The Analysis of Mobile Platform based CNN Networks in the Classification of Sperm Morphology

Omer Lutfu Tortumlu, Hamza Osman Ilhan
Department of Computer Engineering
Yıldız Technical University, 34220 Istanbul, Turkey
omertortumlu@gmail.com, hoilhan@yildiz.edu.tr

***Abstract*—The diagnosis of male factor based infertility is performed by the evaluation of semen specimens in laboratories. Semen samples are investigated in terms of sperm concentration, morphology and motility. These investigations are generally performed manually by experts using microscopes instead of using computer based systems due to their high costs. However, manual observation also known as Visual Assessment (VA), has demonstrated significant subjectivity, including intra-observer and inter-laboratory variations. In this study, two CNN models especially for the possible usage in mobile platforms have been tested in the sperm morphology classification problem to eliminate the human factor in the analysis. In the analysis, three well-known sperm morphology data sets namely, HuSHeM, SMIDS and SCIAN-Morpho have been employed. Due to the data imbalance and scarcity problem of the utilized data sets, data augmentation and epoch analysis are also presented.**

***Keywords*—MobileNet, Sperm Morphology Analysis, Transfer Learning, Abnormal Image Classification**

I. Introduction

Infertility is prevented from having a child after 1 year of regular sexual intercourse. According to WHO reports, 15% - 20% of all couples in the world suffer from some kind of infertility problem [1]. The main reasons for these problems can be seen as male, female, both and unexplained reasons. Semen sample analysis, also known as spermiogram, is the most popular test in the diagnosis of infertility to observe problems related to the male factor. It is performed in two steps; i) The semen sample given is evaluated in terms of physical appearance such as viscosity, color and smell. ii) A specialist measures sperm characteristics in terms of morphology, concentration and motility parameters using computerized or manual evaluation techniques.

In the manual evaluation technique, also known as Visual Assessment (VA) which is cheaper and more practical than computer-based systems, diagnosis of male infertility is carried out by observing the semen sample under microscope. The concentration size, morphological shape, and motility of each sperm has impact in the determination of infertility.

Sperm morphology has an important issue on the diagnosis of infertility. Analysis focuses on dimensional evaluation for sperm head, middle part and tail in sperm morphology based analysis which is performed manually by experts with a high margin of error and objectivity. This observation techniques also known as visual assessment (VA). The variance of the results obtained by different observers was reported as 40% in the literature [2].

Several studies which aim to automatize the analysing procedure have been published about the sperm morphology classification. One of them is the study that uses the dictionary method in the examination of human sperm cells [3]. The dictionary learning algorithm is essentially a learning algorithm that incorporates signal processing [4]. In this study, a system called a dictionary was created by training the system with sperm cells. In the test phase, the algorithm compares the visuals with the resulting dictionary and decides the class of the image. In this study, a novel data set namely HuSHeM (Humen Sperm Head Morphology) has been presented. The results were reported as 92% classification accuracy. However, a manually preprocessing as rotating and cropping to a specific direction is a requirement in this study.

In another study, different sperm morphology data set called as SCIAN-Morpho has been presented [5]. They implemented a gold standard to measure the classification accuracy. However, similar to [3], a manual preprocessing have been applied each image patch before the analysis. They acquired 62% accuracy regarding to low resolution and data imbalance problem in the presented data set.

More recent study, Ilhan et. al used wavelet and descriptor based features in the classification [6], [7]. They introduced another sperm morphology data set named as SMIDS (Sperm Morphology Image Data Set). They fed SVM with wavelet and descriptor based feature sets for the classification of SMIDS. They obtained 82% and 84% accuracies by using Wavelet and SURF features without any preprocessing step, respectively.

Another study published by Ilhan et. al, they applied several preprocessing step to eliminate the noisy parts in the image patches [8]. Then they used SURF and MSER descriptor features to fed a non-linear classification schema, SVM with polynomial kernel. The proposed approach has increased the classification accuracies from 84% to 86% for SMIDS dataset. Similarly, they obtained 85% accuracy for HuSHeM dataset. The most significant novelty

978-1-7281-8073-1/20/$31.00

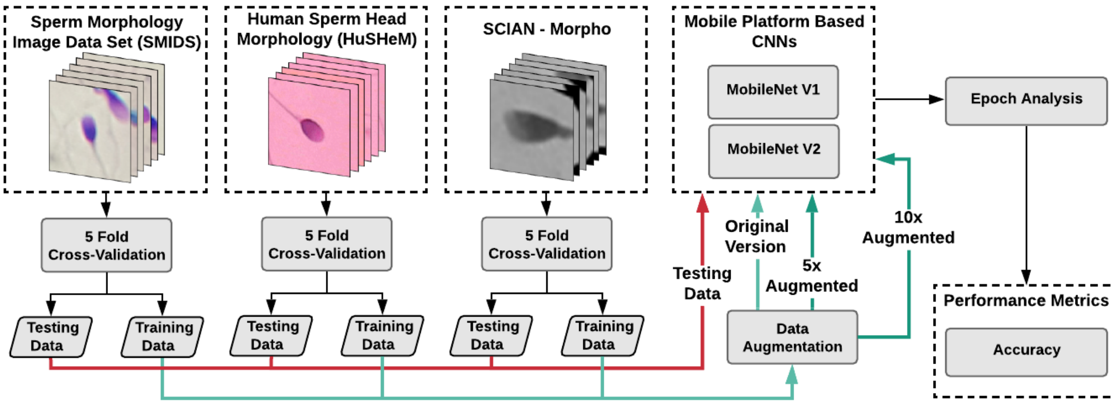

Figure 1: The Flowchart of the study

has been presented as the fully automatic classification approach for any stained sperm morphology image sets. They eliminated the manual preprocessing steps for both data set. A directional masking approach which have been published before [9] was used in this study as an automatic preprocessing step.

In Deep learning based studies, Ilhan et. al implemented MobileNet V2 for the classification of SMIDS similar to this study. But, they aim to train the MobileNet V2 CNN architecture from the stretch and to utilize a classification schema just for SMIDS [10]. They directly used the original images with the differently implemented augmented versions of SMIDS. They emphasized that the excessive times because of training from stretch is a negative side with the 87% accuracy performance.

Another deep learning based study have been performed by Riordon et. al [11]. They used a VGG16 CNN model with a fine tuning approach. However, similar to previous studies performed over HusHem and SCIAN morpho, they manually rotated and cropped each sperm images before the classification. They reported 94% and 64% accuracies for HuSHeM and SCIAN-Morpho data sets, respectively.

In this study, two different version of mobile platform based deep neural networks were utilized for morphological classification of sperm to eliminate the human factor in VA. Tests were performed over three well-known and previously presented datasets which have been stained with different bio-materials causing different colors in images. It is aimed to make morphological analysis autonomously by training these pictures with deep learning networks in mobile based platforms to provide short analyzing times. Eventually, more faster and objective results can be obtained by the proposed approach.

II. Materials and Methods

Two mobile platform based deep learning networks as MobileNet V1 and MobileNet V2 will be used to classify the three sperm morphology datasets. ImageNet based pretrained networks were used in terms of transfer learning idea. In the scope of this study, the effects of Epoch and Data Augmentation sizes were also investigated in order to obtain the highest classification accuracy rates. The flowchart of the study is presented in Figure 1.

This study was implemented with the Keras library running in Python over Ubuntu Operating System (Version 16.04). Keras is a high-level neural network application programming interface (api) written in Python [12]. Training times were measured in a hardware enviroment having Intel Core i5- 6400CPU@2.7GHz, NVIDIA Geforce 970GT.

A. *Dataset Information*

Three datasets as HuSHeM (Human Sperm Head Morphology), SMIDS (Sperm Morphology Image Data Set) and SCIAN-Morpho were used in the evaluation of the performances of the utilized mobile platform based CNNs. HuSHeM data set has four classes as normal, tapered, pyriform and amorphous which is almost equally distributed. SMIDS dataset is in three classes as Normal, Abnormal and Non-Sperm. The sample size is also equally distributed for this data set. SCIAN-Morpho dataset is divided into 5 classes as Normal, Tapered, Pyriform, Small and Amorphous and has a data imbalance problem which has great number of images in Amorphous class. In the training and testing sample organization, K-Fold Cross validation was used with $K = 5$. In this regard, 171, 906 and 2400 image patches were chosen as training samples, while 45, 226 and 600 image patches were chosen as testing samples for each created set of HuSHeM, SCIAN-Morpho and SMIDS, respectively. Following the organisation of the samples into training and testing samples, two versions of the MobileNet were utilized. Additionally, in the training set, data augmentation was also performed in three sizes. The details of utilized datasets are given in Table I.

Table I: The Image Distributions over Classes in Datasets

SMIDS		HuSHeM		SCIAN-Morpho	
Labels	#	Labels	#	Labels	#
Normal	1021	Normal	54	Normal	100
Abnormal	1005	Tapered	53	Tapered	228
Non-Sperm	974	Pyriform	57	Pyriform	76
		Amorphous	52	Small	72
				Amorphous	656
Total	3000	Total	216	Total	1132

B. Epoch Size and Data Augmentation

After the determination of the most successful parameters among the tested ones, the system will be subjected to epoch tests to increase the success rates and the data set will be tried to be expanded with data synthesis methods. In the augmentation step, we employed different size of data augmentation in terms of scaling, resizing, shifting, cropping etc. due to the necessity of large amount of data in deep learning based approaches. We carefully chose the ranges of the applied augmentation techniques in order to avoid disrupting the original normal/abnormal sperm shapes. Otherwise, normal sperm shapes might be detected as abnormal within correctly configured augmentation techniques. In Epoch size analyzing, we tested 10, 30 and 50 epoch sizes with the accuracy rates and processing times.

C. Mobile Platform Based CNNs

In this study, CNN models which can be utilized in the mobile platforms were evaluated in terms of the classification performance. In this respect, two different versions of MobileNet were tested. The main criteria in the evaluation is to define a model provide the highest accuracy with the minimum system resource and time complexity requirement. Both network configuration were set according to parameters presented in Table II. In this section, a brief information of two MobileNet versions will be given.

Table II: Parameter Configuration of the Utilized Networks

Parameter Name	Configuration
Batch Size	8
Activation Function	SoftPlus
Optimizer	Adamax
Loss Function	Categorical Crossentropy
Learning Rate	0.0001
Augmentation Sizes	1x (Original Data), 5x, 10x
Epoch Sizes	10, 30, 50

1) MobileNet V1 [13]: MobileNet is a more suitable model for mobile applications that do not have computing power. It uses in-depth removable evolutions with normal folds of the same depth compared to other networks. The number of parameters has been reduced with the evolution used. The model consists of 91 layers. The default input size for this model is 224x224.

2) MobileNet V2 [14]: Suitable for mobile devices or any device with low computing power, the complexity cost and model size of the network are significantly reduced. In MobileNetV2, a better module is introduced with the reverse residual structure. Non-linear states in narrow layers are eliminated this time. State-of-the-art performances for object detection and semantic segmentation are achieved with MobileNetV2 as backbone for feature extraction. It runs slower than MobileNetV1. The default input size for this model is 224x224.

III. Experimental Results

The performance of the utilized mobile platform based CNN models as Mobilenet V1 and V2 have been evaluated over three tests. The augmentation and epoch test results are presented in Table III and the processing times (learning times) are given in Table IV.

In the Data augmentation analysis, three versions of training sets have been created as 1x (Original Data), 5x and 10x augmented data sets. As a result of these tests, it was aimed to achieve higher performances despite the limited data available and to prevent the networks from memorizing the data. According to obtained results, augmentation of data provided the increments for all data sets in the classification accuracies. The most dramatic increment have been observed over HuSHeM and SCIAN-Morpho datasets for both version of MobileNet. Especially for HuSHeM dataset, data augmentation provided higher classification accuracies when using MobileNet V2. On the contrary, MobileNet V1 was measured more effective with data augmentation approach for SCIAN-Morpho Data set. In the classification of SMIDS, augmentation also increased the accuracy. But, this effect is more slightly due to having enough number of samples in SMIDS. The illustration of the data augmentation effect in the classification of all data sets is given in Figure 2.

Addition to augmentation analysis, the effect of the utilized different epoch sizes on the classification performance were also investigated. Higher epoch size requires more processing times and also might results an over-fitting problem. But, it might have a positive effect in the classification accuracy before reaching the over-fitting zone. On the other hand, lower epoch size might cause an early stopping problem which ends up the learning process before fully validation of the network. Therefore, a reasonable epoch size should be determined in the classification problems. In the sperm morphology analysis, we tested three epoch sizes as 10, 30, and 50. Similar to data augmentation effect, minimum effect was observed for SMIDS due to having enaugh sample size to reach fully validation period. The classification accuracies of HuSHeM and SCIAN-morpho increased more than SMIDS. But for MobileNetV2 in 50 Epoch analysis, a small decrements was also noted due to memorizing the data easier than MobileNet V1. The illustration of the epoch size effect in the classification of all data sets is given in Figure 2.

In the processing time analysis, the training procedure of MobileNet V1 was completed early stages than MobileNet V2

Table III: MobileNet V1 and V2 Performance Analysis in terms of Augmentation and Epoch Tests

Aug/Epo	**MobileNet V1**									**MobileNet V2**								
	HuSHeM			SMIDS			SCIAN			HuSHeM			SMIDS			SCIAN		
	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50	10	30	50
1x	45	54	59	84	84	84	57	58	58	38	44	39	84	85	84	48	54	59
5x	69	72	74	86	**88**	88	62	66	64	62	73	75	85	**87**	87	54	62	62
10x	71	74	**75**	85	86	87	66	65	**67**	72	**77**	77	86	86	87	65	**66**	66

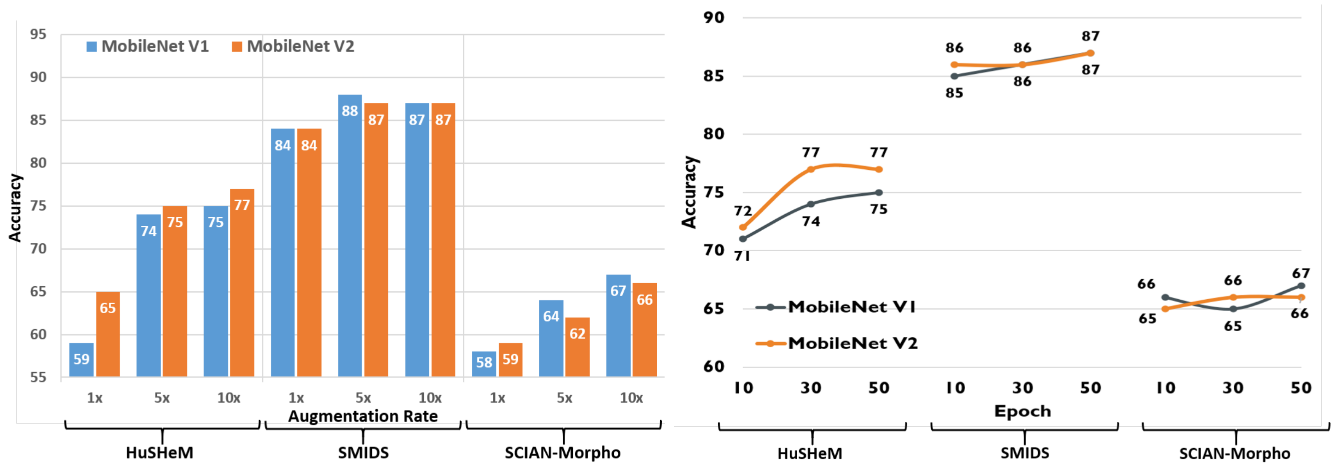

Figure 2: The Effects of the applied data augmentation techniques (left) and different epoch sizes (right) for each data sets

for small size datasets such as HuSHeM and SCIAN-morpho due to having less complex network architecture. But, for large data sets similar to SMIDS, the training process was finished early in MobileNet V2 because of including an optimization procedures in architecture. The training times are presented in Table IV. Only the training times of the cases that obtained highest accuracy scores in the classification listed in bold font type in Table III was presented in Table IV. Times represents the total training times over 5-folds.

Table IV: Training Times for best models

	MobileNet V1		MobileNet V2	
	Max. Acc.	Tr. Time	Max. Acc.	Tr. Time
HuSHeM	75	95	77	72
SMIDS	88	402	87	498
SCIAN-Morpho	67	499	66	378

IV. Conclusion

In this study, three sperm morphology data sets namely, SMIDS, HuSHeM and SCIAN-Morpho were tested on two mobile platform based CNNs, MobileNet V1 and V2. Additionally, in order to eliminate the data scarcity and imbalance problem of data sets, the augmentation versions are tested with the different epoch sizes. It has been observed that MobileNet V2 is more efficient CNN models for the classification of HuSHeM dataset while no significant difference were observed in the classification for SMIDS and SCIAN-Morpho datasets between V1. However, in terms of processing times and resource requirements, the training process of MobileNet V1 resulted in shorter times and consumed less system resources than MobileNet V2 for small size datasets such as HuSHeM and SCIAN-Morpho. For larger datasets such as SMIDS, MobileNet V2 ends up training in shorter times according to optimization layer utilized in architecture. Therefore, utilizing a classification schema for sperm morphology by using MobileNet V2 will be much more efficient way in terms of time aspect due to increasing data in daily life. In the data augmentation tests, the most dramatic effect was observed for the HuSHeM datasets because of having the least sample size. The minimum effect was obtained for SMIDS which has already enough sample size for the classification.

References

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

[2] A. M. Rivera-Montes, A. Rivera-Gallegos, E. Rodríguez-Villasana, A. Juárez-Bengoa, M. de los Ángeles Díaz-Pérez, and M. Hernández-Valencia, "Estimate of the variability in the evaluation of semen analysis," *Ginecologia y obstetricia de Mexico*, vol. 81, no. 11, pp. 639–644, 2013.

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

[4] Dictionary learning information. [Online]. Available: https://team.inria.fr/panama/projects/please/dictionary-learning/

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

[6] H. O. Ilhan, G. Serbes, and N. Aydin, "Dual tree complex wavelet transform based sperm abnormality classification," in *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2018, pp. 1–5.

[7] H. O. Ilhan, I. O. Sigirci, G. Serbes, and N. Aydin, "The effect of nonlinear wavelet transform based de-noising in sperm abnormality classification," in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2018, pp. 658–661.

[8] H. O. Ilhan, G. Serbes, and N. Aydin, "Automated sperm morphology analysis approach using a directional masking technique," *Computers in Biology and Medicine*, p. 103845, 2020.

[9] H. Ilhan, G. Serbes, and N. Aydin, "Automatic directional masking technique for better sperm morphology segmentation and classification analysis," *Electronics Letters*, vol. 55, no. 5, pp. 256–258, 2019.

[10] H. O. Ilhan, I. O. Sigirci, G. Serbes, and N. Aydin, "A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods," *Medical & Biological Engineering & Computing*, pp. 1–22, 2020.

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

[12] Keras info. [Online]. Available: https://keras.io/

[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.