KFUPM
Journal of Undergraduate Research International

Diamond Open Access 130

Article Industrial Engineering

Video Based Sperm Motility Grading and YOLOv8 Detection: A Tracking-Free Approach to Automated Semen Analysis

Abdulrahman Alshaikh and Anis Elgabli*

Cite https://doi.org/10.64589/juri/221335

Submitted: November 27, 2025 **Revised:** April 04, 2026 **Accepted:** April 05, 2026

ABSTRACT

This paper presents a unified deep learning framework for automated sperm counting and motility assessment using Computer-Aided Semen Analysis. Conventional semen analysis is predominantly manual, making it time consuming, subjective, and susceptible to inter-observer variability, whereas existing automated systems often struggle in the presence of debris and morphological abnormalities. To address these limitations, the You Only Look Once version 8 (YOLOv8) Nano object detection model was fine-tuned using carefully annotated clinical data to accurately detect and count only valid sperm cells that are characterized by the presence of both the head and tail, while effectively suppressing non-sperm objects in noisy and unclean semen samples. In addition, a tracking-free approach for motility assessment is proposed by reformulating the problem as a video-level regression task. Instead of relying on computationally expensive and error-prone tracking-by-detection pipelines, the proposed method predicts the overall sperm motility directly from video sequences by modeling global optical flow patterns and temporal dynamics. To achieve this, three spatiotemporal deep learning architectures were developed and systematically compared: a convolutional neural network (CNN)-based optical flow model, CNN–long short-term memory (LSTM) temporal model, and semi-supervised 3D autoencoder designed to learn latent motion representations from limited labeled data. All models were trained and validated on a real-world dataset acquired from a private fertility clinic, reflecting clinically relevant imaging conditions. The experimental results demonstrated that the proposed framework achieved robust sperm detection and reliable motility estimation (with the CNN–LSTM model achieving the best performance). The proposed approach offers an effective alternative to manual analysis.

Keywords: automated semen analysis, semen count, motility, deep learning, YOLOv8

1. INTRODUCTION

Male infertility has emerged as a significant global health concern, with sperm concentration and motility being recognized as the primary biomarkers for clinical diagnosis and fertility prognosis. Driven by the rising prevalence rates, the urgency to counter this issue is further underscored by recent longitudinal assessments reporting a significant decline in sperm concentration over the last five decades[1]. In the wake of significant advancements in automated technologies, the World Health Organization has reaffirmed manual microscopic evaluation as the "gold standard" for semen analysis[2]. This continued reliance is largely attributed to the absence of standardization in conventional Computer-Aided Semen Analysis (CASA) systems, which fail to properly identify sperm cells from other similar cellular artifacts such as leukocytes and cellular debris in a medical sample[3,4]. In response to these limitations, modern deep learning (DL) frameworks have demonstrated superior ability to suppress biological artifacts and debris, ensuring that automated counts are accurate even in noisy or unclean clinical environments[5].

The advent of DL has paved the way for overcoming these shortcomings by facilitating effective learning directly from microscopic images[6] and videos[7]. Within the context of sperm detection, the You Only Look Once (YOLO) model has been recognized as the foremost detection technique, owing to its ability to process data in real time[8]. Although preceding models have successfully detected sperms within scanned images, they often require considerable processing power, thus hindering their use in edge computing[9]. Recent studies have identified YOLO version 8 (YOLOv8) as an improved version that includes an anchor-free detection head and an enhanced fusion technique to effectively boost detection accuracy under noisy conditions. The evolution of YOLO architecture, culminating in the anchor-free detection head of YOLOv8, has significantly improved the ability of DL models to detect small objects in noisy environments[10]. This study focused on the YOLOv8 Nano model[11]. To rationalize the need for this reduced architecture, the goal was to not only accelerate detection processing but also to ensure that this model has the ability to maintain high detection accuracy[12]. Our contribution lies in the fine-tuning of the YOLOv8 Nano architecture using carefully annotated data to enhance its capability to accurately detect and count only valid sperm cells, which are characterized by the presence of both the head and tail, while effectively suppressing non-sperm objects and biological

artifacts. This approach enables rapid inference while preserving high detection accuracy, even in unclean or abnormal semen samples, where debris, cells, or morphological irregularities may otherwise lead to misclassification.

Notably, a critical gap remains in the automation of motility assessments. The prevailing DL paradigm typically relies on "tracking-by-detection," where the system must detect, identify, and link every individual sperm cell across hundreds of consecutive frames to calculate kinematic parameters. Although conceptually straightforward, this approach is computationally expensive and prone to identity-switching errors when the trajectories of the sperms intersect (i.e., occlusion occurs) or when detection fails within a single frame. Furthermore, the training of robust trackers requires datasets with exhaustive frame-by-frame coordinate annotations, which are scarce and expensive to generate. To overcome these bottlenecks, this study developed a unified resource-efficient framework that decouples counting from motility analyses. Instead of relying on error-prone object tracking, a holistic video regression task was used for motility assessment. By analyzing global optical flow patterns and temporal dynamics, the developed model could predict the overall motility percentage directly from video sequences. This approach aligns with the aggregate metrics provided by legacy CASA systems [13]; moreover, it also leverages the feature-learning characteristics of deep neural networks to handle noise and occlusion more effectively.

A custom dataset curated from a private fertility clinic was used in this study, ensuring that the developed models were trained and validated on real-world data characterized by varying contrasts and particulate noise, rather than idealized public datasets. The contributions of this study are threefold, which are as follows:

1. **Robust Counting:** YOLOv8 Nano was fine-tuned to accurately detect sperm cells in high-debris environments without explicit negative-class annotation.
2. **Efficient Motility Grading:** A regression-based motility framework that eliminates the need for frame-level tracking labels is proposed.
3. **Architectural Comparative Analysis:** Three distinct spatiotemporal architectures were evaluated: a convolutional neural network (CNN) with optical flow for explicit motion modeling, CNN–long short-term memory (LSTM) model for capturing long-term temporal dependencies, and semi-supervised 3D Autoencoder designed to extract latent motion features from limited labeled data.

2. METHODOLOGY

All experiments were conducted in a standardized DL environment using Python 3.12.12 and the PyTorch framework to ensure technical precision and full reproducibility. The YOLOv8 models were implemented using the Ultralytics 8.4.12 library. All model training and testing were performed on a Linux-based server equipped with an NVIDIA Tesla T4 GPU to maintain a consistent computational performance across all architectures. The framework was structured around two primary tasks: sperm counting through object detection and sperm motility analysis using a comparative study of three distinct DL architectures.

2.1. Data Collection and Preparation. The data for this study were obtained from a custom set of patient sperm videos and images recorded by a fertility physician. The preparation process was divided into two parts to separately handle the counting and motility.

For the sperm counting, five frames were extracted from 32 videos (160 images in total); then, the images were manually labeled using Roboflow, as shown in Figure 1. Bounding boxes were only placed around valid sperm cells that clearly showed both a head and tail. This is important for helping the model learn the difference between a real sperm and other objects, such as white blood cells. To ensure that the performance of the model was evaluated fairly and prevent data leakage, the dataset was split into training, validation, and test sets in 80:10:10 ratio. Then, data augmentation techniques, such as rotation and flipping, were applied to make the dataset larger and more diverse, increasing the total training pool to 400 images.

A total of 279 videos were used for the motility task. Each video contained 30 frames and was paired with a motility percentage recorded by a doctor. The dataset was partitioned into 80% for training and 20% for testing to maintain consistent evaluation across all motility architectures.

2.2. Sperm Counting Model: YOLOv8. The YOLOv8 model was selected for the sperm counting task in this study because it is a well-known model for real-time object detection and provides high precision for small objects. The model uses convolutional layers integrated within a modified Cross Stage Partial backbone, utilizing Cross Stage Partial with two convolution flow modules to extract rich spatial features from the input images. Unlike classical multistage detection algorithms, YOLOv8 is a single-stage detector that analyzes the entire image in a single pass by concurrently predicting the bounding boxes and computing the confidence scores for the detected sperm cells. The total sperm count was derived by aggregating the individual detections that exceeded a specific confidence threshold, thereby improving the overall reliability of the system for rapid clinical applications.

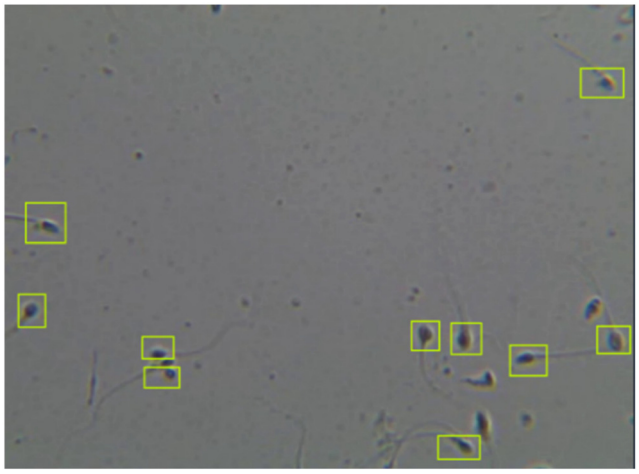

Figure 1. Manually labeled sperm cell image

2.3. Sperm Motility Architectures. To evaluate sperm motility, three different DL architectures were compared, each offering a different approach for processing temporal data. These included a CNN with optical flow (Model A), a CNN–LSTM hybrid model (Model B), and an autoencoder with latent feature regression to analyze compressed representations of movement (Model C). A comparison of these three models aided in determining the strategy that best captured the complex dynamics of sperm cells.

2.3.1. Model A: CNN with Optical Flow. The first approach used a CNN with an optical flow model. This method calculated the motion vectors between consecutive frames to capture temporal information rather than relying solely on static image features. These optical flow maps, which represented the direction and speed of movement, were then processed using a CNN to classify the motility grade. This technique enabled the model to track fast-moving sperm with high accuracy by focusing specifically on movement patterns.

2.3.2. Model B: CNN-LSTM. The second approach used a hybrid CNN–LSTM architecture designed to process raw video sequences directly. In this model, a time-distributed CNN served as a spatial feature extractor that converted individual video frames into dense feature vectors. These vectors were then fed sequentially into the LSTM network. The LSTM component is important for capturing long-term temporal dependencies, enabling the system to remember the historical trajectory of a sperm cell and classify its motility based on the temporal evolution of its movement.

2.3.3. Model C: 3D Autoencoder with Latent Feature Regression. In the third model, a two-stage semi-supervised architecture was used to handle the issue of limited labeled data. This method separated learning from motility predictions.

- **Stage 1: Unsupervised spatiotemporal feature learning:**

The first stage focused on unsupervised spatiotemporal feature learning. By employing a 3D convolutional autoencoder, compressed representations of sperm motion patterns were directly extracted from the raw video data. The architecture of the first stage comprises the following:

  - **Encoder:** The encoder network processes the input video sequence using 3D convolutional layers to extract spatiotemporal features, and downsamples the data into a compact latent vector via global average pooling.
  - **Decoder:** The decoder attempts to reconstruct the original optical flow sequence from a latent vector. The model was trained to minimize the reconstruction loss [mean squared error (MSE)], forcing the latent vector to capture the most essential dynamics of sperm movement.

- **Stage 2: Supervised motility regression:**

Once the autoencoder was trained, the encoder was frozen and used as a feature extractor. The learned latent vectors that encapsulated the high-level motion characteristics of each sample were fed into a separate multilayer perceptron (MLP) regressor. This lightweight regression network was trained to map the latent features directly to the motility percentage, effectively leveraging the robust features learned during the unsupervised stage.

3. EXPERIMENTAL WORK

This section describes the experimental procedures used to implement, train, and test the sperm count and motility models. The following subsections describe the detailed configurations of these models.

3.1. Experimental Setup for Sperm Counting. The sperm counting model was trained by fine-tuning YOLOv8 to the carefully annotated clinical data. The specific configuration is as follows:

- **Model Architecture:** The YOLOv8 Nano model was used for its lightweight architecture.
- **Input Resolution:** Images were processed at a resolution of 640 ×640 pixels.
- **Training Duration:** The model was trained for 50 epochs, which was sufficient to provide loss convergence ability without overfitting.
- **Batch Size:** A batch size of eight was selected.

3.2. Experimental Setup for Sperm Motility. To assess the temporal analysis capabilities of DL frameworks, the aforementioned three models were evaluated on the same dataset. They all shared a common dataset split: training 80% and testing 20%. The detailed configuration is described subsequently.

3.2.1. Model A: CNN with Optical Flow. The first model used a dense optical flow with a 3D CNN. The implementation information includes the following.

- **Input Processing:** The dense optical flow was computed using the Farneback algorithm. The input sequence comprised 30 frames, with each frame resized to 128 ×128 pixels.
- **Normalization:** Flow vectors were normalized to [0, 1] values using MinMax normalization.
- **Architecture:** The model used 3D convolutional layers (32, 64, and 128 filters) with rectified linear unit (ReLU) activation to extract spatiotemporal motion features.
- **From Motion to Prediction:** To translate the complex motion patterns into a single result, a global average pooling 3D layer was used to condense the data, which were then fed into a 64-unit fully connected (dense) layer, followed by a final output neuron to calculate the motility percentage.
- **Regularization:** To prevent overfitting, L2 regularization was performed on the convolutional and dense layers, and a dropout layer (0.5) was applied to the prefinal output.
- **Training Settings:**
  - **Optimizer:** Adam.
  - **Loss Function:** MSE.
  - **Epochs:** 50.
  - **Batch Size:** 8.

3.2.2. Model B: CNN_LSTM:. In the second model, a hybrid CNN and LSTM network was used. Unlike the 3D CNN, this model treated optical flow frames sequentially to determine temporal dependencies. The input shape was (30, 128, 128, 2), which denotes the sequence length (30), spatial dimensions (128 ×128), and flow vectors (2 represents (x, y)).

- **Spatiotemporal Architecture:**
  - **Time-Distributed CNN:** Spatial features of each frame were encoded by attaching TimeDistributed wrappers to the Conv2D layers (16 and 32 filters). Each convolutional block was followed by MaxPooling2D and BatchNormalization to reduce the dimensionality and normalize the activations, respectively.
  - **Sequence Modeling:** To bridge the spatial and temporal components, a global average pooling 2D layer was used to condense the spatial features of each individual frame into a vector. These vectors were then fed into an LSTM layer with 64 units, which enabled the model to learn the temporal dependencies and motion patterns across the 30-frame sequence.
  - **Regression Head:** The LSTM output entered a fully connected dense layer (64 units) with ReLU activation, followed by a dropout layer (0.3) and single neuron output for motility prediction.
- **Training Configuration:**
  - Optimizer: Adam.
  - Loss Function: MSE.
  - Epochs: 50.
  - Batch Size: 8.

3.2.3. Model C: 3D Autoencoder. The third model used a 3D Convolutional autoencoder with a semi-supervised approach. First, the model learned unannotated data to a compressed feature of sperm motion patterns, followed by a regression network in which the learned features were assigned motility scores.

- **Input Configuration:** The input comprised 30 optical flow frames. For the 3D reconstruction process, the spatial resolution was downscaled to 64 ×64 pixels.
- **Autoencoder Architecture:**
  - **Encoder:** To capture the spatiotemporal features, the encoder used three Conv3D blocks (32, 64, and 128 filters). This sequence was followed by MaxPooling3D after each block, which downsampled the spatial dimensions. A GlobalAveragePooling3D layer was subsequently used to extract a compact latent feature vector.
  - **Decoder:** The decoder attempted to reconstruct the original optical flow sequence using mirrored Conv3D and UpSampling3D layers, culminating in a sigmoid activation function.
  - **Unsupervised Training:** In this stage, the model was trained as an autoencoder to reconstruct the input optical flow sequences. By forcing the decoder to rebuild the original data using mirrored Conv3D and UpSampling3D layers, the model learned to condense complex movements into a meaningful latent space. This step is essential because it allows the encoder to capture robust patterns of sperm motion independently before manual labels are introduced for regression.
- **Motility Regression:**
  - After training, the encoder was frozen and used to extract the latent features from the dataset.
  - A separate MLP regressor (dense layers of 64 and 32 units with dropout) was trained on these latent features for 80 epochs to predict the final motility percentage.
  - This two-step process allows the model to leverage robust feature learning even with limited labeled data.

4. RESULTS AND DISCUSSION

The experimental results demonstrated the efficacy of using YOLOv8 for object detection and the comparative strengths of the three models used for motility analysis, which are presented in detail subsequently.

4.1. Sperm Counting Results (YOLOv8). The YOLOv8 model was evaluated using a test dataset to determine its precision in detecting and counting sperm cells of varying densities. The evaluation yielded promising results with the model achieving a mean absolute error (MAE) of 2.9. This low error indicated that, on average, the model prediction deviated from the manual count by an expert by approximately three sperm cells per image. Furthermore, the root mean squared error (RMSE) was calculated to be 3.66, confirming that the model maintained consistent performance without significant outliers or extreme failures in detection.

In terms of total population analysis across the tested batch, the ground-truth manual count identified 315 sperm cells, whereas the YOLOv8 model successfully detected 355 cells. The evaluation yielded precision, recall, and mean average precision at 0.50 intersection over union (mAP@50) of 0.641, 0.714, and 0.668, respectively. The difference in total counts is likely a tradeoff; while the model occasionally missed overlapping or faint sperm cells (affecting recall), it also tended to pick up certain circular biological artifacts or debris as potential detections (affecting precision). Despite these minor misclassifications in noisy clinical samples, the MSE remained low at 13.4, reinforcing the reliability of the system as a tool for clinical screening. To provide a clearer picture of how these metrics were calculated, a detailed breakdown of true positives, false positives, and false negatives is presented in Table 1.

The YOLOv8 model was robust against background noise and other biological artifacts. The sample image (Figure 2) shows several spherical objects (white blood cells) among the sperm. Because of their similarity in size, traditional image processing methods are prone to identifying these round cells as sperm heads. The proposed model demonstrated the ability to differentiate real sperms from these particulate artifacts. The model produced bounding boxes for only objects presenting the typical sperm morphology with confidence scores ranging from 0.69 to 0.80, indicating a filter for circular debris and other non-sperm cells based on the final model results (Figure 2). This narrow-spectrum detection capability is crucial for preventing false positives and obtaining clinically meaningful semen sample counts.

4.2. Model Evaluation for Sperm Motility. To assess sperm motility, the three DL models were evaluated. The objective was to identify the method that can most effectively capture the spatiotemporal dynamics of sperm movement given the limited dataset size. The performance of each model was evaluated

Table 1. *Confusion Matrix Breakdown for Sperm Detection*

Metric	Value	Description
Ground Truth (Instances)	315	Total actual sperm cells identified by an expert
True Positives (TP)	225	Correctly identified sperm cells (recall * instances)
False Positives (FP)	130	Objects wrongly identified as a sperm (debris/artifacts)
False Negatives (FN)	90	Actual sperm cells missed by the model
Total Detections	355	Total number of bounding boxes generated by the model (TP + FP)

using the MSE and RMSE to measure the variance in predictions, as well as the MAE to determine the average deviation from the manual ground truth in percentage points.

4.2.1. Model A: CNN with Optical Flow Results. Model A was evaluated to assess its ability to predict motility based on immediate pixel displacement. This model yielded the second strongest performance among the tested models, achieving a MAE of 10.47%. This indicates that, on average, the motility prediction of the model deviated from the manual expert assessment by approximately 10 percentage points. Moreover, the model recorded MSE and RMSE of 183.02 and 13.53, respectively, reflecting its capability in effectively handling the variance in sperm movement speeds.

From Figure 3, we can observe that the training process exhibits considerable volatility in the initial stages, which is characterized by sharp spikes in the validation loss. After approximately 30 epochs, the model stabilizes, demonstrating successful convergence. The superior performance of this model is due to the Farneback algorithm, which explicitly computes motion features, allowing the network to focus more on the movement dynamics rather than static image noise.

4.2.2. Model B: CNN-LSTM Results. Model B achieved the highest performance among all tested models. The evaluation yielded an MAE of 9.32%, significantly outperforming the pure CNN approach. The MSE was 159.78, with a corresponding RMSE of 12.64.

As shown in Figure 4, the training process exhibits rapid convergence. The CNN–LSTM loss curve drops sharply within the first 10 epochs and stabilizes quickly. This stability suggests that the LSTM component successfully captured long-term temporal dependencies in the sperm movement. By effectively modeling the historical trajectory of the optical flow vectors, the model could distinguish between consistent progressive motility and random motion more accurately than that achieved using only frame-by-frame analysis. The low MAE of 9.32% confirms that combining spatial feature extraction with temporal sequence modeling provides the most robust approach for automated motility assessment.

4.2.3. Model C: Autoencoder with Latent Regression Results. Model C used a semi-supervised approach, achieving an MAE of 11.35%. The MSE was 194.94, with an RMSE of 13.96.

The performance of these two stages is supported by the training curves. Figure 5 shows the unsupervised training of the autoencoder. The reconstruction loss (MSE) decreases smoothly and aligns closely with the validation loss, indicating that the model successfully learned to compress the complex spatiotemporal dynamics of the sperm videos into a compact latent representation without overfitting.

Figure 6 shows the training results of the MLP regressor on these extracted features. The curve demonstrates a significantly rapid convergence, dropping to a low error rate within the first 10 epochs. This suggests that the latent features learned by the autoencoder were highly discriminative and robust, effectively summarizing the motion patterns in a manner that made the regression task straightforward for the MLP.

4.3. Discussion and Comparative Analysis. The comparative evaluation of the three models revealed critical insights

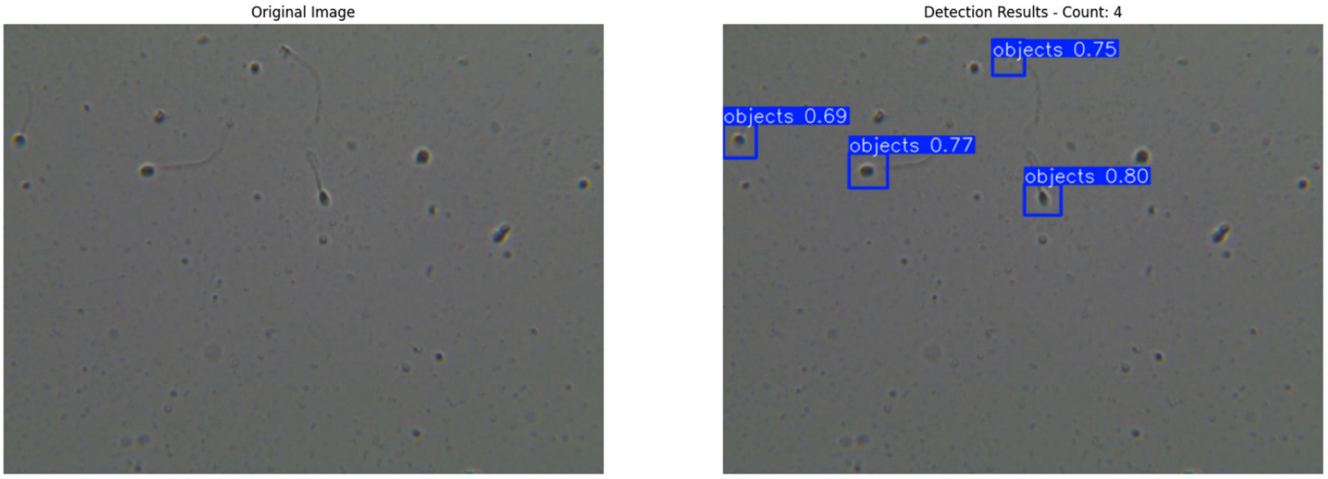

Figure 2. Example of sperm detection: Original and detection results

Table 2. *Performance Comparison of Motility Analysis Models*

Model Architecture	MAE	RMSE	MSE
Model A: CNN + Optical Flow	10.47%	13.53	183.02
Model B: CNN–LSTM	9.32%	12.64	159.78
Model C: Autoencoder + MLP	11.35%	13.96	194.94

into the effective modeling of sperm motility. A summary of the performance metrics is presented in Table 2.

4.3.1. Analysis of Results.

- **Superiority of Spatiotemporal Modeling:** Model B achieved the highest accuracy, with the lowest MAE of 9.32%. This performance demonstrates that when appropriately optimized with global average pooling to prevent overfitting, the LSTM network successfully captures the long-term temporal dependencies of sperm movement.
- **Robustness of Explicit Motion Features:** Model A remained highly competitive, securing the second-best performance with an MAE of 10.47%. This confirms that explicitly calculating motion vectors using the Farneback algorithm provides a clean, noise-free input signal. While it lacks the long-term memory of the LSTM, the direct correlation between optical flow intensity and motility grade proves to be a reliable and computationally efficient method for general motility assessment.
- **Viability of Semi-Supervised Learning:** Model C yielded an MAE of 11.35%. Although it ranked third, its performance validates the potential of unsupervised feature learning. The autoencoder successfully compressed complex video data into meaningful latent representations without requiring frame-level labels. This suggests that in clinical scenarios where expert annotation is scarce or expensive, a semi-supervised approach remains a scientifically valid and effective alternative.

4.3.2. Advantages of tracking-free regression. Beyond individual model performance, this framework establishes the viability of a tracking-free approach to motility analysis. Traditional CASA systems often fail during "identity switching" or presence of occlusions in high-density samples. By framing motility as a global video-level regression task, the proposed method remained robust against these common tracking errors while significantly reducing computational overhead. This efficiency can facilitate high-speed analysis on low-cost clinical hardware and edge devices, offering a practical advantage over more hardware-intensive tracking pipelines.

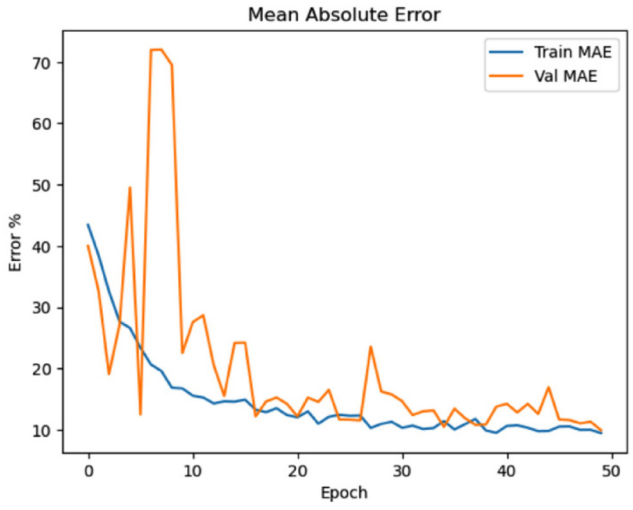

Figure 3. Training and validation MAE for the CNN optical flow model

4.3.3. Strategic advantages of a tracking-free approach. Beyond individual model metrics, this framework demonstrates the clear benefits of moving away from traditional sperm tracking. Standard systems often fail in crowded samples owing to overlapping paths. By analyzing motility as a global, video-level regression task, the proposed method avoids these common errors while significantly lowering the computational burden. This efficiency facilitates high-speed analysis even on budget-friendly clinical hardware.

5. CONCLUSIONS AND FUTURE WORK

5.1. Conclusion. This study explored the application of advanced DL models to enhance the accuracy and efficiency of CASA, thereby addressing the critical limitations of subjective manual evaluation. It primarily focused on two main components of fertility diagnostics: sperm counting and motility. In particular, the YOLOv8 model was fine-tuned for sperm detection, and a comparative analysis was conducted on three different models (CNN with optical flow, CNN–LSTM, and semi-supervised autoencoder) to evaluate temporal movement dynamics. The results demonstrated that YOLOv8 achieved superior performance in sperm counting, with an MAE of 2.9. A significant advantage of this model is its robustness against background noise, as it successfully distinguished sperm cells from non-sperm biological artifacts such as leukocytes and round cells. This capability addresses a major deficiency of traditional threshold-based techniques, which often struggle with noisy, low-contrast clinical samples.

In motility analysis, the comparative study revealed that the CNN–LSTM model was the most effective approach, achieving the lowest MAE of 9.32%. This finding suggests that combining spatial feature extraction with temporal sequence modeling allows the system to capture long-term movement trajectories more effectively than using only frame-by-frame analysis. The CNN with optical flow model also demonstrated a robust performance with an MAE of 10.47%, confirming that explicitly calculating motion vectors provides a clean and reliable signal for motility grading. Meanwhile, the semi-supervised autoencoder showed promising results with an MAE of 11.35%, validating the potential of unsupervised feature learning in scenarios where labeled medical data are scarce. Collectively, these findings reinforce the feasibility of integrating AI-driven tools into clinical practice to offer a standardized, objective, and reproducible alternative to manual analysis.

5.2. Future Work. Although the proposed models demonstrated a strong baseline performance, several avenues for technical optimization and clinical expansion remain unexplored. Future research should prioritize the systematic evaluation of different iterations within the YOLOv8 family. Although the current study utilized the nano variant for its real-time efficiency, subsequent work should benchmark medium and large versions

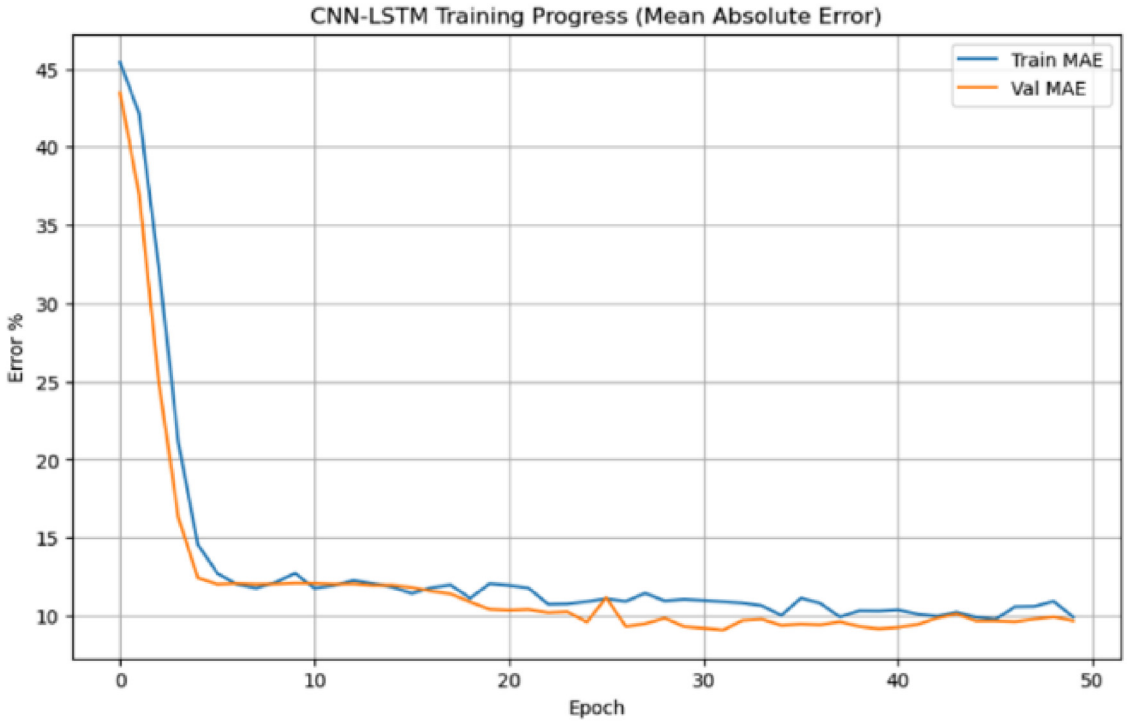

Figure 4. Training and validation MAE for the CNN-LSTM model

to quantify the tradeoffs between computational latency and detection precision. This comparative analysis will serve as a foundation for a cost-benefit analysis, aiding in the identification of the optimal architecture for specific use cases, such as deploying high-accuracy models on cloud servers versus lightweight models for edge devices and tablet-based microscope interfaces.

To further address the challenges of detecting small, densely packed sperm cells, the integration of Slicing Aided Hyper Inference, which is an open-source framework that has been designed to significantly improve the detection and segmentation of small or dense objects in high-resolution images, will be crucial. This technique involves partitioning high-resolution microscopic images into overlapping slices during the inference stage, allowing the model to detect small objects with greater visual fidelity before merging the results. Concurrently, this pipeline is enhanced using super-resolution preprocessing. By employing DL-based upscaling techniques to increase the resolution of raw video feeds before analysis, the system can recover fine morphological details from lower-quality optical hardware, thereby improving the overall robustness of both counting and motility assessments.

Beyond these technical optimizations, the next phase of this research should prioritize external validation and patient-level generalization. Because the current dataset was derived from a single clinical source, future studies should incorporate multicenter data to ensure the robustness of the framework across different laboratory settings and optical hardware. We also intend

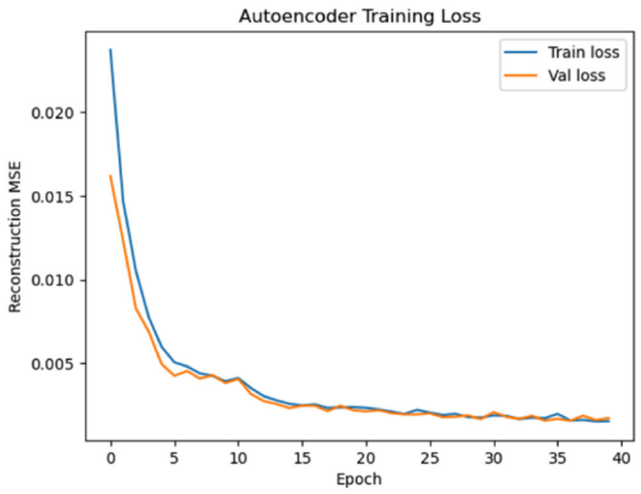

Figure 5. Autoencoder Training Loss (MSE)

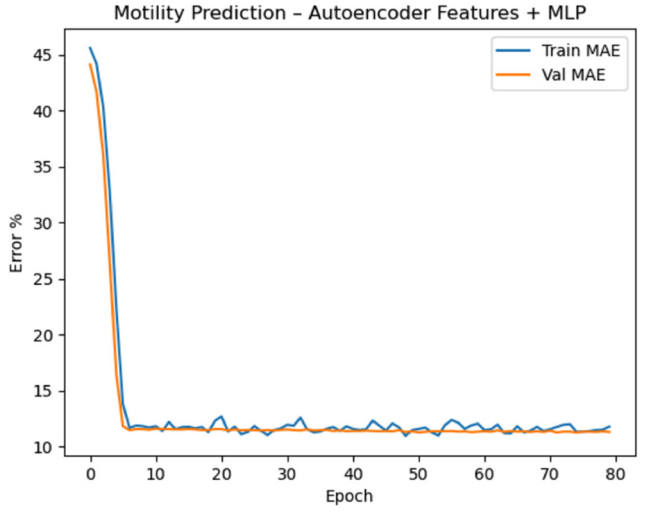

Figure 6. Motility prediction error (MAE) for the MLP regressor trained on extracted latent features

to conduct a more rigorous statistical analysis, including the calculation of confidence intervals for all performance metrics and a direct comparison with established computational baselines. By expanding the dataset to include a wider diversity of patient samples, we aim to transition this exploratory proof-of-concept into a clinically validated tool that is ready for real-world diagnostic deployment.

Finally, future iterations should aim to create a comprehensive diagnostic tool by incorporating sperm morphology classification to detect head and tail defects directly in the detection pipeline. Long-term studies comparing the outputs of the AI system with manual counts will be essential to validate the broader applicability of these advanced computer vision techniques in routine fertility diagnostics.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Abdulrahman Alshaikh – *Undergraduate Student, Department of Industrial and System Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0009-0009-9597-1064
Email: s202031920@kfupm.edu.sa

Corresponding Author

Anis Elgabli – *Research Mentor, Department of Industrial and System Engineering, Interdisciplinary Research Center for Communication Systems and Sensing (IRC-CSS), King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0000-0001-5012-2370
Email: anis.elgabli@kfupm.edu.sa

ACKNOWLEDGEMENTS

The authors express their sincere gratitude to the Department of Industrial and Systems Engineering and Interdisciplinary Research Center for Communication Systems and Sensing (IRC-CSS). at King Fahd University of Petroleum and Minerals for their unwavering support throughout this study.

REFERENCES

(1) Levine, H. et al. Temporal trends in sperm count: a systematic review and meta-regression analysis of samples collected globally in the 20th and 21st centuries. *Hum. Reprod. Update* **29**, 157–176 (2023).

(2) Boitrelle, F. et al. The sixth edition of the WHO manual for human semen analysis: a critical review and SWOT analysis. Life **11**, 1368 (2021).

(3) Hürland, M. et al. The use of machine learning methods to predict sperm quality in Holstein bulls. *Theriogenology* **197**, 16–25 (2023).

(4) Finelli, R. et al. The validity and reliability of computer-aided semen analyzers in performing semen analysis: a systematic review. *Transl. Androl. Urol.* **10**, 3069–3079 (2021).

(5) Javadi, S. & Mirroshandel, S. A. A novel deep learning method for automatic assessment of human sperm images. *Comput. Biol. Med.* **109**, 182–194 (2019).

(6) Köse, M. et al. Manual versus computer-automated semen analysis. *Clin. Exp. Obstet. Gynecol.* **41**, 662–664 (2014).

(7) Hicks, S. A. et al. Machine learning-based analysis of sperm videos and participant data for male fertility prediction. *Sci. Rep.* **9**, 16770 (2019).

(8) Zhang, C., Zhang, Y., Chang, Z. & Li, C. Sperm YOLOv8E-TrackEVD: a novel approach for sperm detection and tracking. *Sensors* **24**, 3493 (2024).

(9) Suleman, M. et al. A review of different deep learning techniques for sperm fertility prediction. *AIMS Math.* **8**, 16360–16416 (2023).

(10) Terven, J. & Cordova-Esparza, A. A comprehensive review of YOLO: from YOLOv1 to YOLOv8 and beyond. *arXiv preprint arXiv:2304.00501* (2023).

(11) Li, C. et al. An efficient advanced YOLOv8 framework for sperm motility detection. *J. Assist. Reprod. Genet.* **42**, 3095–3108 (2025).

(12) Guo, Y. et al. Automated deep learning model for sperm head segmentation, pose correction, and classification. *Appl. Sci.* **14**, 11303 (2024).

(13) Choi, Jw. et al. An assessment tool for computer-assisted semen analysis (CASA) algorithms. *Sci. Rep.* **12**, 16830 (2022).