OPEN

Machine Learning-Based Analysis of Sperm Videos and Participant Data for Male Fertility Prediction

Steven A. Hicks[1,2]*, Jorunn M. Andersen[3,5], Oliwia Witczak[3,5], Vajira Thambawita [1,2], Pål Halvorsen[1,2], Hugo L. Hammer[1,2], Trine B. Haugen[3,6] & Michael A. Riegler[1,4,6]

Methods for automatic analysis of clinical data are usually targeted towards a specific modality and do not make use of all relevant data available. In the field of male human reproduction, clinical and biological data are not used to its fullest potential. Manual evaluation of a semen sample using a microscope is time-consuming and requires extensive training. Furthermore, the validity of manual semen analysis has been questioned due to limited reproducibility, and often high inter-personnel variation. The existing computer-aided sperm analyzer systems are not recommended for routine clinical use due to methodological challenges caused by the consistency of the semen sample. Thus, there is a need for an improved methodology. We use modern and classical machine learning techniques together with a dataset consisting of 85 videos of human semen samples and related participant data to automatically predict sperm motility. Used techniques include simple linear regression and more sophisticated methods using convolutional neural networks. Our results indicate that sperm motility prediction based on deep learning using sperm motility videos is rapid to perform and consistent. Adding participant data did not improve the algorithms performance. In conclusion, machine learning-based automatic analysis may become a valuable tool in male infertility investigation and research.

Automatic analysis of clinical data may open new avenues in medicine, though often limited to one modality, usually images[1]. Recently, however, trends have shifted to include data from other modalities, including sensor data and participant data[2,3]. Furthermore, advancements in artificial intelligence, specifically deep learning, have shown its potential in becoming an essential tool for health professionals through its promising results on numerous use-cases[1,4–6].

Male reproduction is a medical field that is gaining increased attention due to several studies indicating a global decline in semen quality during the last decades[7,8] as well as geographical differences[9]. Semen analysis is a central part of infertility investigation, but the clinical value in predicting male fertility is uncertain[10]. Standard semen analysis should be performed according to the recommendations made by the WHO, which includes methods of assessing semen volume, sperm concentration, total sperm count, sperm motility, sperm morphology, and sperm vitality[11]. Sperm motility is categorized into the percentage of progressive, non-progressive, and immotile spermatozoa. Sperm morphology is classified according to the presence of head defects, neck and midpiece defects, principal piece (main part of the tail) defects, and excess residual cytoplasm in a stained preparation of cells. Figure 1 shows an example of a frame extracted from a video of a wet human semen sample. The WHO has established reference ranges for various semen parameters based on the semen quality of fertile men whose partners had a time to pregnancy up to and including 12 months[12]. However, these ranges can not be used to distinguish fertile from infertile men. Manual semen analysis requires trained laboratory personnel, and even when performed in agreement with the WHOs guidelines, it may be prone to high intra- and inter-laboratory variability.

Attempts to develop automatic systems for semen analysis have been carried out for several decades[13]. CASA was introduced during the 1980s after the digitization of images made it possible to analyze images using a computer. A more rapid and objective assessment of sperm concentration and sperm motility was expected by using CASA, but it has been challenging to obtain accurate and reproducible results[13]. The results may be unreliable

[1]Holistic Systems Department, Simula Metropolitan Center for Digital Engineering, Oslo, Norway. [2]Faculty of Technology, Art and Design, OsloMet – Oslo Metropolitan University, Oslo, Norway. [3]Faculty of Health Sciences, OsloMet – Oslo Metropolitan University, Oslo, Norway. [4]Department of Technology, Kristiania University College, Oslo, Norway. [5]These authors contributed equally: Jorunn M. Andersen and Oliwia Witczak. [6]These authors jointly supervised this work: Trine B. Haugen and Michael A. Riegler. *email: steven@simula.no
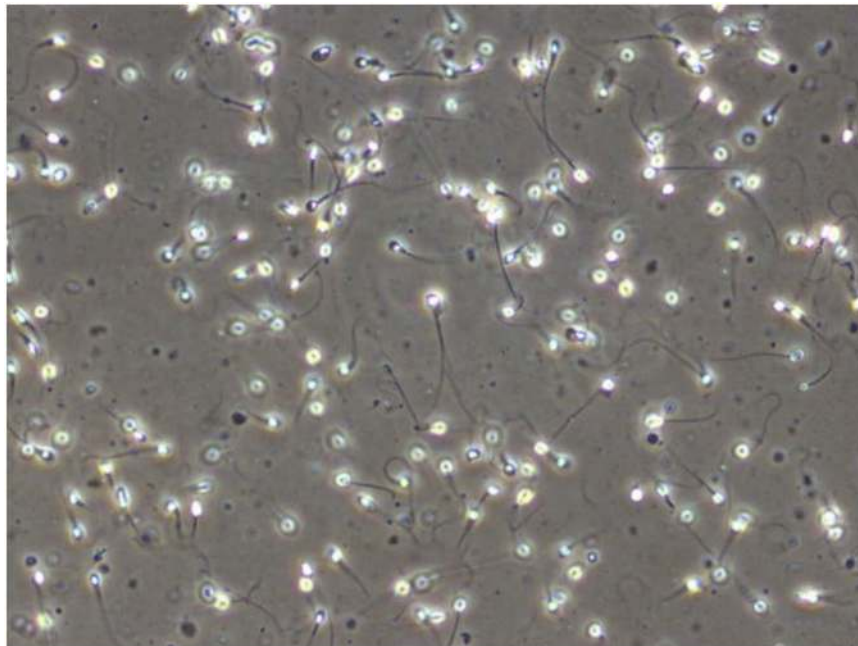
Figure 1. Frame from a microscopic video of a human semen sample showing several spermatozoa (Olympus CX31 phase contrast microscope with heated stage, UEye UI-2210C camera, 400x magnification).

due to particles and other cells than spermatozoa in the sample as well as the occurrence of sperm collisions and crossing sperm trajectories. Better results are obtained when analyzing spermatozoa separated from seminal plasma and re-suspended in a medium. CASA was also developed for assessment of sperm morphology and DNA fragmentation in the sperm. It is claimed that new models can also assess vitality and that some functional tests of a semen sample are possible[13]. However, the assessments require special staining or preparation procedures. Despite its long history as a digitized sperm analyzer, CASA is not recommended for clinical use[11,13]. The technology, however, has been improved, and it has been suggested that using CASA for sperm counting and motility assessment can be a useful tool with less analytical variance than the manual methods[14,15].

Concerning automatic semen analysis in general, Urbano et al.[16] present a fully automated multi-sperm tracking algorithm, which can track hundreds of individual spermatozoa simultaneously. Additionally, it is also able to measure motility parameters over time with minimal operator intervention. The method works by applying a modified version of the jpdaf to microscopic semen recordings, allowing them to track individual spermatozoa at proximities and during head collisions (a common issue with existing CASA instruments). The main contribution made by Urbano et al. is the modified jpdaf algorithm for tracking individual spermatozoa, but by only evaluating the proposed approach on two samples, the generalizability of the method to a larger population is difficult to determine.

Dewan et al.[17] present a similar method, tracking spermatozoa by generating trajectories of the cells across microscopic video sequences. Similar to CASA, object proposals are generated through a greyscale edge detection algorithm, which is then tracked to generate object trajectories. These trajectories are then classified into "sperm" or "non-sperm" entities using a CNN, of which the "sperm" entities are used to estimate three quality measurements for motility (progressive, non-progressive, and immotile), and the concentration of spermatozoa per unit volume of semen. The results seem promising but since the method was evaluated on a closed dataset, it is not possible to directly compare this approach with other methods.

Although not the focus in our work, another essential attribute for semen quality is measuring the number of abnormal spermatozoa present in a semen sample. Ghasemian et al.[18] tried to detect abnormal spermatozoa by individually classifying human spermatozoa into normal or abnormal groups. Shaker et al.[19] did a similar study to predict sperm heads as normal or abnormal by splitting images of sperm heads into square patches and using them as training data for a dictionary-based classifier. A common theme is that all automatic approaches, for both motility and morphology assessment, focus on one modality and do not incorporate other data into the analysis. Additionally, the evaluation is performed on a rather limited or closed data which hinders reproducibility and comparability of the results. In the presented work, we aim to contribute to the field of automated semen analysis in the following three ways: (i) to develop a rapid and consistent method for analyzing sperm motility automatically, (ii) to explore the potential of multimodal analysis methods combining video data with participant data to improve the results of the automatic analysis, and (iii) to compare different methods for predicting sperm motility using algorithms based on deep learning and classical machine learning.

To the best our knowledge, no study has been performed on how deep learning and multimodal data analysis may be used to directly analyze semen recordings in combination with participant/patient data for the automated prediction of motility parameters. Using data from 85 participants and three-fold cross-validation, we observe that the initial results are promising. Thus, machine learning-based automatic analysis may become a valuable tool for the future of male infertility investigation.

Methods

Experimental design.   Our main approach is the use of CNNs to analyze sequences of frames from video recordings of human semen under a microscope to predict sperm motility in terms of progressive, non-progressive, and immotile spermatozoa. The video recordings are then combined with participant data to see how it may improve our methods using the multiple modalities available in our dataset. As there are no related works for which to compare directly, we first trained a series of machine learning algorithms to set a baseline for how well we can expect our deep learning-based algorithms to perform.

The presentation of our methods is divided into three parts. Firstly, we provide a description of the dataset used for both training and evaluation of the presented methods and the statistical analysis. Secondly, we detail how we trained and evaluated the methods based on classical machine learning algorithms. Lastly, we describe our primary approach of using deep learning-based algorithms to predict sperm motility in terms of progressive, non-progressive, and immotile spermatozoa. All experiments were performed following the relevant guidelines and regulations of the Regional Committee for Medical and Health Research Ethics - South East Norway, and the General Data Protection Regulation(GDPR).

Dataset.   For all experiments, we used videos and several variables from the VISEM-dataset[20] [https://data-sets.simula.no/visem/], a fully open and multimodal dataset with anonymized data and videos of semen samples from 85 different participants. In addition to the videos, the selected variables for the analysis included manual assessment of sperm concentration and sperm motility for each semen sample and participant data. Participant data consisted of age, BMI, and days of sexual abstinence. In the experiments, the videos and participant data were used as independent variables whereas the sperm motility values (percentage of progressive, non-progressive sperm motility, and immotile spermatozoa) were used as the dependent variables. We also performed an additional experiment to test the effect of sperm concentration if added as an independent variable to the analysis.

Details on the collection and handling of semen samples have previously been described by Andersen *et al.*[21]. Briefly, the semen samples were collected at a room near the laboratory or at home and handled according to the WHO guidelines[11]. Samples collected at home, were transported close to the body to avoid cooling and analyzed within two hours. Assessment of sperm concentration and sperm motility was performed as described in the WHO 2010 manual[11]. Sperm motility was evaluated using videos of the semen sample, and all samples were assessed by one experienced laboratory technician. $10 \, \mu l$ of semen were placed on a glass slide, covered with a 22×22 mm cover slip and placed under the microscope. Videos were recorded using an Olympus CX31 microscope with phase contrast optics, heated stage (37°C), and a microscope mounted camera (UEye UI-2210C, IDS Imaging Development Systems, Germany). Videos for sperm motility assessment were captured using 400× magnification and stored as AVI files. The recordings vary in length between two to seven minutes with a frame rate of 50 frames-per-second.

Statistical analysis.   For all experiments, we report the MAE calculated over three-fold cross-validation to get a more robust and generalizable evaluation. Furthermore, statistical significance was tested by a corrected paired t-test, where a p-value below or equal to 0.05 was considered significant. Usually, t-test is based on the assumption that samples are independent. However, samples in the folds of cross-validation are not independent. Therefore, a fudge factor is needed to compensate for the not independent samples[22]. The significance test showed that all results with an average MAE below 11 are significant improvements compared to the ZeroR baseline. For ZeroR, which is also commonly known as the null model, the cross-validation coefficient is defined with a $Q2$ value of 0. This means that the ZeroR predictions are equal to the average calculated over the entire training dataset.

Baseline machine learning approach.   For the machine learning baseline, we relied on a combination of well-known algorithms and handcrafted features. To extract features from the video frames, we used the open-source library Lucene Image Retrieval (LIRE)[23]. LIRE is a Java library that offers a simple way to retrieve images and photos based on color and texture characteristics. We tested all available features (more than 30 different ones) with all machine learning algorithms (more than 40 different ones), but in this work, we only report the features that worked best with our machine learning algorithms, which were the Tamura features. Tamura features (coarseness, contrast, directionality, line-likeness, regularity, and roughness) are based on human visual perception, which makes them very important in image representation. Using the Tamura image features, participant data and a combination of both, we trained different algorithms to perform prediction on the motility variables. We performed a total of three experiments per tested algorithm; one using only Tamura features, one using only participant data, and one combining the Tamura features with the participant data through early fusion.

Since the Tamura features are sparse compared to deep features, we used a slightly different approach for selecting frames from the videos. Each video was represented by a feature vector containing the Tamura features of two frames per second (the first and the middle frame) for the first 60 seconds. In total, we had 120 frames per video and a visual feature space consisting of 2160 feature points. These features were then used to train multiple machine learning algorithms using the WEKA machine learning library[24]. We conducted experiments with all available algorithms, but report only the six best performing ones. The reported algorithms are Simple Linear Regression, Random Forests, Gaussian Process, Sequential Minimal Optimization Regression (SMOreg), Elastic Net, and Random Trees. One limitation of these algorithms is that they are only able to predict one value at a time, meaning we had to run them once for each of the three sperm motility variables.

Deep learning approach.   For our primary approach, we use methods based on CNNs to perform regression on the three motility variables. For each deep learning-based experiment, we extracted 250 frame samples (single frames or frame sequences) from each of the 85 videos of our dataset. The reason for only extracting 250 frames
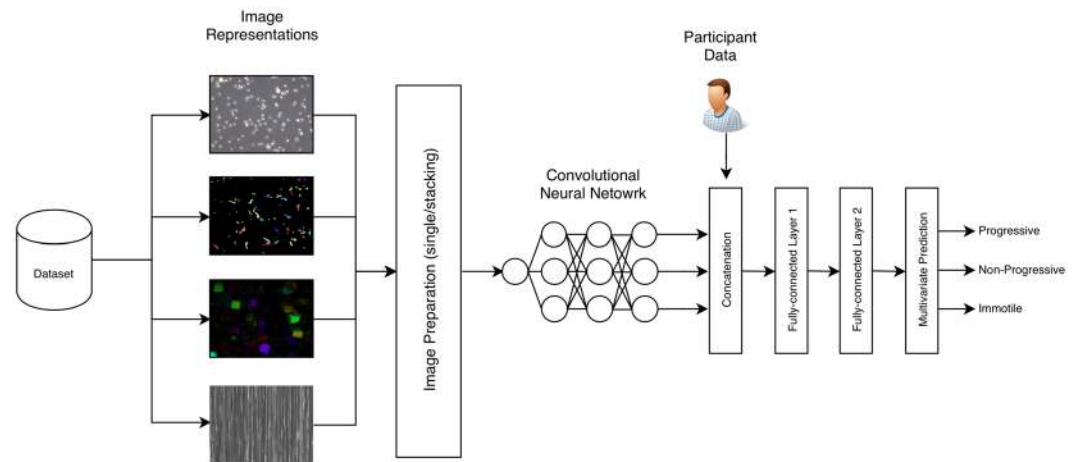
Figure 2. The deep learning pipeline used for all multimodal neural network-based experiments. Starting with our dataset, we extract frame data into four different representations. These four different "images" are sent to the image preparation were we either pass a single image or stacked images to a convolutional neural network (CNN). The CNN is trained to learn a model that captures the spatial or spatial and temporal combined features of sperm motility. This is based on the image representation and preparation (stacking or single frame). The output of the CNN model is then combined with the participant data. This combined vector is passed through two fully-connected layers before performing multivariate prediction on the three motility variables.

per video was due to some videos being too short for collecting more than 250 sequences of 30 frames, which is about 7,500 frames equalling about 2 minutes of video at 50 frames-per-second. This results in a total of 21,250 frames used for training and validation. As we are evaluating each method using three-fold cross-validation, the split between the training and validation datasets is 14,166 and 7,083 frame samples, respectively.

Our deep learning approaches can be split into three groups. Firstly, we analyze raw frames as they are extracted from the videos. The analysis is done by looking at the raw pixel values from a single or a sequence of frames and using these to make a prediction. Secondly, we use optical flow to generate temporal representations of frame sequences to condense the information of the temporal dimension into a single image. The advantages of this representation is that it can model the temporal dependencies in the videos, and it is able to alleviate the hardware costs of analyzing raw frame sequences using CNNs. Lastly, we combine the two previous methods to exploit the advantages of both, by using the visual features of raw video frames together with the temporal information of the optical flow representations.

The baseline for the deep learning approaches are the machine learning algorithms as described above and ZeroR. For each experiment, we predict the percentage of progressive spermatozoa, non-progressive spermatozoa, and immotile spermatozoa for a single semen sample. In contrast to the classical machine learning algorithms, neural networks can predict all three values at once. Figure 2 illustrates a high level overview of the complete deep learning analysis pipeline.

All deep learning-based models were trained using mse to calculate loss and Nadam[25] to optimize the weights. The Nadam optimizer had a learning rate of 0.002, β_1 value of 0.900, and β_2 value of 0.999. We trained each model for as long as it improved with a patience value of 20 epochs, meaning if the mse did not improve on the validation set for 20 epochs, we stopped the training to avoid overfitting. The model used for evaluation was the one which performed best on the validation set, not the model from the last epoch. Furthermore, for each method we trained two models. One model uses only frame data, and the other uses a combination of the frame data and the related participant data (BMI, age, and days of sexual abstinence). To include the participant data in the analysis, we first pass a frame sample through the CNN. Then, we take the output of the last convolutional layer and globally average pool it to produce a one-dimensional feature vector which is concatenated with the participant data. This combined vector is then passed through two fully-connected layers consisting of 2,048 neurons each before being making the final prediction (shown in Fig. 2). In the following few sections, we will describe six different methods used to predict sperm motility; a method using single frames for prediction, a method which stacks frames channel-wise, a method using vertical frame matrices, a method based on sparse optical flow, a method based on dense optical flow, and a method based on two-stream networks.

Single frame prediction. For the single frame-based method, we extracted 250 single frames from each video and used this to train various CNNs models based on popular neural network architectures (such as DenseNet[26], ResNet[27], and Inception[28]). We experimented using transfer learning from the ImageNet[29] weights included with the Keras[30] implementations of the different CNN architectures and found that, in general, using these weights as a base for further training worked better than training from scratch. Note that we did not fine-tune the models, meaning we did not freeze any layers during training. We only report the model which performed best, which in our case was a ResNet-50 model implemented in Keras with a TensorFlow[31] back-end. The frames were resized to 224×224 before being passed through the model, which is the recommended size for the ResNet-based architectures[27].
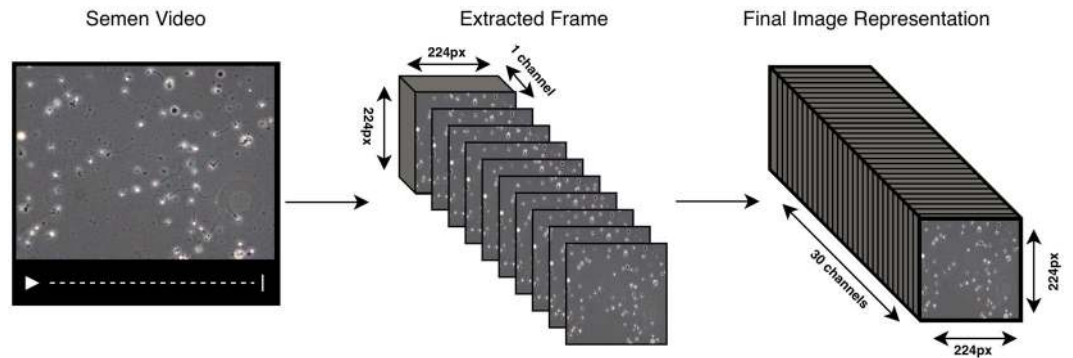
Figure 3. An illustration of how frames are stacked channel-wise after being greyscaled. From a video, a sequence of n frames are extracted and greyscaled. These frames are then stacked channel-wise, meaning each frame occupied one channel-dimension of the final image. The final stacked "image" is then of shape $224 \times 224 \times 30$.

The single frame-based approach is simple and comes with some obvious limitations. Most notably, we lose the temporal information present within the video. Losing the temporal information may be acceptable when measuring attributes that rely on visual clues, such as morphology, but for motility the change over time is an important feature.

Greyscale frame stacking. The Greyscale Frame Stacking method is an extension of the single-frame prediction approach. Here, we extract 250 batches of 30 frames and greyscaled them before stacking them channel-wise (shown in Fig. 3). This results in 21,250 frame samples with a shape of $224 \times 224 \times 30$, which contains the information of 30 consecutive frames. The reasons for greyscaling the frames before stacking them is two-fold. Firstly, seeing as the color of the videos are a feature of the microscope and lab preparation, and not the spermatozoon itself, we assume that this feature may confuse the model in unintended ways. Secondly, greyscaling the frames reduces the size of each frame by three, making stacking 30 frames feasible on less powerful hardware. The motivation behind this approach was to keep the temporal information present in a given frame sequence, yet still, keep the size of the input relatively small.

These extracted frame sequences were used to train a ResNet-50 model implemented in Keras[30]. Note that because we changed the size of the channel dimension, we could not perform transfer learning as we did in the previous method. Apart from this, the model was trained in the same manner as described in the beginning of the Deep Learning Approach section.

Vertical frame matrix. To create the vertical frame matrix, 250 batches of 30 frames were extracted and greyscaled. Each frame was resized to 64×64 before being flattened into a one-dimensional vector. The reason for resizing each frame was to keep the length of the flattened images relatively short. With a size of 64×64, the final vector had a length of 4096. Each vector was then stacked on top of each other which resulted in a matrix with a shape of $30 \times 4096 \times 1$. Examples images using this transformation can be seen in row four of Fig. 4. Similar to the Greyscale Frame Stacking approach described in the previous section, we condense the information of multiple frames into a single image, which we can then pass through a standard two-dimensional CNN. Due to size constraints, the model used for this method was ResNet-18. Otherwise, it was trained in the same way as the previous two methods.

Sparse optical flow. For the Sparse Optical Flow approach, we use Lucas-Kanade's[32] algorithm of estimating optical flow. What makes sparse optical flow "sparse," is that we only measure the difference between a few tracked features from one frame to another. In our case, we use Harris and Stephens corner detection algorithm[33] to detect individual sperm heads (implemented in OpenCV[34] as "goodFeaturesToTrack"). Then, we track the progression of each spermatozoon using Lucas-Kanade's algorithm over a sequence of 30 frames. Similar to the previous methods, sequences were sampled at evenly spaced intervals to maximize differences between optical flow representations. We used a CNN model based on the ResNet-50 architecture implemented in Keras and trained using the same configuration described previously. Examples for the sparse optical flow image representation can be seen in row two of Fig. 4.

Dense optical flow. The Dense Optical Flow approach generates optical flow representations using Gunner Farneback's algorithm[35,36] for two-frame motion estimation. Dense optical flow, in contrast to sparse optical flow, processes all pixels of a given image instead of a few tracked features. For this method, we tried two configurations. The first configuration measures the difference between two consecutive frames. The second configuration adds a stride of 10 frames between selected frame samples. This is done to increase the measured difference between frame comparisons. We collected 250 dense optical flow images and trained one model for each of the two configurations to evaluate the result of this method. For both stride configurations, we train each model using the same architecture (ResNet-50) and training configuration as for the other deep learning methods. Examples for the created image representations using the dense optical flow can be seen in row three of Fig. 4.
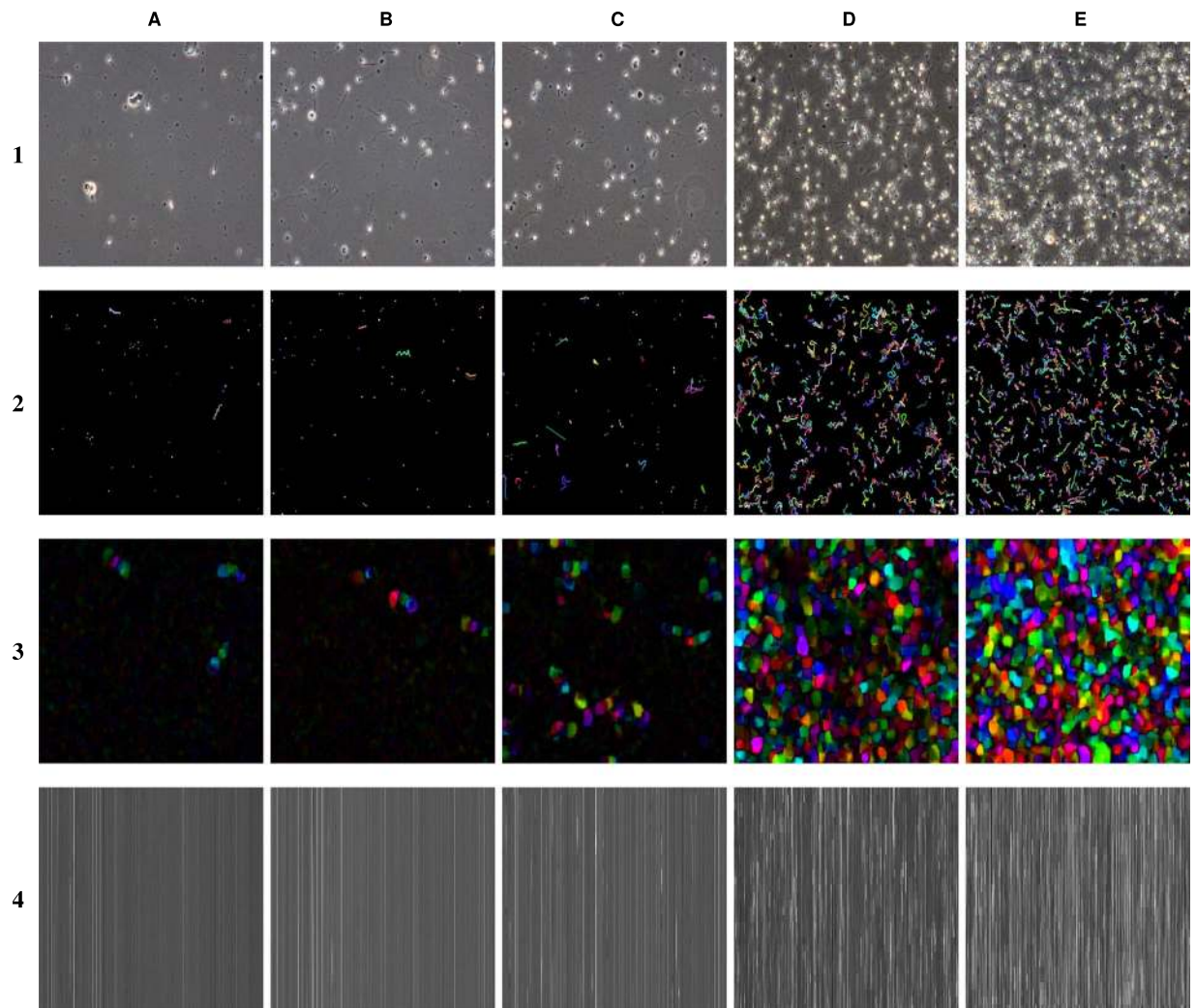
Figure 4. Examples of images from videos of semen samples with different concentrations (columns) and the four image representations used to train the neural network-based algorithms (rows). Sperm concentration; (**A**) 4 per $x106/mL$, (**B**) 33 per $x106/mL$, (**C**) 105 per $x106/mL$, (**D**) 192 per $x106/mL$, and (**E**) 350 per $x106/mL$. Image representation; (1) original video, (2) sparse optical flow, (3) dense optical flow, and (4) vertical frame matrix.

Two-stream network. For the last approach, we combine the two previous methods (visual features of raw frames and the temporal information of optical flow), which is inspired by the work done by Simonyan and Zisserman's[36], where they used a dual-network to perform human action recognition and classification. The model architecture follows a similar structure as described in their article, with the difference being how we input the optical flow representations into the model (we do not stack multiple optical flow representations for different sequences).

Based on this modification, we propose three different methods. Firstly, we use the dual network to analyze one raw video frame in parallel with a Lukas-Kanade sparse optical flow representation of the previous 30 frames. Secondly, we process one raw frame together with a Farneback's dense optical representation. Lastly, we again use one raw frame, but now we combine both the Lukas-Kanade and Farneback's optical flow method by stacking them channel-wise and pass these together through the network. Frames were extracted in the same way as performed for the Single Frame Prediction approach, and the optical flow representations were reused from the Optical Flow-based experiments.

Ethical approval and informed consent. In this study, we used fully anonymized data originally collected based on written informed consent and approval by the Regional Committee for Medical and Health Research Ethics - South East Norway. Furthermore, we confirm that all experiments were performed in accordance with the relevant guidelines and regulations of the Regional Committee for Medical and Health Research Ethics - South East Norway, and the GDPR.

Classical Machine Learning Results				
Method	Progressive	Non-progressive	Immotile	Average Mean Absolute Error
Baseline				
ZeroR	17.260	7.860	13.660	12.927
Participant Data Only				
Elastic Net	15.198	9.525	13.441	12.721
Gaussian Process	15.556	9.762	13.474	12.931
Simple Linear Regression	15.416	9.281	13.601	12.766
SMOreg	15.355	9.441	12.959	12.585
Random Forests	**13.312**	**8.886**	**11.905**	**11.368**
Random Tree	17.801	10.952	14.984	14.579
Tamura Image Features Only				
Elastic Net	14.400	7.750	12.190	11.447
Gaussian Process	13.230	7.260	11.920	10.803
Simple Linear Regression	13.520	8.170	12.690	11.460
SMOreg	**13.220**	**7.260**	**11.920**	**10.800**
Random Forests	13.530	7.400	12.060	10.997
Random Tree	18.700	9.960	16.520	15.060
Tamura Image Features and Participant Data				
Elastic Net	14.130	9.890	11.750	11.923
Gaussian Process	13.700	10.120	11.460	11.760
Simple Linear Regression	13.940	10.240	11.410	11.863
SMOreg	13.710	10.140	11.460	11.770
Random Forests	**13.510**	**10.000**	**11.340**	**11.617**
Random Tree	18.660	13.270	16.960	16.297

Table 1. Prediction performance of the machine learning-based methods in terms of mean absolute error for each of the motility values and the overall average. The best performing algorithm in each category is in bold.

Results and Discussion

A complete overview of the results for each method can be seen in Tables 1 and 2. A chart comparing the results is presented in Fig. 5. Table 1 presents the results for the classical machine learning algorithms trained on participant data, Tamura image features, and a combination of the two. For these results, the Gaussian Process, SMOreg, and Random Forests have a MAE below 11, which according to the paired t-test analysis is significant. One interesting finding is that for all cases where participant data is added, the algorithm performs worse. Although a preliminary result, for BMI this is not in line with the finding in our previous work Andersen *et al.*[21], where BMI was found to be negatively correlated with sperm motility using multiple linear regression. However, the methods are very different and therefore not directly comparable. As future work, we plan to perform an extensive analysis of all methodologies on a new dataset. Another interesting insight gained from this experiment is that the Tamura features seem to be well suited for sperm analysis, which will be interesting to investigate more closely.

Since sperm concentration is an important confounding variable when assessing sperm motility by CASA, we performed additional experiments using the two best-performing algorithms to investigate whether or not it had any influence. For the Random Forest, we achieved a MAE of 11.091 when including sperm concentration, compared to 10.996 when we did not. For SMOReg, the MAE was 10.902 with and 10.800 without. This minor difference in error indicates that our method is not gaining or losing any predictive power when including sperm concentration in the analysis, which can be seen as an advantage compared with CASA systems.

To assess the performance of the deep learning-based methods, we used the best performing classical machine learning approach (SMOreg with a MAE of 10.800) and ZeroR as a baseline. In Table 2, the results for single and multimodal deep learning approaches are shown. For most of the experiments, the deep learning models outperform the best machine learning algorithm (SMOreg) by a margin of one or two points. The two methods which are not significant better than ZeroR are the two-stream neural networks, which combined the two optical flow representations in a custom network.

We hypothesize that this is related to the fact that these networks are not able to learn the association between the temporal information of the optical flow and the visual data of the raw frame. Similar to the machine learning algorithms, all methods which combined the participant data with the videos performed worse than those without, leading to the same conclusion as previously discussed. Thus, in our study, adding patient data does not improve the results compared to using only video data, regardless of the algorithms used. If these findings also apply to other patient data needs to be further investigated.

The best performing approaches were a near tie between the method Channel-wise Greyscale and Dense Optical Flow using a stride of 1 or 10 (see Fig. 5). The Channel-wise Greyscale approach achieved a MAE of 8.786, which is two points lower than that of the best performing classical machine learning algorithm (see Table 2). The two Dense Optical Flow methods have the same performance as the Channel-wise Greyscale approach but using one-tenth of the image size, which makes them faster and less computational resource demanding.

Deep Learning Results				
Method	Progressive	Non-progressive	Immotile	Average Mean Absolute Error
Raw Frame Data Approach				
Single Frames (ResNet50)	13.162	8.024	10.967	10.718
Single Frames (ResNet50) + PD	13.659	8.196	12.293	11.383
Channel-wise Greyscale	**10.498**	**7.037**	**8.822**	**8.786**
Channel-wise Greyscale + PD	11.599	7.849	10.132	9.860
Vertical Frame Matrix	11.149	8.218	9.418	9.595
Vertical Frame Matrix + PD	11.182	8.199	9.274	9.552
Optical Flow Approach				
Sparse Optical Flow	11.573	7.263	10.155	9.664
Sparse Optical Flow + PD	12.214	7.760	10.802	10.259
Dense Optical Flow (stride = 1)	**10.191**	**7.114**	**8.914**	**8.740**
Dense Optical Flow (stride = 1) + PD	10.795	7.856	8.745	9.132
Dense Optical Flow (stride = 10)	10.319	7.546	8.782	8.882
Dense Optical Flow (stride = 10) + PD	11.386	7.825	9.734	9.648
Two Stream Network Approach				
Two Stream Sparse	15.888	8.187	13.326	12.467
Two Stream Sparse + PD	16.435	8.197	13.172	12.601
Two Stream Dense (stride = 1)	14.583	7.393	11.996	11.324
Two Stream Dense (stride = 1) + PD	18.166	8.570	15.983	13.940
Two Stream SP + DE (stride = 1)	**11.848**	**7.070**	**10.823**	**9.917**
Two Stream SP + DE (stride = 1) + PD	17.304	8.066	13.783	13.051

Table 2. Prediction performance of the deep learning-based methods in terms of mean absolute error for each of the motility values and overall mean. Note that for each method, we trained two models, one with participant data and one without. Methods which used participant data under training are marked with (+PD). For the methods which use dense optical flow, stride represents the number of frames skipped when comparing the difference of two frames.
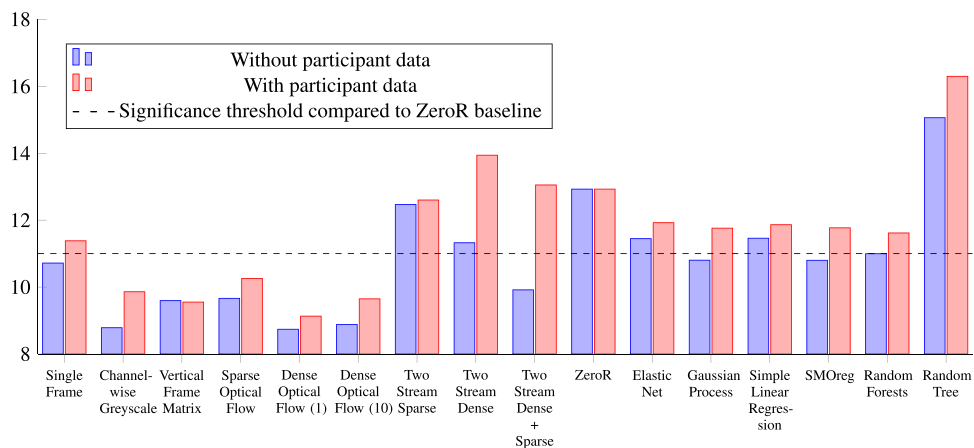
Figure 5. The different machine learning-based algorithms (classical and deep learning) used to predict semen quality in terms of progressive, non-progressive, and immotile spermatozoon. The stippled line represents the threshold for the results to be considered significant compared to the ZeroR baseline. The y-axis does not start at 0 to better highlight the differences. For the methods which used dense optical flow, stride values, how many frames are skipped when comparing two frames, are presented with a *1* or *10* indicating the number of skipped frames. Dense Optical Flow (1) and Channel-wise Greyscale are the best-performing ones but, several of our proposed methods are below the significance threshold.

It is important to point out that the 250 frames used in the analysis were extracted evenly distributed across the entire video length. This means that if there were a noticeable reduction in sperm motility after a certain amount of time, it would be taken into account by the algorithm. The results also support this assumption as the deep learning methods outperformed all classical machine learning methods. This is one of the advantages of the deep learning-based methods presented here.

In terms of time needed for the analysis, all presented methods perform the prediction within five minutes, including data preparation which takes most of the time. This is considerably faster than manual sperm motility assessment would be. The classical machine learning methods are faster to train, but in terms of application of the model, the speed is comparable with the deep learning methods.

Conclusion and Future Work

Overall, our results indicate that deep learning algorithms have the potential to predict sperm motility consistently and time efficiently. Multimodal analysis methods combining video data with participant data did not improve the prediction of sperm motility compared to using only the video data. However, it is possible that multimodal analysis using other participant data could improve the prediction. Our results indicate that the deep learning models can incorporate time into their analysis, and therefore are able to predict motility values better than the classical machine learning algorithms. In the future, deep learning-based methods could be used as an efficient support tool for human semen analysis. The presented methods can easily be applied to other relevant assessments such as automatic evaluation of sperm morphology.

Efficient analysis of long videos is a challenge, and future work should focus on how to combine the different modalities of time, imaging, and patient data. The dataset used in this study is also shared openly to ensure comparability and reproducibility of the results. Furthermore, we hope that the methods described in this work will inspire to further development of automatic analysis within the field of male reproduction.

Data availability

The dataset used for all experiments is publicly available at https://datasets.simula.no/visem/ for non-commercial use. The data is fully anonymized (no keys for re-identification are stored).

References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. medicine* **25**, 44 (2019).
2. Boll, S., Meyer, J. & O'Connor, N. E. Health media: From multimedia signals to personal health insights. *IEEE MultiMedia* **25**, 51–60 (2018).
3. Riegler, M. *et al*. Multimedia and medicine: Teammates for better disease detection and survival. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 968–977, https://doi.org/10.1145/2964284.2976760 (ACM, 2016).
4. Hannun, A. Y. *et al*. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Medicine* **25**, 65–69, https://doi.org/10.1038/s41591-018-0268-3 (2019).
5. Esteva, A. *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nat.* **542**, 115–118, https://doi.org/10.1038/nature21056 (2017).
6. Pogorelov, K. *et al*. Efficient disease detection in gastrointestinal videos – global features versus neural networks. *Multimed. Tools Appl.* **76**, 22493–22525, https://doi.org/10.1007/s11042-017-4989-y (2017).
7. Carlsen, E., Giwercman, A., Keiding, N. & Skakkebæk, N. E. Evidence for decreasing quality of semen during past 50 years. *Br. Med. J.* **305**, 609–613 (1992).
8. Levine, H. *et al*. Temporal trends in sperm count: a systematic review and meta-regression analysis. *Hum. Reproduction Updat.* **23**, 646–659 (2017).
9. Jørgensen, N. *et al*. East–west gradient in semen quality in the nordic–baltic area: a study of men from the general population in denmark, norway, estonia and finland. *Hum. Reproduction* **17**, 2199–2208 (2002).
10. Tomlinson, M. Uncertainty of measurement and clinical value of semen analysis: has standardisation through professional guidelines helped or hindered progress? *Androl.* **4**, 763–770 (2016).
11. World Health Organization, Department of Reproductive Health and Research. *WHO laboratory manual for the examination and processing of human semen* (Geneva: World Health Organization, 2010).
12. Cooper, T. G. *et al*. World health organization reference values for human semen characteristics. *Hum. Reproduction Updat.* **16**, 231–245, https://doi.org/10.1093/humupd/dmp048 (2010).
13. Mortimer, S. T., van der Horst, G. & Mortimer, D. The future of computer-aided sperm analysis. *Asian journal andrology* **17**, 545 (2015).
14. Dearing, C. G., Kilburn, S. & Lindsay, K. S. Validation of the sperm class analyser casa system for sperm counting in a busy diagnostic semen analysis laboratory. *Hum. Fertility* **17**, 37–44 (2014).
15. Dearing, C., Jayasena, C. & Lindsay, K. Can the sperm class analyser (sca) casa-mot system for human sperm motility analysis reduce imprecision and operator subjectivity and improve semen analysis? *Hum. Fertility* 1–11 (2019).
16. Urbano, L. F., Masson, P., VerMilyea, M. & Kam, M. Automatic tracking and motility analysis of human sperm in time-lapse images. *IEEE Transactions on Med. Imaging* **36**, 792–801, https://doi.org/10.1109/TMI.2016.2630720 (2017).
17. Dewan, K., Rai Dastidar, T. & Ahmad, M. Estimation of sperm concentration and total motility from microscopic videos of human semen samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2018).
18. Ghasemian, F., Mirroshandel, S. A., Monji-Azad, S., Azarnia, M. & Zahiri, Z. An efficient method for automatic morphological abnormality detection from human sperm images. *Comput. methods programs biomedicine* **122**, 409–420 (2015).
19. Shaker, F., Monadjemi, S. A., Alirezaie, J. & Naghsh-Nilchi, A. R. A dictionary learning approach for human sperm heads classification. *Comput. biology medicine* **91**, 181–190 (2017).
20. Haugen, T. *et al*. Visem: A multimodal video dataset of human spermatozoa. In *Proceedings of the ACM Multimedia Systems Conference (MMSYS)*, https://doi.org/10.1145/3304109.3325814 (ACM, 2019).
21. Andersen, J. M. *et al*. Body mass index is associated with impaired semen characteristics and reduced levels of anti müllerian hormone across a wide weight range. *PloS one* **10**, e0130210 (2015).
22. Nadeau, C. & Bengio, Y. Inference for the generalization error. In *Proceeding of the Advances in neural information processing systems (NIPS)*, 307–313 (2000).
23. Lux, M., Riegler, M., Halvorsen, P., Pogorelov, K. & Anagnostopoulos, N. Lire: open source visual information retrieval. In *Proceedings of the ACM Multimedia Systems Conference (MMSYS)*, **30** (2016).
24. Hall, M. *et al*. The WEKA data mining software: an update. *SIGKDD Explor.* **11**, 10–18 (2009).
25. Dozat, T. *Incorporating nesterov momentum into adam*. (2015).
26. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

27. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition*. 770–778 (2016).
28. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567* (2015).
29. Deng, J. *et al.* ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2009).
30. Chollet, F. *et al. Keras: Deep learning library for theano and tensorflow*. https://keras.io (2015).
31. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283 (2016).
32. Lucas, B. D. & Kanade, T. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) - Volume 2*, 674–679 (Morgan Kaufmann Publishers Inc., 1981).
33. Harris, C. & Stephens, M. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference*, 147–151 (1988).
34. Bradski, G. *The OpenCV Library. Dr. Dobb's J. Softw. Tools* (2000).
35. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In Bigun, J. & Gustavsson, T. (eds) *Image Analysis*, 363–370 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2003).
36. Simonyan, K. & Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 568–576 (2014).

Author contributions

S.A.H., M.A.R. and T.B.H. conceived the experiment(s), S.A.H., V.T. and M.A.R. conducted the experiment(s), all authors analysed the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.A.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.