# Sperm DNA Fragmentation Prediction from Videos Using Deep Learning

Koki Hayashi*, Kento Morita*, Hiroki Takeuchi† and Tetsushi Wakabayashi*

*Department of Information Engineering, Graduate School of Engineering, Mie University, Japan
†Department of Obstetrics and Gynecology, Graduate School of Medicine, Mie University, Japan
E-MAIL: 425M525@m.mie-u.ac.jp

Abstract:

Sperm DNA fragmentation (SDF) is a key indicator of male infertility. However, standard diagnostic methods such as the TUNEL-FACS (Terminal deoxynucleotidyl transferase dUTP Nick End Labeling combined with Fluorescence-Activated Cell Sorting) assay are invasive and require specialized equipment. This study proposes a non-invasive deep learning framework that estimates SDF values directly from sperm motion videos captured using standard phase-contrast microscopy. Each video was divided into overlapping 8-frame clips using a sliding window to capture temporal dynamics. Two model architectures, 3D ResNet and TimeSformer, were evaluated under various preprocessing strategies designed to enhance sperm visibility and suppress background noise. To obtain the most effective video-level prediction method, we conduct a comparative analysis of three aggregation methods: mean, median, and best aggregation. TimeSformer with spatial preprocessing and median aggregation achieved the best practical performance. Although the best aggregation method achieved relatively high accuracy, the overall performance remains insufficient for clinical application. This result suggested that intelligent clip selection could enhance prediction reliability. The proposed method presents a stain-free, non-invasive, and promising alternative to conventional SDF testing.

Keywords:

Video analysis; Deep learning; Medical imaging; Classification; Reproductive health

## 1 Introduction

Infertility is a significant global health concern, with male-related factors contributing to approximately half of all cases. According to the World Health Organization, an estimated 17.5% of the adult population experience infertility at some point in their lives [1]. Among the various contributing factors, sperm DNA fragmentation (SDF), characterized by DNA damage within sperm cells, has been shown to negatively correlate with fertilization rates, embryo development, and pregnancy outcomes [2, 3, 4].

One of the most widely used techniques for assessing SDF is the terminal deoxynucleotidyl transferase dUTP nick end labeling (TUNEL) assay combined with flow cytometry, known as TUNEL-FACS [5]. While this method provides clinically validated accuracy, it is invasive, requires chemical staining, and depends on specialized laboratory equipment. These limitations restrict its practicality for routine screening. Moreover, the sperm cells tested cannot be used in assisted reproductive technologies, rendering the method unsuitable for real-time sperm selection [6].

To overcome these challenges, recent studies have explored non-invasive approaches based on artificial intelligence and label-free imaging techniques. For instance, Noy et al. [7] proposed a method that combines quantitative phase imaging with convolutional neural networks to predict SDF levels from unstained static sperm images. Their model achieved over 0.9 of sensitivity and specificity, demonstrating strong clinical potential. However, the reliance on static imaging and the need for specialized hardware may limit its widespread adoption.

In this study, we propose a practical and accessible approach that leverages standard phase-contrast microscopy, which is commonly available in fertility clinics, along with deep learning to predict SDF values from sperm motion videos. Previous studies have suggested that dynamic features of motile sperm are indicative of overall sperm quality [9], and deep learning has shown promise in the analysis

of medical video data [8].

We evaluate two deep learning architectures, namely 3D convolutional neural networks (3D CNN) and Transformer, using microscopic sperm motion videos. To improve input quality and emphasize motile sperm, we implement four preprocessing strategies, including resizing frames to 224×224 pixels, spatial cropping into non-overlapping clips, median filtering, and threshold-based binarization to extract sperm heads, and background subtraction to eliminate static components.

The prediction performance of deep learning models are assessed for both regression and binary classification tasks. For regression, prediction accuracy is evaluated using the root mean squared error (RMSE). Binary classification is conducted by applying the clinically relevant threshold of 10 to the predicted continuous SDF values, in order to distinguish between low and high levels of DNA fragmentation. The proposed non-invasive framework has the potential to reduce cost, time, and patient burden, and may serve as an effective pre-screening tool prior to in vitro fertilization (IVF).

## 2 Subjects and Materials

This study aims to develop a non-invasive deep learning framework that predicts sperm DNA fragmentation (SDF) from microscopic sperm motion videos, using corresponding clinically measured SDF values as ground truth. The dataset was collected from 224 adult male patients, aged between 20 and 55 years, who provided semen samples during standard clinical evaluations or infertility treatments. The study protocol was approved by the institutional ethics committee of Mie University Hospital, and informed consent was obtained from all participants.

To ensure data quality, samples were excluded if sperm were not visible under the microscope or exhibited minimal motility. Specifically, the average pixel intensity change was computed for each video, and 39 videos were excluded based on an average motion value below a predefined threshold, ensuring the exclusion of samples with insufficient sperm motility. As a result, a total of 185 videos were retained in the experiment.

All videos were acquired using a standard phase-contrast microscope, with a resolution of 672 by 522 pixels and a frame rate of 15 frames per second. Each video had a duration of 3 seconds, and one video was collected per subject. This acquisition setting was designed to reflect the constraints and conditions typical of real-world clinical environments, without requiring specialized equipment.

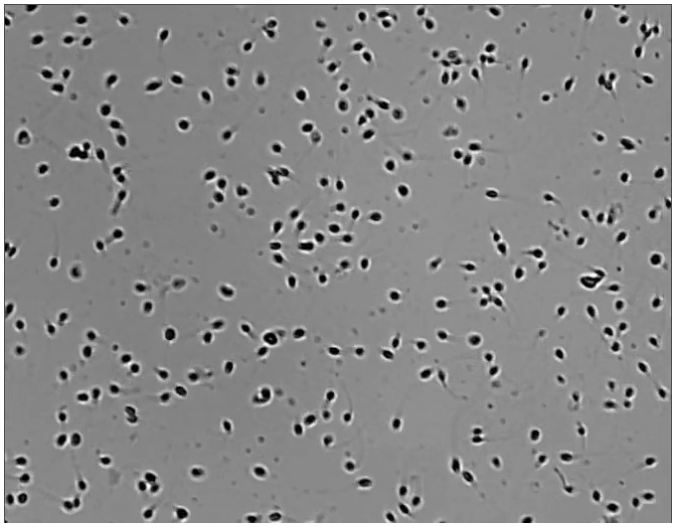

FIGURE 1. Example frame from a raw video recorded under phase-contrast microscopy (672 by 522 pixels).

SDF values were measured using the TUNEL-FACS method [5]. The resulting SDF scores ranged from 0 to 100 and were treated as continuous variables for regression analysis.

The dataset of 185 videos was partitioned into training (129 videos), validation (28 videos), and test (28 videos) subsets. The division was performed at the subject level to avoid data leakage, ensuring that no individual's data appeared in more than one subset. This approach supports a reliable evaluation of the model's generalization capability.

## 3 Proposed Method

This study presents a deep learning framework for predicting SDF values from microscopic sperm motion videos. Two model architectures, 3D ResNet and TimeSformer [10], are proposed along with four preprocessing strategies designed to capture both spatial and temporal features of sperm motility.

### 3.1 Data Preprocessing

Each input video had a resolution of 672×522 pixels, recorded over 3 seconds at 15 frames per second, resulting in a total of 45 frames. To enhance temporal variability and augment the training data, a sliding window technique

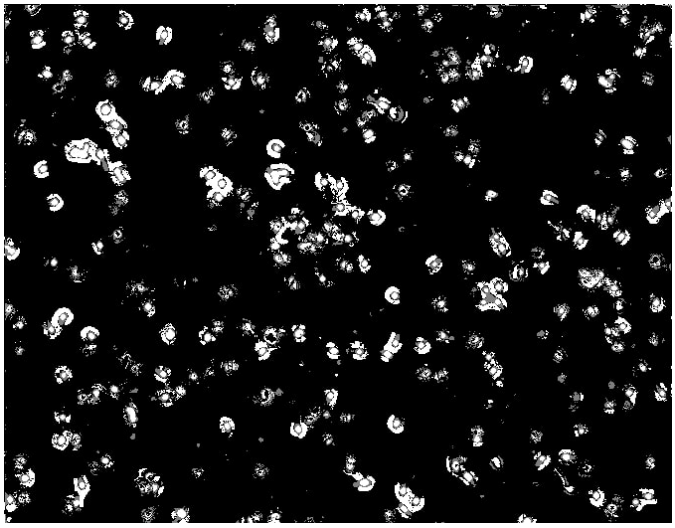

FIGURE 2. Example frame after applying Preprocessing D.

was applied to extract 8-frame clips with a stride of 4 frames. This process yielded 10 clips per video, each with a shape of 3 channels, 8 frames, and 672×522 pixels.

As the pre-processing for the extracted clip, four preprocessing strategies were explored:

- Preprocessing A: Resizing the original frame to 224×224 pixels size using bilinear interpolation.
- Preprocessing B: Dividing the original clip into six non-overlapping 224×224 pixels clips.
- Preprocessing C: Same as B, with additional median filtering and binarization using a threshold of 25 to isolate sperm heads.
- Preprocessing D: Same as C, with background subtraction to remove static components prior to binarization.

### 3.2 Model Architecture

For the comparison, the proposed method employs two model architectures, 3D ResNet and TimeSformer, and evaluates their performance. Each model receives input clips of 8 frames of 3-channel 224×224-pixel images and predicts a single continuous value representing the SDF. Both models were initialized with weights pre-trained on the Kinetics-400 video dataset and subsequently fine-tuned on the SDF dataset.

### 3.3 Training Procedure

The models were trained using mean squared error (MSE) as the loss function and the Adam optimizer. The batch size was set to 16. Learning rates were 0.0001 for 3D ResNet and 0.00005 for TimeSformer. Early stopping was applied when validation loss did not improve for 20 consecutive epochs.

### 3.4 Clip Aggregation and Video-Level Prediction

From each video, 10 overlapping clips were generated. In Preprocessing B, C, and D, these clips were extracted from each of the six spatially segmented regions, increasing clip diversity. To obtain a single prediction per video, the following aggregation methods are compared:

- Mean: The average of all clip-level predictions.
- Median: The median of all clip-level predictions.
- Best: The prediction closest to the actual SDF label.

The "best" aggregation is not applicable in real-world scenarios, as it assumes access to the ground truth. However, it provides an upper bound for performance estimation and reveals the potential of clip selection techniques.

### 3.5 Binary Classification

For binary classification, video-level predictions are thresholded at 10 based on expert clinical recommendations:

- Class 0 (Low SDF): SDF value less than or equal to 10, indicating high fertility potential.
- Class 1 (High SDF): SDF value greater than 10, indicating low fertility potential.

This threshold reflects a clinically meaningful distinction between fertile and subfertile patients and is consistent with existing fertility evaluation guidelines.

## 4 Experiment

We evaluated the proposed framework using two model architectures (3D ResNet and TimeSformer), four preprocessing strategies (A–D), and three video-level aggregation methods (mean, median, and best). The purpose of these experiments was to investigate the effect of each design choice on both regression and binary classification performance.

TABLE 1. Model performance under different preprocessing and aggregation conditions (excluding "Best" aggregation).
(A: Resize, B: Spatial crop, C: Spatial crop followed by binarization, D: Spatial crop followed by background subtracted binarization.)

| Model | Preprocessing | Aggregation | RMSE | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| 3D ResNet | A | Mean | 9.34 | 0.545 | 0.8 | 0.649 |
| | A | Median | 9.44 | 0.565 | 0.867 | 0.684 |
| | B | Mean | 8.37 | 0.565 | 0.867 | 0.684 |
| | B | Median | 8.37 | 0.545 | 0.800 | 0.649 |
| | C | Mean | 9.00 | 0.536 | 1.00 | 0.698 |
| | C | Median | 9.00 | 0.536 | 1.00 | 0.698 |
| | D | Mean | 9.00 | 0.536 | 1.00 | 0.698 |
| | D | Median | 8.71 | 0.519 | 0.933 | 0.667 |
| TimeSformer | A | Mean | 9.73 | 0.556 | 1.00 | 0.714 |
| | A | Median | 9.73 | 0.556 | 1.00 | 0.714 |
| | B | Mean | 8.31 | 0.476 | 0.667 | 0.556 |
| | B | Median | 8.26 | 0.476 | 0.667 | 0.556 |
| | C | Mean | 8.96 | 0.536 | 1.00 | 0.698 |
| | C | Median | 8.96 | 0.536 | 1.00 | 0.698 |
| | D | Mean | 9.22 | 0.536 | 1.00 | 0.698 |
| | D | Median | 9.03 | 0.536 | 1.00 | 0.698 |

TABLE 2. Model performance with "Best" aggregation (reference upper bound).

| Model | Preprocessing | Aggregation | RMSE | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| 3D ResNet | A | Best | 7.30 | 0.625 | 1.00 | 0.769 |
| | B | Best | 6.14 | 0.652 | 1.00 | 0.789 |
| | C | Best | 9.00 | 0.536 | 1.00 | 0.698 |
| | D | Best | 5.22 | 0.625 | 1.00 | 0.769 |
| TimeSformer | A | Best | 7.21 | 0.577 | 1.00 | 0.732 |
| | B | Best | 3.93 | 0.938 | 1.00 | 0.968 |
| | C | Best | 3.73 | 0.600 | 1.00 | 0.750 |
| | D | Best | 2.02 | 0.938 | 1.00 | 0.968 |

### 4.1 Experimental Setup

The dataset, comprising 185 videos, was partitioned into training (129 videos), validation (28 videos), and test (28 videos) subsets at the subject level to prevent data leakage. In Preprocessing A, this procedure yielded 10 clips per video. In contrast, the other preprocessing strategies produced 60 clips per video by dividing each frame into multiple spatial regions before temporal sampling. Models were trained on the training set, hyperparameters were optimized using the validation set, and the final evaluation was conducted using only the test set.

### 4.2 Evaluation Metrics

Two evaluation metrics were employed based on the task: regression and binary classification.

For regression, performance was assessed using RMSE, which quantifies the average deviation between the predicted and the ground truth SDF values. Lower RMSE values indicate more accurate predictions.

For classification, a threshold of 10 was used to divide SDF values into two classes: low and high fertility potential. Performance was evaluated using three standard metrics: precision, recall, and F1-score. Precision is the ratio of correctly predicted positive cases to all positive predictions. Recall is the proportion of actual positive cases correctly identified. F1-score is the harmonic mean of precision and recall, representing a balance between the two.

These metrics provide a comprehensive evaluation of model performance for both continuous regression and categorical classification tasks.

### 4.3 Results

Table 1 summarizes the regression and classification performance for all combinations of models, preprocessing strategies, and aggregation methods, excluding the "Best" aggregation, which assumes access to ground truth labels. To estimate an upper bound of model performance, the results using "Best" aggregation are separately presented


FIGURE 3. Scatter plot of predicted vs. ground truth SDF values using mean aggregation.

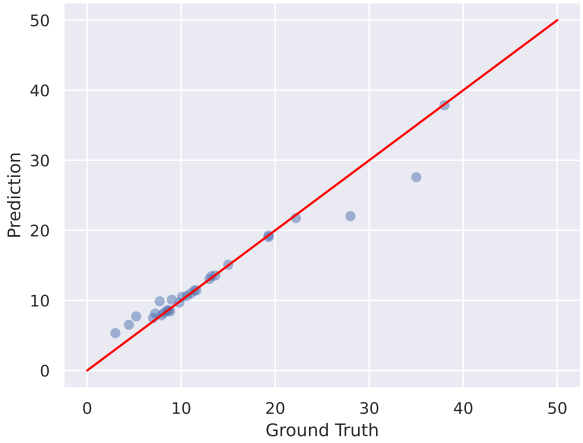


FIGURE 4. Scatter plot of predicted vs. ground truth SDF values using best aggregation.

in Table 2.

In the regression task, both models achieved lower RMSE values when spatial preprocessing (B, C, and D) was applied compared to simple resizing (A). Among the practical configurations, TimeSformer with Preprocessing B and median aggregation achieved the lowest RMSE.

In the classification task, the highest F1-score among feasible settings was 0.714, obtained by TimeSformer with Preprocessing A using either mean or median aggregation. Under the best aggregation setting, which selects the prediction closest to the ground truth and is not applicable in practice, TimeSformer achieved the highest overall performance with a RMSE of 2.02 and a F1-score of 0.968.

Scatter plots of predicted versus ground truth SDF values under mean aggregation and best aggregation are presented in Fig. 3 and 4, respectively. In both figures, the x-axis denotes the ground truth SDF value, and the y-axis represents the model prediction. The diagonal line indicates perfect prediction, facilitating visual assessment of model accuracy.

As seen in Fig. 3, predictions obtained by mean aggregation exhibit considerable variance around the perfect prediction line, particularly in the mid-range SDF values. In addition, a noticeable clustering of predictions is observed around the 10–20% SDF range, indicating a bias of the model toward lower SDF values. Conversely, Figure 4 demonstrates a substantially tighter distribution along the diagonal line, as it assumes the availability of ground truth during aggregation, thereby providing an upper bound on achievable performance.

## 5 Conclusion

This study proposed a deep learning-based framework for estimating sperm DNA fragmentation (SDF) using phase-contrast microscopic videos. We evaluated two model architectures—3D ResNet and TimeSformer—under four different preprocessing strategies and three clip aggregation methods.

In regression tasks, TimeSformer combined with Preprocessing B and median aggregation achieved the best performance among practical settings. In binary classification, strong performance was also observed when using Preprocessing A with both mean and median aggregations. These findings suggest that TimeSformer is effective across both regression and classification tasks. Furthermore, the similarity in results between mean and median aggregation indicates that both methods are suitable for summarizing clip-level predictions.

The best overall performance was observed with TimeSformer using Preprocessing D and the best aggregation method (RMSE of 2.02 and F1-score of 0.968). However, this method relies on selecting the clip with the closest prediction to the ground truth, requiring prior knowledge of the actual SDF value. While not feasible in clinical practice, this result highlights the potential of identifying informative video segments. With further enhancements—such as attention-based or confidence-driven selection

mechanisms—this approach may evolve into a practical clinical tool.

Importantly, the proposed method provides a stain-free, non-invasive alternative to conventional SDF testing, potentially reducing cost, time, and patient burden.

For future work, we plan to extend this framework to a multimodal learning approach by integrating clinical data—such as patient age, body metrics, smoking history, and semen analysis parameters (e.g., sperm concentration, motility, morphology)—with video-based features. This integration is expected to improve prediction accuracy and enhance clinical interpretability by capturing a broader spectrum of factors associated with sperm DNA integrity.

## References


[1] World Health Organization, "1 in 6 people globally affected by infertility: WHO," Geneva, Switzerland, Apr. 2023.

[2] S. W. Andrabi, A. Ara, A. Saharan, M. Jaffar, N. Gugnani, and S. C. Esteves, "Sperm DNA fragmentation: Causes, evaluation and management in male infertility," Arab Journal of Urology, vol. 20, no. 2, pp. 103–122, 2022.

[3] L. Simon, A. Castillo, R. L. Lewis, A. R. Emery, and A. Agarwal, "Sperm DNA damage measured by the alkaline Comet assay as an independent predictor of male infertility," Fertility and Sterility, vol. 97, no. 1, pp. 128–134, 2017.

[4] A. Zini and H. Sigman, "Are tests of sperm DNA damage clinically useful? Pros and cons," Journal of Andrology, vol. 30, no. 3, pp. 219–229, 2009.

[5] R. Sharma, C. Iovine, A. Agarwal, and R. Henkel, "TUNEL assay – Standardized method for testing sperm DNA fragmentation," Andrologia, vol. 52, no. 3, Art. no. e13612, 2020.

[6] A. Agarwal, M. Majzoub, A. Baskaran, and R. P. Selvam, "Clinical utility of sperm DNA fragmentation testing: Practice recommendations based on clinical scenarios," Translational Andrology and Urology, vol. 6, Suppl. 4, pp. S574–S579, 2017.

[7] D. Noy, M. Mutzafi, Y. Golan, A. Arad, and Z. Zalevsky, "Sperm-cell DNA fragmentation prediction using label-free quantitative phase imaging and deep learning," Advanced Intelligent Systems, vol. 5, no. 3, Art. no. 2200231, 2023.

[8] A. Esteva et al., "A guide to deep learning in healthcare," Nature Medicine, vol. 25, pp. 24–29, 2019.

[9] M. Kato, S. Makino, H. Kimura, T. Ota, T. Furuhashi, and Y. Nagamura, "Sperm motion analysis in rats treated with adriamycin and its applicability to male reproductive toxicity studies," Journal of Toxicological Sciences, vol. 47, no. 1, pp. 1–9, 2022.

[10] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," Proceedings of the International Conference on Machine Learning (ICML), vol. 139, pp. 813–824, 2021.