deep learning segmentation approaches, achieving up to 95% average overlapping against hand

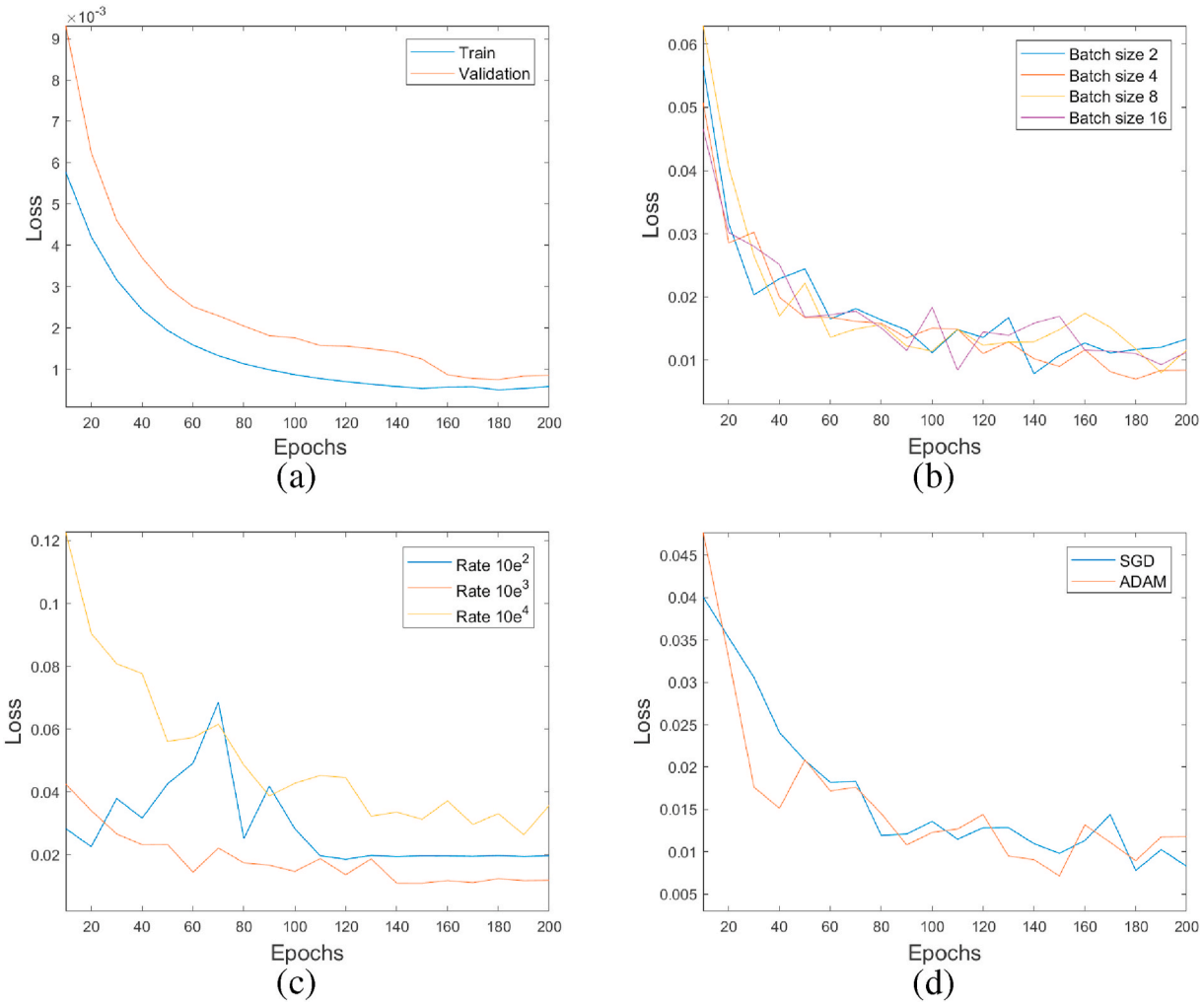

Fig. 5. Typical loss estimation during training of Mask-RCNN for human sperm segmentation Loss line plots in validation and training datasets during the training stage showing how the model is learning. **a)** The loss of the validation decreases using Mask-RCNN and becomes constant after 180 epochs. **b)** Mask-RCNN training is more establed when using batch size 4. **c)** Learning rate affects the performance of Mask-RCNN, and the best performance is obtained with 0.0001 as the initial learning rate. **d)** U-net training shows better performance when using Adam's algorithm for optimization of weights.

segmented masks. First, we evaluated U-net and Mask-RCNN models' performance, using data augmentation and transfer learning. The more accurate results were achieved while using U-net with transfer learning, getting up 0.96, 0.94, and 0.95 of Dice coefficient to segment sperm head, acrosome, and nucleus. Fig. 9 shows an image gallery with some segmentation results using the best approach presented in the previous section (U-net with transfer learning), considering head, acrosome, and nucleus segmentation. For each segmentation (head, acrosome, and nucleus), we present the best, average, and worst results regarding the Dice coefficient as an evaluation metric.

From the previous section, we can observe the significant impact of transfer learning in segmenting sperm parts. Using a very similar domain dataset contributes to having an average Dice coefficient more remarkable than the one obtained without pre-training using the U-net model (Fig. 7). The significant variability of the pre-training dataset in imaging modality, image magnification, cell type, and obviously, the dataset's size substantially impacts the achieved promising results.

Table 5 shows significant improvement concerning the state-of-the-art methods for the segmentation of human sperm parts. In this sense, the use of the U-net model with transfer learning improves the segmentation rates for sperm head, acrosome, and nucleus achieved by methods proposed by Shaker et al. [14] and Movahed et al. [15]. In particular, Shaker et al. proposal achieved segmentation rates of 91% for sperm head, 82% for acrosome and 87% for nucleus, using the same dataset SCIAN-SpermSegGS and Dice coefficient as the evaluation metric. It is essential to mention that these mean values for the Dice coefficient were recalculated according to the original individualized Dice coefficients because there was not reported standard deviation in the original article [14]. While Movahed et al. reported 90%, 77%, and 79% for sperm head, acrosome, and nucleus, respectively, using a modified version of the SCIAN-SpermSegGS dataset and the same evaluation metric. As it can be observed, the U-net model with transfer learning achieves 95% average Dice coefficient evaluated with SCIAN-SpermSegGS. This table showed that the U-net approach for segmenting sperm cell parts achieves a higher Dice coefficient and less standard deviation (which translates to less dispersion) concerning the related methods' results. In summary, the best deep learning segmentation approach evaluated in this research work achieves up to 95%, which is significantly better than the state-of-the-art results (87%). U-net results' statistical significance compared to Shaker's and Movahed's results were evaluated using the T-student test with $\alpha = 0.05$ significance level. According to this test, the improvements to head, acrosome, and nucleus segmentation were statistically significant.

There are specific situations in test images that impact the U-net model's performance using transfer learning (see Fig. 10). The deep learning segmentation models are not naturally immune to object position problems. In this sense, those sperm heads close to each other were not segmented by the model. Therefore, the model could not learn

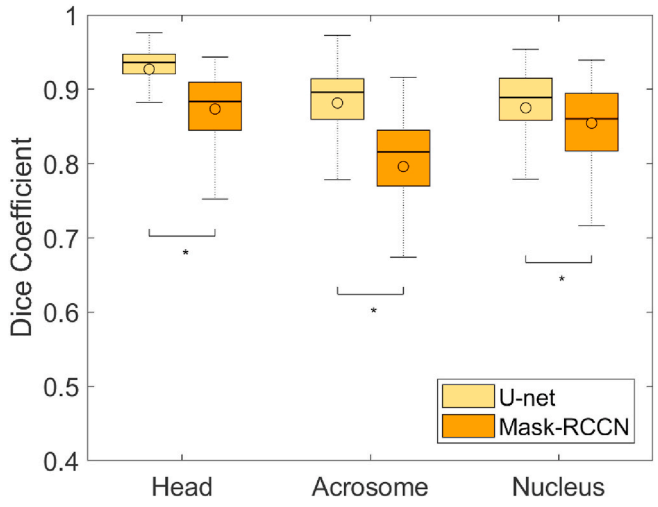

Fig. 6. Dice coefficient of deep learning models for human sperm segmentation. According to the Dice coefficient, U-net (yellow) and Mask-RCNN (orange) performance using data augmentation strategy over the SCIAN-SpermSegGS dataset. We show the median value (horizontal line) and the sample mean (o) for each box. The edges are the 25th and 75th percentiles on each box, and the whiskers extend to the most extreme data points that are not outliers. Statistically significant differences between U-net and Mask-RCNN using Wilcoxon rank-sum test are indicated (*$p < 0.05$).

Table 3
Mean Dice coefficient (with std) for head, acrosome and nucleus using U-net and Mask-RCNN from scratch.

	U-net	Mask-RCNN
Head	0.9271 ± 0.034	0.8736 ± 0.045
Acrosome	0.8824 ± 0.054	0.8039 ± 0.065
Nucleus	0.8750 ± 0.057	0.8544 ± 0.051

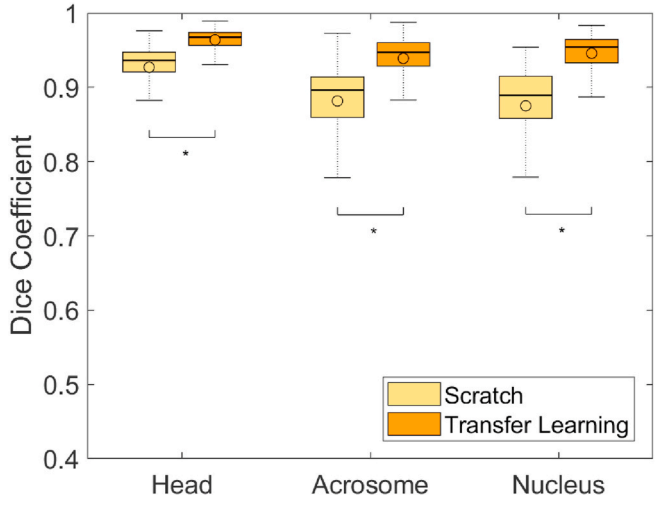

Fig. 7. Dice coefficient of the impact of transfer learning for human sperm segmentation using U-net. According to the Dice coefficient, U-net's performance from scratch (yellow) compared to performance using transfer learning (orange). We show the median value (horizontal line) and the sample mean (o) for each box. The edges are the 25th and 75th percentiles on each box, and the whiskers extend to the most extreme data points that are not outliers. Statistically significant differences between U-net from scratch and U-net with transfer learning using Wilcoxon rank-sum test are indicated (*$p < 0.05$).

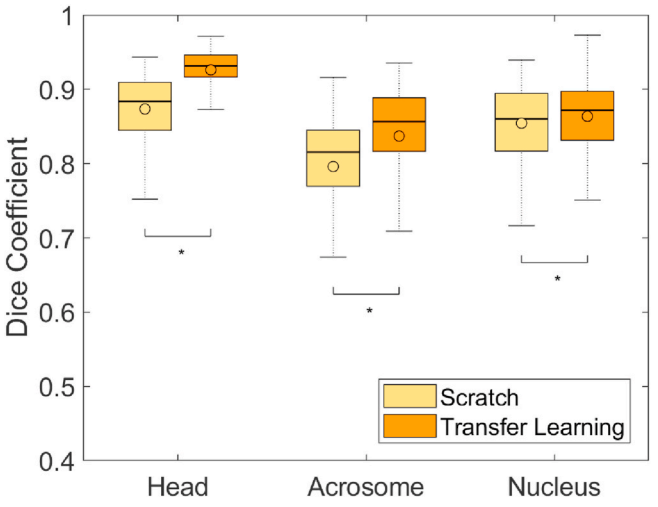

Fig. 8. Dice coefficient of the impact of transfer learning for human sperm segmentation using Mask-RCNN. According to the Dice coefficient, the performance of Mask-RCNN from scratch (yellow) compared to performance using transfer learning (orange). The edges are the 25th and 75th percentiles on each box, and the whiskers extend to the most extreme data points that are not outliers. We show the median value (horizontal line) and the sample mean (o) for each box. Statistically significant differences between Mask-RCNN from scratch and Mask-RCNN with transfer learning using Wilcoxon rank-sum test are indicated (*$p < 0.05$).

Table 4
Mean Dice coefficient (with std) for head, nucleus and acrosome using U-net and Mask-RCNN using transfer learning. Statistically significant differences between U-net and Mask-RCNN using Wilcoxon rank sum test are indicated (*$p < 0.05$).

	U-net	Mask-RCNN
Head*	0.9638 ± 0.016	0.9262 ± 0.029
Acrosome*	0.9403 ± 0.031	0.8453 ± 0.062
Nucleus*	0.9458 ± 0.026	0.8636 ± 0.044

to deal with differences like the closeness of two heads. Unfortunately, there are not enough heads with these characteristics in the original dataset to meet the validity condition to be clinically evaluated. The network fails to learn from these few cases provided as input. This fact is further accentuated by the lack of nearby cells in the dataset used for transfer learning. Considering that a model's learning capability depends on the diversity of training data, this is a reason to understand why the model does not perform well in Fig. 9a–c. Also, due to the type of staining, higher values for nucleus and acrosomes were not achieved. We believe that another staining technique could have highlighted the elements within the sperm head more specifically. Besides, enzymes were present in the image that could make it difficult to see and distinguish between the parts of sperm cells that are not distinguishable through U-net using transfer learning. Only one sperm acrosome (Fig. 9d) was not segmented at all, and none sperm nucleus for which the U-net model with transfer learning fails.

The vast majority of sperm cells for which segmentation using the U-net model with transfer learning is successful, i.e., a high Dice coefficient value is achieved, benefit from the similarity between the cells in the pre-training dataset and the original dataset.

5. Summary and conclusions

This paper has presented a robust experimental framework to evaluate two known image segmentation approaches using deep learning applied to human sperm part segmentation. In particular, the U-net [9], and Mask-RCNN [27] models were evaluated to assess their

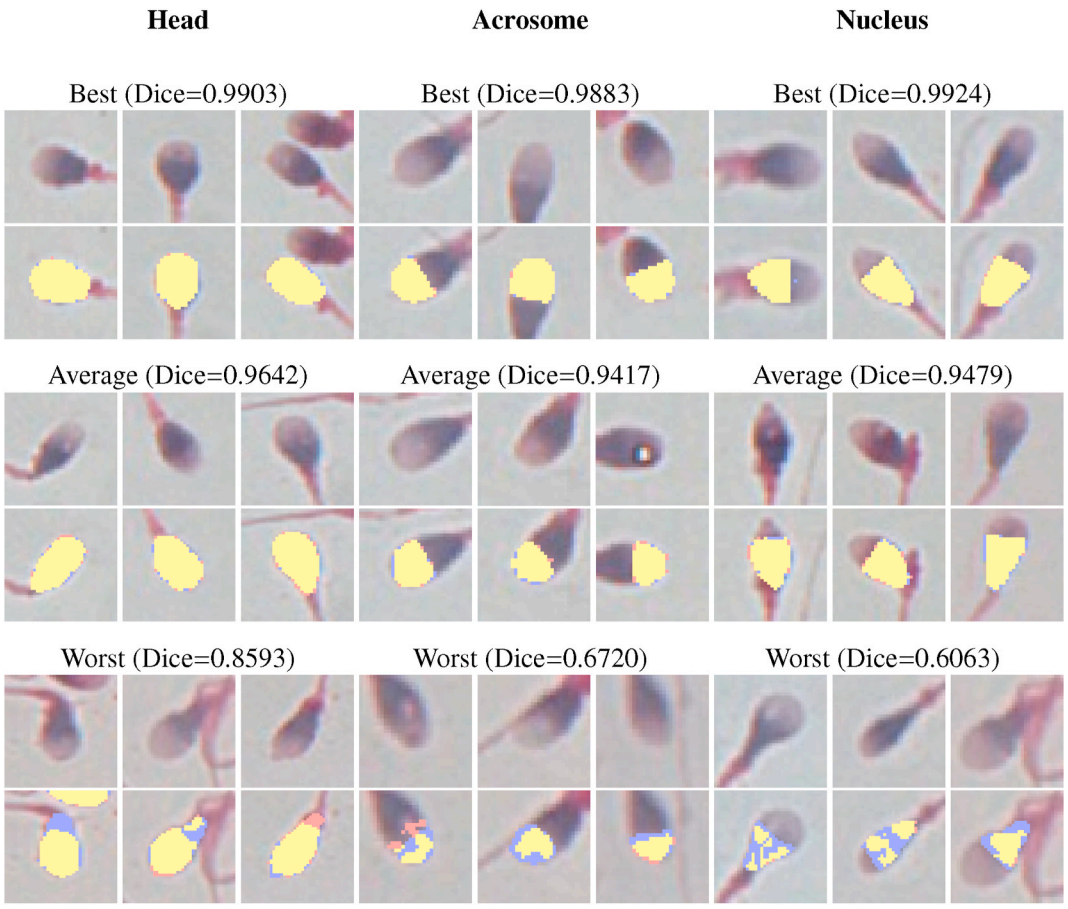

Fig. 9. Results of head, acrosome and nucleus segmentation. Upper row shows representatives for best results, middle row for average results, and last row for worst results. For each part (head, acrosome, and nucleus), we present the original (first row) and result obtained using U-net with transfer learning. The blue color represents the gold-standard, red presents U-net with transfer learning segmentation result, and yellow the overlap between gold-standard and U-net model.

Table 5
Mean Dice coefficient (with std) for the head, acrosome, and nucleus segmentation comparing U-net with transfer learning, Shaker's and Movahed's method. Mean Dice coefficient (with std) for Shaker's method were recalculated according to the Dice coefficient for each sperm cell shared by the first author of [14]. Mean Dice coefficient (with std) for Movahed's method is strictly as reported in Ref. [15]. Statistically significant differences between the U-net approach against state-of-the-art methods using T-student test are indicated (*$p < 0.05$).

	U-net TL	Shaker et al.	Movahed et al.
Head*	0.9638 ± 0.016	0.9127 ± 0.054	0.9040 ± 0.024
Acrosome*	0.9403 ± 0.031	0.8207 ± 0.099	0.7730 ± 0.004
Nucleus*	0.9458 ± 0.026	0.8651 ± 0.053	0.7880 ± 0.028

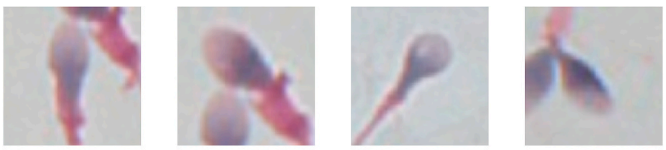

Fig. 10. Challenging cases for segmentation human sperm parts using U-net with transfer learning Some particular sperm cells are not segmented at all while using the best deep learning approach evaluated in this paper for segmentation of sperm head (a–c) and acrosome (d).

performance on the public dataset SCIAN-SpermSegGS [13] which has 210 sperm cells including hand-segmented masks for the head, acrosome, and nucleus (among other parts) of each sperm.

Due to the limited size of the original dataset, a data augmentation strategy was designed, combining a series of isometric geometric transformations, which ensure that the sperm cell's morphological characteristics are preserved, given the ultimate goal of achieving a morphological classification from the segmentation results. We applied a grid search to find the best set of hyperparameters to adapt both models, such as optimization method, initial learning rate, number of epochs, and batch size. Besides, seeking to take full advantage of the enlarged dataset's size, we used a 5-fold cross-validation scheme.

We performed two different experiments evaluating both architectures. The first experiment regards the original architecture without pre-trained weights and using the data augmentation strategy to assess the adapted model's performance employing a specific and limited dataset. The second experiment aims to assess the impact of transfer learning in each model using the 2018 Data Science Bowl dataset, which has more than 37,000 segmented cells.

Our experimental evaluation shows that U-net with transfer learning was the best configuration for this task achieving up to 95% overlapping against hand-segmented masks for sperm head (0.96), acrosome (0.94), and nucleus (0.95), using Dice coefficient as the evaluation metric.

We conclude that the impact of transfer learning is substantial, significantly improving the results of state-of-the-art methods (Shaker et al. [14], and Movahed et al. [15]) with a higher Dice coefficient, less dispersion, and fewer cases where the model failed to segment sperm parts. These results represent a promising advance in the ultimate goal of performing computer-assisted morphological sperm analysis. In this

sense, the next step is to take the results of sperm parts segmentation using transfer learning and the adaptation of the U-net model as an input to the current classification methods, under the hypothesis that a more precise segmentation would allow improving the classification accuracy in such a way that the whole framework could be considered as a support tool in the clinical field.

This paper suggests several directions for future research. First, we plan to explore and experiment with optimizing the U-net architecture limited to a specific task: segmentation of human sperm heads instead of using the whole architecture that considers multi-scale schemes in a problem with no multi-scale. Second, as computer-assisted morphological sperm analysis is currently an active research area in the machine learning community, we plan to continue researching this area. Finally, as said before, we are interested in assessing if the results presented in this work could improve the current morphology classification methods. The use of deep learning tools for segmentation and morphology classification does not have, as a whole, transcendental results in research.

Declaration of competing interest

We certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript titled Deep learning for human sperm segmentation towards an accurate morphological sperm analysis.

Acknowledgments

The authors thank P. Roman and J. Saavedra for their support with the processing units. In addition, the authors thank Fariba Shaker for sharing her method's Dice coefficient values. Violeta Chang was funded by ANID (PAI 77180012 and FONDECYT 11190851). Ruth Marín was partially funded by DIINF (DIINF_PO2019).

References

[1] A. Agarwal, A. Mulgund, A. Hamada, M. Chyatte, A unique view on male infertility around the globe, Reprod. Biol. Endocrinol. 13 (1) (2015) 1477–7827.
[2] J. MacLeod, R. Gold, The male factor in fertility and infertility – sperm morphology in fertile and infertile marriage, Fertil. Steril. 2 (1) (1951) 394–414.
[3] A. Domar, A. Broome, P. Zuttermeister, M. Seibel, R. Friedman, The prevalence and predictability of depression in infertile women, Fertil. Steril. 58 (6) (1992) 1158–1163.
[4] Who, World Health Organization - Laboratory, Manual for the Examination and Processing of Human Semen, fifth ed., Cambridge University Press, 2010.
[5] D. Katz, J. Overstreet, S. Samuels, P. Niswander, T. Bloom, E. Lewis, Morphometric analysis of spermatozoa in the assessment of human male fertility, J. Androl. 7 (4) (1986) 203–210.
[6] C. Brazil, Practical semen analysis: from a to z, Asian J. Androl. 14 (2010) 14–20.
[7] V. Chang, L. Heutte, C. Petitjean, S. Härtel, N. Hitschfeld, Automatic classification of human sperm head morphology, Comput. Biol. Med. 84 (2017) 205–216.
[8] J. Riordon, C. McCallum, D. Sinton, Deep learning for the classification of human sperm, Comput. Biol. Med. 111 (2019) 103342.
[9] O. Ronneberger, P. Fischer, T. Brox, U-net, Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, 9351, 2015, pp. 234–241.
[10] J.W. Johnson, Adapting Mask-RCNN for Automatic Nucleus Segmentation, arXiv E-Prints, 2018 arXiv:1805.00500.
[11] S. Fujita, X.-H. Han, Cell detection and segmentation in microscopy images with improved mask R-CNN, in: Proceedings of the Asian Conference on Computer Vision, ACCV, 2020, pp. 58–70.
[12] C. Ling, M. Halter, A. Plant, M. Majurski, J. Stinson, J. Chalfoun, Analyzing u-net robustness for single cell nucleus segmentation from phase contrast images, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2020, pp. 4157–4163.
[13] V. Chang, J. Saavedra, V. Castañeda, L. Sarabia, N. Hitschfeld, S. Härtel, Gold-standard and improved framework for sperm head segmentation, Comput. Methods Progr. Biomed. 117 (2) (2014) 225–237.
[14] F. Shaker, S. Monadjemi, A. Naghsh-Nilchi, Automatic detection and segmentation of sperm head, acrosome and nucleus in microscopic images of human semen smears, Comput. Methods Progr. Biomed. 132 (2016) 11–20.
[15] R. Movahed, E. Mohammadi, M. Orooji, Automatic segmentation of sperm's parts in microscopic images of human semen smears using concatenated learning approaches, Comput. Biol. Med. 109 (2019) 242–253.
[16] M. Hesamian, W. Jia, X. He, P. Kennedy, Deep learning techniques for medical image segmentation: achievements and challenges, J. Digit. Imag. 32 (4) (2019) 582–596.
[17] J. Wang, L. Perez, The Effectiveness of Data Augmentation in Image Classification Using Deep Learning, arXiv: 1712, 2017, pp. 1–8, 04621.
[18] L. Taylor, G. Nitschke, Improving deep learning with generic data augmentation, in: 2018 IEEE Symposium Series on Computational Intelligence, SSCI, 2018, pp. 1542–1547.
[19] H. Chan, R.K.L. Hadjiiski, C. Zhou, Deep learning in medical image analysis, Adv. Exp. Med. Biol. 1213 (2020) 3–21.
[20] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks?, in: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2, 2014, pp. 3320–3328.
[21] P. Mazzeo, A. Argentieri, F. De Luca, P. Spagnolo, C. Distante, M. Leo, P. Carcagnì, Nucleus segmentation across imaging experiments: the 2018 data science Bowl, Nat. Methods 16 (2019) 1247–1253.
[22] F. Renard, S. Guedria, N. De Palma, N. Vuillerme, Variability and reproducibility in deep learning for medical image segmentation, Sci. Rep. 10 (13724) (2020) 2045–2322.
[23] M. Browne, Cross-validation methods, J. Math. Psychol. 44 (1) (2000) 108–132.
[24] F. Sultana, A. Sufian, P. Dutta, Evolution of image segmentation using deep convolutional neural network: a survey, Knowl. Base Syst. 201–202 (2020) 106062.
[25] G. Fan, H. Liu, Z. Wu, Y. Li, C. Feng, D. Wang, J. Luo, W. Wells, S. He, Deep learning–based automatic segmentation of lumbosacral nerves on CT for spinal intervention: a translational study, Am. J. Neuroradiol. 40 (6) (2019) 1074–1081.
[26] L. Wong, L. Shi, F. Xiao, J. Griffith, Fully automated segmentation of wrist bones on T2-weighted fat-suppressed MR images in early rheumatoid arthritis, Quant. Imag. Med. Surg. 9 (4) (2019) 579–589.
[27] K. He, G. Gkioxari, P. Dollár, R. Girshick, R.-C.N.N. Mask, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2980–2988.
[28] S. Ren, K. He, R. Girshick, J. Sun, R.-C.N.N. Faster, Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.
[29] L. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302.
[30] G. Rote, Computing the minimum hausdorff distance between two point sets on a line under translation, Inf. Process. Lett. 38 (3) (1991) 123–127.