Computers in Biology and Medicine 111 (2019) 103342

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Deep learning for the classification of human sperm

Jason Riordon, Christopher McCallum, David Sinton*

Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, ON, M5S 3G8, Canada

ARTICLE INFO

Keywords:
Convolutional neural network
Deep learning
Fertility
Sperm selection
Sperm diagnostics
Sperm head classification
Transfer learning

ABSTRACT

Background: Infertility is a global health concern, and couples are increasingly seeking medical assistance to achieve reproduction. Semen analysis is a primary assessment performed by a clinician, in which the morphology of the sperm population is evaluated. Machine learning algorithms that automate, standardize, and expedite sperm classification are the subject of ongoing research.
Method: We demonstrate a deep learning method to classify sperm into one of several World Health Organization (WHO) shape-based categories. Our method uses VGG16, a deep convolutional neural network (CNN) initially trained on ImageNet, a collection of human-annotated everyday images, which we retrain for sperm classification using two freely-available sperm head datasets (HuSHeM and SCIAN).
Results: Our deep learning approach classifies sperm at high accuracy and performs well in head-to-head comparisons with earlier approaches using identical datasets. We demonstrate improvement in true positive rate over a classifier approach based on a cascade ensemble of support vector machines (CE-SVM) and show similar true positive rates as compared to an adaptive patch-based dictionary learning (APDL) method. Retraining an off-the-shelf VGG16 network avoids excessive neural network computation or having to learn and use the massive dictionaries required for sparse representation, both of which can be computationally expensive.
Conclusions: We show that our deep learning approach to sperm head classification represents a viable method to automate, standardize, and accelerate semen analysis. Our approach highlights the potential of artificial intelligence technologies to eventually exceed human experts in terms of accuracy, reliability, and throughput.

1. Introduction

A fundamental challenge in infertility diagnosis and treatment lies in performing rapid and consistent analysis of information-rich sperm images. Clinicians routinely evaluate the morphological characteristics of sperm populations, either as part of semen analysis of fixed, stained sperm as a first screen for infertility diagnosis, or during selection of unlabeled, live sperm for intracytoplasmic sperm injection (ICSI) – a popular assisted reproduction technology (ART) where an individual sperm cell is selected and injected into an egg. In both applications, quick and accurate classification of sperm into either normal or abnormal categories is critical. This assessment, however, is highly subjective and time-consuming [1]. While Computer-Assisted Semen Analysis (CASA) [2] brought a level of automation and standardization to the field, wide variability persists in how sperm cells are assessed. New methods are needed to standardize, automate, and accelerate the sperm classification process.

Traditional machine learning approaches have achieved important advances in sperm classification but have relied on manual extraction of cell features where cell shape-based descriptors, such as the head area, perimeter, or eccentricity, are fed into a classifier [3–7]. Notably, Chang et al. achieved an average true positive rate of 58% in classifying sperm heads within a dataset where at least 2 out of 3 experts agree [5]. In this case, the authors classified sperm into 5 categories listed in the *Laboratory Manual for the Examination and Processing of Human Semen* published by the World Health Organization (WHO) [8] using a cascade ensemble of support vector machine (CE-SVM) classifiers. CE-SVM requires extracting a series of shape-based descriptors, some rather intuitive (e.g., area, perimeter, eccentricity) and others more abstract (e.g., Zernike moments, Fourier descriptors, geometric Hu moments). To classify sperm heads into one of five categories, a two-stage approach is employed. In the first stage, an SVM is trained to filter out amorphous sperm and distinguish the remaining four categories at high accuracy. In the second stage, one of four SVMs (one for each non-amorphous class, and trained to be an expert on distinguishing that particular class from amorphous) is used to do a double-check and confirm that class assignment is correct.

Recently, Shaker et al. improved on this approach via an adaptive patch-based dictionary learning (APDL) approach, where class-specific dictionaries are trained from square patches extracted from sperm

[*] Corresponding author.
E-mail addresses: jason.riordon@utoronto.ca (J. Riordon), christopher.mccallum@utoronto.ca (C. McCallum), sinton@mie.utoronto.ca (D. Sinton).

https://doi.org/10.1016/j.compbiomed.2019.103342
Received 27 November 2018; Received in revised form 21 June 2019; Accepted 22 June 2019
0010-4825/

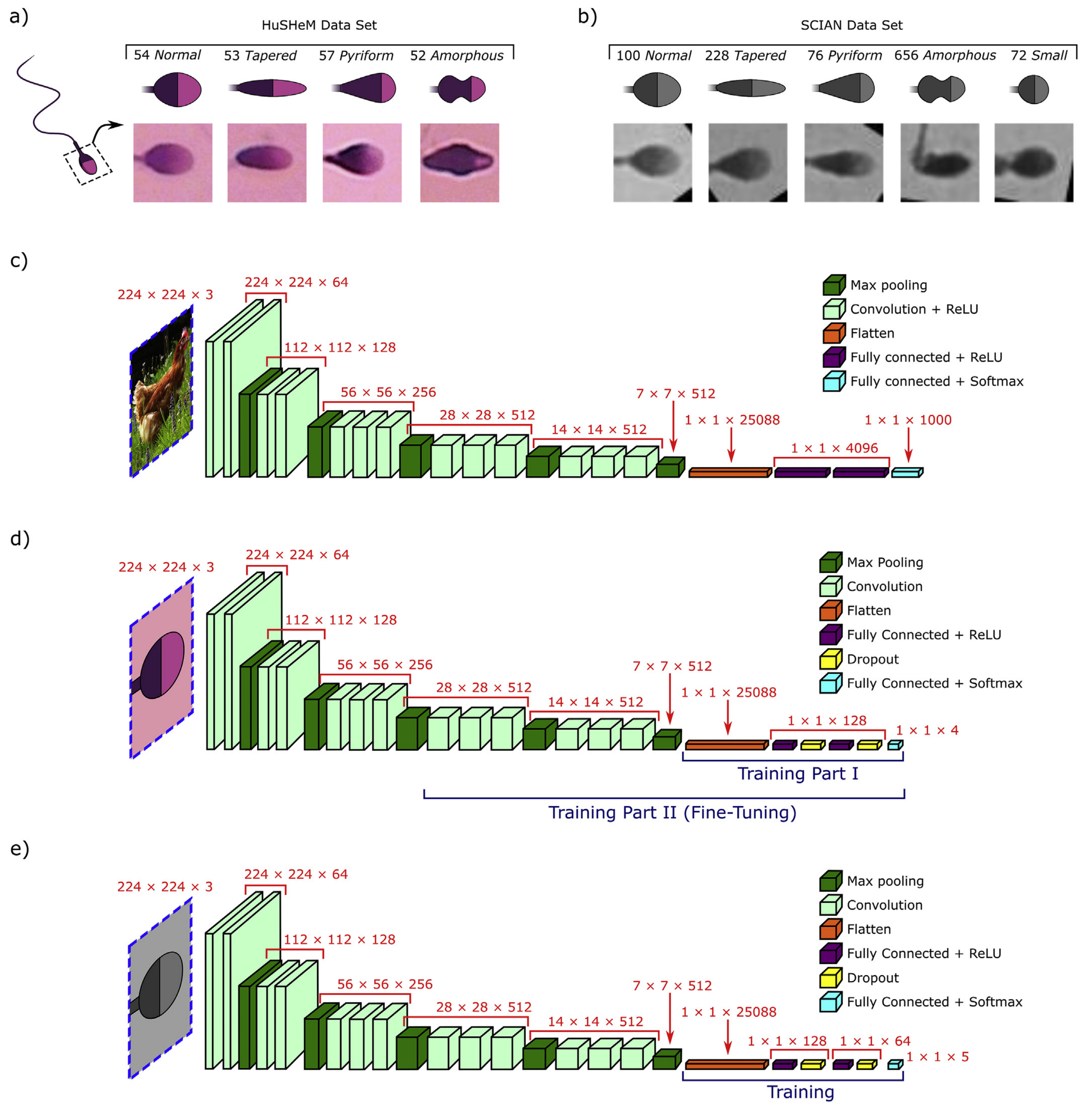

Fig. 1. Description of datasets and deep learning architecture. a) Sample images from the HuSHeM and b) SCIAN datasets as reported by Shaker et al. [9,11] and Chang et al. [3], respectively. HuSHeM images are reproduced from Ref. [11] and licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Schematics of WHO classes adapted from Ref. [8]. c) Original VGG16 convolutional neural network as proposed by Simonyan et al. [16]. VGG16 convolutional neural network optimized for sperm head classification for the d) HuSHeM and e) SCIAN datasets.

images [9]. This approach is a twist on the sparse dictionary learning strategy that Yang et al. successfully applied to face recognition [10]. In sparse dictionary learning, a sparse (i.e., loosely coupled) representation of elements is inferred, which can reasonably approximate any given signal as a linear combination of a small number of elements. This dictionary is learned from input data by minimizing a cost function. Yang et al. modified this approach by using class-specific dictionaries, and Shaker et al. further modified this approach by using small patches taken from sample images rather than images in their entirety. Test image patches were reconstructed from class-specific dictionaries and evaluated to determine the best matching class. They achieved an average true positive rate of 62% on the SCIAN dataset and an average true positive rate of 92.3% on their own dataset for cases where all three experts agree on class (HuSHeM) [11]. Collectively, these works demonstrate the vast potential of machine learning approaches in performing sperm classification – an arduous task where subtle sperm shapes must be rapidly discerned, and thus traditionally limited to expert clinicians.

Deep learning has emerged more recently as a form of machine learning that can efficiently leverage forms of structured data, such as sequences (e.g., ordered DNA sequences) or spatially structured data such as images (arrays of pixel values). Deep neural networks are typically composed of many layers, with specific architectures tuned to desired applications. Buoyed by the rise of low-cost GPUs and advances in neural network architectures, deep learning now offers several advantages over traditional machine learning approaches, notably (i) the ability to process raw images without the need for significant preprocessing or manual feature extraction, (ii) the potential for higher classification accuracy, and (iii) the simple visualization of learned features [12,13]. In related fields, such as microfluidic flow cytometry, deep convolutional neural networks (CNNs) have demonstrated significant improvement in cell classification over other machine learning approaches [14,15]. In CNNs, convolution layers perform a convolution operation on the input data and output the result to the next layer. Each node perceives only a small section of the preceding layer, allowing large image inputs to be processed effectively. These early layers perform feature extraction, and a series of fully-connected layers in the latter part of the network perform classification [16]. In the fertility field, deep convolutional neural networks are only beginning to have an impact, with Thirumalaraju et al. reporting at a recent conference a deep learning strategy capable of 89% accuracy in identifying healthy sperm based on clinician annotations, and ultimately 100% accuracy in identifying normal and abnormal samples as a whole [17]. Recently, Javadi and Mirroshandel showed how deep learning could be applied to low resolution images and detect abnormalities in sperm morphology [18].

In this work we show that a deep CNN is competitive with previous machine learning approaches in classifying human sperm according to WHO criteria. Rather than train a deep CNN model from scratch, we use transfer learning, whereby we retrain VGG16 [16] initially trained on the ImageNet [19] database, a collection of human-annotated everyday images (objects, animals, etc.). We then train a classifier using labeled sperm head images from freely-available datasets (HuSHeM [11] and SCIAN [3]) and fine-tune the network. Sperm head images are classified into WHO categories: Normal, Tapered, Pyriform, Small, and Amorphous. Our approach is highly effective, producing an average true positive rate of 94.1% on the HuSHeM dataset (matching the APDL approach and exceeding the CE-SVM approach) and 62% on the partial-agreement SCIAN dataset (matching earlier machine learning approaches). Our retrained network is capable of successful classification even in cases with few example sperm images. The VGG16 transfer-learning approach demonstrated here is computationally efficient by not requiring a full deep neural network or massive dictionaries be learned from scratch. Our work highlights the potential of deep learning methods to eventually exceed human experts in terms of accuracy, reliability, and throughput.

2. Methods

2.1. Dataset characteristics

Two datasets were used in this work to evaluate our VGG16-based approach: the human sperm head morphology (HuSHeM) dataset [11] (see sample images, Fig. 1a) and the Laboratory for Scientific Image Analysis Gold-standard for Morphological Sperm Analysis (SCIAN) dataset [3] (see sample images, Fig. 1b). Both datasets are publicly available and published as reference databases upon which to evaluate and baseline classification algorithms. Information on how samples were acquired and imaged are available here [3,9]. The HuSHeM dataset consists of 216 RGB images of stained sperm heads (54 Normal, 53 Tapered, 57 Pyriform, and 52 Amorphous), each 131 pixels × 131 pixels and taken at 100 × magnification. All images within the dataset were initially classified and labeled by 3 experts [9]; only images with 3-expert consensus were kept. The SCIAN dataset consists of 1132 greyscale images of stained sperm heads (100 Normal, 228 Tapered, 76 Pyriform, 656 Amorphous, and 72 Small), each 35 pixels × 35 pixels and taken at 63 × magnification. In this set, images were initially labeled by 3 experts, and samples with at least 2-out-of-3 expert agreement were kept. Within this "at least 2-out-of-3 agreement" dataset with a total of 1132 images, 384 images (35 Normal, 69 Tapered, 7 Pyriform, 11 Small, and 262 Amorphous) had full agreement. Full details on dataset characteristics are reported elsewhere [3,9,11]. Throughout this work, both datasets were kept fully separate, enabling VGG16 performance to be directly compared to earlier approaches on these same datasets.

2.2. Image processing

Images were processed before use. HuSHeM dataset sperm heads were aligned and cropped to 70 × 70 pixels using *ImageJ*. SCIAN dataset images were rotated to align sperm heads (i.e., rotated so heads face the same direction) and converted to RGB.

2.3. Dataset partitioning and rebalancing

Each dataset was randomly partitioned into 5 segments, where four groups (80% of all data) formed the training set, and the remaining group (20% of all data) formed the fully independent test set. To be clear, in every run the model was trained solely on the training set – the test set is used to evaluate the model. For 5-fold cross validation, this entire training/evaluation process was repeated five times from the beginning, each time rotating which 20% segment of the full dataset constitutes the testing set. This 5-fold cross-validation configuration was employed to provide a direct comparison with work by Shaker et al. [9], and avoid overfitting. The model trained on the training set from the "at least 2-out-of-3 expert agreement" SCIAN dataset, was evaluated on not only the corresponding test set with at least 2-out-of-3 expert agreement, but also on the subset of test images with full agreement. Thus, we evaluated how such a model performs on the "easy-to-predict" cells within the test set. Such an approach was employed by Chang et al. [5], allowing for direct results comparison.

Dataset partitioning was accomplished by randomly allocating 20% of images to each of 5 groups by shuffling a master list of images within each category and saving each image file within corresponding folders. These groups were unchanged throughout each series of five-fold cross-validations. Specific training set and test set image allocation for each fold is presented in Appendix A: Supplementary Material.

To rectify the imbalance between classes and avoid class bias, images within classes with fewer examples were duplicated as needed to have the same number of images in each class. For example, the Small training group in Fold 1 of the SCIAN dataset has 57 distinct images, each duplicated 9 or 10 times (randomly decided) for a total of 524 images, and thus matching the image total of the largest category, Amorphous. This process was used to produce balanced training and testing groups, in keeping with deep learning best practices [20]. To produce a learning curve that quantifies the impact of data volume on overall accuracy, training sets with fewer distinct images (but the same number of total images) were randomly generated using a similar process. Throughout this work, the true positive rates calculated during training (i.e., for each epoch) use balanced testing sets, whereas final true positive rates (at the end of training) are calculated using original testing sets.

2.4. Deep learning approach

A VGG16 convolutional neural network pre-trained on ImageNet was retrained to classify sperm images into WHO-assigned categories. The original VGG16 network is shown in Fig. 1c. The initial layers of this CNN are used to associate information spatially. This early part of the network is left intact, and we focus on retraining a strong nonlinear neural network classifier to interpret extracted features and enable classification of our set of sperm images. Our model is written in Python

Table 1
Network hyperparameters for the HuSHeM [11] and SCIAN [3] datasets.

Hyperparameter	HuSHeM	SCIAN
Maximum rotation (°)	0	2.5
Vertical/horizontal shift (%)	0	5
Vertical/horizontal image flipping	No	Yes
Dense Layer (DL) 1 size	128	128
Dropout rate of DL 1	0.4	0.05
DL 2 size	128	64
Dropout rate of DL 2	0.4	0.2
Learning rate during initial training	10^{-5}	10^{-5}
Learning rate during fine-tuning	10^{-6}	–
Batch size	64	64
Steps per epoch	25	200
Number of epochs for initial training	100	100
Number of epochs for fine-tuning	100	–

(v.3.6.5) in Keras (v2.1.4) [21] with TensorFlow (v1.7.0) [22]. The deep learning algorithms employed here were adapted from instructional code from Pedersen's series of Keras/TensorFlow tutorials [23] using Jupyter. Given the differences between datasets, a separate network configuration was applied to each dataset. The final hyperparameters used for each configuration are shown in Table 1. In both cases, the last two fully-connected layers were removed and replaced with two new fully-connected layers with ReLU activation functions and trained from scratch. Two dropout layers, which randomly set input units to 0 every iteration, were added after each of these fully-connected layers. A softmax layer of nodes corresponding to the number of classes was then added to each configuration. In the case of the HuSHeM configuration, it was beneficial to perform a second round of training (i.e., fine-tuning) to retrain the last 6 convolutional layers of the network as shown in Fig. 1d. A series of data augmentations were also implemented to enhance the dataset – in the case of the SCIAN dataset, images were randomly flipped, rotated up to 2.5° and shifted vertically and horizontally between 0 and 5% prior to being input to vastly expand the effective size of the training set. Retraining the network requires ~30 min and ~2 h per fold for the HuSHeM and SCIAN datasets, respectively, with the NVIDIA GeForce GTX 1080 GPU used in this work. The final configurations in Table 1 represent the strongest set of three series of runs after preliminary testing of combinations of hyperparameters (see Appendix A: Supplementary Information for additional tested configurations). For example, we found that for the SCIAN dataset, training over 20,000 steps (100 epochs × 200 steps per epoch) led to a stronger performance than 25,000 steps (100 steps × 250 steps per epoch), and so we settled on the former (62% vs 60% average true positive rate for the test set). These training parameters, when applied to retraining the VGG16 convolutional neural network shown in Fig. 1c, led to the networks shown Fig. 1d and e for the HuSHeM and SCIAN datasets, respectively.

2.5. *Evaluating performance*

Algorithm performance was evaluated via several metrics, namely average true positive rate (the percentage of correct predictions in each actual class, averaged over all classes), average positive predictive value (the percentage of correct predictions among predictions for that class, averaged over all classes), F-Score (the harmonic mean of the true positive rate and the positive predictive value, averaged over all classes), and accuracy (the percentage of correct predictions overall). In the case of the SCIAN full consensus dataset, where not all randomly distributed folds have examples in each category, categories with no examples were omitted in calculating the average true positive rate.

2.6. *Saliency mapping*

Saliency maps were prepared to map gradients and highlight regions of interest where small changes in input lead to large changes in output. A technical description of saliency mapping methodology has been presented in detail by Simonyan et al. [24] In summary, image-specific class saliency maps show which pixels have the greatest impact on classification (i.e., where the smallest change in value leads to the greatest classification score change). Ideally, areas with the highest pixel values are regions corresponding to sperm heads, and not unimportant background objects. For each dataset, we performed saliency mapping on three sample images from each class as well as on an averaged image (formed by averaging all images in each HuSHeM class, or ~45 images of each SCIAN class). Saliency maps were contrast-enhanced and visualized via the "16-color" lookup table in ImageJ.

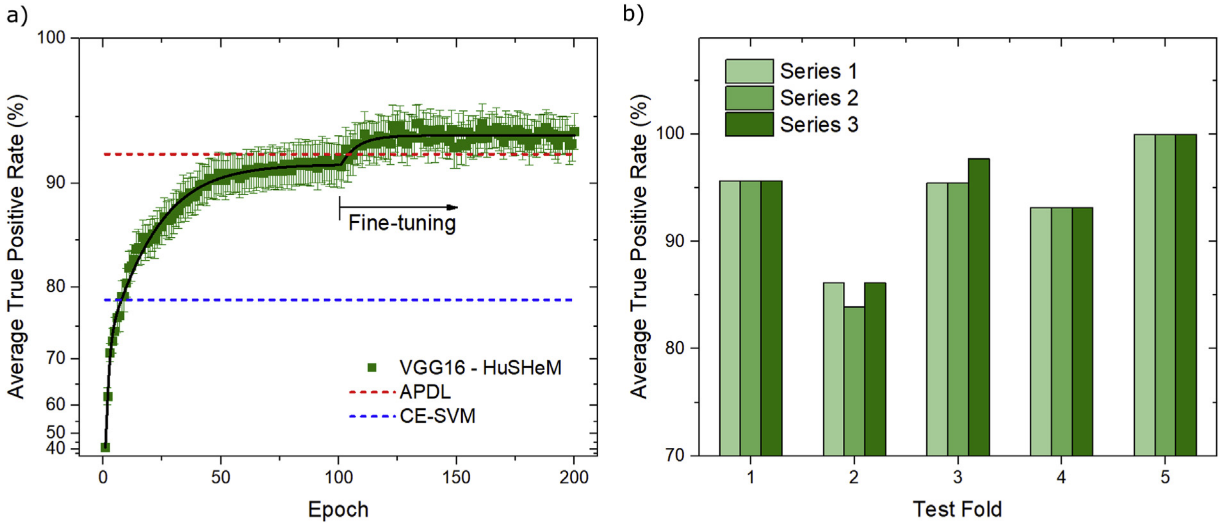

Fig. 2. VGG16 performance using the HuSHeM dataset [11]. a) Progression of testing average true positive rate of VGG16 during training as compared to the final true positive rate of APDL [9] and CE-SVM [5] approaches. Data are fitted to a saturation function for the initial training period (epochs 1–100) and the fine-tuning period (epochs 101–200). Each data point and error bar represent the average and standard error of 3 series of 5-fold validations (15 runs total). b) Average true positive rate measured on test sets at the end of each of 15 runs. Images were not re-allocated over the course of these runs – the same images are kept within each fold for each series. Thus, the images used to train and test are always the same for a given test fold.

3. Results and discussion

3.1. Training and testing on HuSHeM dataset

The performance of the deep learning model with the HuSHeM dataset is shown in Fig. 2. Fig. 2a shows how the testing average true positive rate improves over time during training for the HuSHeM dataset, both during initial training of the last two fully-connected layers (epochs 1–100) and during the fine-tuning stage (epochs 101–200). After initially training the classifier, the fine-tuning stage (where earlier layers are unlocked and slowly retrained) achieves a final true positive rate of 94.1%. This result represents a vast improvement over the performance of the CE-SVM method (94.1% vs 78.5% true positive rate), where shape descriptors (i.e., head size, ellipticity, Zernike moments) are used rather than images directly [9]. Further details on training set and test set true positive rate and loss during training are presented in Appendix A: Supplementary Material. Our method produces a true positive rate greater than that of APDL method (94.1% vs 92.3% true positive rate) [9]; however, given the high variability between fold compositions (datasets are small, and thus fold-to-fold variability is high), such an improvement cannot be considered statistically significant (*t*-test results in insignificant mean difference). This variability between folds is visualized in Fig. 2b. Each series of the same fold is consistent – most variability is between folds, where some folds have easier-to-interpret images than others. Our high fold-to-fold variability is consistent with the findings of Shaker et al., where standard deviation on accuracy (over 5 folds) was 7% compared to our 5% (over 3 series of 5 folds). This variability further justifies the need for larger datasets to discriminate between classification approaches. Nonetheless, the high true positive rate of our deep learning method on the HuSHeM dataset showcases the power and versatility of a retrained VGG16 network. Here, we used a fixed number of steps and epochs for all runs for a given dataset. However, an alternate approach is to auto-stop training based on performance criteria, such as loss. We performed a second set of 15 runs with identical hyperparameters, using an auto-stop function and found similar true positive rates to the fixed training period approach employed here (see Appendix A: Supplementary Information).

The 5-fold cross validation approach was employed here to directly compare with Shaker et al. [9], which used the same data distribution scheme. However, an alternate method is to separate data into training, validation, and test categories, with the latter data segment only being tested once optimization is complete. For comparison, we performed a new series of tests using such a data distribution by randomly allocating the existing 20% data segments to training (3 segments, 60%), validation (1 segment, 20%) and test (1 segment, 20%) sets. Hyperparameter optimization was performed solely on training and validation data over a series of 48 runs, with results detailed in Appendix A: Supplementary Information. The optimal configuration was then used to create a new model and evaluated on the test data. After independently training the same model three times, the average true positive rate for the test set was 96.2%. This result demonstrates that employing either 5-fold cross-validation or a train-validation-test configuration, our deep learning approach is robust and performant. An off-the-shelf VGG16 neural network pre-trained on non-medical images, with only a few tweaks and new example images, is proven capable of making highly accurate predictions without requiring computationally expensive full network training from scratch or learning and applying the massive dictionaries typical of sparse representation [25].

The confusion matrix produced following the training performed in Fig. 2 – where each cell value reflects how often each actual class is correctly or incorrectly identified – is presented as Table 2. Each class is correctly identified by the model over 92% of the time, with Normal cells most often correctly identified (96.4% true positive rate) and Pyriform cells least often correctly identified (92.3% true positive rate). Normal and Pyriform cells were also the best and worst performers, respectively, using the APDL method [9]. Among incorrect classifications, the most common mistake is misidentifying Tapered cells for Pyriform cells (5.5% of actual Tapered cells), and vice-versa (5.5% of actual Pyriform cells). The overall accuracy, average true positive rate, average positive predictive value, and average F-score for the HuSHeM dataset are listed in Table 3.

Table 2
Confusion matrix showing how often images from each class (Normal, Tapered, Pyriform, and Amorphous) are correctly or incorrectly predicted in the testing sets for the HuSHeM dataset [11]. Each cell value is the average of 15 runs, each normalized within respective ground truth (actual) categories. The green shading intensity within each cell is scaled to the maximum value within the matrix (shading insignificant outside the diagonal).

		Predicted			
		Normal	*Tapered*	*Pyriform*	*Amorphous*
Actual	*Normal*	96.4	0.0	3.6	0.0
	Tapered	0.0	94.5	5.5	0.0
	Pyriform	1.7	5.5	92.3	0.6
	Amorphous	0.0	3.0	3.8	93.2

3.2. Training and testing on SCIAN dataset

The deep learning model was also trained and evaluated on the SCIAN dataset set, with outcomes shown in Fig. 3. True positive rate (on test set) improves over time, as shown in Fig. 3a. In this case, fine-tuning was not beneficial and thus only improvement during the training of the two fully-connected layers is shown. After 100 epochs, the final average true positive rate was 62%, which is consistent with the performance of the CE-SVM (58%) [5] and APDL methods (62%) [9]. Despite the 5-fold larger number of training examples in the SCIAN dataset vs the HuSHeM dataset, the model does not perform as well. This reduced performance is attributed to a few key factors: the relative low resolution of SCIAN images, the use of 5 classes instead of 4, and the fact that the SCIAN dataset had a “at least 2 out of 3 experts agree” criterion, as opposed to the more stringent “3 out of 3 experts agree” condition. The latter human factor effectively sets an upper bound to the effectiveness of *any* modeling approach that is reliant upon expert assessment. Fig. 3b shows the consistency of each fold, with a standard deviation slightly lower than for the HuSHeM dataset (4% vs 5%). We also evaluated the model using an early stopping approach, as in the previous section, and found a similar true positive rate to our fixed training period method (see Appendix A: Supplementary Information). We also compared the standard 5-fold cross-validation approach to an alternate train-validation-test configuration as discussed in the previous section and obtain a final true positive rate on the test set of 59% (over 3 replicates).

The confusion matrix produced by VGG16 when applied to the SCIAN dataset is presented as Table 4. The Small category performed best (77.9% true positive rate), and the Amorphous category performed worst (38.3% true positive rate). A similarly low true positive rate for the Amorphous class was also reported using the CE-SVM method (30% true positive rate) [5]. As expected by the overall reduced performance as compared to the HuSHeM dataset, the classification mismatches for SCIAN are more striking than in Table 2. Interestingly, the Pyriform and Tapered classes that were problematic in the case of the HuSHeM dataset (cells of values 5.5 and 5.5 in Table 2) were among the most easily differentiated classes in SCIAN.

In the case of the SCIAN dataset, the VGG16 model trained on the full SCIAN dataset (the 2-out-of-3 expert training set) was also evaluated on the subset of SCIAN data with 3-expert agreement. All HuSHeM data have 3-expert agreement, but only a small portion of SCIAN has 3-expert agreement. Table 5 shows how the method again performs similarly to the CE-SVM method [5], reaching an average true positive rate of 72%. When the same model trained on data with at least 2-out-of-3 expert agreement is applied to both the corresponding 2-out-

Table 3
Comparison of sperm head classification approaches on the HuSHeM dataset [11].

Model	**Accuracy (%)**	Average True Positive Rate (%)	Average Positive Predictive Value (%)	Average F-Score (%)
CE-SVM	78.5	78.5	80.5	78.9
APDL	92.2	92.3	93.5	92.9
VGG16	94.0	94.1	94.7	94.1

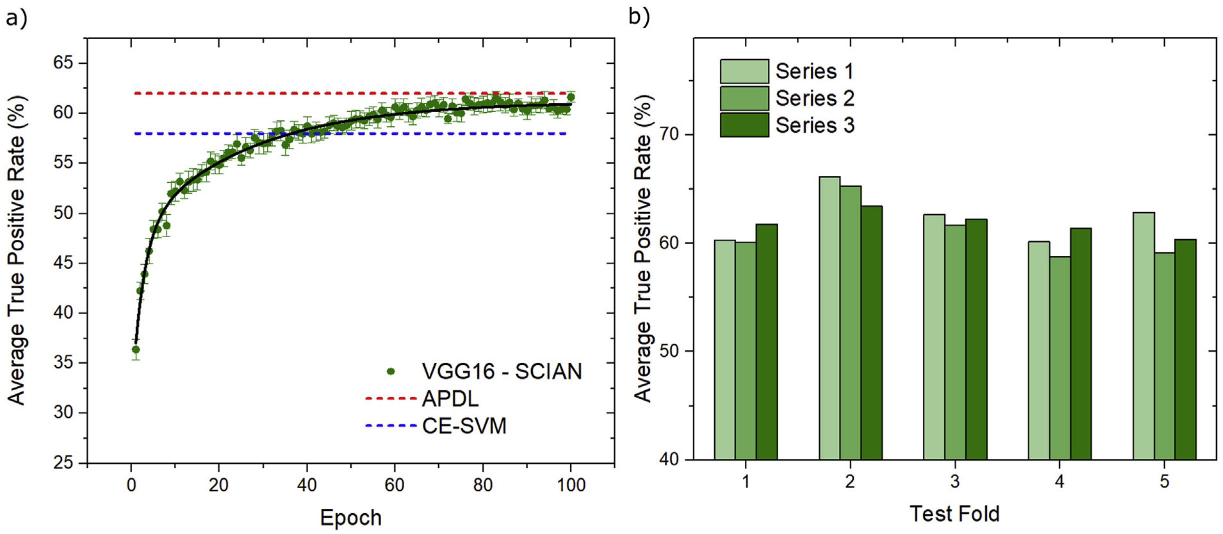

Fig. 3. VGG16 performance using the SCIAN dataset [3]. a) Progression of testing average true positive rate of VGG16 during training as compared to testing true positive rate of APDL [9] and CE-SVM [5] approaches. Each data point and error bar represent the average and standard deviation of 3 series of 5-fold cross validations (15 runs total). Data are fitted to a saturation function. b) Average true positive rate measured on test sets at the end of each of the 15 runs (3 series × 5 folds).

Table 4
Confusion matrix shows how often images from each class (Normal, Tapered, Pyriform, Amorphous, and Small) are correctly or incorrectly predicted in the testing sets for the SCIAN dataset [3]. Each value is the average of 15 runs and is expressed as a percentage of the sum of each ground truth (actual) category. The green shading intensity within each cell is scaled to the maximum value within the matrix.

				Predicted		
		Normal	*Tapered*	*Pyriform*	*Amorphous*	*Small*
Actual	*Normal*	67.0	5.0	12.0	12.3	3.7
	Tapered	10.5	56.6	1.8	20.9	10.1
	Pyriform	11.2	1.4	68.9	16.5	1.9
	Amorphous	18.5	10.3	15.5	38.3	17.5
	Small	5.3	5.6	1.5	9.8	77.9

Table 5
Comparison of average true positive rates for the SCIAN dataset for the CE-SVM, APDL and VGG16 approaches [3,5,9]. The same model trained on full training sets (with at least 2-out-of-3 expert agreement) was tested on corresponding test sets (with at least 2-out-of-3 expert agreement), as well as subsets of the latter where all three experts agreed.

Model	At Least 2-out-of-3 Expert Agreement (%)	3-out-of-3 Expert Agreement (%)
CE-SVM	58	74
APDL	62	–
VGG16	62	72

of-3 expert agreement test set as well as the 3-out-of-3 expert agreement test set subset, there is a clear improvement in the latter. This improvement further highlights the need for accurately labeled data and the potential of the approach. These findings indicate the importance of both the quantity of data and the quality of the labels in the success of deep learning approaches.

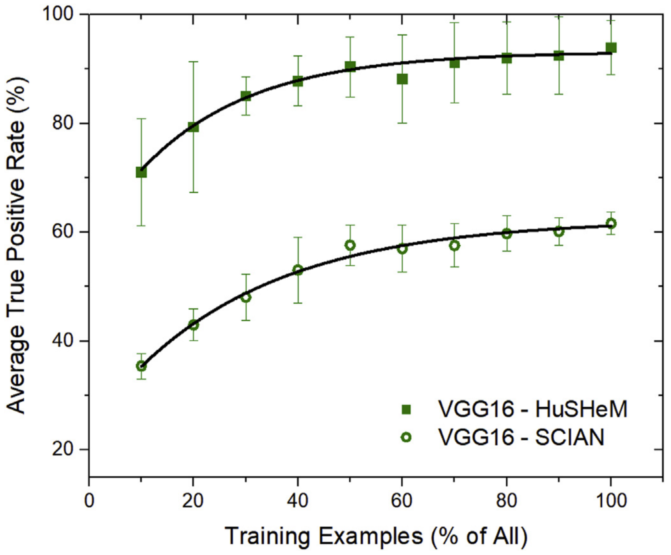

Fig. 4. VGG16 learning curves for (a) the HuSHeM dataset [11] and (b) the SCIAN dataset [3]. Each data point and error bar represent the average and standard deviation on the test set true positive rate following 5-fold cross validation, except for the "100%" case where 3 series of 5-fold cross validations were performed. The training and testing set for each iteration were balanced as described in **Methods**.

3.3. Learning curves

The learning curves for VGG16 for both the HuSHeM and SCIAN datasets are presented in Fig. 4. Learning curves for both datasets follow

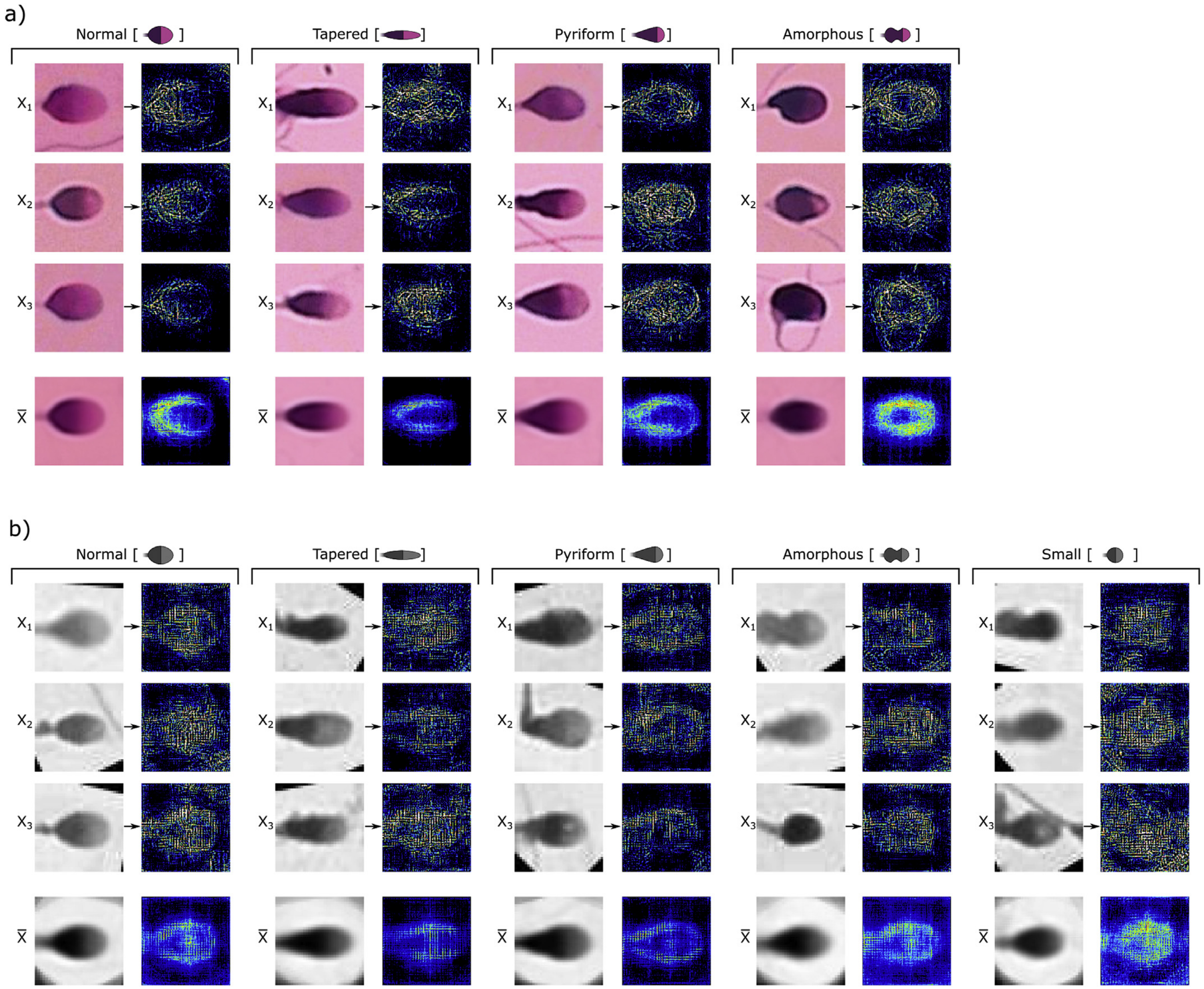

Fig. 5. Saliency maps for example images (X_1, X_2, and X_3) from (a) the HuSHeM dataset [11] and (b) the SCIAN dataset [3]. HuSHeM images are reproduced from Ref. [11] and licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Average images and saliency maps are also shown for each class ($\bar{X}$). In the case of the HuSHeM dataset, all images from the balanced training set are used to produce class averages. In the case of the SCIAN dataset, a sample of ~45 images per class are used to produce each class average. Additional information on saliency mapping can be found here [24].

the same saturation trend, where the impact of additional data is progressively reduced. Based on the saturation curve fits, one could expect only 0.3% and 1.2% further improvement from training on infinitely more data for the HuSHeM and SCIAN datasets, respectively. Thus, the model does not suffer from a lack of data. The transfer learning approach employed here means that much of what the algorithm needs to interpret an image is already in place, and, thus, can quickly get up to speed on classifying new sperm head examples (i.e., no need to train a network from scratch). It is encouraging that the model performs reasonably well (71%) with the lower bound “10% of All” value. For this test, the network was typically trained on only 16 sperm head images total: a mere 4 distinct images per class. This result highlights the effectiveness of the transfer learning approach employed here - harnessing the pre-trained image feature extraction ability of VGG16 to interpret a new dataset requires only minimal training. Specifically, by leveraging an extensively pre-trained network, the data requirements for subsequent application to sperm image analysis are minimal.

3.4. Saliency maps

The saliency maps of examples from (a) the HuSHeM dataset and (b) the SCIAN dataset are shown in Fig. 5. Saliency maps allow visualization of image classification models by mapping the gradient of the output to the input image [24]. These maps show how significantly the output class changes with small changes in the input image, and thus highlight regions of interest. Fig. 5 shows how the model excels at recognising the edges of sperm cells, and in some cases the inner regions of sperm as well. Outer pixels are, for the most part, ignored. Overall, these saliency images confirm that relevant features of sperm heads are recognised and valued, and thus offer insight into what is otherwise a black box neural network approach.

3.5. Robustness and generalizability

While the retrained VGG16 network approach works well on the datasets tested here, applying the model to a wider range of larger datasets (sourced from more donors and more clinics) is an important next step in evaluating generalizability. Notably, the small datasets used herein led to a large variability in loss and validation true positive rate between runs, and thus the performance of the approach here is constrained, at least in part, by the small size of the dataset. Nonetheless, this work has shown that employing a large network like VGG16, which has been pre-trained on 1000s of everyday images and retraining only part of the network is a powerful approach that leads to high true positive rates even on small datasets such as the HuSHeM and SCIAN datasets and is competitive with earlier approaches. The retrained VGG16 approach demonstrated here shows the potential of deep learning in sperm head classification, and could lead to improved automation, standardization and acceleration of semen analysis, particularly once trained with a larger data set.

4. Conclusions

Our work demonstrates that deep learning represents an effective means to classify sperm. Our retrained VGG16 CNN approach where images are input directly (and features extracted automatically) shows improvement in true positive rate over a CE-SVM approach using the HuSHeM dataset (94.1% vs 78.5%), and a similar performance when used on the SCIAN dataset (62% vs 58%). Notably, our approach does not require pre-extraction of shape descriptors, relying uniquely on image inputs. We also baseline our approach to an APDL method and demonstrate similar true positive rates (94.1% vs 92.3% on HuSHeM and 62% vs 62% on Partial Agreement SCIAN) without requiring the learning of massive dictionaries inherent to this approach. The VGG16 network, initially trained on a vastly different dataset of images, can be retrained to recognise subtle shape and size variations in sperm head images, requiring only a few examples. Whereas learning curves suggest that more data would not significantly improve performance, more data would facilitate the evaluation and direct comparison of classification algorithms. Saliency maps provide a strong indication that the model values sperm-relevant pixels appropriately. Overall, the strong performance of our model demonstrates the power and versatility of the deep learning approach, and the potential of convolutional neural networks to modernise sperm assessment in fertility clinics. The potential for AI technologies, as proven with reference to human expert data, could exceed human experts in terms of accuracy, reliability, and throughput.

Acknowledgements

Funding: This work was supported by the Natural Sciences and Engineering Council of Canada through the CHRP program [CHRP 508388-17]; the Canadian Institutes of Health Research [CPG 508388-17]; the Discovery Grants program [RGPIN-2015-06701]; the Discovery Accelerator program [477898-2015-RGPAS]; a Steacie Memorial Fellowship [492246-2016]; and the Canada Research Chairs program [230931]. The authors thank Professor Scott Sanner for helpful deep learning guidance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.103342.

References

[1] M. Freund, Standards for the rating of human sperm morphology. A cooperative study, Int. J. Fertil. 11 (1966) 97–180.

[2] R.P. Amann, D. Waberski, Computer-assisted sperm analysis (CASA): capabilities and potential developments, Theriogenology 81 (2014) 5–17, https://doi.org/10.1016/j.theriogenology.2013.09.004 e3.

[3] V. Chang, A. Garcia, N. Hitschfeld, S. Härtel, Gold-standard for computer-assisted morphological sperm analysis, Comput. Biol. Med. 83 (2017) 143–150, https://doi.org/10.1016/j.compbiomed.2017.03.004.

[4] W.J. Yi, K.S. Park, J.S. Paick, Parameterized characterization of elliptic sperm heads using Fourier representation and wavelet transform, Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering towards the Year 2000 and beyond (Cat. No.98CH36286), vol. 2, 1998, pp. 974–977, , https://doi.org/10.1109/IEMBS.1998.745610.

[5] V. Chang, L. Heutte, C. Petitjean, S. Härtel, N. Hitschfeld, Automatic classification of human sperm head morphology, Comput. Biol. Med. 84 (2017) 205–216, https://doi.org/10.1016/j.compbiomed.2017.03.029.

[6] J. Li, K.K. Tseng, H. Dong, Y. Li, M. Zhao, M. Ding, Human sperm health diagnosis with principal component analysis and K-nearest neighbor algorithm, 2014 International Conference on Medical Biometrics, 2014, pp. 108–113, , https://doi.org/10.1109/ICMB.2014.26.

[7] F. Shaker, S.A. Monadjemi, J. Alirezaie, Classification of human sperm heads using elliptic features and LDA, 2017 3rd International Conference on Pattern Recognition and Image Analysis, IPRIA, 2017, pp. 151–155, , https://doi.org/10.1109/PRIA.2017.7983036.

[8] World Health Organization (Ed.), WHO Laboratory Manual for the Examination and Processing of Human Semen, fifth ed., World Health Organization, Geneva, 2010.

[9] F. Shaker, S.A. Monadjemi, J. Alirezaie, A.R. Naghsh-Nilchi, A dictionary learning approach for human sperm heads classification, Comput. Biol. Med. 91 (2017) 181–190, https://doi.org/10.1016/j.compbiomed.2017.10.009.

[10] M. Yang, L. Zhang, J. Yang, D. Zhang, Metaface learning for sparse representation based face recognition, 2010 IEEE International Conference on Image Processing, 2010, pp. 1601–1604, , https://doi.org/10.1109/ICIP.2010.5652363.

[11] F. Shaker, Human Sperm Head Morphology Dataset, HuSHeM, 2017, p. 1, https://doi.org/10.17632/tt3yj2pf38.1.

[12] P. Eulenberg, N. Köhler, T. Blasi, A. Filby, A.E. Carpenter, P. Rees, F.J. Theis, F.A. Wolf, Reconstructing cell cycle and disease progression using deep learning, Nat. Commun. 8 (2017), https://doi.org/10.1038/s41467-017-00623-3.

[13] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.

[14] N. Nitta, T. Sugimura, A. Isozaki, H. Mikami, K. Hiraki, S. Sakuma, T. Iino, F. Arai, T. Endo, Y. Fujiwaki, H. Fukuzawa, M. Hase, T. Hayakawa, K. Hiramatsu, Y. Hoshino, M. Inaba, T. Ito, H. Karakawa, Y. Kasai, K. Koizumi, S. Lee, C. Lei, M. Li, T. Maeno, S. Matsusaka, D. Murakami, A. Nakagawa, Y. Oguchi, M. Oikawa, T. Ota, K. Shiba, H. Shintaku, Y. Shirasaki, K. Suga, Y. Suzuki, N. Suzuki, Y. Tanaka, H. Tezuka, C. Toyokawa, Y. Yalikun, M. Yamada, M. Yamagishi, T. Yamano, A. Yasumoto, Y. Yatomi, M. Yazawa, D. Di Carlo, Y. Hosokawa, S. Uemura, Y. Ozeki, K. Goda, Intelligent image-activated cell sorting, Cell 175 (2018) 266–276, https://doi.org/10.1016/j.cell.2018.08.028 e13.

[15] J. Riordon, D. Sovilj, S. Sanner, D. Sinton, E.W.K. Young, Deep learning with microfluidics for biotechnology, Trends Biotechnol. 37 (2019) 310–324, https://doi.org/10.1016/j.tibtech.2018.08.005.

[16] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 1409.1556, (2014).

[17] P. Thirumalaraju, C.L. Bormann, M. Kanakasabapathy, F. Doshi, I. Souter, I. Dimitriadis, H. Shafiee, Automated sperm morpshology testing using artificial intelligence, Fertil. Steril. 110 (2018) e432, https://doi.org/10.1016/j.fertnstert.2018.08.039.

[18] S. Javadi, S.A. Mirroshandel, A novel deep learning method for automatic assessment of human sperm images, Comput. Biol. Med. 109 (2019) 182–194, https://doi.org/10.1016/j.compbiomed.2019.04.030.

[19] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, , https://doi.org/10.1109/CVPR.2009.5206848.

[20] N. Japkowicz, The class imbalance problem: significance and strategies, In Proceedings of the 2000 International Conference on Artificial Intelligence, ICAI, 2000, pp. 111–117.

[21] F. Chollet, Keras, https://keras.io, (2015).

[22] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: a system for large-scale machine learning, Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, USA, 2016, p. 21.

[23] M.E.H. Pedersen, https://github.com/Hvass-Labs/TensorFlow-Tutorials, (2019).

[24] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, (2013) 1312.6034 , Accessed date: 14 November 2018http://arxiv.org/abs/1312.6034.

[25] L.L. Magoarou, R. Gribonval, Learning Computationally Efficient Dictionaries and Their Implementation as Fast Transforms, (2014) 1406.5388 , Accessed date: 14 November 2018http://arxiv.org/abs/1406.5388.