Curricular Unit: **Deep Learning**

Group Members: **Daniel Correia (20200665),**

Joana Rafael (20200588),

Ricardo Santos (20200620)

NOVA
IMS
Information
Management
School

Project Report

A Deep Learning approach to sperm-cell image classification

This document is accompanied by:

1. **SCIAN and HuSHem datasets**
2. **CNN Model selection**
3. **CNN Model Assessment**
4. **Image Classification APP**

github Repository: github.com/RicardoSantos0/Deep_Learning_Project

Abstract

Sperm analysis is one of the first steps when studying the fertility problems of a couple. The morphology of human sperm cells can vary widely and high counts of abnormal sperm morphologies can be the underlying cause of fertility issues. The World Health Organization recognizes 11 classes of sperm-cell heads, according to their morphology (size and shape), the most common being Normal, Tapered, Pyriform, and Amorphous. Here, we take advantage of 2 publicly available datasets (the SCIAN and HuSHeM datasets) to tackle the laborious problem of classification of sperm cell heads. We explore diverse convolutional neural network architectures in an attempt to find the most promising architecture for this particular classification task. Our results confirm the high degree of inter-expert variability in the laborious task of classification of sperm head analysis. Our achieved overall accuracy of 39% on unseen data and low precision and recall scores for the different classes showcases the scarcity and poor resolution of publicly available data for the present problem.

I. Introduction

Over 30% of the cases of infertility are related to men. A low production or the production of a low-quality spermatozoon, the male reproductive cells, can be the cause of human infertility (Maduro et al. 2002). These spermatozoa can be immobile and/or of abnormal shape. Abnormalities in the head of the spermatozoa are among the most common and play a major role in male infertility. Hence, a crucial step for male fertility diagnosis is the microscopy observation of sperm morphology to detect and classify head morphological defects and abnormalities. However, manual examination of sperm samples is a laborious and repetitive task that requires a long time to perform. Furthermore, there is a high degree of subjectivity in this task (from the laboratory technician analyzing the sample), as detecting and classifying very small variations in the head's shape can be difficult. Therefore, designing an automatic and accurate model to perform this task is of paramount importance, as it would save long hours of laborious work and increase the homogeneity in the classification of samples.

In this work, we develop a deep learning model to extract features from sperm images from two open-source datasets (the SCIAN-MorphoSpermGS[1] dataset and the Human Sperm head Morphology dataset – HuSHeM[2]) and perform a classification of the cell in 4 different categories (normal, tapered, pyriform and amorphous), which are recognized by the World Wealth Organization; this classification is attributed according to the morphological characteristics and size of the head of the cell.

II. Background

II.1. A brief historical overview of Convolutional Neural Networks (CNN)

CNNs have been successfully used in problem-solving since the '90s (LeCun et al., 1990), where architectures like LeNet-5 have seen application in e.g. digit reading (LeCun et al., 1998). From the 2000s onward, CNNs have seen modest increases in use and application in tasks now associated with computer vision such as face detection (Garcia and Delakis, 2004) or sex phenotyping in embryos (Feng Ning et al., 2005).

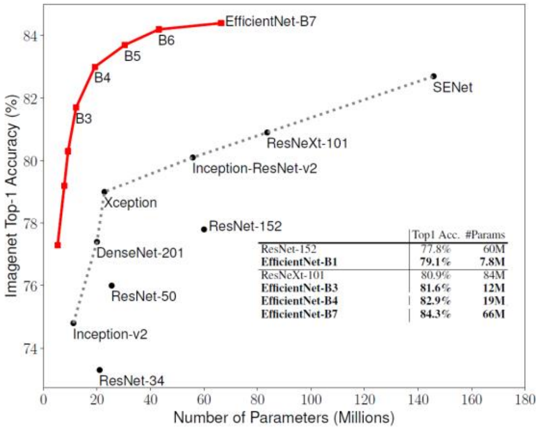
The breakthrough for CNNs occurred when the model later coined AlexNet (Krizhevsky et al., 2017) won the 2012 ImageNet Classification challenge by a considerable margin[3]. The interest garnered by the results pushed CNN architectures to the forefront of computer vision. Other noteworthy historical models are Very Deep Convolutional Networks -VGGs (Simonyan and Zisserman, 2015) and Residual Networks - ResNets (He et al., 2015) whose use and deployment were, and still is, popular among practitioners.

[1] Publicly available dataset with 1132 images of sperm cells (image size 35x35px) available at https://cimt.uchile.cl/gold10/ (last visited in 31/03/2021).

[2] Publicly available dataset with 216 images of sperm cells (cell size 131x131px) available at https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:77214;jsessionid=D919EC63361317E338A7AE6A0E5327C8 (last visited in 31/03/2021).

[3] 2012 results for Large Scale Image Recognition Challenge, available at http://www.image-net.org/static_files/files/ilsvrc2012.pdf (last visited in 31/03/2021)

As of the current date, the Efficient Net architecture is a popular and effective alternative to solve deep learning problems. As showcased in Figure 1, EfficientNets have been very successful in reaching high scores in the ImageNet Classification Challenge with fewer parameters than its predecessors.



	Top1 Acc.	#Params
ResNet-152	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
EfficientNet-B4	**82.9%**	**19M**
EfficientNet-B7	**84.3%**	**66M**

Figure 1: Tradeoff between the number of parameters and accuracy for ResNets, Efficient Nets, and inception models. Top-1 accuracy to be understood as the accuracy that the model assigns the highest probability to the class with the real label. Adapted from *(Tan and Le, 2020)*

As of the time of writing, the model that holds the highest scores on ImageNet is Meta Pseudo Models, a novel semi-supervised architecture from GoogleBrain where:

i. One model (the teacher) learns from data,
ii. A second model (student) learns from the output of the teacher model,
iii. The student's performance in the real world updates the weights on the teacher.

The models used as teacher and student are both Efficient Nets (that still rely on Convolutional layers) (Pham et al., 2021).

II.2. Computer Vision and Sperm Cell Morphology

The advances in the capabilities of Computer Vision have not gone unnoticed in the biological sciences. Ever since AlexNet popularized the use of deep learning in image classification, researchers worldwide have sought to find opportunities for computer vision to automate necessary, although laborious, tasks that require visual analysis.

The use of Computer Vision to detect and/or classify sperm cells has very diverse approaches (Ilhan et al., 2020; McCallum et al., 2019; Valiuškaitė et al., 2020). In the scope of our work, we will focus on use cases of classification of sperm cells based on head morphology:

i. Riordon *et al.* looked to classify different morphologies of a sperm's head by using the VGG16 architecture with 3 variations of weights: a) transfer learning with the weights of the ImageNet dataset, b) by training the model from scratch with the HuSHeM dataset, c) by training the model from scratch with the SCIAN dataset. Overall, the authors report an average True Positive Rate (TPR) of over 90% on HuSHeM and roughly 63% TPR on the SCIAN dataset (Riordon et al., 2019).
ii. Ilhan *et al.* used different deep and non-deep learning approaches to distinguish normal from abnormal sperm cells in a custom, not publicly available dataset, called SMIDS. In their work, the deep models with convolutional layers (VGG19, InceptionV3, and Mobile Nets) averaged 87% overall accuracy, severely outperforming non-deep models (Ilhan et al., 2020).

iii. Iqbal *et al*. proposed a 4 main kernel components architecture called *Morphological Classification of Human Sperm Heads (MC-HSH)*. The 53 layers convolutional model uses Leaky-ReLU as an activation function, uses channel-wise concatenation and element-wise addition to increase the effectiveness of model classification, and uses L2 regularization in the dense layers to prevent overfitting. In general, the model outperformed Riordon's results with VGG (Iqbal et al., 2020).

III. Methodology

III.1. Dataset Description, Partition, and Augmentation

The SCIAN dataset is a publicly available and gold-standard dataset for the morphological classification of human sperm heads. The images present in this dataset were classified by 3 different specialists in 5 categories: Normal, Tapered, Pyriform, Amorphous, and Small. The average vote of the labels was taken, yielding a highly unbalanced dataset comprising 1132 images (100 Normal, 228 Tapered, 76 Pyriform, 656 Amorphous, and 72 Small). These images' size is 35x35 pixels. We also used the HuSHeM dataset, a balanced dataset comprising 216 images of size 131x131 pixels. This dataset's images were classified into 4 distinct classes: Normal, Tapered, Pyriform, and Amorphous. All images were taken at a magnification of 63x under a microscopy setup and, in the case of the SCIAN dataset, show the same direction even if not necessarily the same orientation.

Due to the low sample size, and in the hope of increasing generalization ability, we opted to use both the SCIAN and HuSHeM datasets, eliminating the class *Small*. To accommodate the difference in magnitude of the images, we opted to downscale all images to a 32x32 pixel resolution. Whenever possible, we opted to use greyscale (more information will be provided in the respective model). We started by dividing our entire dataset by classes and further divided it into 3 folders, corresponding to the training set, validation set, and test set, with a distribution of 763 (60%), 383 (30%), and 130 (10%) images, respectively.

To tackle the issue of the highly imbalanced dataset and the scarcity of images, we performed two rounds of image augmentation in our training data, from here on referred to as *offline* and *online* augmentation. With *offline* data augmentation, we created physical variations of the training set by performing rotations, flipping, and mirroring them. All classes were increased to 6400 images per class. With *online* augmentation, we used the IMAGEDATAGENERATOR method from Keras to increase the diversity of our dataset while in the training process of the CNN. Here, provided with more options, we probabilistically rotated them, translated them (left or right and/or up or down), flipped them (vertically and/or horizontally), brightened or darkened them, and zoomed in up to 20%. Finally, all images from the training, validation, and test sets were converted to vectors, and the pixel's values were rescaled to be in the 0 to 1 interval.

III.2. Proposed Deep CNN Architectures

In total, we opted to test our model with 4 different architectures: one designed by us and 3 others inspired by ImageNet classification winners:

Model 1: Conv-layer (with 32 4x4 filters, same padding, and ReLU activation function), followed by a second Conv-layer (with 64 2x2 filters, same padding, and ReLU activation function). Before flattening, the dimensionality was reduced via max-pooling (filter 2x2). Following 3 fully connected layers (with 576, 1024, and 256 nodes respectively) with the ReLU activation function. The final layer was a 4 node dense layer with softmax activation function. All layers had BatchNormalization.

Model 2: Architecture inspired by the network developed by Alex Krizhevsky that won ImageNet in 2012: 5 Conv-Layers followed by 4 Dense Layers. The filter size of the first layer was reduced to account for the difference in proportion between the intended input (224 x 224 x 3 px) of AlexNet and our input is, at this stage (32 x 32 x 3 px).

Model 3: To minimize the issues stemming from our low sample size, we opted to also look for opportunities to make use of architectures (and weights) that have shown results in other multiclass classification problems: Model 3 was based on the ResNet50 architecture with the frozen ImageNet weights and the addition of 3 trainable Dense Layers, where the last has a softmax activation function.

Model 4: Model 3 was based on the Efficient Net B7 architecture with the frozen ImageNet weights and the addition of 3 trainable Dense Layers, where the last has a softmax activation function.

All models were trained with the following Callbacks: EarlyStopping on 20 epochs without decreases in val_loss, ModelCheckpoint on best val_accuracy, and ReduceLRonPlateau by 0.5 on every 5 consecutive epochs without decreases in val_loss.

The performance of all models was assessed by measuring the model's top-1 accuracy, precision (fraction of examples classified as positive that are truly positive), and recall (True Positive Rate) of every model. The model with the 'best' results was selected and retrained with some variations with the hopes of increasing performance:

i. RGB vs Grayscale input color scheme,
ii. Introduction of L2 regularization in the model's convolutional layers to help prevent overfitting.

The performance of our final model was then reviewed via top-1 accuracy and the area under the curve (AUC) of both Precision-Recall and Receiver Operator Characteristic (ROC) curves. To calculate the AUC on this multi-classification problem we used a *one vs all approach*.

The predictions on test data of the best performing model were then created as input for our SpermApp.py[4]. For a given labeled image, this app compares the predicted label and the true label and returns whether the model got the prediction right or not. This app is built to, in the future, be fed a new image directly and allow it to take the trained model to predict new labels, to be useful in a real-life context.

IV. Results

IV.1. Model Selection

The results of our model selection are presented in Table I. The results reveal better overall generalization power in model 2 – based on the AlexNet architecture. Model 4, which uses Efficient Net, displayed promising results, on a first look, promising results. However, upon further inspection, we noticed that the accuracy and recall values were due to the model only predicting "Amorphous" sperm heads. As a result, we selected model 2 to take to additional model assessment endeavors.

[4] Sent with the remaining files of the project under the name (Head_SpCells_App.py).

Table I: Results obtained when training our model. The best performing model (as in, with the best generalization ability) was model 2.

Performance Metric	Model 1	Model 2	Model 3	Model 4
Validation Accuracy	0.33	0.39	0.32	0.55
Validation Precision (weighed average)	0.34	0.41	0.39	0.30
Validation Recall (weighed average)	0.33	0.39	0.32	0.55

IV.2. Model Fine Tuning

Table 2 shows the results of the model with 2 different efforts for model optimization:

i. Model 2 with L2 regularization (equal to 0.01) on every convolutional layer.
ii. Model 2 with input shape adapted to receive grayscale images and L2 regularization (equal to 0.01) on every convolutional layer.

Table II: Performance of the tested variations of our AlexNet derivated network. Model 2.a.: Originally trained model with RGB input, Model 2.b: Originally trained model with Grayscale input, Model 2.c: Model RGB and L2 Regularization on Conv layers, Model 2.d: Greyscale and L2 Regularization.

Performance Metric	Model 2.a	Model 2.b	Model 2.c	Model 2.d
Validation Accuracy	0.39	0.38	0.33	0.35
Validation Precision (weighed average)	0.41	0.40	0.37	0.36
Validation Recall (weighed average)	0.39	0.38	0.33	0.35
Time (s/epoch) – GPU: NVIDIA GTX 1660 TI – 5 GB	25	20	100	55

The original model ended up being the one with the "best" results, even if these are not significantly superior to other model options. When tested with our test set, the model's predictive abilities were fairly similar to the results obtained with the validation set:

Table III – Performance of our final model Model on test data.

Performance Metric	Model 2.a
Test Accuracy	0.39
Test Precision (weighted average)	0.41
Test Recall (weighted average)	0.39

As shown in figure 2a, the model's ROC Curve for each class is very close to the diagonal (which represents the performance of a random classifier), in some classes even below it with an area under the curve (AUC score) close to 0.50 in most cases.

Additionally, the generally horizontal lines for each class, seen in the Precision-Recall Curve from Figure 2b, highlight, once more, the poor predictive performance of the obtained CNN; horizontal curves in Precision-Recall Curves are typical of a random or baseline classifier.

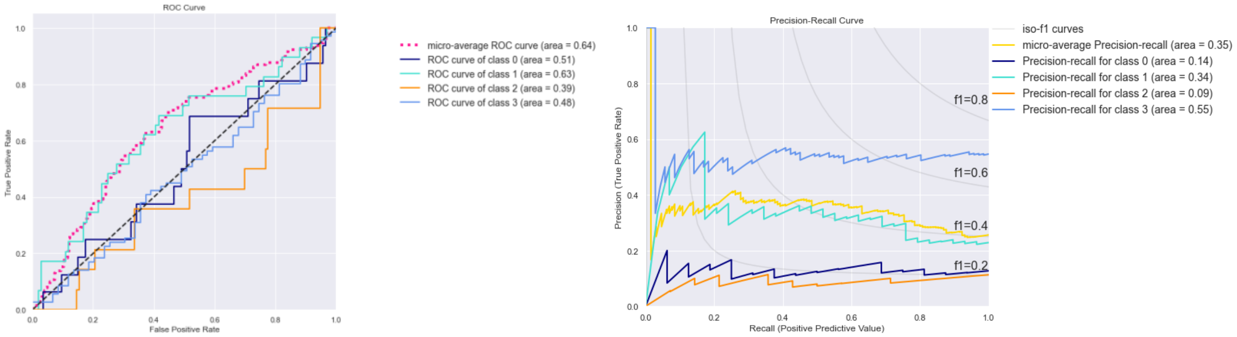
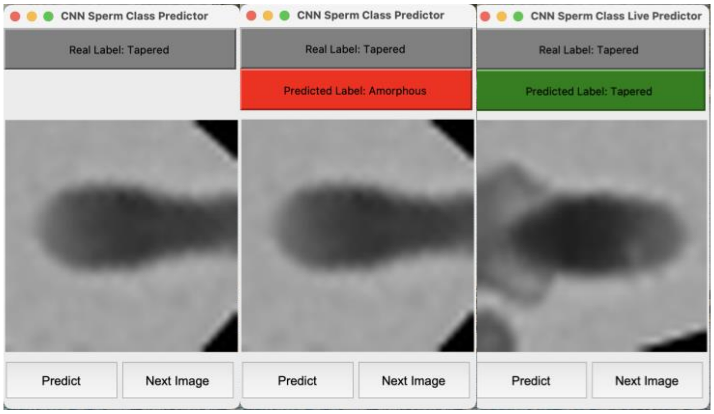


Figure 2: Tradeoff between the number of parameters and accuracy for ResNets, Efficient Nets, and inception models. Top-1 accuracy to be understood as the accuracy that the model assigns the highest probability to the class with the real label. Adapted from *(Tan and Le, 2020)*

IV.2. Visualizing the Classifier's Prediction


CNN Sperm Class Predictor
CNN Sperm Class Predictor
CNN Sperm Class Live Predictor
Predict
Next Image
Predict
Next Image
Predict
Next Image

Real Label: Tapered

Real Label: Tapered

Predicted Label: Amorphous

Real Label: Tapered

Predicted Label: Tapered

Figure 3: Developed head Sperm Classification App. On the left, the interface of the app; middle figure represented an incorrectly predicted label; figure on the right represented a correct predicted image.

To be able to visualize the prediction ability of our model, we have developed a small visual and intuitive app to make predictions of newly seen data. Ideally, this app could be used to visually observe the classification of the model of newly fed data, in real-time.

V. Discussion

In essence, the results display our model's difficulty to distinguish sperm cell head morphologies from one another. Even though we opted to train our model by resorting to CNN architectures that have shown to be effective in other applications, the performance of our classifier on unseen data was below our initial expectations. There are many possible reasons for these results and, in this section, we will look to discuss the possible reasons that may be behind the showcased poor performance, even when comparing with the work of other authors:

V.1. Scarcity and low resolution of Image Data

Our model was trained on a mixed batch of data stemming from the HuSHeM (a balanced 216 images dataset where each picture is originally 131x131px) and the Partial Agreement of the SCIAN dataset (a severely unbalanced 1132 images of sperm cells where the entire image size is 35x35px). In total, we worked with 1348 original images, with 708 of them belonging to a single class (*Amorphous*). This issue propagated to several problems:

a. The limitations of the low-resolution dataset led us to opt to downscale all input images to 32 x 32 px.
b. All augmented training data was derived from a minute amount of samples – thus biasing the sample.
c. The imbalanced validation set with a low sample size made it difficult to assess the results of the loss function on the validation data.
d. The combination of the low resolution of input images and other differences in the design of our experiment when compared to the designs made by other authors (see refer to section V.2.) hindered our ability to reach successful results with deeper, more modern architectures such as the Residual Net 50 and the Efficient Net due to vanishing gradient.

V.2. Differences in experimental design

When comparing the choices (when it comes to image manipulation and preprocessing) we made throughout our project with the choices made by other authors that have previously worked with these datasets, we noted some highlightable differences (Iqbal et al., 2020; Riordon et al., 2019) that may help explain the reported differences in performance:

a. In both quoted studies, each dataset was split and trained separately from one another.
b. Both authors opted to use 5-Fold Cross-validation on each dataset.
c. In Riordon's work, one of the five folds had been specifically designed to only include the "easy to predict results" on the test set (SCIAN Dataset). Easy to predict results to be understood as unanimously voted classifications.
d. Both authors rotated the images so that they were facing the same orientation. The importance of rotating the image may have been one of the keys to our model's low performance in comparison to others. The differences between sperm head cells are barely noticeable when the cells are perfectly lined up. The introduction of more variability (either as a result of augmentation or just by not aligning them automatically) could have contributed towards decreasing the model's discriminative capabilities.

V.3. The Sperm Head Morphology Classification Task

There are inherent difficulties to the task of labeling sperm heads. Case in point, the SCIAN Dataset we used included pictures where the head morphology was not consensual between experts. The introduction of non-consensual images as training data exacerbates the difficulty of a CNN model to learn.

VI. Conclusion

We looked to understand if and how we could develop a neural network that would be able to automate an important, but laborious and difficult classification task. For that, we combined the most prominent publicly available sperm image sets: HuSHeM and SCIAN Partial Agreement, and fed them as training data to different architectures of CNNs: one chosen by us and others following architectures that won the ImageNet challenge. The results we obtained with the selected model proved to be sub-optimal to the task at hand, highlighting the difficulty underlying the classification of biological and, specifically, microscopic images, often scarce and with low resolution. For future work, we will, with most priority, invest in preprocessing steps of the images, potentially allowing us to increase the capacity of deep neural networks to recognize patterns in the shape of sperm cells, and improving the generalization ability of the model.

VII. References

Feng Ning, Delhomme, D., LeCun, Y., Piano, F., Bottou, L., Barbano, P.E., 2005. Toward automatic phenotyping of developing embryos from videos. IEEE Trans. Image Process. 14, 1360–1371. https://doi.org/10.1109/TIP.2005.852470

Garcia, C., Delakis, M., 2004. Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection. IEEE Trans. PATTERN Anal. Mach. Intell. 26, 16.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. ArXiv151203385 Cs.

Ilhan, H.O., Sigirci, I.O., Serbes, G., Aydin, N., 2020. A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods. Med. Biol. Eng. Comput. 58, 1047–1068. https://doi.org/10.1007/s11517-019-02101-y

Iqbal, I., Mustafa, G., Ma, J., 2020. Deep Learning-Based Morphological Classification of Human Sperm Heads. Diagnostics 10, 325. https://doi.org/10.3390/diagnostics10050325

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90. https://doi.org/10.1145/3065386

LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D., 1990. Handwritten Digit Recognition with a Back-Propagation Network. Proc Adv. Neural Inf. Process. Syst. 396–404.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-Based Learning Applied to Document Recognition. Proc. IEEE 86, 2278–2324.

McCallum, C., Riordon, J., Wang, Y., Kong, T., You, J.B., Sanner, S., Lagunov, A., Hannam, T.G., Jarvi, K., Sinton, D., 2019. Deep learning-based selection of human sperm with high DNA integrity. Commun. Biol. 2, 250. https://doi.org/10.1038/s42003-019-0491-6

Pham, H., Dai, Z., Xie, Q., Luong, M.-T., Le, Q.V., 2021. Meta Pseudo Labels. ArXiv200310580 Cs Stat.

Riordon, J., McCallum, C., Sinton, D., 2019. Deep learning for the classification of human sperm. Comput. Biol. Med. 111, 103342. https://doi.org/10.1016/j.compbiomed.2019.103342

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv14091556 Cs.

Tan, M., Le, Q.V., 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. ArXiv190511946 Cs Stat.

Valiuškaitė, V., Raudonis, V., Maskeliūnas, R., Damaševičius, R., Krilavičius, T., 2020. Deep Learning Based Evaluation of Spermatozoid Motility for Artificial Insemination. Sensors 21, 72. https://doi.org/10.3390/s21010072