Predicting Sperm Motility and Morphology using Deep Learning and Handcrafted Features

Steven Hicks[1,3], Pål Halvorsen[1,3], Trine B. Haugen[3], Jorunn M. Andersen[3], Oliwia Witczak[3], Hugo L. Hammer[3], Duc-Tien Dang-Nguyen[4],
Mathias Lux[2], Michael Riegler[1]
[1]SimulaMet, Norway; [2]Alpen-Adria-Universität Klagenfurt, Austria;
[3]Oslo Metropolitan University, Norway; [4]University of Bergen, Norway;

ABSTRACT

This paper presents the approach proposed by the organizer team (SimulaMet) for MediaEval 2019 Multimedia for Medicine: The Medico Task. The approach uses a data preparation method which is based on global features extracted from multiple frames within each video and then combines this with information about the patient in order to create a compressed representation of each video. The goal is to create a less hardware expensive data representation that still retains the temporal information of the video and related patient data. Overall, the results need some improvement before being a viable option for clinical use.

1 INTRODUCTION

In this paper, we detail the approach of the Medico Task organization team (SimulaMet) as part of MediaEval 2019. The Medico Task explores the challenge of using multimedia data to make the daily work of medical doctors more efficient [10]. This year's task focuses on automatically predicting sperm quality, in terms of motility and morphology, based on a microscopic video recordings of human semen. The task provides a dataset consisting of 85 videos and associated patient data, which will be used to make predictions on semen quality. The videos are taken from the VISEM dataset [4], which is an open-source dataset with barely any restrictions regarding usage. More details on the 2019 Medico Task and the provided dataset can be found in the overview paper [6] and the paper on VISEM [4].

The proposed approach is based on handcrafted features extracted form multiple frames within each video and associated patient/sensor data which is fed into a deep learning model for the prediction. This approach is used to solve the *prediction of motility task* and the *prediction of morphology task*, which are the tasks required in order to participate in the competition. The prediction of motiltiy task involves predicting three quality metrics (progressive, non-progressive, and immotile) tied to the movement of the sperm in a given semen sample. The prediction of morphology task focuses on predicting more visual features of the sperm, namely, predicting defects which may be present in the head, tail, or midpiece.

2 APPROACH

As the organization team, our main goal for this year's Medico Task was to present a baseline approach which utilize all the available data in the dataset to make a prediction on sperm motility and morphology (as the tasks require). Furthermore, we wanted our approach to be computationally efficient so to not require expensive hardware or a complex setup procedure. Therefore, we decided to base our contribution on handcrafted features which were extracted from multiple frames within each video in the provided dataset. These handcrafted features are combined with each category of associated patient data in order to train a deep learning model to make a prediction on each of the sperm quality features. In the following few section, we will describe the approach in more detail. This includes a description on how the data was prepared, information about the model architecture, and how each model was trained.

2.1 Data Preparation

To prepare the data for training and evaluation, we start by extracting the first two frames of each second for 60 seconds for each video. The result is a sub-sample of each video containing 120 frames from the first minute of each of the 85 videos in the dataset. From this step, we extract four different types of global features from each of the frames, namely, Edge Histogram, Tamura, Luminance Layout and Simple Color Histogram [9]. All image features were extracted using the LIRE library Lire [8], which is a popular open-source library for image retrieval. The global image features are concatenated row-wise to represent a single frame. The extracted image feature vectors are then concatenated column-wise into a 225×120 matrix, where the columns represent each frame, and the rows represent the extracted features.

The motivation behind this approach is that each column contains the visual features of a single frame, while the temporal information is represented through the change of feature values across the different columns. To add the information about the patient, we simply concatenate the values to each frame column so that each patient and sensor value will be the same for each frame in a given training sample. We create these video representations using each of the five provided categories of patient information values in the dataset, namely, sex hormones, patient-related data (age, body mass index (BMI), and days of sexual abstinence), fatty acids in serum, fatty acids in spermatozoa, and sperm analysis data. Each of the five video categories were used to train two deep neural networks, one for predicting morphology and one for predicting motility. Note that for the sperm analysis data, we removed the data regarding motility and morphology when used to predict these values in each respective task. This is important as to not feed the ground truth values to the network during training.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

2.2 Model Architecture and Training

As explained in the previous section, the data format used to train our deep neural networks is a matrix of extracted image features from 120 frames within each video and associated patient and sensor data. The exact input shape varies depending on the category of patient information used in the representation (number of variables vary depending on the category), but rest of the network is kept the same. We use a novel convolutional neural network architecture to perform our experiments.

The architecture itself is modeled to analyze the video representations using an ensemble of similar networks (network architecture shown in Figure 1). The network consists of an inital convolution block, after which the output gets passed through multiple residual modules. The output of each network is then concatenated before being passed through two fully-connected layers and then making the prediction. The network is repeated multiple times in order to create an ensemble of networks which each analyze the same input. The number of networks used is a parameter which may be tuned, from which we decided to use three based on some internal testing.

As previously mentioned, the same model architecture was used for both the motility tasks and the morphology task. The model was trained with a learning rate of 0.00001 using Nadam [3] to optimize the weights. We used a mean absolute error (MAE) to calculate loss on a batch size of 16 different samples. Overall, each experiment was trained for a maximum of 5000 epochs, but stopped training if the loss of the model did not improve over the last 300 epochs. Note that the model used for evaluation is the one which achieved the best MAE on the validation set. The hardware used to train all models was a desktop computer running Linux with an Nvidia GTX 1080TI, Intel core i7 processor running at 3.6 gigahertz, and 16 gigabytes of RAM. Models were implemented using the deep learning library Keras [2] with a TensorFlow [1] back-end. Due to the small number of training samples, despite training for many epochs, no experiment took longer than 1 hour to train.

3 RESULTS AND DISCUSSION

Looking at Table 1 and Table 2, we see the results for the Prediction of Morphology Task the Prediction of Motility Task. Overall, the results show that the deep neural networks for both tasks are able to learn something from the data, but the performance is overall quite poor when compared to the ZeroR baseline. For future work, we aim to use features extracted from a deep neural network.

4 CONCLUSION

In this paper, we described the approach submitted by the 2019 Medico organization team (SimulaMet). The presented method used handcrafted features and associated patient data to train a deep learning model to predict sperm quality in terms of motility and morphology. Based on the results, we see that the future of automatic sperm quality prediction is promising, but requires more work before being used in any real-world scenario. The way of representing the video into a single image also allows for future experiments where we want to use grad cam methods, e.g. [5, 7], to explain the important parts of the video and data.

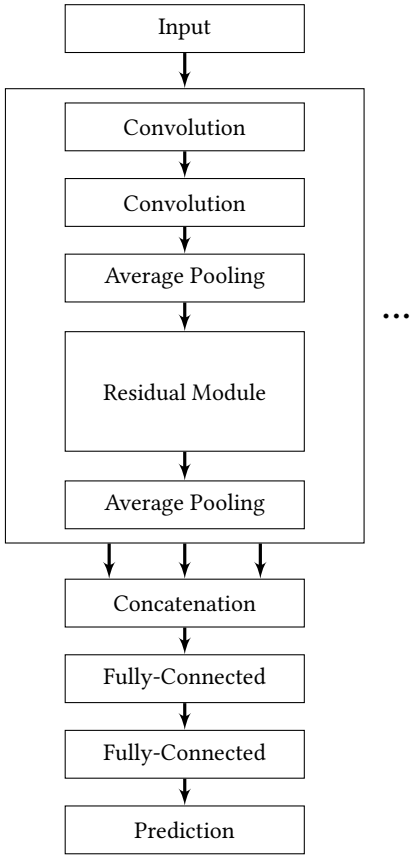

Figure 1: The CNN architecture used to perform the experiments for both the prediction of morphology task and the prediction of motility tasks. The box represents the network which is repeated.

Method	Fold 1		Fold 2		Fold 3		Average	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
None	13.46	17.89	12.39	15.18	13.34	17.34	13.06	16.80
Patient Data	12.72	17.52	12.96	16.94	13.15	16.73	12.94	17.06
Sex Hormones	12.32	16.91	12.14	15.89	12.53	15.67	12.33	16.15
FA Serum	12.76	17.32	10.53	12.78	11.50	15.10	11.60	15.07
FA Spermatozoa	12.04	17.78	10.53	13.12	12.36	15.42	11.64	15.44
Sperm Analysis	12.18	17.35	11.31	15.23	11.16	14.48	11.55	15.69
ZeroR Baseline	13.88	18.68	13.59	16.98	12.09	14.68	13.19	16.86

Table 1: The results of the experiments used to predict sperm morphology in terms of tail progressive, non-progressive, and immotile.

Method	Fold 1		Fold 2		Fold 3		Average	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
None	6.40	8.63	5.84	8.29	5.61	8.27	5.95	8.40
Patient Data	6.02	8.29	5.77	8.27	5.89	8.50	5.89	8.35
Sex Hormones	7.10	8.99	5.88	8.71	6.24	8.62	6.41	8.77
FA Serum	6.04	8.25	5.76	8.19	5.11	7.51	5.63	7.98
FA Spermatozoa	6.18	8.56	5.70	8.10	5.64	8.04	5.84	8.23
Sperm Analysis	6.26	8.42	5.81	8.38	5.96	8.37	6.01	8.39
ZeroR Baseline	5.99	7.95	5.99	8.27	5.82	8.13	5.93	8.12

Table 2: The results of the experiments used to predict sperm motility in terms of tail defects, midpiece defects, and head defects.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] François Chollet and others. 2015. Keras. https://keras.io. (2015).

[3] Timothy Dozat. 2015. Incorporating Nesterov Momentum into adam.

[4] Trine B. Haugen, Steven A. Hicks, Jorunn M. Andersen, Oliwia Witczak, Hugo L. Hammer, Rune Borgli, Pål Halvorsen, and Michael A. Riegler. 2019. VISEM: A Multimodal Video Dataset of Human Spermatozoa. In *Proceedings of the 10th ACM on Multimedia Systems Conference (MMSys'19)*. https://doi.org/10.1145/3304109.3325814

[5] Steven Hicks, Sigrun Eskeland, Mathias Lux, Thomas de Lange, Kristin Ranheim Randel, Mattis Jeppsson, Konstantin Pogorelov, Pål Halvorsen, and Michael Riegler. 2018. Mimir: An Automatic Reporting and Reasoning System for Deep Learning Based Analysis in the Medical Domain. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*. ACM, New York, NY, USA, 369–374. https://doi.org/10.1145/3204949.3208129

[6] Steven Hicks, Pål Halvorsen, Trine B Haugen, Jorunn M Andersen, Oliwia Witczak, Konstantin Pogorelov, Hugo L Hammer, Duc-Tien Dang-Nguyen, Mathias Lux, and Michael Riegler. 2019. Medico Multimedia Task at MediaEval 2019. In *CEUR Workshop Proceedings - Multimedia Benchmark Workshop (MediaEval)*.

[7] S. Hicks, M. Riegler, K. Pogorelov, K. V. Anonsen, T. de Lange, D. Johansen, M. Jeppsson, K. Ranheim Randel, S. Losada Eskeland, and P. Halvorsen. 2018. Dissecting Deep Neural Networks for Better Medical Image Classification and Classification Understanding. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. 363–368. https://doi.org/10.1109/CBMS.2018.00070

[8] Mathias Lux and Savvas A. Chatzichristofis. 2008. Lire: Lucene Image Retrieval: An Extensible Java CBIR Library. In *Proceedings of the 16th ACM International Conference on Multimedia (MM '08)*. ACM, New York, NY, USA, 1085–1088. https://doi.org/10.1145/1459359.1459577

[9] Mathias Lux, Michael Riegler, Pål Halvorsen, Konstantin Pogorelov, and Nektarios Anagnostopoulos. 2016. LIRE: open source visual information retrieval. In *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, 30.

[10] Michael Riegler, Mathias Lux, Carsten Griwodz, Concetto Spampinato, Thomas de Lange, Sigrun L Eskeland, Konstantin Pogorelov, Wallapak Tavanapong, Peter T Schmidt, Cathal Gurrin, and others. 2016. Multimedia and medicine: Teammates for better disease detection and survival. In *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 968–977.