IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 June 2024, accepted 5 August 2024, date of publication 13 August 2024, date of current version 23 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3443179

RESEARCH ARTICLE

Estimation of Multi-Expert Sperm Assessments Using Video Recognition Based Model Trained by EMD Loss

HAYATO NAKAGAWA[1], **TAKURO FUJII**[1], **YASUSHI YUMURA**[2], **AND TOMOKI HAMAGAMI**[3], **(Member, IEEE)**

[1]Graduate School of Engineering Science, Yokohama National University, Yokohama, Kanagawa 240-8501, Japan
[2]Reproduction Center, Yokohama City University Medical Center, Yokohama, Kanagawa 232-0024, Japan
[3]Faculty of Engineering, Yokohama National University, Yokohama, Kanagawa 240-8501, Japan

Corresponding author: Hayato Nakagawa (haya.nakagawa.ynu@gmail.com)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee for Medical and Biological Research Involving Human Subjects under Application No. 2023-27.

ABSTRACT Infertility is a common problem, affecting approximately one in six adults worldwide. Some studies have shown that male factors contribute to infertility in up to 50% of couples. Intracytoplasmic sperm injection (ICSI) is a common treatment for male infertility. This procedure requires a quick and accurate determination of whether sperm are suitable for ICSI. However, this assessment requires expertise and is time-consuming. Several computer-based systems for sperm analysis have been proposed to mitigate the burden on experts. However, there are no systems that can consider both sperm motility and morphology, or that can directly assess sperm suitability for ICSI. To address this problem, we constructed the multi-expert rated sperm video dataset for analysis, that includes motion information and developed an end-to-end sperm grade distribution estimation model using this dataset. Our model predicts a distribution that reflects multiple expert assessments, and thus helps to easily determine the suitability of a given sperm for ICSI. To develop this model, we conducted an exhaustive evaluation of various feature extractors and loss functions. Through this analysis, TimeSformer was identified as the optimal feature extractor from sperm videos, improving on average by 0.1×10^{-2} in MSE, 1.17% in grade distribution accuracy, and 3.41% in grade mode accuracy compared to ResNet, an image recognition model. Moreover, we identified earth mover's distance loss as the most suitable loss function, particularly in segments with lower scores.

INDEX TERMS Medical assistance, automatic sperm analysis, end-to-end, video processing, earth mover's distance.

I. INTRODUCTION

Infertility is a widespread global concern, affecting approximately 17.5% of adults, or about one in six people [1]. To address this issue, affordable and high-quality fertility care is needed.

Some studies suggest that male factors contribute to infertility in up to 50% of couples [2], [3], [4]. Therefore, infertility treatment is important for both males and females. Intracytoplasmic sperm injection (ICSI) is common treatment for male infertility. In ICSI, individual sperm cells are meticulously selected and injected into eggs by experts. In this process, it is important to quickly and accurately classify sperm as either normal or abnormal. This assessment, however, requires expertise and is time-consuming.

Several computer-based systems for sperm analysis have been proposed to mitigate the burden on experts. However, there are no systems that can consider both sperm motility and morphology, or that can directly assess sperm suitability for ICSI. Computer-assisted semen analysis (CASA) systems [5] automated sperm analysis to some extent. This systems provide sperm concentration, motility analysis, morphology

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Moinul Hossain.

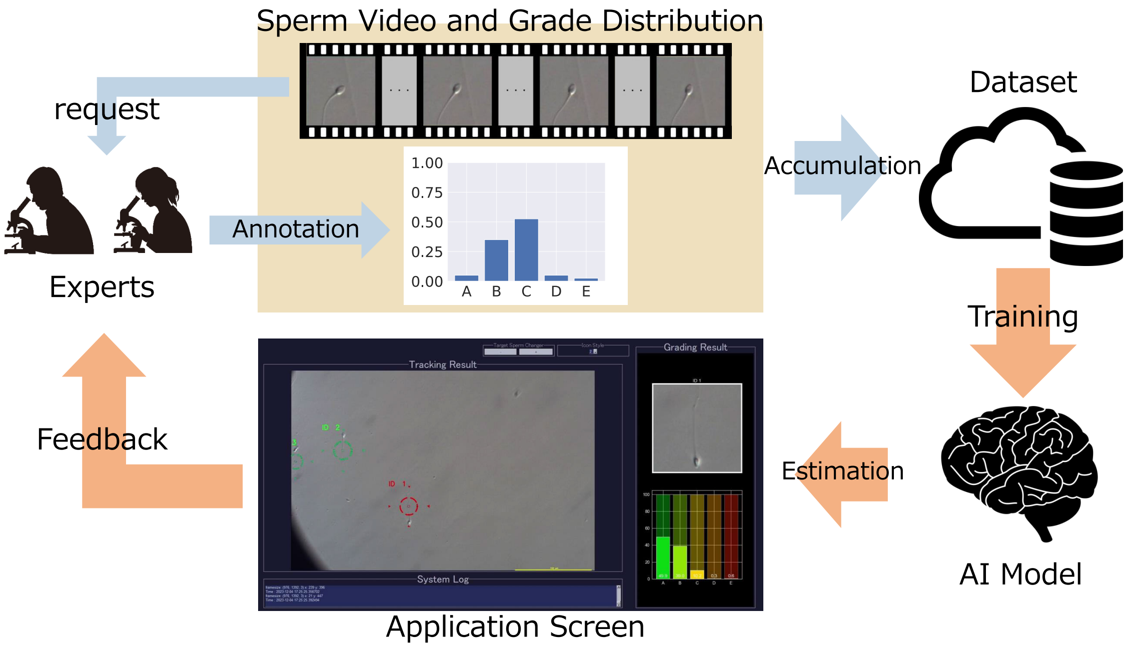


FIGURE 1. **Overview of our sperm grade estimation system. We collected multiple expert evaluations of sperm videos, curated the multi-expert rated sperm video (MERSV) dataset for analysis, and developed an end-to-end sperm grade distribution estimation model trained by this dataset. By referring to the sperm grade distributions predicted by our model, experts can reduce their workload.**

assessment, but they cannot assess sperm considering both motility and morphology and do not assess sperm suitability for ICSI. There are also several studies used machine learning [6], [7], [8], [9], [10], [11], [12]. However, they focused primarily on sperm images and ignored the crucial aspects, motion information and cannot assess sperm suitability for ICSI.

To address this problem, we collected multiple expert evaluations of sperm videos, curated the multi-expert rated sperm video (MERSV) dataset for analysis, that included motion information, and developed an end-to-end sperm grade distribution estimation model trained by this dataset. Figure1 illustrates this process. Owing to the end-to-end inference from the video, this model can consider both motion and morphology. The predicted distribution reflecting the evaluations of multiple experts, can easily help determine the suitability of a particular sperm for ICSI. By referring to predicted sperm grade distributions, experts can efficiently select a sperm from multiple candidates without the need to check all sperm. Furthermore, the collected dataset and predicted sperm grade distribution can be valuable resources for professional development. Therefore, our system can significantly help to reduce the workload for experts.

There are two questions in developing the end-to-end sperm grade distribution estimation model. First, how to extract sperm motion and morphological features from video data. Second, which loss function is appropriate for a model to estimate the grade distribution. To answer these research questions, we conducted a thorough comparison of models based on different feature extractors and loss functions. We considered three video feature extractors: R(2+1)D, SlowFast and TimeSformer, and four loss functions: mean squared error loss (MSE), cross entropy loss (CE), Jensen-Shannon divergence loss (JSD), and earth mover's distance loss (EMD). From the experimental results of the model comparison, TimeSformer emerged as the most promising choice among the video recognition models, and outperform the image-recognition model. This result suggests that image-based analysis is not sufficient for comprehensive sperm analysis, which highlights the importance of video-based analysis. In addition, our experimental results comparing loss functions indicate that EMD performs best, showing superior performance especially for lower scoring segment samples.

Our contributions can be summed up in three points:

- We constructed the multi-expert rated sperm video (MERSV) dataset and proposed an end-to-end sperm grade distribution estimation system based on videos that helps to reduce expert's workload for ICSI.
- We analyzed three feature extractors and found that TimeSformer was the most effective feature extractor for video-based sperm analysis, that outperforms ResNet and highlights the indispensability of video data.
- Earth mover's distance (EMD) loss was identified as the most suitable loss function for the estimating grade distribution and demonstrated its superior performance in lower-scoring segment samples.

The rest of this paper is organised as follows. In SectionII, we briefly present previous work that includes machine learning approaches for sperm assessment, video recognition models and label distribution learning. SectionIII describes details of our compiled MERSV dataset. In SectionIV, we explains details of our proposed sperm grade distribution estimation model. SectionsV shows the experimental setup for comparing of models based on different feature extractors and loss functions and the results. Finally, SectionVI presents the conclusions.

II. RELATED WORK

A. DATASETS AND MACHINE LEARNING APPROACHES FOR SPERM ASSESSMENT

Machine learning have been used on healthcare and medicine [13], [14], [15], [16], [17]. Several machine learning methods have been proposed for sperm morphology analysis too. The methods were validated using three datasets: SMIDS, HuSHeM and SCIAN.

The Sperm Morphology Image (SMIDS) dataset was constructed by the Medical Faculty of Istanbul University using smartphone-based data acquisition [6]. It included 3,000 segmented RGB sperm images labeled as normal(1,021), abnormal(1,005) and non-sperm(974) by an expert.

The Human Sperm Head Morphology (HuSHeM) dataset was curated from sperm samples of 15 patients at the Isfahan Fertility and Infertility Center [7]. This dataset included 216 RGB images of 131 × 131 pixels, which were classified by three experts as Normal(54), Tapered(53), Pyriform(57), and Amorphous(52), with consensus among the three experts.

The Laboratory for Scientific Image Analysis Gold-Standard for Morphological Sperm Analysis (SCIAN) dataset consists of sperm samples from the Medical Faculty of Chile University [8]. It included 1132 grayscale images of sperm head images in 35 × 35 pixels formats, labeled as Normal(100), Tapered(228), Pyriform(76), Amorphous(656), and Small(72). These images were labeled by three experts, and samples were selected for which at least two out of three experts agreed. Both the HuSHeM dataset and the SCIAN dataset are annotated based on the sperm morphological categories provided by the World Health Organization (WHO) [18].

Numerous studies used image recognition models, particularly CNN-based models. Riordon et al. [9] fine-tuned a pre-trained VGG16 from ImageNet. Spencer et al. [10] combined VGG16, VGG19, ResNet-34 and DenseNet-161 using a multi-class meta-classifier. Yüzkat et al. [11] proposed six CNN models and conbined their decisions with soft-voting. Non-CNN methods were also introduced. Ilhan et al. [12] presented a computational framework using multi-stage cascade-connected preprocessing techniques, region-based descriptor features, and non-linear kernel SVM-based learning.

However, these approaches do not consider sperm movement and, focus exclusively on sperm images. Therefore, we compiled the MERSV dataset that can be analyzed including movement information. Using this dataset, an end-to-end sperm grade distribution estimation model was developed.

B. VIDEO RECOGNITION MODELS

Deep learning is used for video data analysis such as human action recognition [19], [20], [21], [22], [23], [24], [25], surveillance system [26], [27] and visual speech recognition [28], [29]. Video recognition models, mainly used for action recognition, have evolved in parallel with image recognition models such as ResNet [30]. These models can be broadly classified into two categories: CNN-based and transformer-based models.

CNN-based video recognition models began with R3D [19], which extended 2D convolution to 3D. However, owing to the large number of parameters, R3D's performance is limited. R(2+1)D [20] addressed this problem by pseudo-representing a 3D convolution by a 2D + 1D convolution, resulting in a reduction of parameters and improved performance. SlowFast [21] further improved performance by extracting features from videos with different frame rates. It incorporates a slow pathway that focuses on shape, and a fast pathway, that emphasizes motion.

Transformer-based video recognition models began with ViViT [22], which was inspired by the transformer-based image recognition model ViT [31]. TimeSformer [23] introduced divided space-time attention, which was superior to various self-attention methods in terms of computational complexity and accuracy, and achieved improved accuracy.

We focus on commonly used video recognition models such as R(2+1)D and SlowFast, which are advanced CNN-based video recognition models, and TimeSformer, an advanced transformer-based video recognition model. These models are used as video feature extractors in this study.

There are alternative methods for extracting features from a video using video coding techniques that involve motion estimation. In their work, Kumar et al. [32] proposed the K-MCSP algorithm for motion estimation, which incorporates a non-linear function as opposed to MCSP, which uses a linear function. This modification results in reduced computational complexity while minimizing PSNR variation during reconstruction. We didn't employ this method because it requires learning the acquisition of features to be extracted from a video based on a grade distribution that reflects multiple expert assessments in the training data.

In addition to frame features, there are studies that have achieved high accuracy by incorporating new task-specific features. Cao et al. [33] proposed e-TSN, a TSN network for hand gesture recognition that incorporates hand skeletal features instead of optical flow. In this study, in addition to the video frame features of the detected sperm, the speed calculated from the position information obtained during detection is added as a feature.

C. LABEL DISTRIBUTION LEARNING

Geng introduced the concept of a label distribution which includes different degrees of description across multiple labels, and investigated the best algorithm [34].

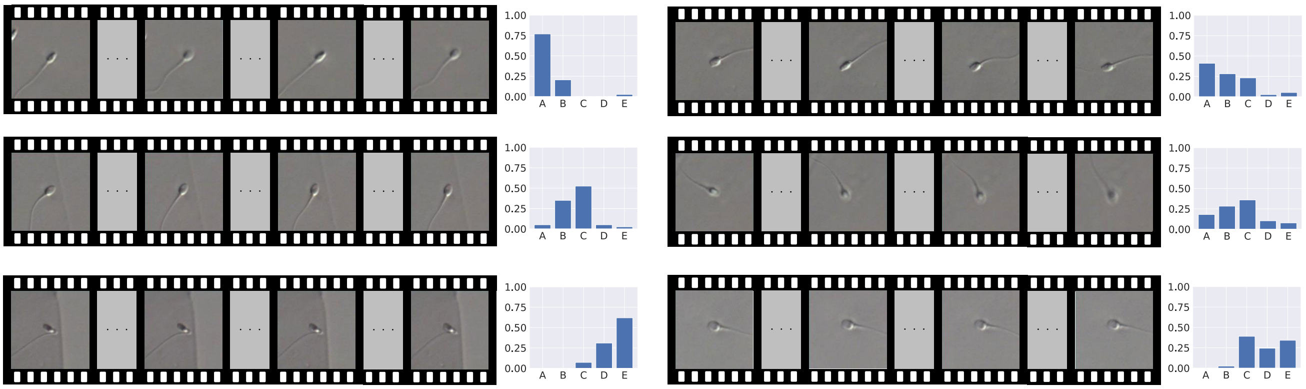
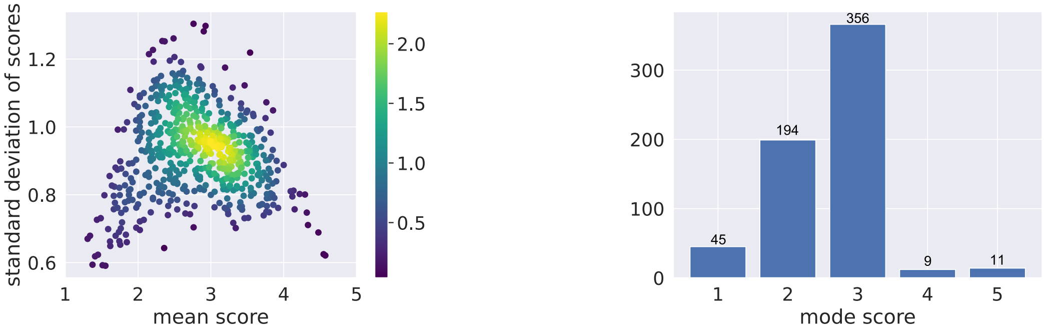


FIGURE 2. Six examples from our (MERSV) dataset. It shows 4 out of a 16 video frames (left) and their grade distribution (right) in each sample. Top samples, good; middle samples, medium; bottom samples show examples of concentrated bad ratings. Samples on the left side are examples of low variability in ratings, while samples on the right side are examples of high variability in ratings.



(a) (b)

FIGURE 3. Statistical analysis of the results of multiple experts ratings (grade distribution). Left (a): Joint histogram of the mean and standard deviations of grade scores. Right (b): Histograms of the grade mode score. Grade mode score is the most frequently evaluated grade score. These graphs show that most of the samples had grade distributions concentrated in B(2) and C(3).

Gao et al. [35] subsequently demonstrated the effectiveness of end-to-end learning with KL-Divergence for tasks with label distributions.

In our context, the grade distribution serves as the estimation target. This distribution is a special case of a label distribution whose labels have an ordinal relationship. Consequently, we propose using earth mover's distance for estimating the grade distribution, as this metric considers the distances between labels within the distribution.

III. MULTI-EXPERT RATED SPERM VIDEO (MERSV) DATASET

A. DATASET CONSTRUCTION

We compiled the multi-experts rated sperm video (MERSV) dataset for precise sperm assessment. The dataset included 615 videos recorded under a microscope, with each video having a grade distribution determined by annotations from approximately 40 experts. The study was approved by Ethics Committee for Medical and Biological Research Involving Human Subjects No. 2023-27.

1) SPERM VIDEO RECORDING METHOD

The videos were filmed in the semen of patients who had undergone ICSI treatment with consent. The number of patients is 615, which corresponds to the number of videos. Consent has been obtained from all patients for the collection and utilization of data. Sperm suspensions for video recording were prepared using density-gradient centrifugation followed by the swim-up method to obtain sperm with good motility. Video recording was conducted using an Olympus IX73 microscope. Due to variations in the thickness of the sperm suspension, the focus and light source were adjusted accordingly for each recording. Other settings are consistent with those described in [36].

2) DETAILS OF RECORDED SPERM VIDEO

The videos were recorded at a rate of 15 frames per second (fps). Each frame had a resolution of 1392 × 976 pixels. Each frame was annotated using a 150 × 150 bounding box to focus on a single sperm. We performed this bounding box annotation by tagging the target sperm in the first frame, and then tracking it throughout the video using template-matching techniques.

3) SPERM EVALUATION BY MULTIPLE EXPERTS

The experts assessed the sperm using a five-grade grading system: A(good), B(better), C(middle), D(worse), and E(bad). We refer to these labels as "grade class labels." By combining the experts ratings, we obtain a distribution, which we refer to as "grade distribution."

Figure2 shows examples of sperm videos and their grade distributions. The sufficiency of the video data quantity is discussed in SectionVI-A. While we cannot publish this dataset now, we plan to make it public in the future.

B. STATISTICS OF GRADE DISTRIBUTION

The total number of expert ratings collected was 24,533. For statistical analysis, we represented grade class labels as categorical variables using assigned numerical grade scores to them: A,1; B,2; C,3; D,4; and E,5. The means and standard deviations of the grade distributions are shown in Figure3(a). We refer to the most frequently evaluated grade class label/score as the "grade mode class label/score." Histograms illustrating the grade mode class scores are shown in Figure3(b). These graphs show that most of the samples had grade distributions concentrated in B(2) and C(3). When training a model with this dataset, data imbalance may occur. Therefore, evaluation results for each grade mode class label should be examined.

C. GRADE DISTRIBUTION ESTIMATION TASK

The grade distribution prediction task is referred to as the "grade distribution estimation task." As mentioned in SectionIII-A, grade distribution consists of grade class labels that have been annotated by multiple experts. The grade distribution was normalized by dividing it by the number of experts, to ensure that the sum of its values was 1. Therefore, each value in the grade distribution represents the probability of each grade class (y_i represents the probability of the i-th grade class). Grade distribution can be viewed as a special version of both single and multi-label annotations. Up to this point, the grade distribution shares similarities with the label distribution [34]. However, there are ordering relationships between the class labels within the grade distribution.

IV. SPERM GRADE DISTRIBUTION ESTIMATION MODEL

A. MODEL STRUCTURE

We propose a sperm grade distribution estimation model for automated sperm assessment and provide an overview of the model in Figure4. This model predicts grade distribution from detected frames and sperm positions.

Our proposed sperm grade estimation model is based on a video recognition model architecture. More explicitly, we examined three video recognition model architectures: R(2+1)D [20], SlowFast [21], and TimeSformer [23]. The video recognition models were used as feature extractors (Backbone). The last layer (Head) outputs the grade distribution based on the extracted features and sperm speed. To achieve this, we replaced the last layer of the video recognition model with five neurons, followed by Soft-max activation. The mathematical representation of the sperm grade distribution estimation model's process flow is as follows.

$$\hat{\boldsymbol{y}} = H(B(\boldsymbol{x}), \boldsymbol{v}) \tag{1}$$

Here, $\boldsymbol{y}$ represents the predicted grade distribution, H stands for the model head, B denotes the video feature extractor in the model, $\boldsymbol{x}$ represents the detected sperm video frames, and $\boldsymbol{v}$ corresponds to the sperm speed.

In addition, we prepared an image-based model as a baseline for comparison. We believe that video information is necessary for sperm analysis, as image information alone is not adequate, and demonstrate this in later experiments.

B. LOSS FUNCTION

There is a lack of research that clarifies the appropriate loss function for grade distribution estimation tasks. Four loss functions were prepared to determine the best loss function for this task: cross entropy loss (CE), mean squared error loss (MSE), Jensen-Shannon divergence loss (JSD) and earth mover's distance loss (EMD).

1) CROSS ENTROPY LOSS (CE)

CE is widely used as a training loss in classification tasks.

$$\mathcal{L}_{CE}(\boldsymbol{p}, \boldsymbol{q}) = -\sum_{i=1}^{n} p_i \log(q_i) \tag{2}$$

The true grade distribution is denoted by $\boldsymbol{p}$, the ith true grade class probability is denoted by p_i. The predicted grade distribution is denoted by $\boldsymbol{q}$, the ith predicted grade class probability is denoted by q_i, and the number of grade classes is denoted by n.

2) MEAN SQUARED ERROR (MSE)

MSE is widely used as a training loss function for regression tasks.

$$\mathcal{L}_{MSE}(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{n}\sum_{i=1}^{n} (p_i - q_i)^2 \tag{3}$$

3) JENSEN-SHANNON DIVERGENCE (JSD)

JSD measures the difference between two probability distributions. JSD is based on the Kullback-Leibler divergence (KLD), but it differs in that it is symmetric and always has a finite value.

$$\mathcal{L}_{JS}(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2}\mathcal{L}_{KL}\left(\boldsymbol{p}, \frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right) + \frac{1}{2}\mathcal{L}_{KL}\left(\boldsymbol{q}, \frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right) \tag{4}$$

$$\mathcal{L}_{KL}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \tag{5}$$

These three loss functions lack the inter-class relationships between score buckets.

4) EARTH MOVER'S DISTANCE (EMD)

EMD is defined as the minimum cost of transporting the mass from one distribution to another. The EMD is also known as the Wasserstein distance. It is used as a loss function

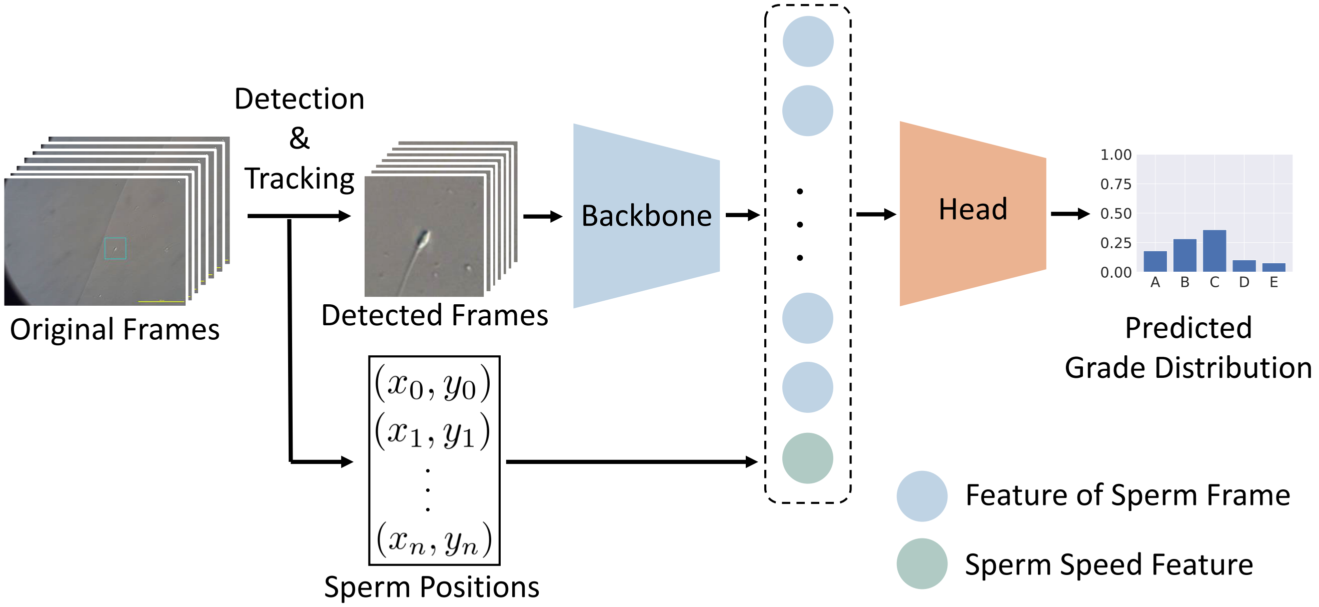


FIGURE 4. Overview of sperm grade distribution estimation model. This model predicts grade distribution from detected frames and sperm positions. In our model, video recognition model is used as feature extractor (Backbone) and the last layer (Head) outputs the grade distribution based on extracted features and sperm speed.

in various scenarios, such as order-class classification [37], [38], [39], adversarial training [40], and modality alignment learning [41]. In classification tasks where labels have strong relationships, EMD-based losses yield better results than other loss functions [37]. EMD can be solved exactly in closed form if the sums of the distributions are equal and the class space can be represented by a one-dimensional embedding. If the sums of the distributions are equal and the class space can be represented by a one-dimensional embedding, then an exact closed-form solution can be obtained [42]. The graded distributions considered in this study satisfy these conditions. Because the grade classes have an ordering relation of $1(A) < 2(B) < 3(C) < 4(D) < 5(E)$, the ground distance matrix of the grade class labels has a one dimensional embedding. In Addition, the two distributions, $\boldsymbol{p}$ and $\boldsymbol{q}$ have equal mass.: $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i$. Consequently, EMD can be computed exactly and in closed-form. As in [37], we use the Euclidean distance between the CDFs, which allows easier optimization with gradient descent.

$$\mathcal{L}_{EMD}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{n} (CDF_i(\boldsymbol{p}) - CDF_i(\boldsymbol{q}))^2 \tag{6}$$

V. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETTINGS

1) DATASETS

We used the MERSV Dataset described in SectionIII. For evaluation, a stratified 3-fold cross-validation was performed, considering the grade mode class as a label. This ensured that the distribution of the number of grade mode classes in the training samples remained the same in the validation samples. In each fold, the number of training data was 492, while for the test data, it was 123 and there is no confusion between videos of the same patient in train and test data. As each video differs in length, only the initial second, equivalent to 16 frames, is utilized. Additionally, fair sampling was performed such that the number of each grade mode class in training samples remained the same as that in validation samples. To align with the pretrained models, we utilize upsampled frames/images of size 224 × 224 from 150 × 150. Data augmentations, such as random rotation and color jittering, were also applied to improve the robustness of our model.

2) IMPLEMENTATION DETAILS

The sperm grade distribution estimation models presented in this study were implemented using PyTorch [43]. We used a pretrained image/video recognition model as the image/video feature extractor (Backbone). Specifically, ResNet [30] was used as the image feature extractor. In this image-based model, predictions were obtained from all 16 frames and evaluated individually. For the video feature extractor, we used R(2+1)D [20], SlowFast [21] and TimeSformer [23]. In these video-based model, predictions were generated by analyzing frames extracted at evenly spaced intervals, specifically 8 out of a total 16 frames. The last fully-connected layer is randomly initialized. In training, we used the stochastic gradient descent (SGD) optimizer in all experiments with a learning rate of 1e-3, momentum of 0.9, and weight decay of 5e-4. The models were trained for 1,000 epochs. The batch size was different for each model (32:ResNet, 16:SlowFast, 8:R(2+1)D, TimeSformer).

TABLE 1. Performance of the grade distribution models for different backbone models and loss function settings. For a more detailed analysis, we compared backbone models (Table2) and loss functions (Table3).

BackBone	Loss Function	EMD($\times 10^{-2}$) (↓)	MSE($\times 10^{-2}$) (↓)	JSD($\times 10^{-1}$) (↓)	CE (↓)	HI (↑)	Macro HI (↑)	Macro F1 (↑)
ResNet	EMD	1.620	1.242	2.120	1.540	0.7880	0.7542	0.3293
	MSE	1.652	1.223	2.054	1.536	0.7911	0.7370	0.3442
	JSD	1.897	1.611	1.334	1.512	0.7729	0.7512	0.3328
	CE	3.853	5.926	7.190	1.489	0.5646	0.4460	0.2706
R(2+1)D	EMD	1.696	1.303	1.763	1.529	0.7897	0.7203	0.3027
	MSE	1.712	1.276	1.899	1.531	0.7915	0.7054	0.3218
	JSD	1.982	1.800	1.332	1.511	0.7604	0.7072	0.3236
	CE	4.348	6.418	5.684	1.484	0.5414	0.4543	0.3266
SlowFast	EMD	1.447	1.150	1.527	1.523	0.8012	0.7382	0.3589
	MSE	1.478	1.143	1.637	1.526	0.8031	0.7550	0.3694
	JSD	1.614	1.369	1.142	1.509	0.7873	0.7320	0.3502
	CE	3.881	6.380	5.997	1.480	0.5420	0.4454	0.3373
TimeSformer	EMD	**1.230**	**1.000**	1.261	1.517	**0.8157**	0.7700	**0.4113**
	MSE	1.292	1.017	1.282	1.518	0.8141	**0.7720**	0.4060
	JSD	1.564	1.463	**0.969**	1.505	0.7822	0.7579	0.3943
	CE	3.889	6.122	5.090	**1.475**	0.5512	0.4829	0.4016

3) EVALUATION METRICS

We employed multiple evaluation metrics to accurately ascertain the performance among models and loss functions. To evaluate distribution differences, We used EMD, MSE, JSD and CE. For a fair evaluation of the loss functions, the histogram intersection (HI) [44], which is not used as a loss function, was used as an evaluation metric. The HI is a similarity measure that quantifies the degree of overlap between two histograms. We defined HI as "grade distribution accuracy".

$$HI(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{n} \min(p_i, q_i) \tag{7}$$

The grade distribution estimation task includes the grade mode classification aspect. To evaluate this aspect under the label imbalance, we used the macro F1 score (Macro F1). We also show the (grade mode) accuracy score which provides a clear understanding of the quality of our models' predictions, however, accuracy cannot accurately evaluate performance under the label imbalance. Therefore, we did not utilize accuracy in our detailed analysis. In Addition, to evaluate the robustness of grade mode class of the true distribution in the presence of distributional differences, a macro histogram intersection (Macro HI) was used. We define the Macro HI as the mean HI of each grade mode class.

B. RESULTS

Table1 shows the performance of the grade distribution estimation models for different backbone models and loss function settings. Lower values of EMD, MSE, JSD, and CE indicate higher performance, whereas higher values of HI(Macro HI) and MacroF1 indicate higher performance. We analyzed these results by comparing the backbone models and the loss functions.

TABLE 2. Comparison between backbone models in terms of ranking for each metric and overall average.

BackBone	ResNet	R(2+1)D	SlowFast	TimeSformer
EMD Rank	2.50	4.00	2.00	**1.50**
MSE Rank	2.50	4.00	2.00	**1.50**
JSD Rank	4.00	2.75	2.25	**1.00**
CE Rank	4.00	3.00	2.00	**1.00**
HI Rank	3.00	3.50	2.00	**1.50**
Macro HI Rank	3.00	3.75	2.25	**1.00**
Macro F1 Rank	3.25	3.75	2.00	**1.00**
Average	3.18	3.54	2.07	**1.21**

TABLE 3. Comparison between loss functions in terms of ranking for each metric and overall average.

Loss Function	EMD	MSE	JSD	CE
EMD Rank	**1.00**	2.00	3.00	4.00
MSE Rank	1.75	**1.25**	3.00	4.00
JSD Rank	2.25	2.75	**1.00**	4.00
CE Rank	3.25	3.75	2.00	**1.00**
HI Rank	1.75	**1.25**	3.00	4.00
Macro HI Rank	**1.25**	2.00	2.75	4.00
Macro F1 Rank	2.50	**1.75**	2.75	3.00
Average	**1.96**	2.11	2.50	3.43

1) COMPARISON BETWEEN BACKBONE MODELS

For the each loss function, we ranked the models and compared them based on their average rankings across the different evaluation metrics.

Table2 shows the average rankings of each metrics and their averages. For all metrics, the average rankings of the models exhibited similar trends. TimeSformer consistently outperformed all other models in all metrics, while R(2+1)D the lowest in almost all metrics. TimeSformer improved on average by 0.1×10^{-2} in MSE and 1.17% in HI (grade distribution accuracy) compared to ResNet, an image-based model. The superior performance of TimeSformer can be attributed to its effective feature extraction from the video

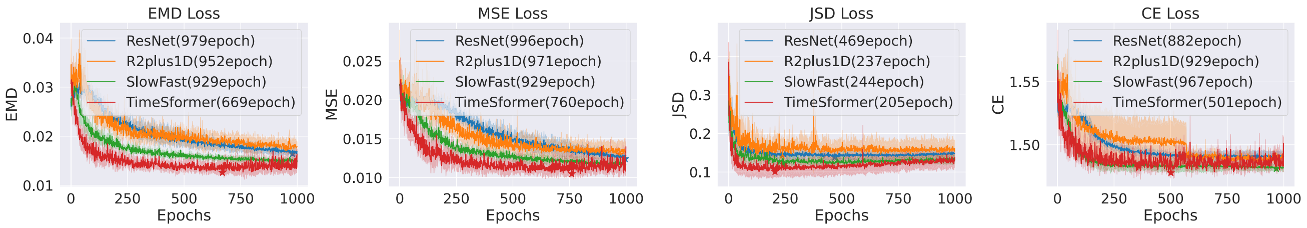


FIGURE 5. Learning curves of each loss functions in test data. The number within the parentheses indicates the epoch at which the model performs best.

TABLE 4. Accuracy in grade mode class. Note that this accuracy is assessed under label imbalance.

Backbone	ResNet				R(2+1)D				SlowFast				TimeSformer			
Loss Function	EMD	MSE	JSD	CE	EMD	MSE	JSD	CE	EMD	MSE	JSD	CE	EMD	MSE	JSD	CE
Accuracy	0.6289	0.6444	0.6322	0.6508	0.6423	0.6602	0.6276	0.6813	0.6715	0.6748	0.6715	0.6976	0.6862	0.6699	0.6488	0.6878

TABLE 5. Top 25/50/75% HI results for different backbone models and loss function settings. For a more detailed analysis, we compared backbone models (Table6) and loss functions (Table7).

BackBone		ResNet				R(2+1)D				SlowFast				TimeSformer			
Loss Function		EMD	MSE	JSD	CE	EMD	MSE	JSD	CE	EMD	MSE	JSD	CE	EMD	MSE	JSD	CE
Top 25% HI	A	0.843	0.841	0.804	0.462	0.790	0.803	0.757	0.660	0.847	0.859	0.823	0.513	0.838	0.870	0.857	0.642
	B	0.865	0.876	0.852	0.667	0.879	0.874	0.827	0.666	0.868	0.878	0.865	0.640	0.877	0.885	0.861	0.685
	C	0.864	0.870	0.879	0.694	0.876	0.875	0.872	0.624	0.878	0.882	0.881	0.641	0.889	0.888	0.865	0.629
	D	0.828	0.756	0.817	0.425	0.809	0.827	0.795	0.409	0.840	0.849	0.853	0.436	0.801	0.831	0.779	0.391
	E	0.763	0.748	0.783	0.437	0.644	0.642	0.671	0.297	0.698	0.703	0.756	0.456	0.758	0.778	0.807	0.471
	Macro	0.833	0.818	0.827	0.537	0.800	0.804	0.785	0.531	0.826	0.834	0.835	0.537	0.833	0.851	0.834	0.563
Top 50% HI	A	0.779	0.781	0.728	0.373	0.734	0.740	0.656	0.486	0.749	0.776	0.721	0.462	0.769	0.785	0.768	0.545
	B	0.808	0.817	0.775	0.568	0.810	0.822	0.750	0.591	0.822	0.826	0.796	0.559	0.837	0.831	0.790	0.590
	C	0.810	0.815	0.815	0.607	0.822	0.825	0.809	0.538	0.822	0.830	0.826	0.550	0.845	0.843	0.809	0.538
	D	0.718	0.674	0.773	0.393	0.626	0.609	0.742	0.385	0.730	0.743	0.638	0.385	0.757	0.710	0.731	0.373
	E	0.656	0.598	0.665	0.289	0.610	0.531	0.578	0.272	0.569	0.601	0.679	0.271	0.690	0.681	0.760	0.289
	Macro	0.754	0.737	0.751	0.446	0.72	0.705	0.707	0.454	0.738	0.755	0.732	0.445	0.780	0.770	0.772	0.467
Top 75% HI	A	0.688	0.691	0.622	0.306	0.657	0.656	0.552	0.418	0.668	0.664	0.656	0.362	0.737	0.754	0.661	0.457
	B	0.743	0.745	0.677	0.468	0.717	0.736	0.645	0.481	0.756	0.757	0.709	0.438	0.757	0.750	0.691	0.454
	C	0.743	0.748	0.739	0.516	0.760	0.763	0.728	0.465	0.760	0.765	0.741	0.484	0.779	0.775	0.744	0.470
	D	0.673	0.626	0.686	0.358	0.599	0.593	0.646	0.344	0.657	0.638	0.626	0.347	0.579	0.584	0.599	0.334
	E	0.498	0.493	0.539	0.222	0.496	0.462	0.517	0.166	0.453	0.424	0.527	0.201	0.672	0.646	0.715	0.244
	Macro	0.669	0.661	0.652	0.374	0.646	0.642	0.618	0.375	0.659	0.650	0.652	0.366	0.705	0.702	0.682	0.392

data. However, R(2+1)D legs behind ResNet. Video-based models have the potential to perform better than image-based models because they can use not only shape but also motion information. However, if the video-based model fails to extract features, as is the case in R(2+1)D, its performance is inferior to that of image-based model.

2) COMPARISON BETWEEN LOSS FUNCTIONS

For each model, we ranked the loss functions and compared them based on their average rankings across different evaluation metrics.

Table3 shows the average rankings for each metric and overall averages. For all loss functions, the average ranking was the highest when the metric was identical to the loss function. EMD had the best average ranking, indicating that it was the most appropriate function for the grade distribution estimation task. Conversely, CE was the lowest for almost all metrics, suggesting that it was not suitable for the grade distribution estimation task. However, in some cases, CE demonstrated superior performance compared to other loss functions with respect to MacroF1, indicating its suitability for classification tasks. JSD outperformed CE, however, lagged behind EMD or MSE in average rating.

3) OBSERVATIONS DURING TRAINING

Figure5 shows the learning curves of test data for each loss function. The numbers within parentheses indicate the epochs in which the model performed best. These graphs confirm that all models exhibited well-converged learning. Additionally, it is evident that TimeSformer consistently attained the lowest value in the earliest epochs for all loss functions, indicating efficient and rapid convergence during training.

4) ACCURACY IN GRADE MODE CLASS

We also assessed the accuracy of our models in grade mode (grade mode accuracy), and Table4 shows the results. Note

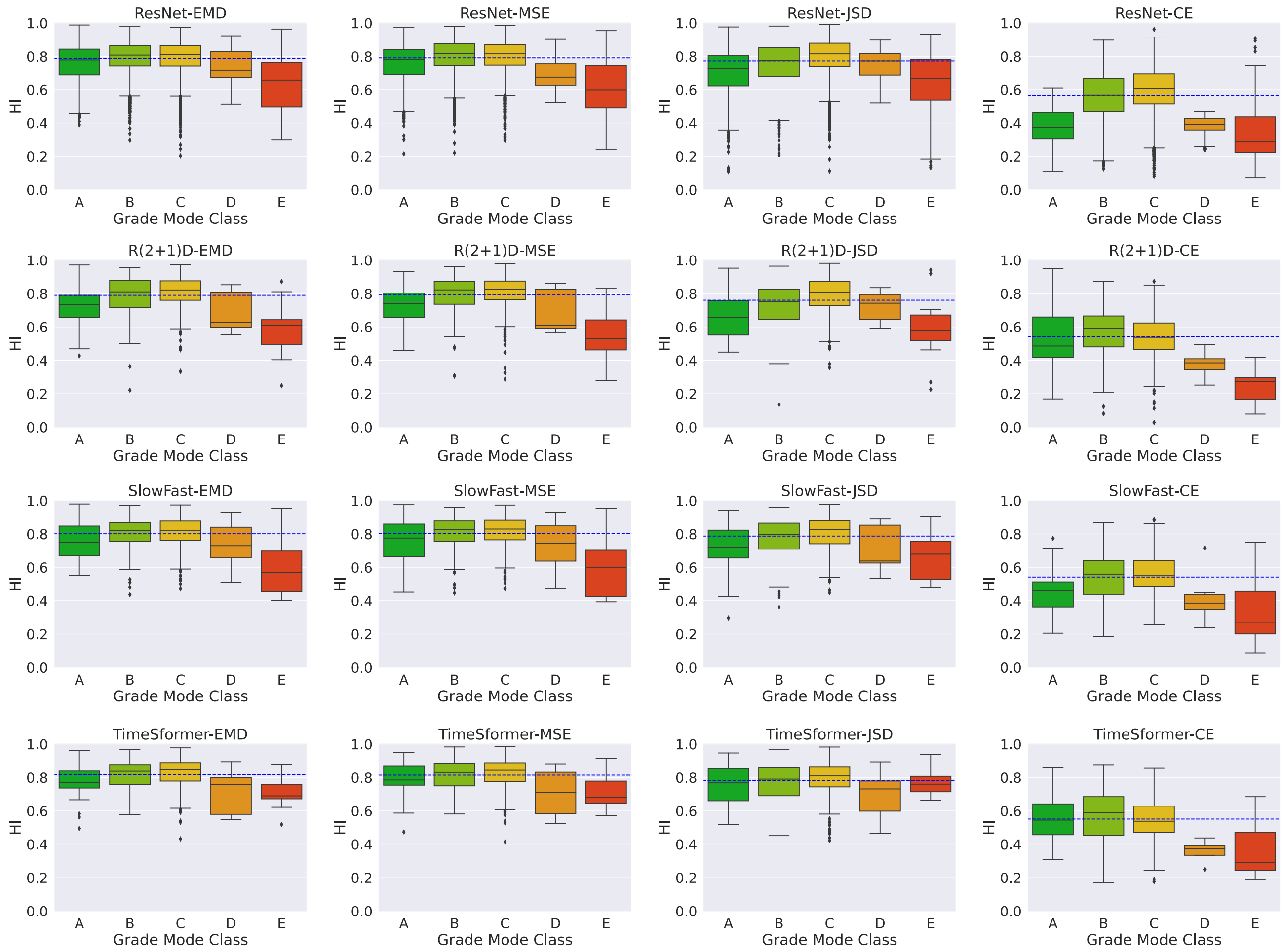


FIGURE 6. Distributions of HI for each grade mode class label. For a more detailed analysis, we focused on specific segments of the HI score distribution (Table5). The blue dotted line shows the average HI score.

that this accuracy is assessed under label imbalance. The accuracy ranges from 65% to 70%. On average, TimeSformer shows a 3.41% improvement compared to ResNet and video recognition models outperform ResNet, an image-based model, in almost all loss functions. These results reinforce the need for video-based analysis of sperm.

C. ANALYSIS OF HI DISTRIBUTION IN TEST DATA

We also compared the HI distributions in the test data. Figure6 shows the HI distributions for each grade mode class label and the blue dotted line shows the average HI. To analyze these distributions, we focused on specific segments of the HI score distribution, namely, the higher (Top25%), middle (Top50%), and lower (Top75%) segments. Table5 lists the Top25/50/75% HI results, which we analyzed by comparing backbone models and loss functions below.

1) COMPARISON BETWEEN BACKBONE MODELS

For each loss function, we ranked the backbone models and compared them based on their average rankings from the results in Table5. Table6 lists the average rankings for each metric and overall averages.

TABLE 6. Comparison of the Top25/50/75% of HI between backbone models.

Loss Function	ResNet	R(2+1)D	SlowFast	TimeSformer
Top25% Macro HI Rank	2.75	3.75	2.00	**1.00**
Top50% Macro HI Rank	2.50	3.25	2.75	**1.00**
Top75% Macro HI Rank	2.25	3.25	3.00	**1.00**
Average	2.50	3.42	2.58	**1.00**

Analyzing Table6, it is found that TimeSformer performs exceptionally well across all segments of the HI score, indicating its superiority in this dataset, which is consistent with the findings in SectionV-B. SlowFast outperformed ResNet in the Top25% segment, however, it performed worse than ResNet in the Top50% and 75% of the segments. This discrepancy suggests that SlowFast has difficulty effectively extracting generic features from video data. R(2+1)D consistently lagged behind ResNet in all the segments. This result is consistent with our observations in SectionV-B, suggesting that R(2+1)D cannot effectively extract the appropriate features from the video data.

2) COMPARISON BETWEEN LOSS FUNCTIONS

For each models, we ranked the loss functions and compared them based on their average rankings from the results in Table5.

Table7 lists a summary of the average rankings for each loss function across the various metrics. Upon analyzing the overall average, it is found that EMD was the best among the loss functions. Specifically, EMD outperformed the other loss functions in the Top50% and Top75% HI score segments. This can be attributed to ability of EMD to consider the relationships between grade classes and prevent fatal mistakes in the grade distribution estimation task. Particularly in the medical domain, it is important to guarantee the worst score from the perspective of reliability. Therefore, EMD is a suitable choice for practical applications. Conversely, CE performed the worst among all the loss functions. Similar to the findings presented in SectionV-B, these results reinforce the unsuitability of CE for the grade distribution estimation task.

TABLE 7. Comparison of the top25/50/75% of HI between loss functions.

Loss Function	EMD	MSE	JSD	CE
Top25% Macro HI Rank	2.25	**1.75**	2.00	4.00
Top50% Macro HI Rank	**1.25**	2.50	2.25	4.00
Top75% Macro HI Rank	**1.00**	2.25	2.75	4.00
Average	**1.50**	2.17	2.33	4.00

From the analysis so far, that the best Backbone is TimeSformer, and the best loss function is EMD. We visualized the latent space of our model in this setting and discussed the sufficiency of the quantity of video data in SectionVI-A.

VI. CONCLUSION

In this study, we curated the MERSV dataset and proposed a model for end-to-end sperm grade distribution estimation from videos to contribute to reduce experts workload in sperm selection for ICSI. Based on our experimental results, TimeSformer is the most promising of the video recognition models and outperforms the image recognition model ResNet. This result indicates that image-based analysis is insufficient for comprehensive sperm analysis, which highlights the importance of video-based analysis. We also identified the EMD as the most suitable loss function for the grade distribution estimation task, demonstrating its superior performance in lower-scoring segment samples.

However, our study has three limitations. Firstly, we cannot evaluate our models on multiple databases because there are no datasets with sperm videos and multi-expert annotations. In adddition, our dataset cannot be published now. Therefore, we will keep collecting new data and prepare to publish our dataset in the future. Secondly, the video model has a higher number of parameters than the image model, which may make it difficult to use in the clinics and reduce the inference speed (see SectionVI-B). Therefore, we will also work on model compression or build a new effective model in the future. Thirdly, although we assessed the performance of our proposed model in estimating grade distributions, we are unable to gauge its effectiveness in reducing the workload for experts. Hense, we aim to introduce this system into clinical environments and validate its effectiveness.

APPENDIX

A. VIDEO FEATURE DISTRIBUTIONS AND DATASET SIZE DISCUSSION

We checked the distribution of video features in our dataset using 768 dimensional features extracted from the TimeSformer trained by EMD. Figure7 shows that 2-dimensional video features compressed from 768 dimensions by PCA. The proportion of the variance for the first principal component (PC1) was 69%, while for the second principal component (PC2), it was 4.7%. The color of the points in this plot shows that the grade mode score of each sample. In this plot, moving towards the upper right causes the samples to become reddish, while moving towards the lower left causes the samples to become bluish. This confirms that TimeSformer can extract appropriate features from a video.

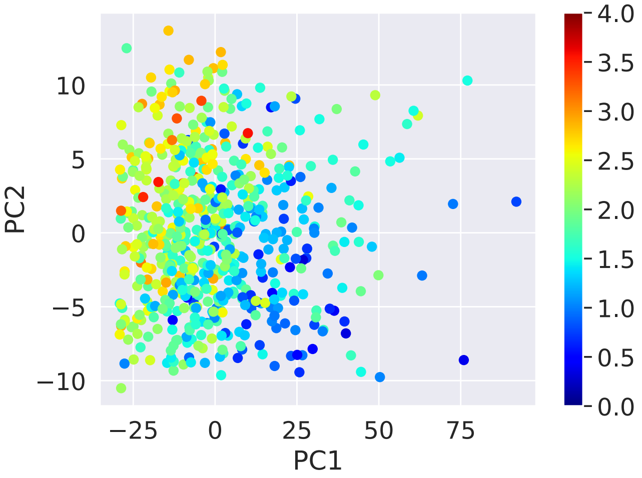


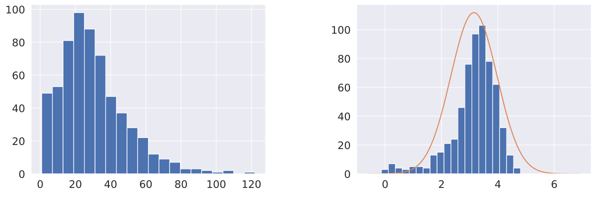
FIGURE 7. Feature distribution of TimeSformer-EMD compressed by PCA.

Figure8(a) shows the histogram of PC1 with an added bias to prevent negative values from appearing, while Figure8(b) presents the logarithmically transformed version of the PC1. The mean of logarithmically transformed PC1 (μ) was 3.15, while for the standard deviation (σ) it was 0.82. The orange line in this plot represents a normal distribution with a mean of μ and a standard deviation of σ. If the distribution of PC1 approximates the orange line, the population mean can be estimated.

The number of samples (n) is 615 and we can get the 95% confidence interval range (μ_r) in the following equation.

$$\mu_r = 1.96 \times \sqrt{\frac{\sigma^2}{n}} = 1.96 \times \sqrt{\frac{0.82^2}{615}} \fallingdotseq 0.065 \qquad (8)$$

μ_r is somewhat small relative to μ, which suggests that the number of samples in our dataset is adequate.



(a) (b)

FIGURE 8. PC1 Distribution. Left(a): The histogram of PC1 with an added bias to prevent negative values from appearing. Right(b): The histogram of the logarithmically transformed version of the PC1.

B. MODEL INFERENCE TIME

We measured the inference speed and show the results and the number of parameters in Table8. The measurement of inference speed was conducted with a batch size of 8 using Intel(R) Core(TM) i9-10940X CPU for use in a medical setting.

TABLE 8. Model inference time with a batch size of 8.

Backbone	Parameters [M]	Inference Time [ms]
ResNet	23.5	110.7
R(2+1)D	31.3	2594
SlowFast	33.7	2853
TimeSformer	121.3	4180

REFERENCES

[1] WHO. (2023). *Infertility Prevalence Estimates, 1990–2021*. [Online]. Available: https://www.who.int/publications/i/item/978920068315

[2] S. Pandruvada, R. Royfman, T. A. Shah, P. Sindhwani, J. M. Dupree, S. Schon, and T. Avidor-Reiss, "Lack of trusted diagnostic tools for undetermined male infertility," *J. Assist. Reproduction Genet.*, vol. 38, no. 2, pp. 265–276, Feb. 2021.

[3] C. A. Holden, R. I. McLachlan, R. Cumming, G. Wittert, D. J. Handelsman, D. M. de Kretser, and M. Pitts, "Sexual activity, fertility and contraceptive use in middle-aged and older men: Men in Australia, telephone survey (MATeS)," *Human Reproduction*, vol. 20, no. 12, pp. 3429–3434, Dec. 2005.

[4] A. Isidori, M. Latini, and F. Romanelli, "Treatment of male infertility," *Contraception*, vol. 72, no. 4, pp. 314–318, 2005.

[5] J. C. Lu, Y. F. Huang, and N. Q. Lü, "Computer-aided sperm analysis: Past, present and future," *Andrologia*, vol. 46, no. 4, pp. 329–338, May 2014.

[6] H. O. Ilhan, I. O. Sigirci, G. Serbes, and N. Aydin, "A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods," *Med. Biol. Eng. Comput.*, vol. 58, no. 5, pp. 1047–1068, May 2020.

[7] F. Shaker, S. A. Monadjemi, J. Alirezaie, and A. R. Naghsh-Nilchi, "A dictionary learning approach for human sperm heads classification," *Comput. Biol. Med.*, vol. 91, pp. 181–190, Dec. 2017.

[8] V. Chang, A. Garcia, N. Hitschfeld, and S. Härtel, "Gold-standard for computer-assisted morphological sperm analysis," *Comput. Biol. Med.*, vol. 83, pp. 143–150, Apr. 2017.

[9] J. Riordon, C. McCallum, and D. Sinton, "Deep learning for the classification of human sperm," *Comput. Biol. Med.*, vol. 111, Aug. 2019, Art. no. 103342.

[10] L. Spencer, J. Fernando, F. Akbaridoust, K. Ackermann, and R. Nosrati, "Ensembled deep learning for the classification of human sperm head morphology," *Adv. Intell. Syst.*, vol. 4, no. 10, Oct. 2022, Art. no. 2200111.

[11] M. Yüzkat, H. O. Ilhan, and N. Aydin, "Multi-model CNN fusion for sperm morphology analysis," *Comput. Biol. Med.*, vol. 137, Oct. 2021, Art. no. 104790.

[12] H. O. Ilhan, G. Serbes, and N. Aydin, "Automated sperm morphology analysis approach using a directional masking technique," *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103845.

[13] M. Yang, X. Huang, L. Huang, and G. Cai, "Diagnosis of Parkinson's disease based on 3D ResNet: The frontal lobe is crucial," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 104904.

[14] S. Xue and C. Abhayaratne, "Region-of-Interest aware 3D ResNet for classification of COVID-19 chest computerised tomography scans," *IEEE Access*, vol. 11, pp. 28856–28872, 2023.

[15] J. Liao, X. Li, Y. Gan, S. Han, P. Rong, W. Wang, W. Li, and L. Zhou, "Artificial intelligence assists precision medicine in cancer treatment," *Frontiers Oncol.*, vol. 12, Jan. 2023, Art. no. 998222.

[16] J. Qiu, L. Li, J. Sun, J. Peng, P. Shi, R. Zhang, Y. Dong, K. Lam, F. P.-W. Lo, B. Xiao, W. Yuan, N. Wang, D. Xu, and B. Lo, "Large AI models in health informatics: Applications, challenges, and the future," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 12, pp. 6074–6087, Dec. 2023.

[17] L. Xiao and Y. Zhang, "AI-driven smart pharmacology," *Intell. Pharmacy*, vol. 1, no. 4, pp. 179–182, Dec. 2023.

[18] *WHO Laboratory Manual for the Examination and Processing of Human Semen*, 6th ed., WHO, Geneva, Switzerland, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240030787

[19] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3154–3160.

[20] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[21] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6202–6211.

[22] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826.

[23] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" 2021, *arXiv:2102.05095*.

[24] M. C. Schiappa, N. Biyani, P. Kamtam, S. Vyas, H. Palangi, V. Vineet, and Y. Rawat, "A large-scale robustness analysis of video action recognition models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14698–14708.

[25] Z. Xing, Q. Dai, H. Hu, J. Chen, Z. Wu, and Y.-G. Jiang, "SVFormer: Semi-supervised video transformer for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18816–18826.

[26] H. Deng, Z. Zhang, S. Zou, and X. Li, "Bi-directional frame interpolation for unsupervised video anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2633–2642.

[27] A. Aich, K.-C. Peng, and A. K. Roy-Chowdhury, "Cross-domain video anomaly detection without target domain adaptation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2578–2590.

[28] K. R. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5152–5162.

[29] A. Koumparoulis and G. Potamianos, "Accurate and resource-efficient lipreading with EfficientNetv2 and transformers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8467–8471.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] R. Kumar, K. Kumar, S. Mahajan, C. Chakraborty, and A. K. Pandit, "Implementation of K-multi constraint shortest paths (K-MCSP) for video compression," *Multimedia Tools Appl.*, vol. 81, no. 24, pp. 35379–35410, Oct. 2022.

[33] Y. Cao, J. Li, C. Chakraborty, L. Qin, L. Tao, and X. Shao, "Temporal segment neural networks-enabled dynamic hand-gesture recognition for industrial cyber-physical authentication systems," *IEEE Syst. J.*, vol. 17, no. 4, pp. 5315–5326, Dec. 2023.

[34] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.

[35] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017.

[36] G. Lo Monte, F. Murisier, I. Piva, M. Germond, and R. Marci, "Focus on intracytoplasmic morphologically selected sperm injection (IMSI): A mini-review," *Asian J. Androl.*, vol. 15, no. 5, pp. 608–615, Jul. 2013.

[37] L. Hou, C.-P. Yu, and D. Samaras, "Squared Earth Mover's distance-based loss for training deep neural networks," 2016, *arXiv:1611.05916*.

[38] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a Wasserstein loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2053–2061.

[39] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[40] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 70, D. Precup and Y. W. Teh, Eds. Sydney, NSW, Australia: International Convention Centre, Aug. 2017, pp. 214–223.

[41] Y. Ling, Z. Zhong, D. Cao, Z. Luo, Y. Lin, S. Li, and N. Sebe, "Cross-modality Earth mover's distance for visible thermal person re-identification," in *Proc. AAAI*, 2023, pp. 1631–1639.

[42] E. Levina and P. Bickel, "The Earth mover's distance is the Mallows distance: Some insights from statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV()*, vol. 2, Oct. 2001, pp. 251–256.

[43] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037. [Online]. Available: https://github.com/pytorch/pytorch/tree/main

[44] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 5, pp. 1568–1577, Mar. 2017.

HAYATO NAKAGAWA was born in Nagoya, Aichi, Japan, in 1998. He received the B.E. degree in electrical and computer engineering from Yokohama National University, Kanagawa, Japan, in 2022, where he is currently pursuing the master's degree. His research interests include current deep learning advances in the field of image/video processing and medical applications of machine learning.

TAKURO FUJII was born in Okayama, Japan, in 2000. He is currently pursuing the master's degree with Yokohama National University. His research interests include medical computer vision, natural language processing, and vision-and-language.

YASUSHI YUMURA received the degree from the School of Medicine, Yokohama City University, in 1993, and the Ph.D. degree, in 2009. He joined the Department of Urology, Yokohama City University, in 1995. In 2012, he became the Director of the Reproduction Center, Yokohama City University Medical Center. He has been the Vice Director of Yokohama City University Medical Center, since 2020, and has been a Professor with the Reproduction Center, since 2022.

TOMOKI HAMAGAMI (Member, IEEE) received the B.S. degree in electrical engineering and the Ph.D. degree from Chiba University, in 1988 and 1999, respectively. From 1988 to 2000, he was with Intelligent System Laboratories, SECOM Company Ltd., as a Senior Researcher. From 2001 to 2004, he was with the Graduate School of Engineering, Chiba University, as an Assistant Professor. In 2001, he moved to Yokohama National University, as an Associate Professor, in 2004, where he was made a Professor, in 2008. His research interests include artificial intelligence, machine learning, reinforcement learning, multiagent systems, and smart systems with the IoT and CPS.