MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Goh *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 20 (2024) 347-359

RESEARCH ARTICLE

Multimodal Convolutional Neural Networks for Sperm Motility and Concentration Predictions

Voon Hueh Goh[a], Muhammad Asraf Mansor[a], Muhammad Amir As'ari[a,b]*, Lukman Hakim Ismail[a]

[a]Department of Biomedical Engineering and Health Sciences, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia; [b]Sport Innovation and Technology Centre (SITC), Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

Abstract Semen analysis is an important analysis for male infertility primary investigation and manual semen analysis is a conventional method to assess it. Manual semen analysis has been revealed with accuracy and precision limitations due to noncompliance to guidelines and procedures. Sperm motility and concentration are the main indicators for pregnancy and conception rate hence they were selected for parameters prediction. Convolutional neural network (CNN) has benefited computer vision application industry in recent years and has been widely applied in computer vision research tasks. In this paper, three-dimensional CNN (3DCNN) was designed to extract motion and temporal features, which are vital for sperm motility prediction. For sperm concentration, since two-dimensional CNN (2DCNN) is efficient in recognizing and extracting spatial features, well-established Residual Network (ResNet) architecture was adopted and customized for sperm concentration prediction. Multimodal learning approach is a technique to aggregate learnt features from different deep learning architecture that adopted other forms of modalities, which could provide deep learning model with better insights on their tasks. Hence, a multimodal learning deep learning architecture was designed to receive both image-based (frames extracted from video samples) and video-based (stacked frames pre-processed from video samples) input that could provide well-extracted spatial and temporal features for sperm parameters prediction. The results obtained using the proposed methodology have surpassed other similar research works who used deep learning approach. For sperm motility, its best achieved average mean absolute error (MAE) was 8.048, and sperm concentration obtained a competent Pearson's correlation coefficient (R_P) value of 0.853.

Keywords: Sperm parameters prediction, Semen analysis, 3DCNN, ResNet18, Multimodal learning.

***For correspondence:**
amir-asari@utm.my

Received: 2 Nov. 2023
Accepted: 4 Feb. 2024

Introduction

Infertility, is a medical condition where sexually active and non-contracepting couples are unable to successfully achieve clinical pregnancy, as defined by WHO [1]. Infertility could happen in both men and women, however, half of the failure in childbearing was contributed by infertile men [2]. Semen analysis is one of the primary and important analysis required to study the probability of male causing infertility among an infertile couple, then only treatment planning options are available for conception. Generally, the parameters that will be included in the analysis are sperm concentration, total sperm count, sperm motility, sperm morphology, semen volume, semen viscosity, pH values of semen sample and sperm vitality [3]–[6]. Sperm motility and concentration are significantly related to pregnancy rate and time to pregnancy; hence they act as better conception predictor than other parameters [3], [7], [8]. Sperm movement is due to the flagellar beating of sperm tail, and the movement can be categorized and graded as progressively motile, non-progressively motile and immotile spermatozoa [3], [9]. Sperm concentration is defined as the total number of spermatozoa per unit volume [3]. Human evaluation on

sperm motility and concentration are subjective and might be over or underestimated [10], [11]. Mostly semen analysis is performed manually by experienced laboratory technicians comply to guidelines provided by WHO [3]. It is a time and human resource consuming process as it requires intensive training and regular participation in quality assurance programs. Besides, it has been revealed with limitations such as lack of standardization in methodologies and tools used for analysis which affected the result's accuracy and precision [12]–[14].

This stimulated the development of automated semen analyzer such as CASA (Computer Aided Semen Analyzer) few decades ago and a recent model named as SQA (Sperm Quality Analyzer) [15]. CASA working principle is based on image processing algorithms on the captured microscopic images to detect motile and immotile spermatozoa [16], while SQA is based on electro-optical signals generated by moving spermatozoa, the detection of these signals is then interpreted by proprietary algorithms for sperm concentration prediction [16]–[18]. The expected advantages of automated semen analyzer are uniform standardization, faster, higher precision, lesser human error and training resources to operate the system compared with human evaluation method. However, these devices' credibility's on accuracy and precision were questioned based on the correlation of predicted results with manual measurement. Besides, the algorithms have issues dealing with semen samples that contains lots of debris and round cells, which confuse the algorithm to correctly identify target and non-target entity, hence affecting the accuracy [17], [19]. From several comparison studies of both models using manual evaluation as gold standard, generally SQA series presents closer results with manual evaluation than CASA, however, there are also contradict findings from different research stated that sperm count and motility assessment using SQA models were poorer and not comparable with manual methods except sperm count results of samples with high concentration groups [13]–[16], [18]–[21]. Other than that, these automated semen analyzers still required some training to operate the system despite being identified as automated system. Although CASA and SQA are available in the market, CASA is currently not recommended for routine clinical use due to its methodological procedures that received criticisms, and SQA proper guidelines was not being discussed in WHO manual.

Artificial intelligence and deep learning have taken off rapidly in recent decades, which video analysis has also achieved remarkable achievements such as activity recognition [22] and video classification [23]. For sperm motility prediction, Thambawita *et al*. introduced a novel method that was which is the combination of autoencoders and CNN for sperm morphology classification and sperm motility prediction. The general idea is to pretrain an autoencoder which its decoder will form an image with embedded learnt features, then pass it through pretrained ResNet34 for final motility and morphology prediction [24]. A similar work by the same author is by using ResNet34 that adopted video-based input, obtained by stacking an RGB frame, followed by 2 dense optical flow frames generated with stride 1 and 10 respectively [25]. Both approaches experimented by Thambawita *et al*. achieved MAE of 9.427 and 8.825 respectively [24], [25]. In Hicks *et al*.'s work, the authors presented several machines and deep learning algorithms they have experimented for motility prediction. The best method is by using pretrained ResNet50 that adopt image-based input, which is dense optical flow frames extracted with stride 1 and this approach achieved an average MAE of 8.740 [10]. Nevertheless, the approaches explored by both authors were using 2DCNN classification architectures which could be less efficient in extracting temporal information from either image-based or video-based input. Rosenblad *et al*. Battacharjee *et al*. have proposed using 3DCNN in their work to capture temporal information [26], [27]. Rosenblad *et al*. established a consecutive 3DCNN architecture and preprocessed the input into 15 consecutive grayscale frames stacked as video-based modality. It has achieved motility prediction with an MAE of 8.83 [26]. On the other hand, Battacharjee *et al*. incorporated ResNet18 into 3DCNN architecture forming a ResNet18-3DCNN architecture, which aim to classify a 3D video-based input into one of the motility classes (progressive, non-progressive, immotile). The input was generated by stacking 50 consecutive grayscale frames extracted from video sample [27]. Although the work proposed has achieved 100% accuracy on motility classification, the work does not predict the proportion/percentage of each motility classes of a sample but classify entire sample into one class, which the output might be less useful for further evaluation by the clinician.

In previous studies that have been discussed afore, those research works focused on using deep learning algorithms for sperm motility prediction. This following work presented an approach using a classical ANN architecture to predict sperm concentration [28]. The input was based on a full absorption spectrum obtained using a UV-visible spectrophotometer, where the light absorption values were used to quantify sperm concentration. The information obtained was then fed into a simple ANN architecture with 711 input variables as the first layer, 12 neurons in the first hidden layer, 20 neurons in the second hidden layer, and followed by the last output variable (711:12:20:1). It was reported to achieve 93% accuracy with clinical measurements using a manual approach. The equation defined by the author was the percentage ratio between predicted concentrations and manual evaluation values

(100 × prediction/manual value) [28]. If a sample's predicted concentration exceeds clinical evaluation would have achieved an accuracy of more than 100%, this would contribute to controversial accuracy evaluation. All in all, since the modality adopted in this paper was vision input (video-based and image-based modalities extracted from VISEM's video samples), 2DCNN is more suitable for concentration prediction than ANN as demonstrated by Lesani *et al*.'s work.

The aforementioned approaches for parameters predictions were in unimodal learning approach [24]–[26], [28]. Nevertheless, aside from using either video-based or image-based data as the main modalities for unimodal learning, Hicks *et al*. and Bhattacharjee *et al*. have also demonstrated multimodal learning techniques in their research work by feeding in additional tabular data from VISEM dataset [10], [27]. Hicks *et al*. utilized multimodal learning concept in deep learning architecture where the input types were image-based input and tabular data, it was reported to have achieved MAE of 9.132 [10]. From the results, multimodal learning did not show improvement as hypothesized where the accuracy should improve by providing additional modalities when compared with the unimodal learning approach, but they degrade instead. In Bhattacharjee *et al*.'s study where 3DCNN architecture was used for motility classification, it achieved 100% accuracy in both validation and test sets for unimodal learning approach, while multimodal learning approach test sets results decreased to 88.89% [27]. From the reported results, both research works' findings showed that not only their proposed architecture and modalities could not improve the performance by learning the association between both modalities, but by providing additional modalities the performance degrades. This could give a hint that tabular data is not a suitable modality to provide better insights for motility prediction. Nonetheless, in this paper, a multimodal network would be designed to adopt suitable multimodalities. First was 3D input (video-based) that well represented temporal and movement information, the second type was 2D input (image-based) which could help generate feature maps to classify targets (sperm) from other non-targets. In short, multimodal learning is a plausible approach to improve the model's insights however the types of modalities should be considered wisely.

In this paper, it aimed to formulate a multimodal deep learning methods in sperm parameters prediction by adopting image-based and video-based input, which the model's accuracy would be compared with other related research works. The rest of the paper is structured as follows. Section 2 discussed the methodologies to preprocess the suitable modalities and architecture design. Section 3 presented the experimental settings. Section 4 showed the results and discussion with a descriptive analysis on the accuracies achieved. Finally, the conclusion and suggestions for future improvements were proposed in Section 5.

Methods

Dataset

The recorded semen videos are obtained from an online multimodal video dataset, VISEM [29] which contains different data sources such as videos, biological analysis data, and participant data that are collected from 85 anonymized participants. Semen samples collection and handling method are according to WHO guidelines as described by Andersen *et al*. [30]. Generally, semen samples collected were analyzed within two hours and evaluation approach are as defined in WHO manual as well. Videos were recorded using Olympus CX31 phase contrast microscope, with heated stage at 37°C, and a mounted camera (UEye UI-2210C, IDS Imaging Development System). Semen videos were captured using 400× magnification with a frame rate of 50 fps.

Preprocessing and Modalities Preparation

The 3DCNN was foreseen to capture temporal information effectively, given its modality is in video-based form. To assist the model differentiates better whether the motions captured were from sperm or non-sperm entities, a classification 2DCNN (ResNet18) could learn and provide insights which functioned as a classifier. Besides it can also be used for sperm concentration regression as the temporal feature was not necessary but need only spatial information. To sum it up, a multimodal network would acquire final feature maps learnt from 3DCNN and ResNet18 as inputs, then underwent a late fusion mechanism for motility and concentration prediction tasks. Hence, suitable video-based and image-based modalities were generated, both modalities were denoted as D1 and D2 respectively.

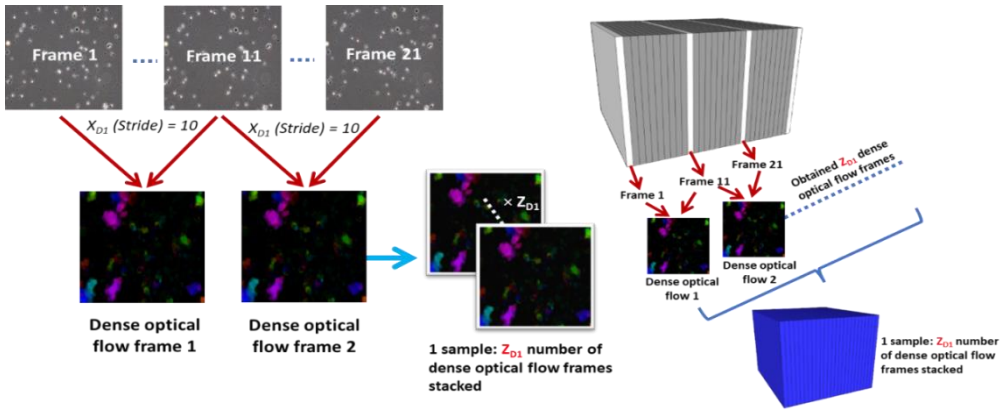

Figure 1. Generating dense optical flow frames from extracted images and stacked them to form video-based input

D1 Input Generation (Video-based modality)

D1 video-based input was prepared for unimodal 3DCNN architecture in the form of stacked dense optical flow frames. One dense optical flow frame was generated by computing the pixel intensity changes between 2 consecutive frames extracted from microscopic video, using Gunnar Farneback's Optical Flow algorithm (tools available in OpenCV), the intensity changes denotes motion occurs. The general idea is to generate N number of dense optical flow frames for each video. Then, as inspired by Thambawita *et al*., dense optical flow frames generation with stride 1, (e.g. extract 1st and 2nd microscopic video's frame to generate one dense optical flow frame) would provide better results than generating dense optical flow frames using stride 10 (e.g. extract 1st and 11th microscopic video's frame to generate one dense optical flow frame) or other stride numbers [25]. This process could be summarized in Figure 1. Next, a Z number of dense optical flow frames generated were stacked together, after resizing them to a dimension of (3 × 144 × 144), forming a single training/validation sample with a dimension of (3 × 144 × 144 × Z).

D2 Input Generation (Image-based modality)

D2 image-based input was prepared for unimodal 2DCNN (ResNet) in the form of RGB images. The D2 sample was the first image of the microscopic frame sequences used for D1 sample generation. For example, if a D1 sample with configuration is as follows: X_{D1} = 10 and Z_{D1} = 8; then RGB images extracted as D2 samples would be 1st frame, 81st frame, 161st frame and the list went on with a skipping factor or stride of X_{D1} × Z_{D1}, in this context was 10 × 8. This process was summarized in Figure 2. The default input size for ResNet18 architecture was 224 × 224, but to preserve the information embedded within an extracted image, all D2 images were downsized with a ratio of 0.8. The original image size was (3 × 640 × 480) and after downsizing was (3 × 512 × 384).

Different Sets with Varying Stride and Depth

Table 1 shows the types of sets which would be trained with varying configuration X_{D1} and Z_{D1}, then identify which configuration type would provide the best validation accuracy. The total samples generated were kept constant for all sets to accurately analyse and deduce the effect of varying D1 input's depth and stride number on the performance.

Ground Truth Scaling

For motility prediction, the ground truth was the percentage of each spermatozoa category within each sample, hence the summation of all categories (progressive, non-progressive, immotile) would be equal to 100%. The output algorithm used for motility prediction was Softmax where its output within the range of 0 to 1. Therefore, the data scaling for motility percentages ground truth was dividing them by 100. While for concentration values (×10^6/mL), its data distribution characteristics were very much similar to prices and not in linear distribution. It was observed in price prediction-related regression tasks, the ground truth would undergo log transformation to minimize skewness to be closer to normal distribution [31], [32]. Thus, log transformation was utilized as scaling approach to reduce the ground truth skewness.

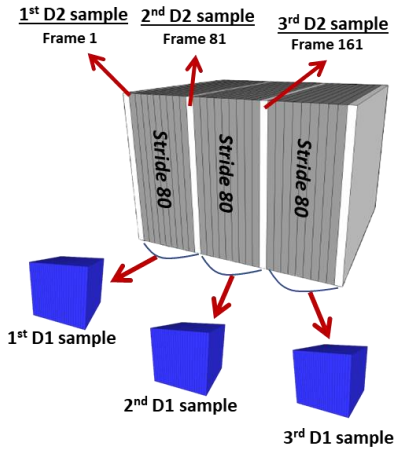

Figure 2. Selection of D2 sample from frame sequences used to generate D1 sample

Table 1. Different Configuration in Dataset Generation

Set	X_{D1} (Stride Number)	Z_{D1} (Depth)	N (Total Frames per Video)	Total Samples Generated (D1, D2) $*N/Z_{D1} \times 85$	D1 Dimension
A	1	8	272	2890	3 × 144 × 144 × 8
B	1	11	374	2890	3 × 144 × 144 × 11
C	1	14	476	2890	3 × 144 × 144 × 14
D	10	8	272	2890	3 × 144 × 144 × 8
E	10	11	374	2890	3 × 144 × 144 × 11
F	10	14	476	2890	3 × 144 × 144 × 14

Deep Learning Architectures

The following subsections will briefly describe the structure of unimodal (for early stage pretraining) and multimodal deep learning architectures. The general idea was to pretrain 3DCNN and ResNet18 for motility and concentration prediction respectively. To acquire ResNet18's learnt features for sperm and non-sperm entities classification, a multimodal network was designed to assemble partially pretrained 3DCNN and ResNet18 to finetune the motility prediction performance. As explained earlier, stacked dense optical flow frames (D1 input) only capture motion hence it needs assistance to better identify if the motion was from moving sperm, other cells or just a drifting effect in liquid samples. Hence, this induced the idea of a multimodal network that combines both 3DCNN and 2DCNN to improve motility prediction. Finally, a multimodal network with finalized finetuned model parts and parameters were assembled for multiple-output regression, that is obtaining motility and concentration prediction with one run.

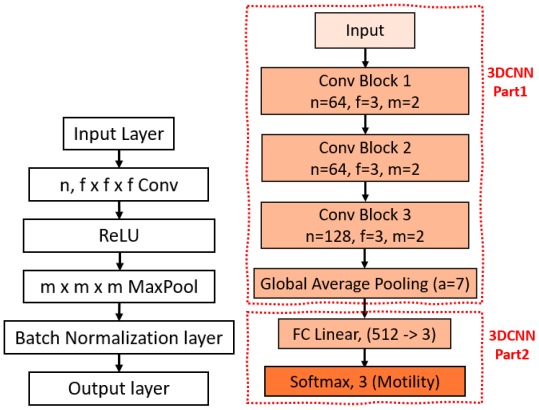

Figure 3. 3DCNN Structure

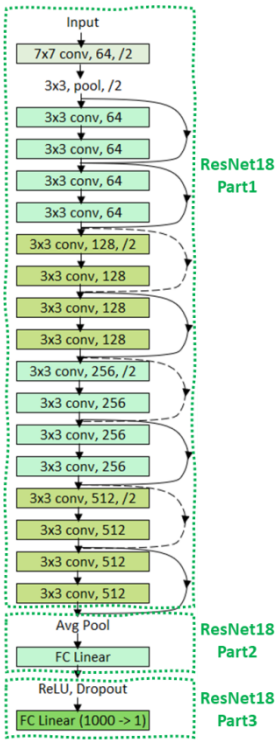

Figure 4. Modified ResNet18 for Concentration Prediction

3DCNN

This 3DCNN model was designed for sperm motility prediction which consists of several basic convolutional building blocks and complete architectures were summarized in Figure 3. The complete architecture was divided into Part1 and Part2, which was to assist the implementation and configuration of the following multimodal network design that requires parameters/weight freezing.

ResNet18

The original ResNet18 architecture was preserved, and then a fully connected layer for concentration regression was added, as shown in Figure 4. Similar to 3DCNN, ResNet18 was separated into three parts for assembly process in the late development stage.

Multimodal Network with Single Output (Motility)

The complete architecture design of multimodal network for motility prediction which combines pretrained 3DCNN and ResNet18 is as shown in Figure 5(a). Pretrained weights of 3DCNN model Part1 was imported to finetune the motility prediction performance. ResNet18 Part1's weights were frozen and not allowed for backpropagation during the training process, as the parameters were later shared with concentration prediction tasks (Section 2.3.4). ResNet18 Part2 (fully connected layer with 1000 outputs) was proceeded for training to allow partial parameters tuning for best motility accuracy. The output from both model parts were then concatenated forming a fully connected layer with a total of 1512 units, then slim down to 1000 neurons before the final motility output. Before concatenation, the outputs of both model parts were subjected to batch normalization. As suggested by several research works, batch normalization should be inserted between linear transformation and non-linear unit when two outputs of different modalities were to be joined together in one similar embedding [33], [34].

Multimodal Network with Multiple Output (Concentration & Motility)

The complete architecture design of a multimodal network for concentration and motility prediction was presented in Figure 5(b). It utilized the best-learned parameters from ResNet18 and Multimodal Network with Single Output for concentration and motility regression tasks respectively. This model did not require training and hence all weights were not subjected to back propagation, just simply model assembly from the best-learned weights.

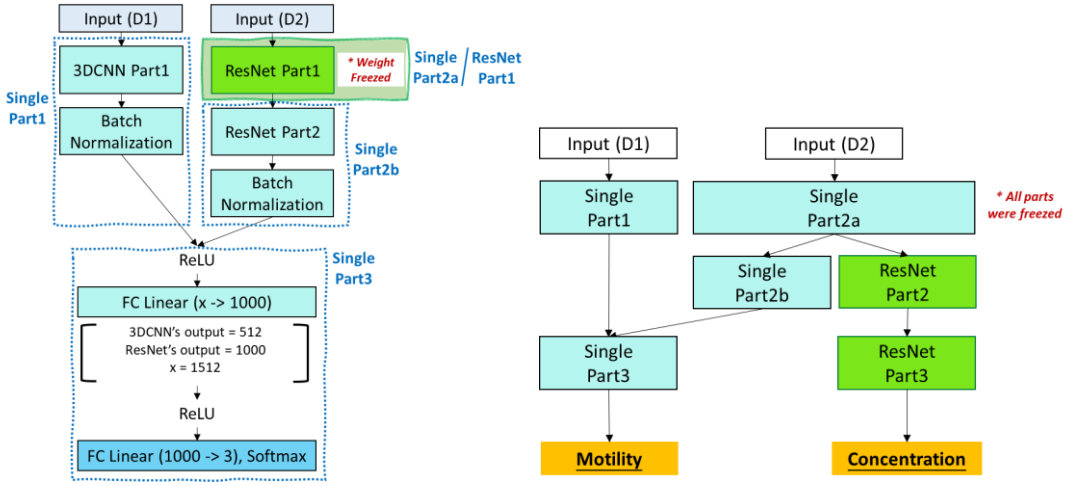

Figure 5. The proposed (a) Multimodal Network with Single Output (Motility). (b) Multimodal Multimodal Network with Multiple Outputs (Motility & Concentration)

Table 2. Hyperparameters for Implementation Setup

Model Type	Initial Learning Rate	Training Epochs	Loss Function
ResNet18	3×10^{-4}	100	RMSE
3DCNN	1×10^{-6}	Stopped if not improved after 10 epochs	MSE
Multimodal Network with Single Output	3×10^{-5}	40	MSE

Experimental Settings

All models used Adam optimization algorithm with different initial learning rate as shown in Table 2, and same learning rate scheduler "StepLR" which decay with gamma 0.95 every 10 epochs. Batch size for all models were standardized as 30. Unimodal 3DCNN would stopped training if the validation accuracy did not improve after 10 epochs. Whereas the other models were trained with maximum epochs as shown in Table 2 and the epoch checkpoint that achieved lowest validation error were saved. For models (Unimodal 3DCNN, Multimodal Network with Single Output) that tackled motility prediction, the loss function adopted was mean squared error (MSE) as suggested by other similar research works [10], [24]–[27], whereas the loss function of ResNet18 for concentration prediction was root mean squared error (RMSE), which also has been practiced by Lesani *et al*. [28] in his research work for same parameter prediction. All sets prepared in subsection 2.2.3 would undergo three-fold cross validation as practiced by other research works. MAE was selected as the evaluation metric for sperm motility prediction and to allow the performance comparison with existing research works that used the same VISEM dataset [10], [24]–[27], [35]. In addition, MAE was also used as one of the evaluation metrics for concentration regression to compare the performance amongst different sets generated in this research study. For sperm concentration prediction, since there were limited research works that explored the deep learning approach, Pearson's correlation coefficient (R_P) and Spearman rank correlation (R_S) were chosen as metrics to compare with other related studies. Both performance metrics were used in the comparison studies of sperm concentration prediction performances of automated semen analyser (CASA, SQA series) with manual semen analysis [18], [19], [36].

Results and Discussion

Concentration Prediction (Unimodal ResNet18)

Table 3 represents the performance of the proposed method (modified ResNet18 for concentration prediction) compared with the commercialized products which were CASA and SQA-Vision in terms of R_P and R_S. Overall, the proposed approach was comparable with the commercialized products as most of the sets achieved correlation values above 0.8, which was a strong linear relation between predicted concentration and actual concentration.

Table 3. Comparison of Concentration Prediction between Previous Research Works and Proposed Modified ResNet18

Previous research works / Datasets Type	Performance metrics	
	R_P	R_S
[36]	0.970 (CASA)	-
[18]	0.958 (SQA-Vision)	0.978 (SQA-Vision)
[19]	-	0.950 (CASA, SQA-V-Gold)
A (X_{D2} = 8)	0.843	0.821
B (X_{D2} = 11)	0.825	0.828
C (X_{D2} = 14)	0.835	0.805
D (X_{D2} = 80)	0.807	0.822
E (X_{D2} = 110)	0.796	0.833
F (X_{D2} = 140)	0.853	0.831

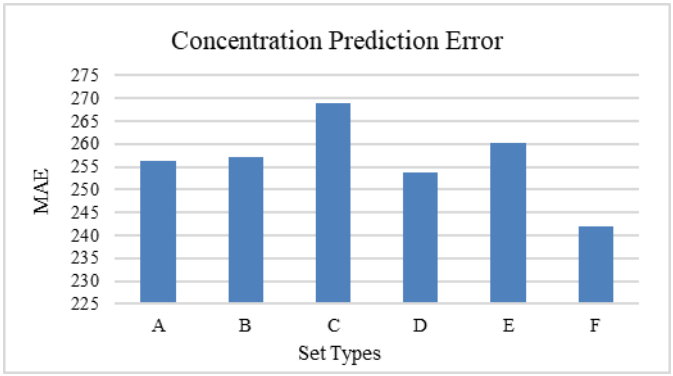

Figure 6. Concentration Prediction Error (MAE) of Different Set

By using R_P, it was observed Set F had the highest RP correlation value, which was 0.853, while Set E was the lowest with 0.796. On the other hand, Set E had topped the others with the highest RS correlation value of 0.833 while Set C achieved the lowest RS value which was 0.805. From Figure 6, it was observed that generally Set A to C had higher MAE than Set D to F. This observation reflected the changes in the number of skipping frames used to extract the D2 input, X_{D2}. X_{D2} denoted the time interval between 2 consecutive frames which were extracted as D2 samples within a video, this ranged from 0.16 seconds to 2.8 seconds corresponding to Set A to F. ResNet18 was supposed to learn spatial and target cells' features that were embedded within image-based input. If the dispersion of target cells of was observed in a similar distribution in all the frames extracted from one video sample due to small X_{D2} value (Set A to C), then it could not be an effective dataset for deep learning model to learn as the model could not regularize well enough to predict the test/validation set. Hence, sets generated with larger X_{D2} value regularized better, therefore lower MAE and better accuracy.

Motility Prediction (Unimodal 3DCNN)

Table 4 presented the MAE of motility prediction by the proposed method on the prepared sets from subsection 2.2.3 and the set achieved lowest MAE was Set C with MAE of 8.506. Results showed that by using only unimodal learning approach, the combination of 3DCNN and video-based stacked dense optical flow frames has already surpassed most of the similar research works that employed deep learning model and unimodal image-based or video-based modality. It showed that Set A to C have already performed better than all other previous studies. This indicated that the proposed method (3DCNN + D1 input) for motility prediction were effective on temporal features learning. On the other hand, Set D to F performed slightly weaker than Hicks et. al.'s work [10] which used dense optical flow frames as their image-based input, but still surpassed Thambawita et. al.'s work (autoencoders + image-based input) [24], Thambawita et. al. (ResNet34 + video-based input) [25] and Rosenblad et. al. (3DCNN + stacked grayscale frames as video-based input) [26].

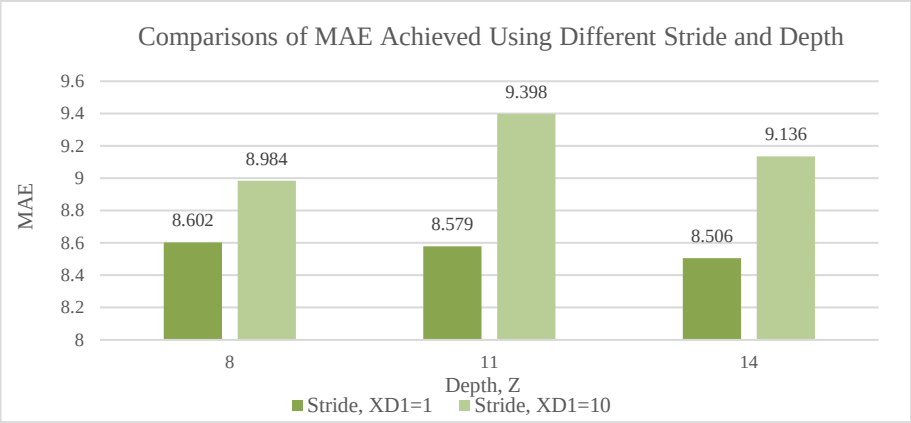

Figure 7. Comparison of MAE Achieved Using Different Stride and Depth

Table 4. Comparison of Motility Prediction between Previous Research Works and Proposed 3DCNN Unimodal Learning Approach

Previous research works / Dataset type	Modality	MAE			
		PR	NPR	IM	Average
[7]	Feature vectors	13.220	7.260	11.920	10.800
[7]	Image-based (dense optical flow frames)	10.191	7.114	8.914	8.740
[21]	Image-based	-	-	-	9.427
[22]	Video-based	-	-	-	8.825
[23]	Video-based	10.160	7.410	8.920	8.830
Set A (X_{D1}=1, Z_{D1}=8)	Video-based	10.322	6.919	8.567	8.602
Set B (X_{D1}=1, Z_{D1}=11)		10.254	7.021	8.463	8.579
Set C (X_{D1}=1, Z_{D1}=14)		9.974	6.959	8.583	8.506
Set D (X_{D1}=10, Z_{D1}=8)		10.620	7.075	9.257	8.984
Set E (X_{D1}=10, Z_{D1}=11)		11.187	7.310	9.697	9.398
Set F (X_{D1}=10, Z_{D1}=14)		10.780	7.150	9.479	9.136

The reason that indirectly caused Set D to F to have a slightly higher error was due to the stride number, X_{D1} used to generate dense optical flow frames. This is due to similar concept as explained in subsection 4.2 on concentration prediction. The higher error trend indicated that precision of moving targets' location has better impact on achieving higher accuracy than having longer duration of information (Z_{D2}). As shown in Figure 7, a trend where MAE decreased as the depth, Z_{D1} increased was observed in Set A to C (X_{D1} = 1). This indicated that with deeper depth in video-based modality, D1 data provided more information for the deep learning model to learn, hence better results than shallower depth. For sets generated with X_{D1} = 10, due to the precision of moving targets' location as addressed afore, the impact of depth on motility prediction is ignored as similar trend was not observed.

Table 5. Comparison of Motility Prediction between Proposed Multimodal Learning Architecture (3DCNN + ResNet18) with Other Previous Works

Previous research works / Dataset type	Modality	MAE			
		PR	NPR	IM	Average
[7]	Feature vectors	13.220	7.260	11.920	10.800
[7]	Frame-based	10.191	7.114	8.914	8.740
[21]	Frame-based	-	-	-	9.427
[22]	Video-based	-	-	-	8.825
[23]	Video-based	10.160	7.410	8.920	8.830
Set A	Video-based	9.365	7.038	7.740	8.048
Set B		10.025	6.721	7.899	8.215
Set C		10.403	6.701	7.997	8.367
Set D		10.588	7.268	8.670	8.842
Set E		11.323	7.513	9.379	9.405
Set F		11.108	7.201	9.220	9.176

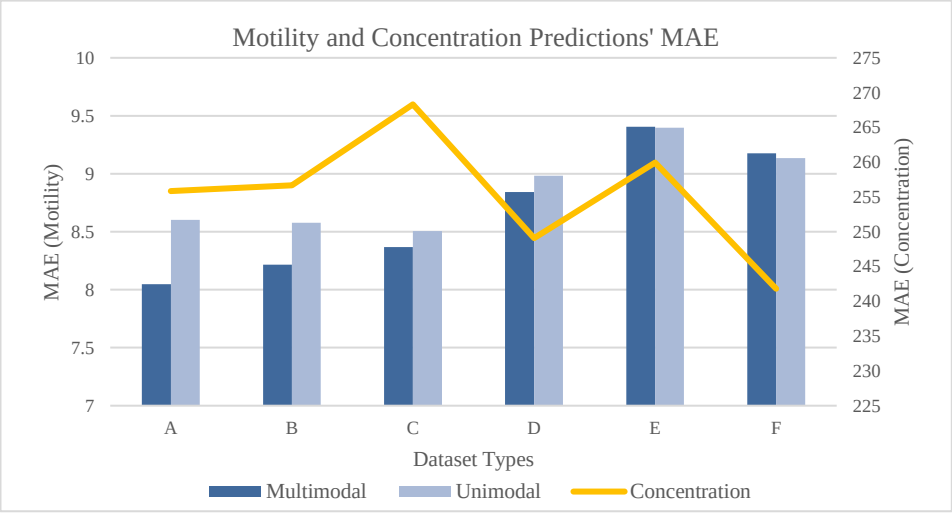

Figure 8. Motility and Concentration Predictions' MAE

Motility and Concentration Prediction Using by Multimodal Learning Approach (3DCNN + ResNet18)

Table 5 represents the performance of proposed multimodal network for motility prediction (3DCNN + ResNet18) compared with other similar research works. It showed that using the multimodal learning approach with appropriate modalities which are stacked dense optical flow frames as video-based modality, and RGB image-based modality, it could reduce the error by 6.450% at maximum when compared with unimodal learning approach, which was achieved by Set A with MAE of 8.048. From Set A to C, the error decrement trend for motility prediction was in reversed with unimodal learning approach (Figure 7), where Set A was the one with the lowest MAE of 8.048 and Set C with a relatively higher MAE of 8.367, despite the latter had deeper depth and theoretically should have lower MAE. During multimodal training, the features learnt from ResNet18 which was pretrained for concentration prediction was imported and it affected the motility prediction.

Table 6. Validation of Results Obtained by Final Assembled Model (Multimodal Network with Multiple Outputs)

Dataset Types	MAE (Motility)	R_P (Concentration)		
	Multimodal, Single output	Multimodal, Multiple output	ResNet18	Multimodal, Multiple output
A	8.048	8.065	0.843	0.843
B	8.215	8.379	0.825	0.825
C	8.367	8.260	0.835	0.835
D	8.842	8.791	0.807	0.807
E	9.405	9.352	0.796	0.796
F	9.176	9.141	0.853	0.853

From Figure 8, the motility prediction error decrement trend was in sync with concentration prediction error decrement trend as denoted by the line graph, where the error decreased from Set C to Set A. Therefore, instead of adding the depth of video-based modality to improve the accuracy, the spatial features learnt from ResNet18 assisted better and had a bigger impact than the depth of video-based modality on motility prediction. This observation also indirectly proved that a multimodal learning approach with a suitable modality would improve the model's insights on their learning tasks.

Validation of Final Assembled Model (Multimodal Network with Multiple Outputs)

Since this model was just merely combining the model parts and finetuned parameters which means it did not require any backpropagation, hence this section only evaluate whether this proposed architecture obtained the same best-optimized accuracy as previous pretrained model parts (ResNet18 and Multimodal Network with Single Output). From the outcome acquired by Set A to F as shown in Table 6, it showed both concentration and motility prediction of this assembled model was on par with the best accuracy achieved by ResNet18 and Multimodal Network Single Output, the model assembly process was successful.

Conclusion and Future Recommendations

In this paper, a multimodal deep learning architecture consist of 3DCNN and ResNet18 which can run the regression for sperm parameters prediction has been successfully developed, specifically the motility and concentration prediction. Video-based and image-based input were generated as the modalities for this multimodal learning network. The video-based input was in the form of stacked dense optical flow frames that accentuates motion features, whereas image-based input which were extracted from microscopic videos act as the main modality for concentration prediction. The features learnt from ResNet18 was then imported to finetune multimodal learning network for motility prediction as it could help the model to identify target and non-target entities. The results obtained from these proposed architectures were compared and analysed with other similar research works. For motility prediction, the multimodal network was proven to have surpassed all other studies that use the deep learning approach with the lowest error, which is MAE of 8.048, whereas for concentration prediction it achieved a comparable performance with commercialized products, with Pearson's correlation coefficients of 0.853.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

The authors would like to express their gratitude to Universiti Teknologi Malaysia (UTM) and Ministry of Higher Education Malaysia for supporting this research.

Funding Statement

The authors would like to thank the Ministry of Higher Education under Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UTM/02/1) for funding this research.

References

[1] World Health Organization. (2020). Infertility. Sep. 2020.
[2] M. C. Inhorn and P. Patrizio. (2014). Infertility around the globe: New thinking on gender, reproductive technologies and global movements in the 21st century. *Hum. Reprod. Update*, *21*(4), 411-426. Doi: 10.1093/humupd/dmv016.
[3] World Health. (2010). Examination and processing of human semen.*World Health*, *10*, 286.
[4] N. Punjani *et al.* (2023). Changes in semen analysis over time: A temporal trend analysis of 20 years of subfertile non-azoospermic men. *World J. Mens. Health*, *41*(2), 382. Doi: 10.5534/wjmh.210201.
[5] F. Xianchun, F. Jun, D. Zhijun, and H. Mingyun. (2023). Effects of ureaplasma urealyticum infection on semen quality and sperm morphology. *Front. Endocrinol. (Lausanne).*, 14. Doi: 10.3389/fendo.2023.1113130.
[6] J. Huang, H. Chen, N. Li, and Y. Zhao. (2023). Emerging microfluidic technologies for sperm sorting. *Eng. Regen.*, *4*(2): 161-169. Doi: 10.1016/j.engreg.2023.02.001.
[7] B. Ducot, A. Spira, D. Feneux, and P. Jouannet. (1988). Male factors and the likelihood of pregnancy in infertile couples. 11. Study of clinical characteristics — practical consequences*. J. Androl.*, *11*(5), 395-404. Doi: 10.1111/j.1365-2605.1988.tb01012.x.
[8] K. P. Nallella, R. K. Sharma, N. Aziz, and A. Agarwal. (2006). Significance of sperm characteristics in the evaluation of male infertility. *Fertil. Steril.*, *85*(3), 629-634. Doi: 10.1016/j.fertnstert.2005.08.024.
[9] N. Kumar and A. Singh. (2015). Trends of male factor infertility, an important cause of infertility: A review of literature. *J. Hum. Reprod. Sci.*, *8*(4), 191-196. Doi: 10.4103/0974-1208.170370.
[10] S. A. Hicks *et al.* (2019). Machine learning-based analysis of sperm videos and participant data for male fertility prediction. *Sci. Rep.*, 9(1), 1-10. Doi: 10.1038/s41598-019-53217-y.
[11] T. G. Cooper *et al.* (2009). World Health Organization reference values for human semen characteristics. *Hum. Reprod. Update*, *16*(3), 231-245. Doi: 10.1093/humupd/dmp048.
[12] J. Auger *et al.* (2000). Intra- and inter-individual variability in human sperm concentration, motility and vitality assessment during a workshop involving ten laboratories. *Hum. Reprod.*, *15*(11), 2360-2368. Doi: 10.1093/humrep/15.11.2360.
[13] J. Lammers, S. Chtourou, A. Reignier, S. Loubersac, P. Barrière, and T. Fréour. (2021). Comparison of two automated sperm analyzers using 2 different detection methods versus manual semen assessment. *J. Gynecol. Obstet. Hum. Reprod.*, *50*(8). Doi: 10.1016/j.jogoh.2021.102084.
[14] M. J. Tomlinson. (2016). Uncertainty of measurement and clinical value of semen analysis: has standardisation through professional guidelines helped or hindered progress? *Andrology*, *4*(5), 763-770. Doi: 10.1111/andr.12209.
[15] A. Agarwal and R. K. Sharma. (2007). Automation is the key to standardized semen analysis using the automated SQA-V sperm quality analyzer. *Fertil. Steril.*, *87*(1), 156-162. Doi: 10.1016/j.fertnstert.2006.05.083.
[16] J. F. Moruzzi, A. J. Wyrobek, B. H. Mayall, and B. L. Gledhill. (1988). Quantification and classification of human sperm morphology by computer-assisted image analysis. *Fertil. Steril.*, *50*(1), 142-152. Doi: 10.1016/s0015-0282(16)60022-5.
[17] S. T. Mortimer, G. Van Der Horst, and D. Mortimer. (2015). The future of computer-aided sperm analysis. *Asian J. Androl.*, *17*(4), 545-553x. Doi: 10.4103/1008-682X.154312.
[18] K. M. Engel, S. Grunewald, J. Schiller, and U. Paasch. (2019). Automated semen analysis by SQA Vision ® versus the manual approach—A prospective double-blind study. *Andrologia*, *51*(1), 1-10. Doi: 10.1111/and.13149.
[19] J. Lammers, C. Splingart, P. Barrière, M. Jean, and T. Fréour. (2014). Double-blind prospective study comparing two automated sperm analyzers versus manual semen assessment. *J. Assist. Reprod. Genet.*, 31(1), 35-43. Doi: 10.1007/s10815-013-0139-2.
[20] O. M. Yis. (2020). Comparison of fully automatic analyzer and manual measurement methods in sperm analysis and clinical affect. *Exp. Biomed. Res.*, *34*: 224-230. Doi: 10.30714/j-ebr.2020463605.
[21] T. G. Cooper and C. H. Yeung. (2006). Computer-aided evaluation of assessment of 'grade a' spermatozoa by experienced technicians. *Fertil. Steril.*, *85*(1), 220-224. Doi: 10.1016/j.fertnstert.2005.07.1286.
[22] S. Ji, W. Xu, M. Yang, and K. Yu. (2013). 3D Convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, *35*(1), 221-231. Doi: 10.1109/TPAMI.2012.59.
[23] D. Brezeale and D. J. Cook. (2008). Automatic video classification: A survey of the literature. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, 38(3): 416-430. Doi: 10.1109/TSMCC.2008.919173.
[24] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, and T. B. Haugen. (2019). Extracting temporal features into a spatial domain using autoencoders for sperm video analysis. *arXiv*, 2670, 3-5.
[25] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, and T. B. Haugen. (2019). Stacked dense optical flows and dropout layers to predict sperm motility and morphology. *arXiv*, *10*, 9-11.
[26] J. M. Rosenblad, S. Hicks, H. K. Stensland, T. B. Haugen, P. Halvorsen, and M. Riegler. (2019). Using 2D and 3D convolutional neural networks to predict semen quality. *CEUR Workshop Proc.*, *2670*, 27-29,
[27] Priyansi, B. Bhattacharjee, and J. H. Rahim. (2021). Predicting semen motility using three-dimensional convolutional neural networks, 1-8.

[28] A. Lesani *et al.*, "Quantification of human sperm concentration using machine learning-based spectrophotometry," *Comput. Biol. Med.*, vol. 127, no. August, p. 104061, 2020, doi: 10.1016/j.compbiomed.2020.104061.
[29] "Simula Visem." .
[30] S. Hicks *et al.*, "Predicting sperm motility and morphology using deep learning and handcrafted features," *CEUR Workshop Proc.*, vol. 2670, no. October, pp. 27–29, 2019.
[31] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, vol. 2017-Decem, pp. 319–323, 2018, doi: 10.1109/IEEM.2017.8289904.
[32] S. Lessmann and S. Voß, "Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy," *Int. J. Forecast.*, vol. 33, no. 4, pp. 864–877, 2017, doi: 10.1016/j.ijforecast.2017.04.003.
[33] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "CNN-based sensor fusion techniques for multimodal human activity recognition," *Proc. - Int. Symp. Wearable Comput. ISWC*, vol. Part F1305, pp. 158–165, 2017, doi: 10.1145/3123021.3123046.
[34] K. Wang, M. Bansal, and J. M. Frahm, "Retweet wars: Tweet popularity prediction via dynamic multimodal regression," *Proc. - 2018 IEEE Winter Conf. Appl. Comput. Vision, WACV 2018*, vol. 2018-Janua, pp. 1842–1851, 2018, doi: 10.1109/WACV.2018.00204.
[35] T. B. Haugen *et al.*, "VISEM: A multimodal video dataset of human spermatozoa," *Proc. 10th ACM Multimed. Syst. Conf. MMSys 2019*, pp. 261–266, Jun. 2019, doi: 10.1145/3304109.3325814.
[36] M. J. Tomlinson *et al.*, "Validation of a novel computer-assisted sperm analysis (CASA) system using multitarget-tracking algorithms," *Fertil. Steril.*, vol. 93, no. 6, pp. 1911–1920, 2010, doi: 10.1016/j.fertnstert.2008.12.064.