ADVANCED INTELLIGENT SYSTEMS
Open Access
www.advintellsyst.com

Hi-LabSpermTracking: A Novel and High-Quality Sperm Tracking Dataset with an Advanced Ensemble Detection and Tracking Approach for Real-World Clinical Scenarios

Abdulsamet Aktas, Gorkem Serbes, Hakkı Uzun, Merve Huner Yigit, Nizamettin Aydin, and Hamza Osman Ilhan*

Sperm motility, a critical factor in diagnosing male infertility, requires computer-based solutions due to the limitations of manual evaluation methods. This study introduces the Hi-LabSpermTracking dataset, comprising 66 videos (60 s each, 10 fps) collected from 14 patients and meticulously annotated by experts. Unlike similar datasets, these uninterrupted, long-duration videos enable continuous tracking of individual sperm cells, each assigned a unique ID throughout the video, supporting both sperm detection and tracking tasks. Experimental evaluations employ you only look once v8 (YOLOv8), real-time detection transformer, and simple online and realtime tracking with a deep association metric across three scenarios. In Scenario I (sperm detection), the YOLOv8n model achieves 98.9% mAP50 and 97.9% F1-score. In Scenario II (sperm tracking), performance metrics include 83.88% mAP50, 87.63% F1-score, 72.27% higher order tracking accuracy (HOTA), and 77.88% multiple object tracking accuracy (MOTA). Scenario III simulates real-world challenges by separating training and testing videos. Ensemble methods are applied, with the proposed mean ensemble achieving superior results: 86.55% mAP50, 87.87% F1-score, 66.66% HOTA, and 76.42% MOTA. The Hi-LabSpermTracking dataset enables robust sperm tracking research, while the mean ensemble method amplifies accuracy by uniting model strengths.

1. Introduction

According to the Infertility Prevalence Estimates study conducted by the World Health Organization, ≈48 million to 186 million couples globally are affected by infertility.[1] Approximately 50% of all infertility cases are attributed to male factors. In sperm analysis, parameters such as sperm morphology, concentration, DNA integrity, and motility play a critical role. However, in clinical assessments, sperm concentration and motility are mostly taken into account in the diagnosis of male infertility. Sperm concentration measures the number of sperm cells per unit volume of ejaculate. According to the handbook published by the World Health Organization to standardize testing methods in laboratories, the presence of 15 million sperm per milliliter is considered normal.[2] Low sperm concentration (oligospermia) may be an indicator of conditions such as genetic anomalies,[3] functional disorders in spermatogenesis,[4] undescended testis,[5] and even azoospermia.[6]

While sperm concentration provides valuable insights into potential underlying conditions, it does not fully capture the functional capacity of sperms. To gain a

A. Aktas
Department of Computer Engineering, Faculty of Technology
Marmara University
34840 Istanbul, Turkey
E-mail: abdulsamet.aktas@marmara.com.tr

A. Aktas, H. O. Ilhan
Department of Computer Engineering, Faculty of Electrical and Electronics
Yildiz Technical University
34220 Istanbul, Turkey

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202500115.

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.202500115

G. Serbes
Department of Biomedical Engineering, Faculty of Electrical and Electronics
Yildiz Technical University
34220 Istanbul, Turkey

H. Uzun
Department of Urology, Faculty of Medicine
Recep Tayyip Erdoğan University
53020 Rize, Turkey

M. H. Yigit
Department of Biochemistry, Faculty of Medicine
Recep Tayyip Erdoğan University
53020 Rize, Turkey

N. Aydin
Department of Computer Engineering, Faculty of Computer and Informatics
Istanbul Technical University
Istanbul 34467, Turkey

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

more comprehensive understanding of male fertility, it is also crucial to evaluate sperm motility. The assessment of sperm motility is based on the percentage of sperm cells with progressive movement. Sperm motility is classified into four categories: Grade A, B, C, and D, which are defined as progressively motile, motile, nonprogressively motile, and immotile, respectively.[2] According to the same handbook, at least 40% of the total sperm count in the sperm concentration should be motile, and at least 4% of the measured motile sperm cells should be of the progressive type (Grade A+B). Low sperm motility (asthenozoospermia) may indicate conditions such as testicular damage,[7] sperm mitochondrial dysfunction,[8] or varicocele.[9]

Currently, sperm analysis is performed using two different techniques. The first of these, visual assessment, is a labor-intensive method with low efficiency. Sperm cells are examined under a microscope with the aid of medical materials such as the Makler chamber and hemocytometer, allowing for counting and motility analysis processes to be conducted. However, the results obtained through this method are significantly dependent on the experience and expertise of the observer. To automate this process, another analysis technique known as CASA (computer-assisted sperm analysis) systems has been developed. Compared to visual assessment, CASA systems provide the ability to perform more accurate, consistent, and objective analyses. These systems recognize and localize sperm cells in microscopic images and videos of semen samples, allowing for the tracking of multiple sperm over a period to calculate sperm concentration and motility. However, these systems utilize high-grade microscopes integrated with high-performance computers for data processing and storage. The use of these costly systems also necessitates the need for professional technicians. For these reasons, the widespread adoption of CASA systems on a global scale is challenging.[10]

Due to the aforementioned limitations, the need for affordable, accessible, accurate, and unbiased diagnosis of male infertility continues to persist worldwide.[1] In response to this need, various studies have been conducted in recent years in the field of sperm morphology.[11–16] On the other hand, studies focusing on sperm motility analysis and sperm tracking—both critical components of infertility diagnosis—remain limited. One of the primary reasons for this gap is the scarcity of large, diverse, and well-labeled datasets that capture the motility and tracking characteristics of sperm cells. However, assembling such datasets presents significant challenges related to manual labeling and the need for controlled laboratory environments, making the process both time-consuming and expensive. While traditional approaches—such as visual assessment and CASA—have long been employed to quantify sperm motility, these methods alone cannot fully address the growing demand for reliable and scalable solutions. Consequently, there is a pressing need to develop more advanced automated technologies that provide effective, fast, accurate, cost-effective, and objective assessments of sperm motility. Compared to traditional methods, the latest deep learning approaches offer cost-effective and unbiased solutions for sperm morphology analysis.[11–16] Similarly, employing deep learning-based advanced detection and tracking algorithms for sperm motility measurement is considered a promising strategy to overcome these existing deficiencies.[17–22]

This study presents a novel approach for the detection and tracking of sperm cells in semen samples. Utilizing the Hi-LabSpermTracking dataset that we developed, the results of a convolutional neural network (CNN)-based you only look once v8 (YOLOv8) architecture and a transformer-based real-time detection transformer (RT-DETR)[23] architecture have been combined and evaluated in three different scenarios, including theoretical and practical usage. The aim of this study was to design a system that combines the fast feature extraction capability of YOLOv8 with the contextual analysis power of RT-DETR, enabling a more comprehensive examination of object relationships within the scene. While YOLO provides rapid and reliable primary detection, the in-depth analysis offered by RT-DETR enhances the system's comprehensiveness. The analysis of detection and tracking metrics has been conducted in this study, employing the simple online and real-time tracking with a deep association metric (DeepSORT) tracking algorithm for monitoring sperm cells. This combination presents an effective solution for applications related to infertility diagnosis by improving detection and tracking performance.

The rest of the article is organized as follows: Section 2 provides a detailed review of publicly available datasets and an overview of existing sperm detection and tracking approaches in the field of sperm motility. Section 3 describes the materials and methods, including the dataset description, data acquisition approach, and a brief summary of the applied experimental scenarios. The obtained results are discussed in Section 4, while the significance of these results and the contributions of this study are addressed in Section 5. Finally, the conclusion of the study is given in Section 6, followed by a dedicated discussion of the study's limitations and future work in Section 7.

2. Related Works

Accurate detection of sperm cells is a fundamental requirement for effectively tracking these cells over a certain period. A review of the literature reveals that there are a large number of open-access datasets available for the classification of sperm morphology, but limited datasets focusing on the detection of sperm cells and especially on tracking processes. The presence of diverse and open-access datasets is of critical importance for ensuring the accuracy and generalizability of deep learning-based systems. While this section provides a detailed review of the open-access sperm datasets and detection-tracking methods in the current literature, it also systematically evaluates the advancements made in the field and identifies existing research gaps.

2.1. Open-Access Sperm Analysis Datasets

One of the major challenges encountered in the development of sperm analysis systems using deep learning methods is the insufficient availability of labeled and open-source datasets required for training models. A review of the literature reveals that existing open-source datasets for sperm analysis generally focus on sperm morphology and are primarily used in

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

classification problems. This section presents the open-source datasets found in the literature in chronological order.

The dataset created by Yi et al. contains 300 images with a resolution of 640 × 480, obtained from 12 different individuals.[24] The images have been labeled by experts and categorized into four different classes: normal, small, elongated, and megalo. These four-type labeled images are mostly used in classification problems. Another dataset, prepared by Ramos et al.[25] consists of 590 images with a resolution of 512 × 512. This dataset contains 160 normal and 430 abnormal sperm images and is used in two-class classification problems. Moreover, the dataset created by Abbiramy et al. contains 91 images with a resolution of 86 × 100.[26] These images, labeled according to WHO standards as normal and abnormal, were used for classification problems by applying image enhancement methods.

Ghasemian et al. introduced the human sperm morphology analysis dataset (HSMA-DS), which contains a larger number of images compared to previous datasets. This dataset consists of a total of 1457 images with a resolution of 576 × 764, obtained from samples of 235 patients.[27] The dataset was created by labeling the structures of sperm samples such as head and tail as normal or abnormal. Another important dataset is the human sperm head morphology (HuSHeM) dataset developed by Shaker et al. This dataset contains 216 images with a resolution of 131 × 131 and has been used in many studies to benchmark systems developed on sperm morphology.[11,13,16,28,29] Although the HuSHeM dataset encompasses four distinct sperm head abnormalities, it only provides classification data and does not support detection or tracking tasks.

Another dataset created with a similar structure to the HuSHeM dataset is the SCIAN-MorphoSpermGS dataset developed by Chang et al.[30] This dataset consists of a total of 1132 images with 35 × 35 resolution belonging to five different sperm abnormality classes and has been used in many studies to benchmark various deep learning and machine learning methods.[11,13,28,29] Javadi et al. modified the HSMA-DS dataset to create the MHSMA dataset, which includes a total of 1540 images with a resolution of 128 × 128.[31] Mohammed et al. on the other hand, conducted classification studies on normal and abnormal sperm heads using their own dataset, which comprises 384 sperm head images with a resolution of 20 × 20.[32]

The Sperm Morphology Image Data Set (SMIDS), developed by Ilhan et al. contains relatively more images compared to other datasets.[33] This dataset consists of a total of 3000 images with a resolution of 190 × 170, categorized into three classes: normal, abnormal, and nonsperm. Similar to the HuSHeM and SCIAN-MorphoSpermGS datasets, the SMIDS dataset is also widely used in various studies for method comparison and performance evaluation.[11,13,16,34]

Regarding the sperm video datasets, Chen et al. introduced a new one named as the sperm videos and images analysis (SVIA) dataset.[35] This dataset comprises 101 short videos, each ranging from 1 to 3 s in duration, which have been manually annotated. The dataset is divided into three subsets: Subset-A, Subset-B, and Subset-C. Subset-A includes 101 video clips with a total of 125 000 object locations. Subset-B contains 10 videos with 451 ground-truth segmentation masks, while the Subset-C consists of cropped images categorized into two classes: impurity and sperm.

Yuzkat et al. developed the semen specimen video dataset (SeSViD) for use in sperm detection tasks.[21] SeSViD comprises 12 videos collected from 12 different patients aged between 19 and 37 years. Originally recorded at 30 fps, the videos were downsampled to 5 fps for annotation purposes. The initial 1920 × 1200 resolution videos were rescaled to 640 × 640 resolution to facilitate use in deep learning-based detection algorithms. Comprising ≈1250 frames, this dataset can be utilized in both theoretical and practical testing scenarios. In theoretical applications, frames are randomly mixed and split into training and test sets, while in more practical scenarios, entire videos are kept separate for training and testing.

The VISEM dataset, introduced by Haugen et al. was developed from data collected from 85 patients over the age of 18.[36] The original videos, each 2–7 min in duration, were recorded at a resolution of 640 × 480. From these videos, 250 frames per video were extracted, and paired with expert-provided semen_analysis_data, allowing the dataset to be framed as a regression task. In total, the dataset comprises 21 250 images and was evaluated using threefold cross-validation. Building on this work, Vajira et al. created the VISEM-Tracking dataset,[19] selecting 20 of the 85 original VISEM videos for detailed annotation. From these selected videos, 30 s clips were extracted, resulting in a total of 29 196 labeled frames. The resolution (640 × 480) remained the same as in the original VISEM dataset. Unlike its predecessor, VISEM-Tracking includes bounding box coordinates and identity information for individual sperm cells. This richer labeling scheme makes the VISEM-Tracking dataset more useful, as it supports both detection and tracking tasks.

Table 1 presents an overview of widely used open-access sperm datasets in the literature. Most existing datasets primarily emphasize sperm morphology and are commonly utilized for classification tasks.[24–28,30,32,33,35,37] Although the SeSViD dataset is geared toward sperm detection and offers labeled detection data, it does not include identity annotations necessary for tracking individual sperm cells.[21] Meanwhile, the VISEM dataset treats the problem as a regression task rather than focusing on detection and tracking.[36] While VISEM-Tracking dataset provides detailed bounding box and identity annotations,[19] it is limited to 20 videos—each only 30 s long. Such short recordings may be insufficient for analyzing long-term sperm motility and behavior, which are critical for a more accurate assessment of motility and the effectiveness of tracking algorithms. Furthermore, it has been determined that some of the videos in the dataset (e.g., 14, 15, 21, 23, 35, 47, 54, 82) are created by merging two different videos, while others (e.g., 29, 52, 60) are formed by merging three different videos. This situation makes it impossible to consistently analyze the trajectory of a sperm cell throughout the entire video. Additionally, the videos in the VISEM-Tracking dataset were captured with a 400× magnification and have a resolution of 640 × 480. These technical specifications result in sperm cells being recorded with a narrow field of view, which limits the detailed, wide-angle analysis of their trajectories. Additionally, both VISEM and VISEM-Tracking originate from the same initial dataset, designed to study the effects of excessive weight and obesity on male reproductive function, thus potentially limiting genetic and biological diversity. To achieve more comprehensive and generalizable

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

Table 1. Existing sperm morphology and motility datasets on literature.

Paper	Name—title	Ground truth	Images	Videos	Summary
Yi et al. [24]	–	Classification	300	–	The dataset contains 300 images with a resolution of 640 × 480, collected from 12 distinct individuals.
Ramos et al.[25]	–	Classification	590	–	The dataset consists of 590 images with a resolution of 512 × 512. It includes 160 normal and 430 abnormal sperm images and is utilized for binary (two-class) classification problems.
Abbiramy et al.[26]	–	Classification	91	–	The dataset contains 91 images with a resolution of 86 × 100.
Ghasemian et al.[27]	HSMA-DS: Human Sperm Morphology Analysis DataSet	Classification	1457	–	The dataset consists of a total of 1457 images with a resolution of 576 × 764, containing samples obtained from 235 patients. It was constructed by labeling sperm structures, such as the head and tail, as normal or abnormal.
Shaker et al.[28]	HuSHeM: Human Sperm Head Morphology	Classification	216	–	This dataset contains 216 sperm head images with a resolution of 131 × 131 and is categorized into four classes.
Chang et al.[30]	SCIAN-MorphoSpermGS	Classification	1132	–	This dataset consists of a total of 1132 images with a resolution of 35 × 35, belonging to five different sperm abnormality classes.
Javadi et al.[31]	MHSMA: Modified Human Sperm Morphology Analysis Dataset	Classification	1540	–	This modified dataset contains a total of 1540 images with a resolution of 128 × 128.
Mohammed et al.[32]	–	Classification	384	–	The dataset consists of 384 sperm head images with a resolution of 20 × 20.
Ilhan et al.[10]	SMIDS: Sperm Morphology Image Data Set	Classification	3000	–	This dataset comprises a total of 3000 images with a resolution of 190 × 170, divided into three classes: normal, abnormal, and nonsperm.
Chen et al.[35]	SVIA: Sperm Videos and Images Analysis dataset	Detection, Segmentation, Classification	4041	101 (1–2 sec)	Subset-A consists of 101 video clips and contains a total of 125 000 object locations. Subset-B includes 10 videos, each containing 451 ground-truth segmentation masks. Subset-C consists of cropped images classified into two categories: impurity and sperm.
Yuzkat et al.[21]	SeSViD: Semen Specimen Video Dataset	Detection	1250	12 (20–30 sec)	It contains 12 videos, each ranging between 20 and 30 s. The total number of frames is ≈1250.
Haugen et al.[36]	VISEM	Regression	–	85	The VISEM dataset is based on data obtained from a total of 85 patients over the age of 18. The videos, lasting between 2 and 7 min, are recorded in a resolution of 640 × 480.
Thambawita et al.[19]	VISEM-Tracking	Detection, Tracking, Regression	29 196	20 (30 sec)	The videos in the dataset are 30 s long, with a resolution of 640 × 480, and contain a total of 29 196 frames.
Aktas et al.[37]	Hi-LabSpermMorpho	Classification	49 345	–	The dataset contains 49 345 images belonging to 18 different classes obtained using 3 different staining techniques.
Proposed study	Hi-LabSpermTracking	Detection, Tracking	41 808	66 (60 sec)	The videos in the dataset are 60 s long, with a resolution of 1920 × 1080, and contain a total of 41 808 frames.

insights, larger datasets representing a broader range of health conditions and biological variables would be highly beneficial.

2.2. Overview of Existing Sperm Detection and Tracking Methods

Manual semen analyses are time-consuming, and their accuracy depends heavily on the competence and experience of the clinician. These constraints have underscored the global need for systems that minimize human error and provide automated, objective evaluations. In line with this need, the present study examines sperm analysis systems in the literature using a systematic approach, highlighting recent advancements and identifying current research gaps.

Dewan et al. aimed to predict sperm concentration and total motility in their study.[38] In the study, sperm cells in each frame were detected using a blob detection method, and trajectories representing sperm motilities were created based on these detected sperm cells. Sperm concentration was calculated based on the number of sperm detected in each frame. To determine whether the trajectories correspond to sperm cells, a CNN model developed by the researchers was used. Additionally, the extracted trajectory information was utilized for video stabilization. However, there are some limitations in the study. First, the blob detection method used to identify object proposals can be influenced by residues in sperm samples and lighting conditions. In motility prediction, hard thresholding was applied using a specific threshold value. This approach limits the accuracy of classification by overlooking subtle differences. The dataset used in the study contains a

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

total of 3800 images of 50 × 50 dimensions from more than 100 patients. However, regarding the results of the conducted experiments, metrics such as multiple object tracking accuracy (MOTA) or higher order tracking accuracy (HOTA) were not presented, which created a gap in evaluating the performance of the method.

Thambawita et al. proposed a deep learning-based two-stage method to predict sperm cell motility and morphology.[39] In the first stage, an autoencoder was used to extract temporal features from sperm videos. These features were then converted into so-called "feature-images" by projecting temporal information into a spatial domain. In the second stage, a pretrained ResNet-34 model was fine-tuned using transfer learning on the generated feature-images. Treating motility and morphology analysis as a regression problem, the researchers achieved the lowest mean absolute error (MAE) of 9.427 for motility prediction, and the lowest MAE of 5.606 for morphology prediction.

Hicks et al. aimed to predict sperm motility by utilizing both traditional and modern machine learning techniques in their study.[40] In the study, three different feature extraction methods were used, and the sperm motility prediction was approached as a regression problem. These methods include raw frames, sparse and dense optical flow features, and a combination of both features. A range of models, from simple linear regression to CNN-based deep learning architectures, were evaluated. The most effective result was obtained using the optical flow features on the ResNet-50 network, with an average MAE value of 8.740.

In another study, Thambawita et al. similar to their previous work, investigated the prediction of sperm motility and morphology by leveraging either raw video frame stacks or optical flow-based representations.[41] Unlike Hicks et al.[40] who combined raw frames and optical flow, Thambawita et al. used these modalities separately. In the first raw data processing approach, nine consecutive gray-scale frames were combined to form a stack, while a complex representation that integrated RGB frames and optical flow maps computed at different strides was employed as the second approach. The best performance was obtained using a custom deep neural architecture, incorporating a multilayer perceptron (MLP) with dropout into a ResNet-34 model, when the second approach was followed. This custom network achieved an average MAE of 8.825 for motility and 5.293 for morphology.

Valiuskaite et al. introduced a deep learning method for evaluating sperm head motility based on a region-based CNN (R-CNN) approach.[20] Using a subset of the VISEM dataset (650 manually annotated frames), their method first segmented sperm heads and then tracked them across consecutive frames. Identity assignments were made using a Euclidean distance-based thresholding method. Although the proposed method attained a 91.77% detection accuracy for sperm heads, it was tested on a very small subset of the VISEM dataset and did not address ID switching problem that may occur in overlap scenarios comprehensively. As a result, its practical applicability remains limited. In another study conducted using the VISEM dataset, Ottl et al. used unsupervised tracking algorithms, namely, the Lucas–Kanade and Crocker–Grier methods, to extract displacement features and motion statistics from videos.[42] These features were used to train various neural networks and support vector regression (SVR) models. In the study, the best result was obtained using a linear SVR model trained with the representation of individual displacement features for each sperm cell, and the MAE was reported as 7.31.

Similarly, Adinugroho and Nakazawa tackled sperm motility and morphology as regression problems by introducing a new motion representation technique called MotionFlow.[17] Unlike traditional optical flow methods that compute flow for the entire frame, MotionFlow calculates optical flow vectors for each segmented sperm cell. A semantic segmentation method (an U-Net trained for 49 epochs) was performed on 10 000 images (chosen from 29 196 images) from the VISEM-Tracking dataset to isolate sperm from the background. Subsequently, a ResNet3D-18 model utilized these motion vectors to predict motility and morphology. The approach yielded MAE values of 6.842 for motility and 4.148 for morphology.

Ilhan et al. proposed a hybrid expert system for sperm analysis that offers advantages in terms of cost and modularity.[18] They aimed to make sperm analysis more portable, modular, and practical by using a smartphone-based data collection approach integrated with a microscope. Using their proposed data collection method, they obtained 18 different videos from 6 individuals, and a total of 86 438 sperm trajectories were divided into four different clusters. The videos were stabilized using the speeded-up robust feature-based matching technique, and Kalman filtering was applied for tracking. Compared to an alternative expert system (SQA-Vision), the proposed method reported error rates of 2.1%, 7.4%, and 5.3% for progressive, nonprogressive, and immotile sperm, respectively.

Zhu et al. introduced YOLOv5s-SA, a modified YOLOv5s-based model designed to improve sperm detection performance.[43] Initially, the model reduced the number of parameters while maintaining accuracy by replacing some traditional convolutional layers in the backbone network with depthwise separable convolutional structures. To enhance the model's detection capability, a multiscale fusion module was designed to increase positional information and detailed features in the feature maps. Finally, a spatial attention (SA) mechanism was added in the neck network before the output feature maps, which improved the correlation between feature map channels and allowed for better fusion of detailed features. In this study, ten different sperm videos were used, collected from ten different patients. The videos varied in length from 10 to 130 s and had a resolution of 1280 × 960. The first 100 frames of videos from the first six patients were used to create training and validation sets. 80% of the frames (480 frames) were used for training, and the remaining 20% (120 frames) were used for validation. For the other four patients, 30 randomly selected frames from each video were used as the test set. The proposed YOLOv5s-SA model with a value of 87.5% demonstrated an improvement in average accuracy compared to YOLOv3, YOLOv3-spp, YOLOv5s, and YOLOv5m models, with increases of 18.1%, 15.2%, 6.9%, and 1.9%, respectively.

Yuzkat et al. aimed to improve sperm detection accuracy and consistency by fusing deep learning-based object detection algorithms.[21] The study utilized 12 different videos obtained from 12 infertile patients aged between 19 and 37. Originally, the videos were 1920 × 1200 resolution at 30 fps. However, to facilitate the labeling process, the resolution was downscaled to 640 × 640 and the frame rate was reduced to 5 fps. The videos were labeled by experts as sperm and nonsperm. Using the labeled videos,

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

deep learning models were tested under two different scenarios. In the first scenario, for each sperm sample video, object detection algorithms were trained with different training/test set ratios to observe individual performances. In this scenario, a patient-oriented approach was applied to evaluate the model performances. The videos with the lowest mean average precision (mAP) values were selected as target videos for the second scenario. In the second scenario, to improve detection performance on more challenging videos (those with low mAP values), model training was conducted using all videos (except the challenging ones) without a patient-oriented approach. This allowed for the training of a more generalized model using different videos. Additionally, a combined (fusion) deep learning approach was proposed and applied to enhance the detection performance of the system. In the second scenario, with the proposed fusion method, mAP scores of 86.1% and 84.3% are obtained in the two videos in the test set, which are 1.5% and 0.8% better than YOLOv5 only without fusing, respectively.

Zhang et al. introduced the YOLOv8E-TrackEVD model for sperm detection and tracking, using an augmented version of the VISEM-Tracking dataset combined with the MHSMA dataset.[22] The original VISEM-Tracking dataset, comprising 29 196 images, was divided into 19 530 training, 5580 test, and 2790 validation images. These sets were significantly expanded using the MHSMA dataset, resulting in 459 291 training, 128 707 test, and 65 110 validation images. Augmentation was applied not only to the training set but also to the validation and test sets—an approach that may introduce bias and yield artificially high performance estimates. Although the proposed method reported 74.303% HOTA and 71.167% MOTA on the VISEM-Tracking dataset, the study did not specify which videos were used for these calculations. Because all VISEM-Tracking videos were employed in the training phase, no independent test set remained. Consequently, the reported HOTA and MOTA values may not accurately reflect real-world performance, and the lack of a genuine hold-out set raises concerns about overfitting and the potential for misleading results.

This research delivers an exhaustive overview of current sperm motility studies and their analysis systems, highlighting the gaps in the field. The studies reviewed in sperm detection and tracking methods have been summarized in **Table 2** to provide a clear comparison and detailed insight. To address these gaps, our contributions can be listed as follows: 1) The Hi-LabSpermTracking dataset consists of 66 fully annotated videos, including 36 phone camera-based videos (averaging 60 s each) with 22 865 frames and 30 microscope camera-based videos with 18 943 frames, totaling 41 808 frames. 2) The dataset provides opportunities for sperm detection and tracking using deep learning methods and multiple object tracking (MOT) algorithms, offering not only bounding box annotations but also ID information. 3) The recorded videos are captured as single, uninterrupted 60 s clips without any breaks or skips, providing comprehensive information on sperm trajectories throughout the entire video. This unique aspect distinguishes the dataset

Table 2. Existing sperm detection and motility analysis methods.

Paper	Dataset	Task(s)	Ground truth	Method	Performance
Dewan et al.[38]	Own dataset	Motility estimation	Regression	Drift detection and estimation	Pearson coefficient 0.818 (total motility)
Thambawita et al.[39]	VISEM	Motility estimation	Regression	Autoencoder + ResNet-34 for image feature extraction and motility estimation	MAE of 9.427
Hicks et al.[40]	VISEM	Motility estimation	Regression	OpticalFlow + ResNet50 image feature extraction and motility estimation	MAE of 8.740
Thambawita et al.[41]	VISEM	Motility estimation	Regression	Stacking OpticalFlow + ResNet-34 + MLP for image feature extraction and motility estimation	MAE of 8.825
Valiuvskaite et al.[20]	VISEM	Sperm detection	Detection	Faster R-CNN for sperm head detection	91.77 accuracy
Ottl et al.[42]	VISEM	Motility estimation	Regression	Lucas–Kanade and Crocker–Grier for machine learning algorithm	MAE of 7.31
Adinugroho et al.[17]	VISEM	Motility estimation	Regression	MotionFlow for image feature extraction and ResNet3D-18 for motility estimation	MAE of 6.842
Ilhan et al.[18]	Own dataset	Motility estimation	Tracking	Recursive Kalman Filter and Hungarian Algorithm for motility estimation	Avg. error rate 4.93
Zhu et al.[43]	Own dataset	Sperm detection	Detection	YOLOv5s-SA for sperm detection	87.5 mAP
Yuzkat et al.[21]	SeSVid	Sperm detection	Detection	Fusion model (YOLOv5 + Mask RCNN + Detectron2) for sperm detection	85.2 mAP
Thambawita et al.[19]	VISEM-Tracking	Sperm detection	Detection Regression Tracking	YOLOv5 for sperm detection	22.31 mAP
Zhang et al.[22]	VISEM-Tracking	Sperm detection Sperm tracking	Detection Regression Tracking	YOLOv8E-TrackEVD for sperm detection and tracking	74.303 HOTA 71.167 MOTA
Proposed study	Hi-LabSpermTracking	Sperm detection sperm tracking	Detection Tracking	Ensemble Model (YOLOv8+ RT-DETR+ DeepSORT) for sperm detection and tracking	66.60 HOTA 76.42 MOTA

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

from others (e.g., ref. [19]), making it the only one in the literature to provide such comprehensive and continuous sperm tracking data. 4) Demonstrating suitability for deep learning by conducting detection and tracking using one CNN-based model, one transformer-based model, and the DeepSORT tracking method, resulting in a level of comprehensiveness not previously seen in the literature on sperm detection and tracking. 5) We also demonstrated that our proposed method, which combines the two models using three different ensemble techniques, outperforms existing methods in both detection and tracking tasks. 6) The Hi-LabSpermTracking dataset, generated using a smartphone-based video recording method, enables models trained on it to achieve universal compatibility with various types of microscopes. This approach facilitates sperm motility analysis using a smartphone integrated into a microscope, eliminating the need for additional cameras. 7) Offering the sperm detection and tracking dataset publicly for noncommercial purposes, allowing researchers to validate and compare their algorithms, thus advancing the field.

3. Material and Methods

3.1. Hi-LabSpermTracking Dataset

Hi-LabSpermTracking is a comprehensive video dataset developed to serve as a high-quality resource for deep learning models focused on sperm detection and tracking. The dataset originates from videos obtained from 523 individuals who visited the Andrology Laboratory at Recep Tayyip Erdo$\breve{g}$an University Training and Research Hospital (Rize, Turkey). For each patient, four videos were recorded using a smartphone-based imaging system developed by our research team, and an additional four videos were acquired from the same patient using a camera integrated into a conventional laboratory microscope.[18] The videos captured using the smartphone-based system are particularly significant as they demonstrate the potential for accessible, cost-effective, and portable solutions for sperm motility analysis without the need for advanced laboratory equipment. Moreover, systems developed using these videos have the potential to achieve global compatibility across various microscope types, further enhancing their applicability in diverse clinical and research settings. In total, this process yielded 4184 videos. By providing data from both smartphone and microscope-integrated imaging systems, Hi-LabSpermTracking enables comparative evaluations of cell detection and tracking algorithms, as will be discussed in the Methodology section.

Prior to video collection, all participants received comprehensive information and provided informed consent. Semen samples were obtained following a standardized protocol, involving a 3 day period of sexual abstinence before sample collection between 09:00 and 11:00 a.m. on the fourth day. To mitigate the impact of temperature fluctuations, the samples were maintained either at room temperature (20–37 °C) or in an incubator at 37 °C. A 5 µL aliquot of each semen sample was placed between a glass slide and a coverslip and recorded using a smartphone-based imaging system mounted on a 100× Olympus BX43 microscope. Videos were initially captured at 30 fps. However, minimal changes between consecutive frames led to redundant data and increased labeling time. Consequently, the frame rate was reduced to 10 fps. The resolution remained at 1920 × 1080 pixels.

The sperm cells in the obtained videos were manually annotated by two experts through an intensive and laborious process. Due to the highly demanding and time-consuming nature of this annotation process, it was not feasible to annotate all available videos. Consequently, a total of 66 videos from 14 patients (36 captured with a smartphone and 30 obtained using a microscope-integrated camera) were meticulously annotated to create the Hi-LabSpermTracking dataset. Each sperm in each frame was annotated using bounding boxes. Additionally, a unique ID was assigned to each sperm cell, ensuring that these IDs remained consistent with the same cell across all frames. This configuration makes the dataset a valuable resource for both object detection and tracking tasks. **Table 3** provides more detailed information about the dataset. **Figure 1** illustrates an example frame obtained as a result of this annotation process.

3.2. Sperm Detection Module

In this study, state-of-the-art deep learning methods, YOLOv8 and RT-DETR,[23] were employed to perform sperm detection on the Hi-LabSpermTracking dataset. The aforementioned models were selected on the basis of their complementary strengths and their high compatibility with the objectives outlined in the above-listed points. This section offers a comprehensive analysis of the architectural designs of these models, their inherent advantages, and the underlying rationale for their adoption within the scope of this research.

YOLOv8 is one of the newest versions of the "you only look once" family and is a deep learning model optimized for tasks such as object detection, classification, and segmentation. The model utilizes advanced architectures such as the feature pyramid network and path aggregation network to effectively integrate multiscale feature maps from images. In this way, it has a high performance in detecting both small and large objects. YOLOv8 is a single-stage detector that processes input in a single pass to extract object locations and class information. Unlike two-stage sensors, it adopts an anchor-free design, providing a simpler architecture, reducing computational costs, and enabling significantly faster operation. Its additional advantages include a fast training process, low latency, and strong generalization capability across diverse datasets. However, its performance

Table 3. Summary of quantitative information about the Hi-LabSpermTracking dataset.

Description	Phone camera-based videos	Phone Camera-based test videos	Microscope camera-Based videos
Annotated videos	30	6	30
FPS per video	10	10	10
Annotated frames	18 876	3989	18 943
Frames with at least one sperm	18 876	3989	18 943
Bounding boxes	275 778	116 666	334 849
Unique sperms (with tracking IDs)	588	248	736

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

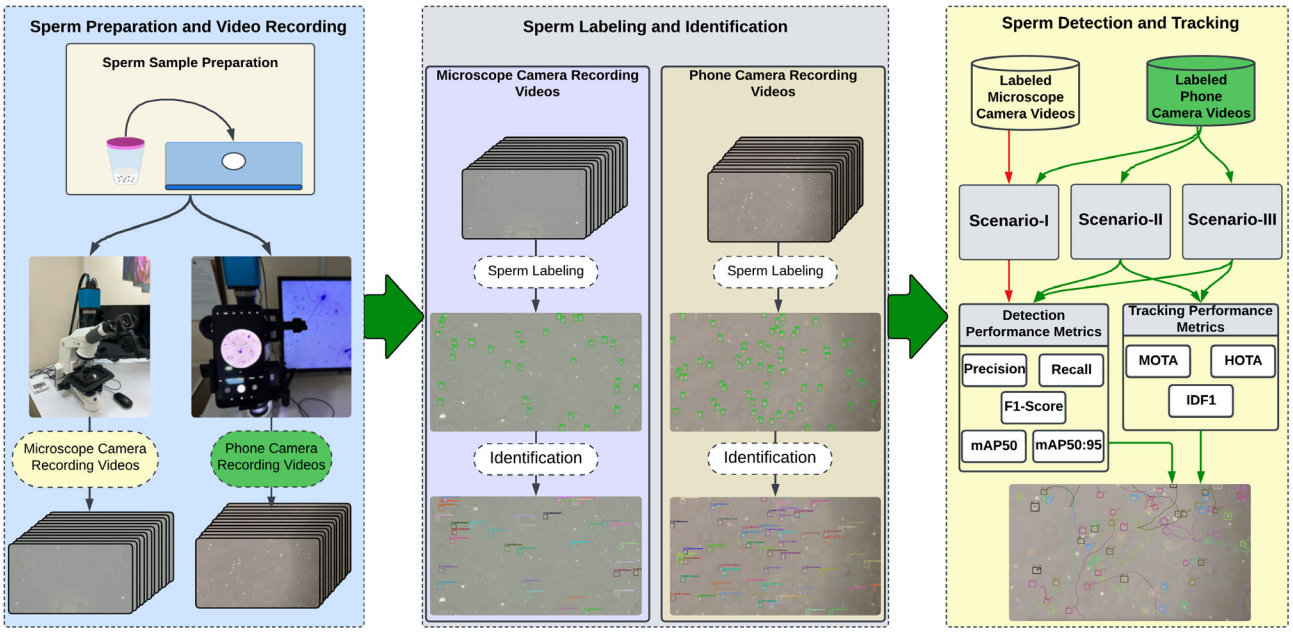

Figure 1. Flow chart of the proposed system.

can be limited in dense and complex scenes or datasets with class imbalance. For instance, YOLOv8 struggles with detecting small and occluded objects in traffic scenarios, which affects its accuracy in densely populated scenes. Furthermore, changes in lighting conditions can also challenge its ability to maintain consistent detection performance.[44] Despite these challenges, YOLOv8 offers significant advantages in terms of speed, making it a highly efficient choice for real-time object detection tasks, especially in scenarios where fast processing is crucial.[45]

YOLOv8 offers several variants optimized for different scales and application requirements. These variants are divided into five categories: nano (n), small (s), medium (m), large (l), and extra large (xl). The main differences between the models are the number of parameters used and the level of architectural complexity. The Nano model stands out with its very low number of parameters and offers a suitable option for resource-constrained environments and real-time applications due to its fast operation and low hardware requirement. On the other hand, the extra large variant, with its high number of parameters and increased architectural complexity, provides high accuracy in more challenging tasks but requires higher hardware capacity. In this context, the nano-variant of the YOLOv8 family was preferred for its resource efficiency and real-time performance advantages.

Another object detection method used in this study is the RT-DETR[23] model, which is based on a transformer architecture. This model provides a balanced performance between accuracy and speed, making it particularly effective for real-time applications. RT-DETR employs an architecture based on attention mechanisms and utilizes the denoising training method to process multiscale features. This approach enhances the robustness of the model against noise, mislabeled data, and complex scenes. The RT-DETR architecture offers models in various sizes, tailored to different application requirements. In this study, the RT-DETR-L model, which provides higher contextual awareness between objects compared to other variants, has been preferred.

An examination of the object detection methods used in this study reveals that the YOLOv8n model performs fast object detection, while the RT-DETR-L model increases contextual awareness between scenes and objects with the attention mechanism. In this regard, combining the outputs of both models allows their strengths to compensate for each other's limitations, thereby optimizing the overall detection performance.

3.3. Sperm Tracking Module

Object tracking is of critical importance, particularly in the field of biomedical imaging, as it involves determining the spatial positions of moving objects over time and continuously monitoring their trajectories. DeepSORT[46] is a widely adopted and highly efficient approach among object tracking algorithms. This algorithm aims to optimize tracking performance by integrating real-time tracking capabilities with deep learning-based feature extraction.

DeepSORT is an improved version of the simple online and realtime tracking (SORT) algorithm. The SORT algorithm basically performs the object tracking task using the Kalman filter[47] and the Hungarian algorithm.[48] However, the SORT algorithm relies solely on geometric features (e.g., bounding box coordinates) obtained from object detection methods for tracking. This approach can lead to matching errors, particularly when dealing with objects that have similar appearances or overlap. To overcome this limitation, the DeepSORT method was developed by integrating a deep learning approach based on visual similarities into the SORT algorithm.[46]

The working principle of the DeepSORT algorithm is based on the detection and tracking association mechanism. In each frame, detected objects are tracked by matching them with the

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

tracking information from the previous frame. This matching process is based on two main factors: visual features and motion estimation. In visual feature extraction, deep learning-based methods are used to extract the embedded features of the objects. These features represent the unique identities of objects and enable correct frame-to-frame matching based on the visual similarity of objects. Motion estimation, on the other hand, is performed using the Kalman filter[47] to predict the spatial positions of the objects. As part of the matching strategy, a cost matrix is created to associate detection and tracking information. This matrix is calculated based on the visual similarities and geometric distance of the objects. The Hungarian algorithm[48] determines the most appropriate matches in this matrix and enables the tracking of objects. In addition, DeepSORT's visual feature-based approach allows the preservation of identity information (IDs) in case objects temporarily disappear or reappear from the frame. This mechanism increases tracking accuracy, making long-term object tracking more reliable.

Thanks to its appearance-based matching mechanism, DeepSORT demonstrates exceptional performance in tracking complex movement patterns and visually similar sperm cells, which are commonly encountered in sperm video analysis. The real-time processing capability of the algorithm enables effective tracking of fast-moving sperm cells in microscopic videos. With features such as deep learning-based visual feature extraction and motion prediction, DeepSORT provides high accuracy and reliability in biomedical image processing, making a significant contribution to the tracking of sperm cells. These capabilities make the algorithm a robust method that can serve as a reference for future research in this field.

3.4. Performance Metrics

3.4.1. Detection Performance Metrics

In sperm detection and tracking problems, the performance of the models has been evaluated using various metrics and approaches. In detection tasks, metrics such as precision (P), recall (R), and F1-score, which represents the harmonic mean of these two metrics, are used to measure the accuracy and effectiveness of the models. Additionally, the mAP metric, which provides a more comprehensive evaluation of the model's overall detection performance, is also included in this study. Precision refers to the rate at which the model makes the correct detections and is calculated using Equation (1). Recall represents the ratio of correctly detected positive samples to the total number of actual positive samples, indicating how well the model captures true positives (TP). Recall is calculated as defined in Equation (2). The F1-score, which evaluates the overall performance of the model by balancing these two metrics, is calculated by taking the harmonic mean of both metrics, as shown in Equation (3).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

$$mAP_{0.5} = \frac{1}{n}\sum_{i=1}^{n} AP_i^{IoU=0.5} \tag{4}$$

$$mAP_{0.5:0.95} = \frac{1}{n \cdot k}\sum_{i=1}^{n}\sum_{j=1}^{k} AP_i^{IoU=0.5+0.05(j-1)} \tag{5}$$

mAP is a metric that is used to evaluate the overall detection performance of a model. This metric is calculated by taking the average of the AP values computed at different intersection over union (IoU) threshold levels. In this study, the metrics mAP50 and mAP50:95, commonly used in the literature, are also used. In Equation (4), mAP50 identifies detections with an IoU of 0.5 or greater as accurate and classifies predictions below this threshold as inaccurate. The mAP50:95 metric, as given in Equation (5), calculates the average of the AP values obtained by varying the IoU threshold between 0.5 and 0.95 in 0.05 steps. In Equation (4) and (5), n denotes the number of classes, while AP_i represents the average precision for the ith class. The k in Equation (5) refers to the number of IoU threshold values.

In Scenario I, two YOLOv8n models were independently trained using videos based on the mobile phone and microscope camera, and detection performance metrics were obtained. To evaluate whether there was a statistically significant difference between the detection performances of these trained models, an analysis of variance (ANOVA) analysis was conducted.[49] Following ANOVA, pairwise comparisons were made in the metrics P, R, F1, mAP50, and mAP50:95 for the test sets of 30 different videos obtained with both models. These comparisons were supported by a posthoc analysis using the Tukey's honestly significant difference (HSD) test.[50] The obtained results are presented in detail in Section 4.

3.4.2. Tracking Performance Metrics

In object tracking scenarios, in addition to detection metrics, MOTA, HOTA, and identity F1-score (IDF1) metrics have been employed to evaluate the tracking performance of the models. MOTA is a widely used metric for measuring the performance of detection and tracking algorithms in MOT tasks. This metric evaluates effectiveness by considering factors such as missed detections, false positives (FP), and identity switches that occur during object tracking. The equation for the MOTA metric is presented in Equation (6). In this equation, the values for FP and false negative (FN) are calculated using an IoU threshold of 0.5. In the equation, t represents the frame indices, where FN_t denotes the number of missed detections, FP_t represents the number of false detections, $\text{identityswitches}(IDSW)_t$ indicates the number of identity switches, and GT_t refers to the number of ground-truth objects in the tth frame. These values are calculated separately for each frame and the total error value is obtained across all frames. The calculated total error value is normalized by dividing it by the total number of ground-truth objects, and this value is subtracted from 1 to obtain the final MOTA metric. In addition to MOTA, the multiple object detection accuracy (MODA) metric measures the accuracy of object detection by focusing on missed detections and FP, while the

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

multiple object tracking precision (MOTP) evaluates the localization precision of detected objects by considering the average distance between detected and ground-truth objects.

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t} \tag{6}$$

Another MOT metric is the IDF1, which is used to assess both the correct matching and identity consistency of tracking algorithms. IDF1 evaluates the identity accuracy of tracked objects, influenced by factors such as false detections and missed objects. This metric is particularly sensitive to IDSW, highlighting the impact of such errors on the algorithm's performance.

$$\text{IDF1} = 2 \cdot \frac{\text{IDP} \cdot \text{IDR}}{\text{IDP} + \text{IDR}} \tag{7}$$

As defined in Equation (7), the IDF1 metric is calculated as the harmonic mean of identity precision (IDP) and identity recall (IDR). IDP represents the ratio of correct matches to the total number of detected objects being tracked, while IDR represents the ratio of correct matches to the total number of objects that should have been correctly matched. The IDF1 metric not only evaluates the success of object detection but also assesses the performance of correctly tracking the identities of these objects. Due to these characteristics, it stands out as an important metric for evaluating tracking algorithms.

Although MOTA and IDF1 are widely used metrics in MOT studies, they have some limitations. MOTA provides an evaluation primarily based on the accuracy of object detection, whereas IDF1 offers more detailed information by focusing on the correct association of object identities. To represent these two metrics in a more balanced way, Luiten et al. developed the HOTA metric.[51] HOTA is calculated as a combination of three different IoU scores and provides a comprehensive approach that evaluates tracking performance across three subtasks: object detection, identity association, and localization. For each of these tasks, IoU scores are calculated and combined to obtain the final HOTA score. Thus, the HOTA metric provides an opportunity for a more balanced and detailed evaluation of tracking algorithms.

In the localization (Loc) step, the spatial alignment between a predicted detection and the ground-truth detection is evaluated. **Figure 2** visualizes the method for calculating the Loc-IoU value. As stated in Equation (8), the localization accuracy (LocA) metric is calculated by averaging the Loc-IoU values of all matching ground-truth and predicted detection pairs in the dataset. In this equation, $|TP|$ represents the total number of TP matches, while

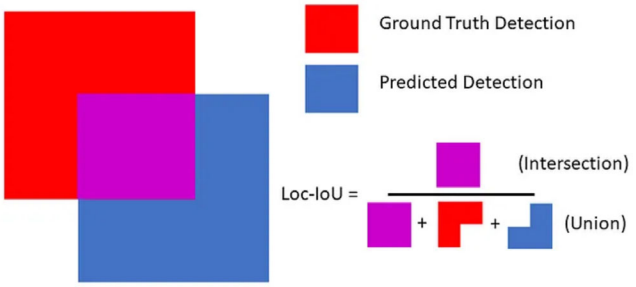

Figure 2. The visualization of the Loc-IoU metric calculation.[51]

$c \in TP$ denotes each correctly matched individual object. LocA serves as a crucial metric for evaluating the spatial accuracy of tracking algorithms.

In the detection step, the alignment between the set of all ground-truth detections and the set of all predicted detections is measured. Detection IoU (Det-IoU), as defined in Equation (9), is also used to evaluate detection accuracy (DetA). When calculating DetA, a specific threshold value of Loc-IoU is used to determine the intersection between ground-truth and predicted detections. However, in this approach, a prediction can intersect with multiple ground truths, or vice versa, leading to potential ambiguities. To handle such situations, the Hungarian Algorithm is applied to ensure a one-to-one matching between the two sets.

The acquired matches align with the symbol TP in Equation (9). Unmatched predictions are defined as $|FP|$, while unmatched ground truths are referred to as $|FN|$. The Det-IoU, calculated based on the total number of TPs, FPs, and FNs in the dataset, provides the overall DetA value. In this way, the algorithm's performance in the detection step is evaluated comprehensively.

Finally, in the association step, the system's ability to assign correct identities to the same detections over time is measured using the ground-truth identities and ground-truth tracks provided in the dataset. As mentioned earlier, this measurement is performed by matching predicted and ground-truth detections using the Hungarian algorithm, and the alignment between all tracks of the predicted detections and all tracks of the ground-truth detections is calculated. As defined in Equation (10), true positive associations (TPA) represent the correct matches between the predicted tracks and ground-truth tracks. Detections in the predicted tracks that are incorrectly matched to another ground-truth track or do not match any track at all are referred to as false-positive associations (FPA). On the other hand, all ground-truth track and detections that do not match any correct prediction are defined as false-negative associations (FNA). Visual examples of TPA, FNA, and FPA definitions are shown in **Figure 3**. The large blue and light blue circles with thick dashed lines in the lower-left corner of Figure 3 represent the GT detections and trajectories. The small black and gray circles with thin dashed lines above them represent the predicted detections and trajectories. Each circle in a different color represents a distinct object. The red square indicates the associated predicted and ground-truth detection pair. To evaluate the quality of the association between two detections, all detections that intersect along the tracks (green squares) represent TPA, while the unmatched detections are shown as FPA (yellow squares) and FNA (brown squares). As indicated in Equation (11), the average Ass-IoU is calculated by matching the ground-truth detections and predicted detections throughout the data set, and this value is then used to calculate the association accuracy (AssA) metric.

$$\text{LocA} = \frac{1}{|TP|} \sum_{c \in TP} \text{Loc-IoU}(c) \tag{8}$$

$$\text{Det-A} = \text{Det-IoU} = \frac{|TP|}{|TP| + |FP| + |FN|} \tag{9}$$

$$\text{Ass-IoU} = \frac{|TPA|}{|TPA| + |FPA| + |FNA|} \tag{10}$$

$$\text{AssA} = \frac{1}{|TP|} \sum_{c \in TP} \text{Ass-IoU}(c) \tag{11}$$

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

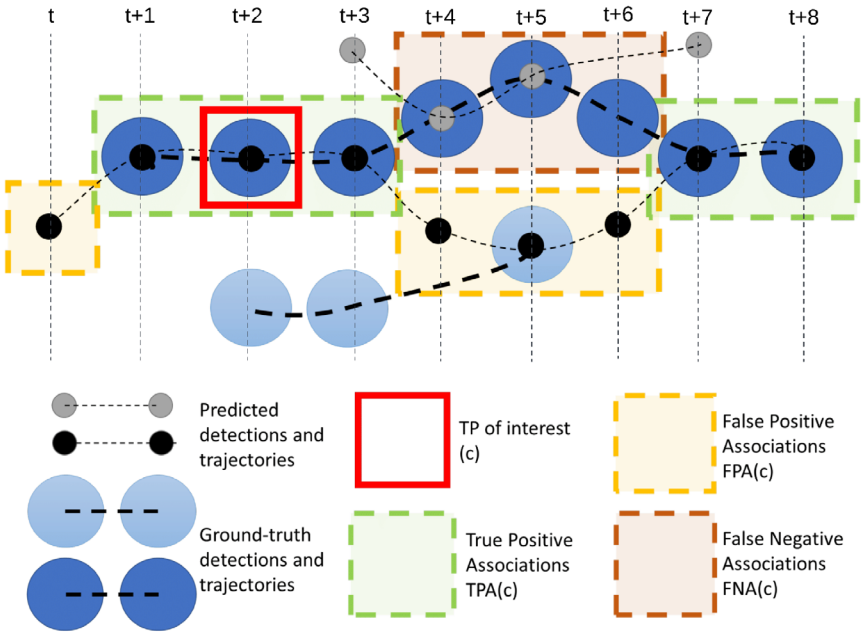

Figure 3. Visual representation of the concepts used in the calculation of AssA metric.[51]

$$= \frac{1}{|TP|} \sum_{c \in TP} \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|} \tag{12}$$

$$\mathrm{HOTA}_\alpha = \sqrt{\mathrm{DetA}_\alpha \cdot \mathrm{AssA}_\alpha} \tag{13}$$

$$= \sqrt{\frac{\sum_{c \in \mathrm{TP}_\alpha} \text{Ass-IoU}_\alpha(c)}{|\mathrm{TP}_\alpha| + |\mathrm{FN}_\alpha| + |\mathrm{FP}_\alpha|}} \tag{14}$$

$$\mathrm{HOTA} = \int_{0<\alpha\leq 1} \mathrm{HOTA}_\alpha \tag{15}$$

$$\approx \frac{1}{19} \sum_{\alpha=0.05}^{0.95} \mathrm{HOTA}_\alpha \tag{16}$$

The localization, detection, and association metrics mentioned above are crucial in the context of the MOT problem. The combination of these three metrics forms HOTA, which provides a comprehensive assessment of a system's performance, offering a sufficient tool to measure the system's performance across three distinct aspects with a single metric value. As shown in Equation (13), HOTA is calculated by taking the geometric mean of the DetA and AssA values, ensuring that both metrics are weighted equally.

In the Hungarian algorithm used for assignment during the calculation of DetA and AssA, a specific Loc-IoU threshold (α) was applied. However, when calculating HOTA, this (α) value is varied from 0.05 to 0.95, allowing for a more general performance evaluation without being restricted to a specific threshold. This approach incorporates LocA as part of the final score. With this calculation, the HOTA score enables a comprehensive evaluation of the model's performance in terms of detection, association, and the spatial accuracy of the detections.

3.5. Experimental Setup

In this study, we trained the YOLOv8-n and RT-DETR-L models for 50 epochs using the AdamW optimizer with a batch size of 4. For the YOLOv8 model, the default loss function provided by the official implementation was employed. A cosine learning rate scheduler was applied throughout the training process to gradually adjust the learning rate and facilitate stable convergence. Additionally, an early stopping mechanism with a patience value of 10 was used to prevent overfitting. No data augmentation techniques were used during training in order to maintain consistency across all experimental scenarios.

For inference, the confidence score threshold was set to 0.5, and the IoU threshold for nonmaximum suppression was fixed at 0.7. These thresholds were chosen based on preliminary experiments to balance precision and recall effectively. A summary of the training and inference settings is provided in **Table 4**.

3.6. Experimental Scenarios

When the studies in the literature are examined, it is seen that there are various studies on sperm detection and tracking in different scenarios. In studies focusing on sperm detection,[20–22,43] it has been observed that datasets were prepared using traditional data splitting methods, where all images were shuffled, a certain percentage was allocated as the test set, and the remaining portion was used for training and validation sets. On the other hand, in studies focusing on sperm tracking,[19,22] it has been observed that, in addition to the aforementioned approach, methods where the training and test sets are derived from completely different videos, in line with real-world use scenarios, have also been applied.

In this study, experiments were conducted using videos obtained by using a mobile phone-based data acquisition method. To

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 4. Training and inference settings for YOLOv8-n and RT-DETR-L models.

Parameter	Value
Epochs	50
Batch size	4
Optimizer	AdamW
Loss function	Default (YOLO)
lr0-lrf	0.01–0.01
Learning rate scheduler	Cosine annealing
Data augmentation	None
Confidence threshold	0.5
IoU threshold	0.7

evaluate the effectiveness of the videos, the YOLOv8n model was trained separately on both microscope camera-based recorded videos and mobile phone videos, and the results of the models were statistically compared. The analysis revealed no significant difference between the two types of data; therefore, subsequent experiments were conducted using mobile phone videos. Based on this information, the dataset created in this study was evaluated under three different scenarios using various performance metrics.

3.6.1. Scenario I

In the first scenario, as in many studies in the literature,[20–22,43] frames extracted from the videos were shuffled, and the dataset was split into 75% training and 25% testing. In this scenario, sperm detection was performed on the dataset and the models' precision, recall, F1 score, mAP50 and mAP50:95 values were presented. In this scenario, the training and test sets are created randomly. Since the frames in the test set are randomly selected, they do not represent a video stream. However, in order to perform sperm tracking and calculate metrics such as HOTA, MOTA, etc., it is necessary to detect sperm in consecutive frames, assign IDs to these sperm, and determine their trajectories. In line with these requirements, a second scenario was developed as an alternative to the first scenario. The flowchart for Scenario I is presented in **Figure 4**.

3.6.2. Scenario II

In the second scenario, the first 45 s of the videos were used as training set and the remaining 15 s were employed as test set. This method was applied for each video and approximately the initial 75% of all videos was reserved as training and the last 25% of all videos was used as test set. Using the YOLOv8n model trained with this scenario, precision, recall, F1-score, mAP50, and mAP50:95 metrics for the last 15 s of each video, as well as tracking metrics such as MOTA, HOTA, and IDSW were calculated using the DeepSORT algorithm. Although the results obtained in the second scenario are more reliable than those in the first scenario, which is widely used in the literature, they are not fully compatible with real-world usage scenarios. For example, analyzing the videos of a patient admitted to the hospital for the first time is not possible in this scenario. The flowchart for Scenario II is presented in **Figure 5**.

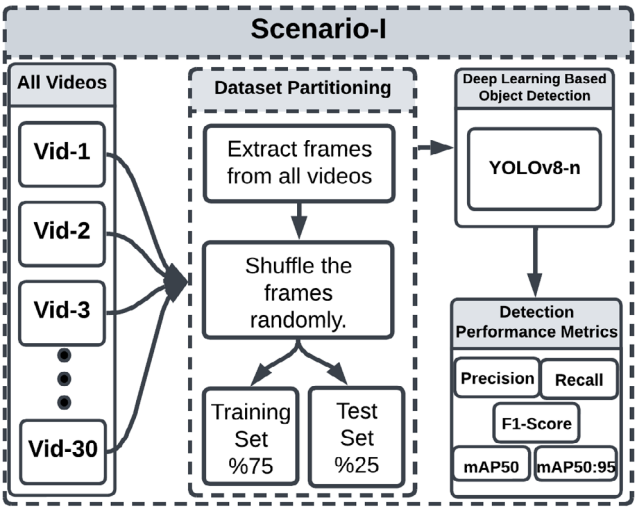

Figure 4. Flowchart of the Scenario I.

3.6.3. Scenario III

Although the scenarios described above are theoretically useful, it is seen that they are limited in terms of practical use in real life. This is because the training and test sets contain different frames of the identical video. However, in real-world scenarios, the model's training set does not contain any frames from patients' videos. Therefore, the third scenario was developed, which represents an approach that is more compatible with real-world applications. In the third scenario, the 30 videos used in the previous two scenarios were included in the training set, and in addition, 6 new videos were added to the dataset for testing. The extended dataset created in this scenario was used to evaluate both the YOLOv8n model and a transformer-based architecture, RT-DETR-L. Upon analyzing the results of the trained individual models, it was observed that the YOLOv8n model performs faster than the RT-DETR-L model. Specifically, in videos containing a high number of sperm, the YOLO model generates more predictions compared to RT-DETR-L. However, when comparing the Precision values, YOLO performs worse than RT-DETR-L. This is primarily due to the YOLO model's tendency to incorrectly predict nonsperm objects as sperm, resulting in false positives. On the other hand, although the RT-DETR-L model is slower and generates fewer predictions, it produces more accurate and reliable results. A review of the literature reveals that CNN- and transformer-based models have frequently been ensembled to improve detection accuracy, particularly in complex scenes.[52,53] In this study, the results of the YOLOv8n (CNN-based) and RT-DETR-L (transformer-based) models were fused using three different ensemble methods and the DeepSORT algorithm was applied for sperm tracking. This approach aimed to combine the strengths of both models to achieve more accurate and reliable detection. At the end of the scenario, the performance metrics for each model

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

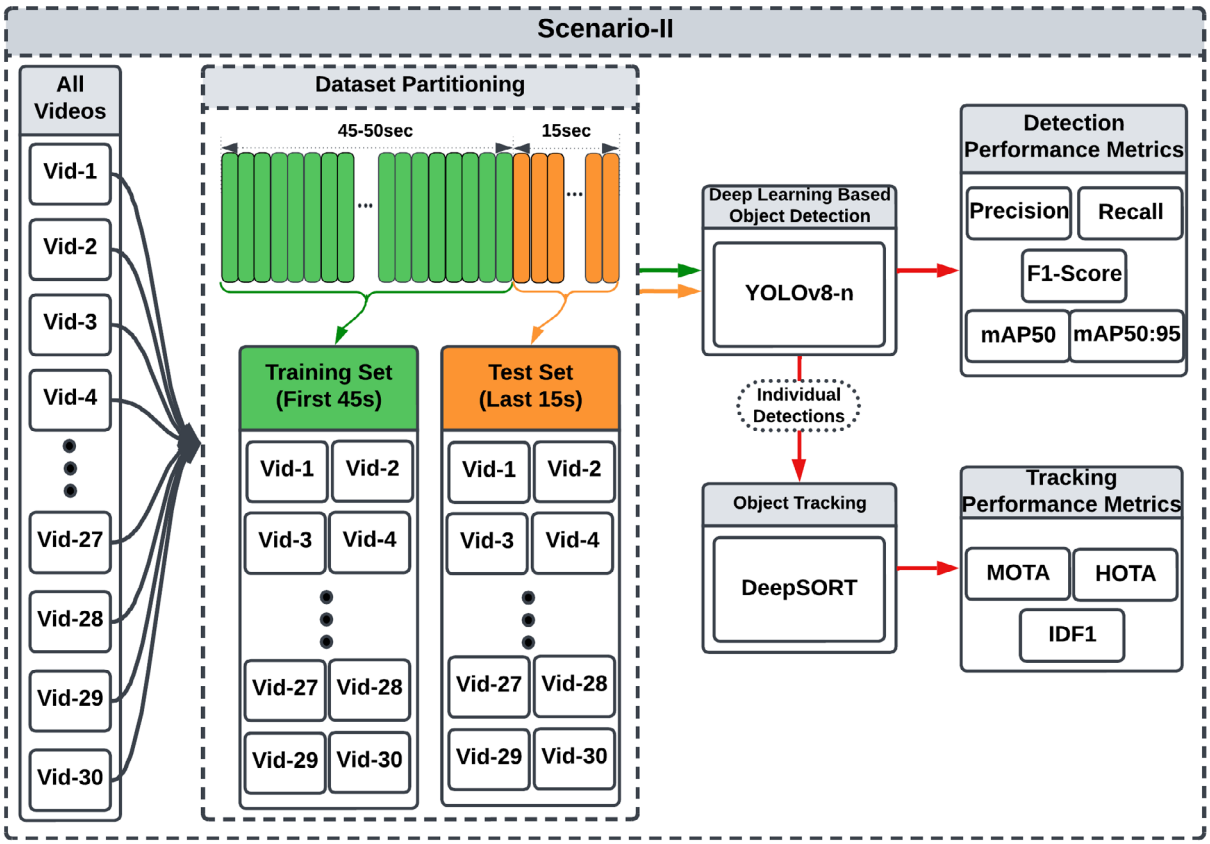

Figure 5. Flowchart of the Scenario II.

and the fused models were thoroughly analyzed and presented. The flowchart for Scenario III is presented in **Figure 6**.

4. Experimental Results

In this study, models were trained and thoroughly analyzed in three distinct scenarios outlined using the Hi-LabSpermTracking dataset. Each scenario involved videos captured by mobile phones from the Hi-LabSpermTracking dataset. To ensure that outputs from mobile phone-recorded videos align with those from microscope camera recordings, the YOLOv8n model was independently trained on each video type for sperm identification. The trained models were then evaluated on their respective test sets, yielding metrics such as precision, recall, mAP50, and mAP50:95. An ANOVA test followed by Tukey's HSD post hoc analysis was conducted to evaluate whether the detection results differed significantly between mobile phone-recorded videos (phone videos) and microscope camera-based-captured videos (microscope videos). The results are summarized in **Table 5**.

The comparative analysis of phone videos and microscope videos is illustrated in Table 5; no significant differences were observed in most metrics, including P, R, F1, and mAP50, as indicated by the *p*-adj values greater than 0.05. However, a statistically significant difference was identified for mAP50:95 (*p*-adj = 0.007), with videos recorded on a mobile phone achieving a higher mean value (0.778) compared to videos captured with a microscope camera (0.734). The positive mean difference (MeanDiff = 0.0441) further indicates that mobile phone recorded videos provide superior performance in capturing a wider range of detection scenarios. In statistical testing, *p*-adj is the adjusted *p*-value and a value below 0.05 suggests a significant difference. Reject indicates that the null hypothesis is rejected when *p*-adj < 0.05. The lower and upper values represent the confidence interval bounds, showing the range where the true mean difference likely lies (from 0.012 to 0.076 in this case). These findings confirm the effectiveness of mobile phone recorded videos and underscore their potential for broader applications in various scenarios within the study.

4.1. Results of Scenario I

A review of the literature revealed that many studies that address the sperm detection problem randomly divide frames from all videos in the dataset into training and test sets and report evaluation metrics for their proposed systems.[18,20–22,43] In this context, all frames from 30 videos in the dataset for Scenario I were combined, and 75% were randomly divided for training, while 25% were used for testing. The test results for Scenario I are provided in column phone videos of Table 5. Upon examining the obtained results, it can be observed that the model achieved 97.77% precision, 98.14% recall, 97.9% F1-score, 98.9% mAP50, and 77.80% mAP50:95 performance in Scenario I. These results indicate a significantly higher performance compared to the sperm detection studies

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

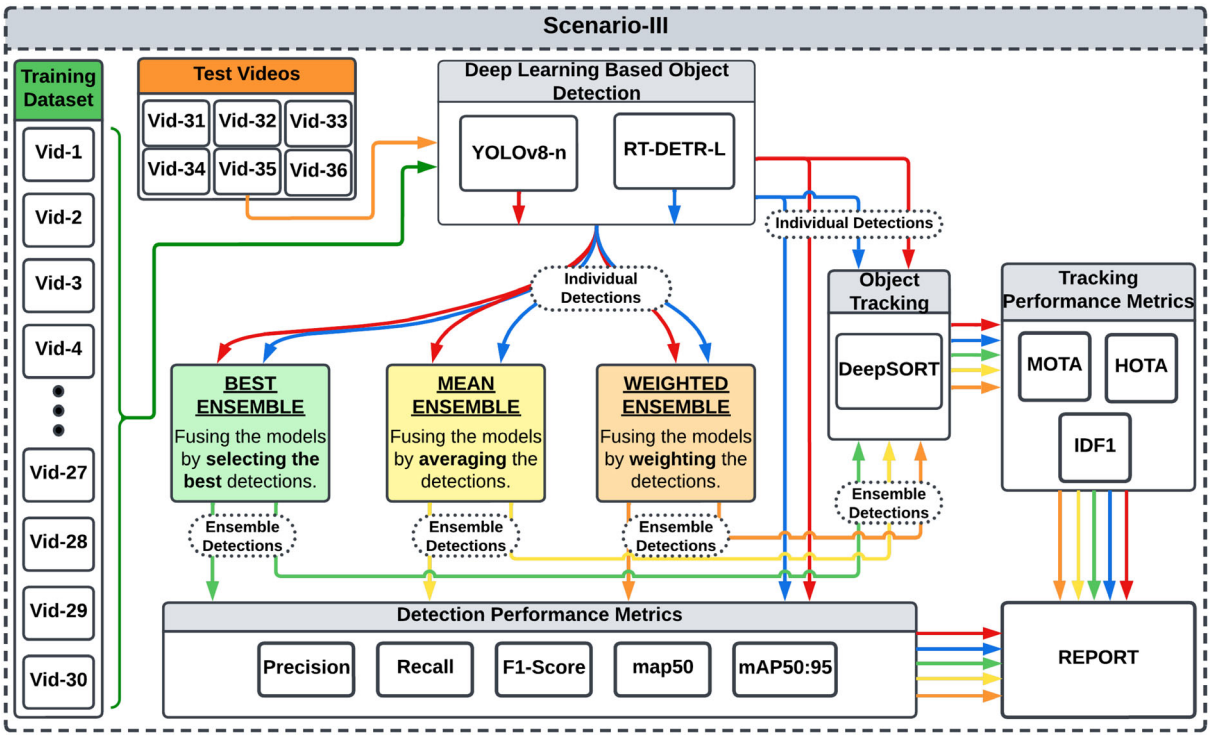

Figure 6. Flowchart of the Scenario III.

Table 5. Tukey's HSD results: evaluation of phone versus microscope camera-based video performance.

Metrics	Phone videos	Microscope videos	MeanDiff	*p*-adj	Lower	Upper	Reject
P	97.77	98.20	−0.0047	0.31	−0.014	0.004	False
R	98.14	97.90	0.0016	0.69	−0.0065	0.009	False
F1	97.90	98.00	−0.0016	0.71	−0.01	0.006	False
mAP50	98.90	98.90	0	0.99	−0.0045	0.004	False
mAP50:95	77.80	73.40	0.0441	0.007	0.0122	0.075	True

presented in Table 2; however, they also highlight several issues related to the partitioning of the dataset, which affect the generalizability and real-world applicability of the model's performance. As the test set was generated by randomly shuffling the frames, it was not possible to compute the tracking metrics. In real-world scenarios, preserving the sequence of frames is essential for evaluating tracking performance, as this allows the model to be assessed on its ability to follow objects across frames. As mentioned in Section 3.6 of the study, objects in consecutive frames must be detected and tracked accurately to be able to calculate MOT metrics.

4.2. Results of Scenario II

The detection and tracking results of the experiments carried out within the scope of Scenario II are presented in **Table 6**. In this scenario, the last 15 s of each video were allocated as the test set, and detection metrics for the videos were obtained using the YOLOv8n model. Sperm detected by the YOLOv8n model were tracked using the DeepSORT algorithm. The averages of the detection and MOT metrics for Scenario II results are summarized in the last row of Table 6. Table 6 displays the GTDets and GTIDs columns, representing the total number of objects and unique IDs for each video. According to the obtained results, the average detection metrics for the 30 videos are 92.81% precision, 84.19% recall, 87.63% F1, 83.88% mAP50, and 63.98% mAP50:95. Furthermore, since the last 15 s of video sequences were used as the test set, MOT metrics for the same videos were also computed. When analyzing the tracking performance within the scope of Scenario II, it is observed that the average values for 30 videos are 72.27% HOTA, 77.88% MOTA, and 86.31% IDF1. The obtained detection and MOT metric performance underscores the potential of the Hi-LabSpermTracking dataset to substantially advance research in the field. However, although this data-partitioning approach is theoretically valid, relying on the last portion of each video as the test set does not fully represent a real-world scenario, in which the model would encounter previously unseen patient videos for the first time.

4.3. Individual Results of Scenario III

The experiments conducted in Scenarios I and II confirmed that the Hi-LabSpermTracking dataset is theoretically well suited for sperm detection and tracking. However, because Scenarios I and II include frames from the same patients featuring the same sperm objects in new orientations and positions in both the

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 6. Scenario II: detection and MOT metrics for the last 15 s (test set) of each video.

Video name	Detection performance					Tracking performance							Video specifications				
	P	R	F1	mAP50	50:95	HOTA	DetA	AssA	MOTA	MOTP	MODA	IDF1	Dets	GTDets	IDs	GTIDs	IDSW
Vid-1	77.57	74.53	75.64	73.48	51.65	62.36	50.13	79.20	53.92	79.48	53.92	76.83	1791	1812	23	13	0
Vid-2	96.02	65.85	77.43	65.62	44.43	60.22	52.71	69.71	66.10	78.07	66.40	76.40	1220	1717	15	13	5
Vid-3	72.98	55.27	60.47	52.21	32.89	43.83	34.28	56.05	37.35	72.78	37.55	61.94	361	514	11	7	1
Vid-4	89.97	85.07	87.28	84.99	68.60	74.02	64.76	85.57	74.80	84.07	74.87	87.03	1434	1480	12	10	8
Vid-5	81.26	85.76	82.96	85.08	54.42	65.01	54.09	78.30	62.76	79.08	62.76	82.03	1601	1493	16	11	0
Vid-6	87.69	76.33	81.39	75.70	51.08	64.06	55.54	75.01	63.81	80.17	64.18	77.08	1934	2194	26	19	8
Vid-7	96.06	72.40	82.31	72.25	43.87	61.55	54.94	69.76	70.37	75.66	70.49	81.48	2023	2626	21	18	3
Vid-8	97.48	82.45	89.07	82.41	66.93	75.67	68.22	84.99	79.84	85.56	80.13	86.86	2040	2441	23	19	7
Vid-9	95.84	80.23	87.23	80.14	57.06	69.11	63.38	76.51	78.69	81.50	78.98	83.39	2666	3107	29	21	9
Vid-10	98.57	92.40	95.31	92.37	66.52	77.64	73.53	84.05	93.26	81.15	93.28	95.55	3473	3648	28	25	1
Vid-11	92.91	96.95	94.85	96.92	76.96	80.58	74.49	88.86	89.15	84.26	89.21	93.54	5909	5586	48	38	3
Vid-12	99.87	92.20	95.83	92.20	78.49	82.86	77.93	88.95	90.93	86.93	90.97	94.92	2672	2933	21	20	1
Vid-13	92.75	75.89	83.25	75.70	49.76	64.85	58.05	73.16	69.43	79.98	69.69	81.83	1882	2290	26	19	6
Vid-14	95.35	78.74	86.04	78.68	66.60	72.27	65.47	80.90	73.26	85.74	73.57	80.19	1603	1956	19	15	6
Vid-15	88.56	58.68	69.69	57.14	30.02	47.34	41.01	54.70	48.92	72.73	49.55	63.52	714	1106	21	11	7
Vid-16	92.74	89.26	90.85	89.02	70.62	74.89	69.46	83.71	82.92	82.83	83.24	86.56	2428	2482	26	18	8
Vid-17	98.75	96.76	97.72	96.76	81.27	86.05	81.24	92.12	95.63	87.12	95.80	96.96	4072	4144	33	28	7
Vid-18	98.86	89.58	93.95	89.57	81.13	84.17	79.35	89.99	88.71	89.58	88.94	91.27	3179	3491	29	24	8
Vid-19	97.40	99.20	98.26	99.19	75.17	81.60	78.27	86.16	97.12	82.24	97.12	98.57	2719	2676	24	20	0
Vid-20	91.41	84.64	87.68	84.40	59.72	70.06	64.21	78.63	77.75	79.64	77.99	86.87	2769	2881	33	21	7
Vid-21	85.06	94.57	89.44	94.41	70.60	73.84	66.27	83.50	76.13	82.68	76.41	85.85	2030	1793	22	13	5
Vid-22	98.98	92.61	95.59	92.61	80.92	85.64	81.05	91.52	94.13	88.03	94.13	96.98	2193	2315	17	16	0
Vid-23	99.78	97.33	98.31	97.33	83.68	86.07	83.46	90.04	98.65	85.83	98.65	99.32	442	444	3	3	0
Vid-24	99.44	97.33	98.08	97.33	86.75	85.99	83.87	88.49	98.34	87.00	98.34	99.17	418	421	3	3	0
Vid-25	96.79	86.90	90.75	86.38	52.80	66.29	63.92	69.26	84.29	77.14	84.57	90.38	630	700	7	5	2
Vid-26	99.41	91.50	95.18	91.50	74.95	81.80	76.07	89.21	92.48	84.55	92.48	96.10	1099	1184	8	8	0
Vid-27	87.53	90.00	88.58	89.85	77.31	77.48	69.29	87.65	77.03	86.94	77.10	85.64	1549	1480	13	10	1
Vid-28	89.96	90.87	90.25	90.72	66.59	73.30	69.04	78.45	80.01	82.07	80.36	87.41	1492	1446	14	10	5
Vid-29	90.46	87.67	88.95	87.53	72.92	76.41	67.63	86.95	78.55	84.42	78.55	89.22	1759	1776	16	12	0
Vid-30	94.82	64.86	76.63	64.79	45.65	63.18	52.61	76.83	61.97	80.91	62.26	76.47	719	1036	10	7	3
Average and combine	**92.81**	**84.19**	**87.63**	**83.88**	**63.98**	**72.27**	**65.81**	**80.61**	**77.88**	**82.27**	**78.05**	**86.31**	**58 821**	**63 172**	**597**	**457**	**111**

training and test sets, these scenarios do not fully represent real-world conditions where a model encounters entirely new patient data for the first time. To address this limitation, Scenario III was designed as a real-world application scenario. In this setup, 30 videos from the dataset were used for just training, and six additional (recorded from four different patients), previously unseen videos were added to evaluate the trained models. **Table 7** presents the GTDets and GTIDs columns, indicating the number of objects and unique IDs in each video. Across these six new videos, there are 166 666 objects and 248 unique IDs. This design more closely approximates real clinical practice by testing sperm detection and tracking models on data from first-time patients.

The results of the individual models trained within the scope of Scenario III are presented in detail in Table 7. The models were trained for 20 epochs using the Hi-LabSpermTracking dataset. After the training process, the models were run on the videos used as the test dataset, and detection and tracking performance metrics specific to each video were computed. It should be noted that the confidence score and IoU threshold were set to 0.5 for the calculation of these metrics. When evaluating the results of the YOLOv8n model, it was observed that the highest detection performance was achieved in the Vid-35 video, which contains fewer objects (1783) compared to other videos, with the following metric values: 99.70% P, 98.57% R, 98.98% F1-score, 98.57% mAP50, and 78.12% mAP50:95. When examining the tracking metrics, it was observed that the Vid-35 video achieved 83.25% HOTA, 99.38% MOTA, and 99.69% IDF1. The objects detected in this video were successfully identified, with no IDSW occurring. In contrast, the lowest performance was

 © 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

observed in the Vid-36 video, which contains a larger number of objects (45 998). In this video, the YOLOv8n model achieved detection performance metrics of 81.80% P, 65.04% R, 72.39% F1-score, 60.47% mAP50, and 21.13% mAP50:95. The tracking metrics were calculated as 49.14% HOTA, 53.16% MOTA, and 65.70% IDF1. In the Vid-36 video, it was noted that the model failed to detect ≈8000 objects and made a total of 84 IDSW. When examining the model's average performance across all videos, the detection metrics achieved were 90.19% P, 83.23% R, 86.37% F1-score, 81.98% mAP50, and 47.43% mAP50:95. For tracking problems, the average values were 64.53% HOTA, 74.54% MOTA, and 82.31% IDF1. A total of 106 157 objects were accurately detected, with 149 IDSW occurring.

Upon examining the results of the RT-DETR-L model presented in Table 7, it is observed that the detection and tracking performance metrics, similar to the YOLOv8n model, reached their best values in the Vid-35 video. The detection performance metrics obtained for Vid-35 are recorded as 98.48% P, 96.47% R, 97.20% F1-score, 96.47% mAP50, and 67.56% mAP50:95. The metrics for the MOT problem demonstrated a highly successful performance with values of 77.67% HOTA, 97.98% MOTA, and 98.99% IDF1. It is noted that the number of detected objects (1769) is very close to the required number of objects (1783), and no IDSW occurred throughout the video. However, as with the other models, the lowest detection and tracking performance was observed in the Vid-36 video, which contains the most number of objects. The analysis of Vid-36 showed detection performance metrics of 79.62% P, 61.87% R, 69.56% F1-score, 57.44% mAP50, and 19.92% mAP50:95. In terms of tracking performance, the model achieved 48.11% HOTA, 48.45% MOTA, and 64.57% IDF1 values. It was observed that the model failed to detect ≈9000 objects and made a total of 113 IDSW during this video. When examining the overall performance of the RT-DETR-L model across the entire test set, the average values for detection problems were 91.04% P, 78.56% R, 84.09% F1-score, 77.28% mAP50, and 42.24% mAP50:95. For the MOT problem, the average values were 62.15% HOTA, 72.82% MOTA, and 81.21% IDF1. It was noted that the total number of objects detected across all videos was 100 181, with a total of 182 IDSW occurring.

When the results of the models were examined (as shown in Table 7), it was observed that both models demonstrated consistent performance across all videos, although there were certain cases where one model had advantages and disadvantages over the other. When the overall performance of the models was evaluated, experimental studies revealed that the YOLOv8n model outperformed the RT-DETR-L model. However, when the results were examined in detail, it was observed that the RT-DETR-L model achieved a higher performance, with a 91.04% P value, compared to the YOLOv8n model. When the total detection (DET) numbers of both models were analyzed, it was noted that

Table 7. Scenario III: detection and MOT metrics for the unseen test videos.

Video name	Detection method	Detection performance					Tracking performance							Video specifications				
		P	R	F1	mAP50	50-95	HOTA	DetA	AssA	MOTA	MOTP	MODA	IDF1	Dets	GTDets	IDs	GTIDs	IDSW
Vid-31	Ensemble	88.69	92.41	90.44	92.32	54.75	66.40	60.30	74.58	80.90	73.45	80.97	88.74	16 288	15 598	65	38	12
Vid-32	Ensemble	79.41	87.24	83.06	84.16	46.10	62.47	52.02	77.64	61.80	74.73	61.87	79.93	21 964	19 297	70	40	14
Vid-33	Ensemble	93.60	90.21	91.82	90.07	56.02	71.17	65.83	79.85	84.95	77.74	85.02	90.13	18 628	19 156	53	36	13
Vid-34	Ensemble	87.73	86.09	86.82	85.88	49.51	65.67	57.67	78.81	72.79	75.20	72.87	84.08	14 855	14 834	56	34	11
Vid-35	Ensemble	98.92	99.01	98.84	99.01	73.50	81.86	79.52	84.91	99.50	81.78	99.50	99.75	1784	1783	3	3	0
Vid-36	Ensemble	80.18	72.71	76.22	67.86	23.67	52.03	46.61	60.46	58.61	66.70	58.77	72.10	43 015	45 998	158	97	73
ine Vid-31	RT-DETR-L	92.34	79.82	85.51	79.48	40.57	62.32	55.13	72.11	74.66	72.98	74.73	84.01	13 845	15 598	62	38	12
Vid-32	RT-DETR-L	85.93	74.96	79.88	72.41	34.95	59.37	50.37	72.36	63.19	73.01	63.29	78.54	18 111	19 297	70	40	20
Vid-33	RT-DETR-L	94.93	77.38	85.16	77.19	46.18	64.15	57.42	72.70	75.65	75.87	75.73	84.08	15 992	19 156	47	36	14
Vid-34	RT-DETR-L	94.93	80.87	87.22	80.69	44.35	61.28	57.98	66.38	76.97	74.11	77.12	77.05	12 986	14 834	62	34	23
Vid-35	RT-DETR-L	98.48	96.47	97.20	96.47	67.56	77.67	74.84	81.64	97.98	78.69	97.98	98.99	1769	1783	3	3	0
Vid-36	RT-DETR-L	79.62	61.87	69.56	57.44	19.92	48.11	42.16	57.55	48.45	65.77	48.69	64.57	37 478	45 998	172	97	113
ine Vid-31	YOLOv8n	92.72	84.95	88.59	84.83	48.14	64.24	56.63	74.80	74.90	72.92	75.00	84.86	14 587	15 598	54	38	16
Vid-32	YOLOv8n	83.98	84.04	83.93	81.88	44.08	62.33	53.47	75.39	66.88	74.60	66.97	79.81	19 916	19 297	71	40	18
Vid-33	YOLOv8n	93.71	83.83	88.40	83.57	48.41	67.04	60.96	76.22	80.10	76.18	80.17	87.49	17 713	19 156	51	36	13
Vid-34	YOLOv8n	89.24	82.95	85.90	82.59	44.70	61.19	56.47	68.71	72.82	74.18	72.94	76.33	14 148	14 834	58	34	18
Vid-35	YOLOv8n	99.70	98.57	98.98	98.57	78.12	83.25	81.36	85.49	99.38	83.45	99.38	99.69	1780	1783	3	3	0
Vid-36	YOLOv8n	81.80	65.04	72.39	60.47	21.13	49.14	43.67	58.15	53.16	67.52	53.34	65.70	38 013	45 998	151	97	84
Average and combine	**Ensemble**	**88.09**	**87.94**	**87.87**	**86.55**	**50.59**	**66.60**	**60.32**	**76.04**	**76.42**	**74.93**	**76.50**	**85.79**	**116 534**	**116 666**	**405**	**248**	**123**
	RT-DETR-L	91.04	78.56	84.09	77.28	42.25	62.15	56.32	70.46	72.82	73.41	72.92	81.21	100 181	116 666	416	248	182
	YOLO	90.19	83.23	86.37	81.98	47.43	64.53	58.76	73.13	74.54	74.81	74.64	82.31	106 157	116 666	388	248	149

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the RT-DETR-L model performed ≈6000 fewer detections than the YOLOv8n model, yet it still had a higher P value. When the Vid-32, Vid-33, and Vid-34 videos were examined, where the RT-DETR-L model achieved a higher P value than the YOLOv8n model, it was observed that the YOLOv8n model's detection numbers were much closer to the GTDets compared to the RT-DETR-L model, and in some cases (e.g., Vid-32), it even performed more detections. Based on the results, it can be concluded that the RT-DETR-L model achieves a higher P rate compared to the YOLOv8n model, indicating a better ability to correctly identify objects (sperm cells). However, it falls short in R, indicating that it detects a smaller proportion of the objects present in the frame compared to YOLOv8n. To address this, the study aimed to integrate the advantages of both models to balance P and R metrics in detection performance, and the impact of this balance on tracking performance was comprehensively analyzed.

4.4. Results of Ensemble Models in Scenario III

In order to leverage the strengths of both the YOLO and RT-DETR models, ensemble methods were employed to enhance detection accuracy, particularly in complex and diverse scenarios. Three different ensemble methods—best, mean, and weighted—were proposed in this study to fuse the detection results from the individual models, aiming to improve overall performance and robustness. Each method employs a distinct strategy for combining detection results. To illustrate the strategies of the proposed ensemble methods, **Figure 7** presents a representative example of sperm detections fused on a sample frame according to each method. In the best method, among the objects detected by the models in each frame, the one with the highest confidence score is considered correct. For example, in Sample-1 presented in Figure 7, a sperm is detected with a confidence score of 65% by the YOLOv8n model, while the same object is detected with a confidence score of 93% by the RT-DETR-L model. When the best ensemble method is applied, the detection from the RT-DETR-L model is accepted as correct instead of the YOLOv8n model's detection, and performance metrics are computed based on this selection.

In the mean method, the average of the coordinate values of the detections made by both models is calculated to form a new bounding box. Performance metrics are then determined using this new bounding box. For example, in the Sperm-4 detection shown in Figure 7, the model bounding boxes are depicted in blue and red, while the new bounding box created by averaging these boxes is shown in yellow. In the final method, weighted, the confidence scores of the models are used to weight the detection results. In this method, the confidence score obtained from the model detection is normalized, ensuring that the sum of the weights is equal to 1. The normalized weights are then applied to the model's coordinate values, resulting in a weighted bounding box. This approach ensures that more successful models have a greater influence in determining the final bounding box.

For objects detected by only one model, such as in Sperm-2 and Sperm-3, since there are no other objects to merge with, these detections are directly included among the detected objects for that frame without any further processing, and performance metrics are calculated accordingly. The IoU method was used for matching the objects detected by the models, with a threshold value set to 0.5. If multiple detection results exist for the same object after the merging process, the most accurate one is selected. After the merging process, the resulting detections were tracked throughout the video using the DeepSORT algorithm.

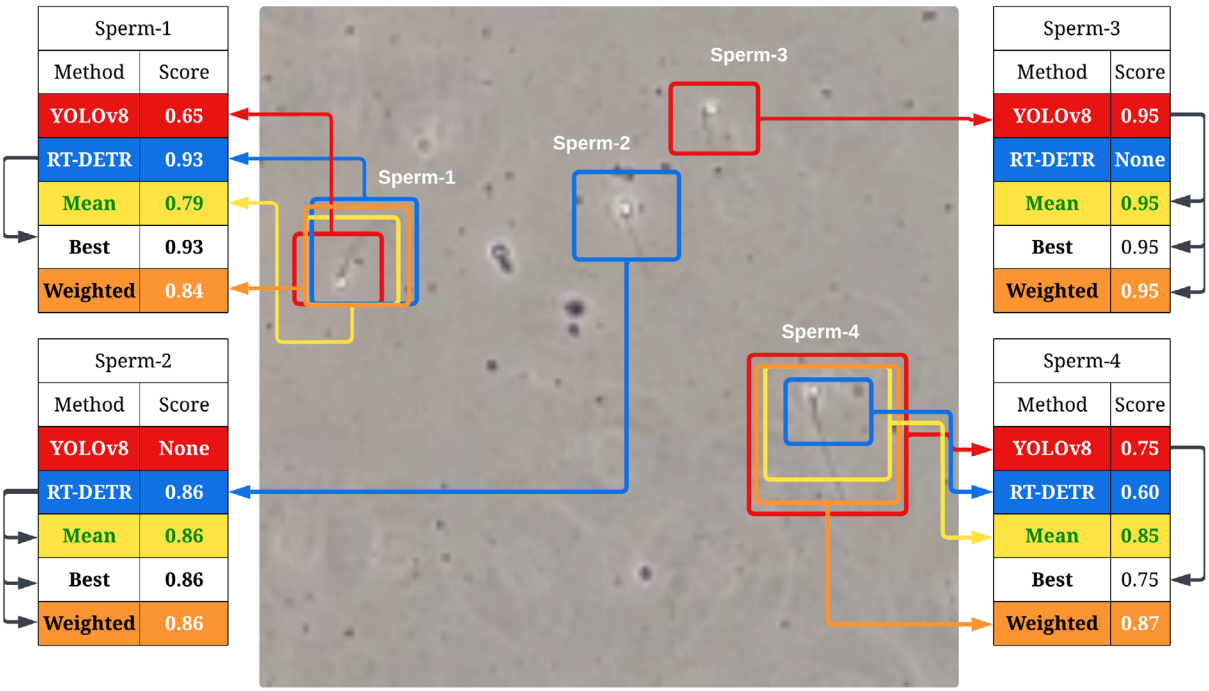

Figure 7. The demonstration of the proposed fusion ideas with deep learning-based detectors.

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The detection and tracking performance metrics for the proposed methods are presented in **Table 8**.

When analyzing the results of the proposed methods presented in Table 8, it is seen that the most successful method in terms of both detection and tracking problems is the mean method, which performs the fusion process by averaging the detections. It is observed that there is not a big difference between the fusion methods, especially in terms of detection metrics. However, although the performance values obtained seem to be close to each other, the IDSW value, which expresses the consistency of the identities assigned to the sperms throughout the video, is lower in the mean method compared to the other methods, so it is understood that this method provides a more consistent identification.

Considering that the mean method is the most successful among the ensemble methods, a comparison between the individual models and the mean model is presented in Table 7. When examining the performance metrics, the mean model achieves higher F1-score (87.87%) by maintaining a balance between precision (88.09%) and recall (87.94%), compared to the individual models. In particular, in the key object detection metrics, mAP50 and mAP50:95, the mean method demonstrates ≈5% and 3% better performance, respectively, compared to the best-performing individual model, YOLOv8n.

When evaluated in terms of tracking performance, the proposed mean ensemble method achieves ≈2% better results in both HOTA and MOTA metrics compared to the RT-DETR and YOLOv8n models. Furthermore, when comparing the total IDSW values across all videos, it is observed that the mean ensemble model performed 59 fewer identity switches than the RT-DETR-L model and 26 fewer than the YOLOv8n model. As a result, the mean method demonstrates ≈5% and 3% higher IDF1 values compared to RT-DETR-L and YOLOv8n models, respectively.

The detection and tracking results of the models on Vid-34 are presented in **Figure 8**. The performance of the methods has been evaluated, particularly based on the IDSW that occur during sperm collisions. In this context, a detailed analysis has been conducted using a selection of sperms from Vid-34. In Figure 8, the GT trajectories and IDs of sperm cells are presented alongside the trajectories and IDs obtained by the proposed mean ensemble method and individual models, enabling a comparison of the results. The expression "frame0:315" at the top of the figure refers to the video segment from the beginning (frame 0) to the 315th frame. When the frame range 0–315 is examined, it was observed that the mean ensemble method assigned the identities 16, 44, and 2 to the sperms, respectively. In other methods, the RT-DETR-L method assigned the identities 25, 64, and 3, while the YOLOv8 method assigned the identities 9, 50, and 1 to the same sperms. A comparison between the methods revealed that the sperms with identities 44, 64, and 50 in Vid-34 represent the same sperm. Furthermore, it was observed that this sperm follows a different trajectory in each method within

Table 8. The results of Scenario III ensemble methods.

Video name	Ensemble method	Detection performance					Tracking performance							Video specifications				
		P	R	F1	mAP50	5–95	HOTA	DetA	AssA	MOTA	MOTP	MODA	IDF1	Dets	GTDets	IDs	GTIDs	IDSW
Vid-31	Best	88.58	92.41	90.46	92.32	51.55	64.24	57.25	73.72	76.59	71.89	76.65	87.13	16 620	15 598	62	38	10
Vid-32	Best	79.13	86.86	82.73	83.70	44.03	60.86	51.84	73.70	61.51	74.50	61.58	77.91	21 978	19 297	66	40	14
Vid-33	Best	93.58	90.03	91.71	89.89	52.56	69.02	64.57	75.12	84.84	76.51	84.91	89.47	18 660	19 156	54	36	14
Vid-34	Best	87.08	85.36	86.13	85.07	48.99	64.20	57.30	73.23	71.64	75.68	71.73	81.95	14 869	14 834	58	34	14
Vid-35	Best	99.03	99.12	98.95	99.12	78.58	83.41	81.48	85.70	99.50	83.64	99.50	99.75	1784	1783	3	3	0
Vid-36	Best	80.59	72.91	76.52	69.00	23.92	50.99	46.64	57.33	57.80	66.97	57.98	71.07	43 099	45 998	162	97	84
Vid-31	Mean	88.69	92.41	90.44	92.32	54.75	66.40	60.30	74.58	80.90	73.45	80.97	88.74	16 288	15 598	65	38	12
Vid-32	Mean	79.41	87.24	83.06	84.16	46.10	62.47	52.02	77.64	61.80	74.73	61.87	79.93	21 964	19 297	70	40	14
Vid-33	Mean	93.60	90.21	91.82	90.07	56.02	71.17	65.83	79.85	84.95	77.74	85.02	90.13	18 628	19 156	53	36	13
Vid-34	Mean	87.73	86.09	86.82	85.88	49.51	65.67	57.67	78.81	72.79	75.20	72.87	84.08	14 855	14 834	56	34	11
Vid-35	Mean	98.92	99.01	98.84	99.01	73.50	81.86	79.52	84.91	99.50	81.78	99.50	99.75	1784	1783	3	3	0
Vid-36	Mean	80.18	72.71	76.22	67.86	23.67	52.03	46.61	60.46	58.61	66.70	58.77	72.10	43 015	45 998	158	97	73
Vid-31	Weighted	88.58	92.41	90.45	92.32	54.68	65.72	59.08	74.58	78.79	73.36	78.86	87.81	16 627	15 598	64	38	12
Vid-32	Weighted	79.46	87.21	83.08	84.13	46.13	62.52	52.07	77.76	61.87	74.74	61.94	79.97	21 979	19 297	69	40	13
Vid-33	Weighted	93.62	90.02	91.73	89.89	56.00	71.37	66.00	80.13	85.05	77.73	85.12	90.30	18 660	19 156	54	36	13
Vid-34	Weighted	87.74	86.01	86.78	85.79	49.60	65.73	57.72	78.87	72.86	75.24	72.93	84.13	14 865	14 834	58	34	11
Vid-35	Weighted	98.92	99.01	98.84	99.01	73.86	82.02	79.74	84.98	99.61	81.90	99.61	99.80	1784	1783	3	3	0
Vid-36	Weighted	80.33	72.67	76.26	67.82	23.72	51.99	46.60	60.37	58.66	66.70	58.84	72.51	43 085	45 998	163	97	85
Average and combine	Best	88.00	87.78	87.75	86.52	49.94	65.45	59.85	73.13	75.31	74.86	75.39	84.55	117 010	116 666	405	248	136
	Mean	88.09	87.94	87.87	86.55	50.59	66.60	60.32	76.04	76.42	74.93	76.50	85.79	116 534	116 666	405	248	123
	Weighted	88.11	87.89	87.86	86.49	50.67	66.56	60.20	76.11	76.14	74.94	76.22	85.75	117 000	116 666	411	248	134

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

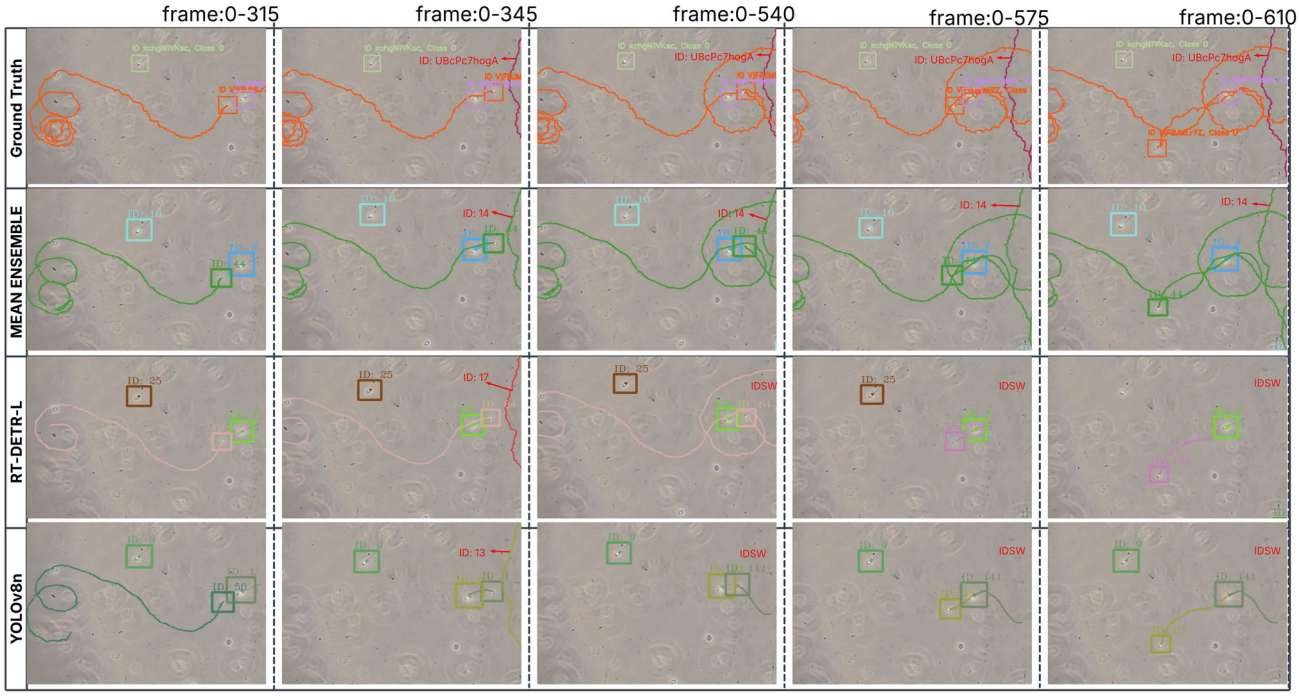

Figure 8. Comparison of detection and trajectory results obtained using YOLO, RT-DETR, and the mean ensemble method on the Vid-34.

the frame range 0–315. The trajectory followed by the mean ensemble method is significantly longer compared to the other individual methods. This is due to the fact that the methods change IDs across frames after assigning an ID to the relevant sperm. Compared to the GT trajectory, the mean ensemble method performs less IDSW than the other methods and obtains a result closer to the actual trajectory of the sperm. Furthermore, the 315th frame represents the moment before the sperms with IDs 44, 64, and 50 collide with those having IDs 2, 3, and 1. To investigate whether IDSW occurred after the sperm collision, the frame range 0–345 was examined. When analyzing frame 0–345, no IDSW was observed in the mean ensemble and RT-DETR-L methods. However, in the YOLOv8n method, an IDSW occurred between the sperm with ID 50 and the sperm with ID 1, and the tracking continued with ID 1. During this process, a new ID was assigned to the sperm with ID 1 in the YOLOv8n method. Furthermore, at the far right of the 345th frame, a trajectory belonging to a sperm moving from bottom to top has been detected by all three methods. The detected sperm is positioned higher than the other sperms under examination, so the ID information is not visible. To track this ID information, the sperm's IDs were manually added to the visual, with values 14, 17, and 13 assigned respectively. As we progress to the frame range 0-540, it is observed that no IDSW occur in the mean ensemble method, while IDSW events are observed in the objects detected by the RT-DETR-L and YOLOv8n models. In the RT-DETR-L method, it is noted that the sperm with ID 17 no longer has its trajectory visible in the frame, while in the YOLOv8n model, the trajectory of the same sperm is no longer drawn due to the change in its ID from 13. Additionally, in the YOLOv8n model, it was observed that the sperm with ID 1 in the previous frame had its ID changed to 141 due to an IDSW and continued with this new ID. When examining the sperm IDs in the frame range 0–575 after the collision, it was found that in the mean ensemble method, no changes occurred in the IDs despite the collision, and the IDs were assigned consistently. In contrast, IDSW were observed in the other individual methods. In the RT-DETR-L model, the sperm previously assigned ID: 64 in previous frames now continues with ID: 182. In the YOLOv8n method, an IDSW event occurred between IDs 141 and 83. When examining the final frame of Vid-34, frame 0–610, it was observed that with the mean ensemble method, the complete trajectories of each sperm were obtained consistently without any IDSW from the beginning to the end of the video. In contrast, for the individual models, IDSW were observed even for immobile sperms, and the trajectories of the moving sperms were not obtained consistently.

The sperm trajectories of Vid-36, which has the highest GTID and GTDET numbers, and Vid-34 are presented in **Figure 9**. The detections and tracking trajectories shown in the figure were obtained using the mean ensemble method. Figure 8 presents a visual fragment illustrating the sperm IDSW event, which is also depicted in Figure 9 as dashed lines. The complete trajectories of each sperm throughout the entire video can be observed in Figure 9. Upon examination of these trajectories, it is evident that some sperm move linearly without any oscillation (e.g., Figure 9b, ID:375), while others exhibit circular motility (e.g., Figure 9a, ID:11), some remain stationary (e.g., Figure 9b, ID:2), and a few move at a slower pace compared to others (e.g., Figure 9b, ID:33). The use of wide-angle videos to create the Hi-LabSpermTracking dataset provides such analyses, enabling insights into the different motility patterns of the sperm.

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

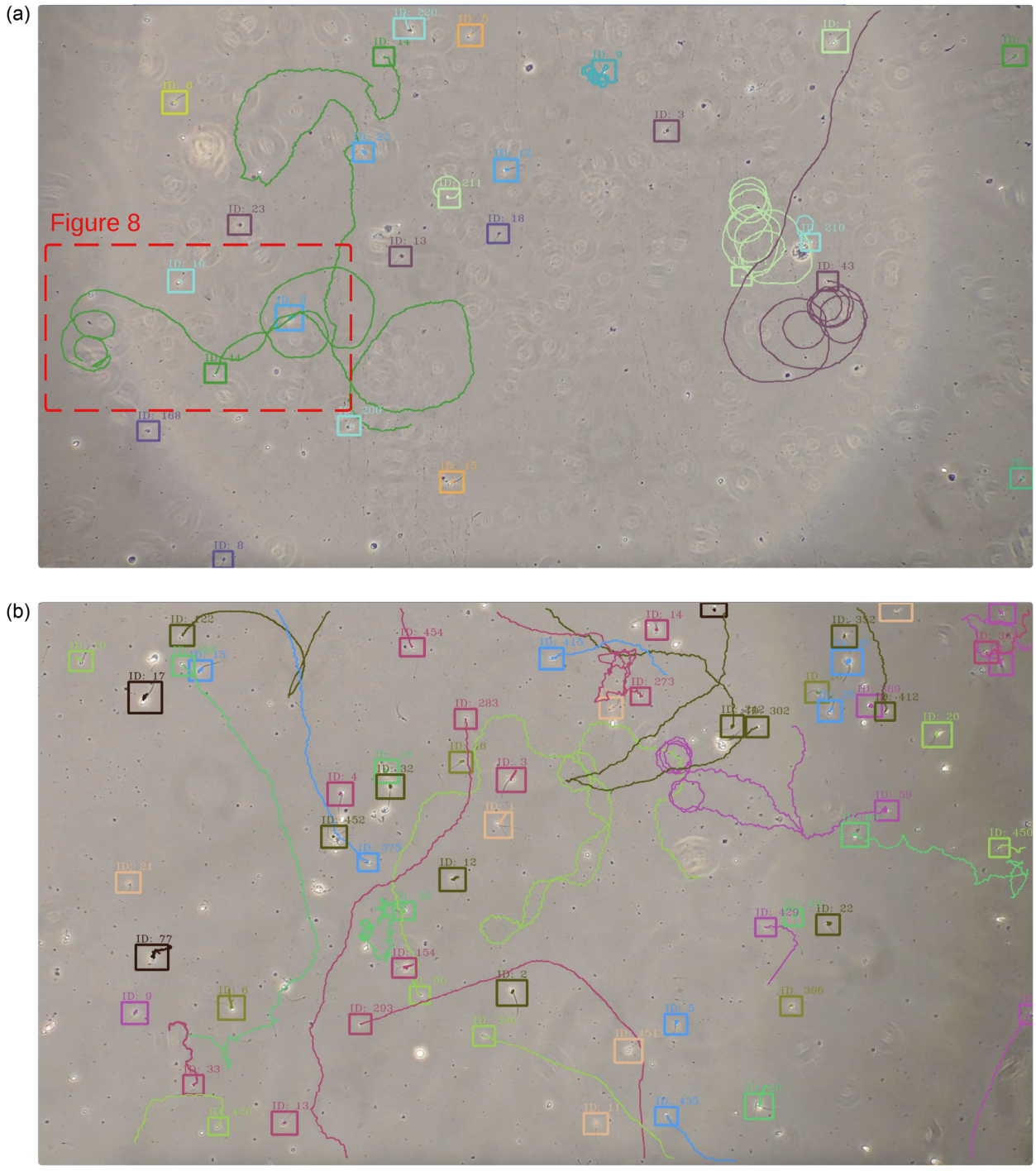

Figure 9. Detection and tracking results on a) Vid-34 and b) Vid-36 using the mean ensemble method (frame: 0–last frame).

5. Discussion

In this study, detections obtained using CNN and transformer-based object detection methods were combined with various ensemble learning techniques, and performance evaluation was conducted for sperm detection and tracking problems using the DeepSORT object tracking algorithm. The experiments carried out in this study were conducted on the novel Hi-LabSpermTracking dataset, which is expected to make significant contributions to the literature, particularly in the field of sperm detection and tracking. This dataset contains a total of 66 videos from 14 different patients and will be made publicly available as part of the study. The performance evaluation of individual and ensemble models for detection problems was conducted using Precision, Recall, F1-score, mAP50, and mAP50:95 metrics.

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

For MOT problems, the performance of the models was analyzed using HOTA, MOTA, and IDF1 metrics.

Table 1 provides a general summary of sperm-related datasets available in the literature. Upon examining the table, it is evident that the majority of the existing datasets were created for the classification of sperm morphology.[24–28,30,32,33,35,37] When examining the datasets developed for sperm detection problems, it is observed that these studies primarily focus on the detection of sperm cells, with the detections labeled using only bounding boxes (bbox).[20–22,35,43] However, for sperm tracking problems, in addition to bbox labels, each sperm needs to be consistently associated with a unique ID across all frames. This additional requirement is critical for the accurate analysis of sperm cell trajectories. In the literature, the only dataset that offers the necessary unique identification and consistent labeling across all frames for sperm tracking problems is the VISEM-Tracking dataset.[19] While this dataset is an important resource in the field of sperm tracking, it has some limitations. First, the fact that the videos in the dataset are only 30 s long significantly limits the scope of the study. Furthermore, it has been determined that some of the videos in the dataset (e.g., 14, 15, 21, 23, 35, 47, 54, 82) are created by merging two different videos, while others (e.g., 29, 52, 60) are formed by merging three different videos. This situation makes it impossible to consistently analyze the trajectory of a sperm cell throughout the entire video. Additionally, the videos in the VISEM-Tracking dataset were captured with a 400× magnification and have a resolution of 640 × 480. These technical specifications result in sperm cells being recorded with a narrow field of view, which limits the detailed, wide-angle analysis of their trajectories. A wider field of view is crucial as it allows for the tracking of sperm trajectories across a larger plane, providing a more comprehensive perspective. This capability is particularly important for reliably obtaining the metrics required for sperm motility classification, such as VCL (curvilinear velocity), VAP (average path velocity), and VSL (straight-line velocity), which are essential for the WHO's Grade A, B, C, and D sperm motility classifications.[18,54] By capturing a broader range of sperm motility, these metrics can be derived with greater accuracy and consistency, thereby enhancing the reliability of motility assessments.

The Hi-LabSpermTracking dataset created within the scope of this study consists of 60-second-long videos that have been recorded without any interruptions or skips. The videos were captured using a phone integrated with a microscope that has 100× magnification. This approach provides a wide field of view, allowing for detailed analysis of the motility trajectories of sperm cells. Hi-LabSpermTracking, its comprehensive labeling and unique ID information, holds significant potential to make important contributions to the literature in both sperm detection and sperm tracking problems. With these features, the dataset provides researchers with the opportunity to study the behavior of sperm cells in greater detail and develop new methods.

Experiments conducted using the Hi-LabSpermTracking dataset were carried out under three different scenarios. In Scenario I, a commonly used method for dataset partitioning from the literature was applied. In this approach, frames from all the videos are combined and randomly shuffled to create the training and test sets.[18,20–22,43] Using this approach, very high values were achieved with YOLOv8n, including 97.77% Precision, 98.14% Recall, 97.9% F1-score, 98.9% mAP50, and 77.80% mAP50:95. However, despite these high results, it has been found that this method is not suitable for real-world problems. The main reason for this is that, due to the random selection of frames for the test set, it is not possible to calculate any metrics related to the sperm's motility trajectories. This demonstrates that the method is functional for detection problems but inadequate for tracking problems. To eliminate this issue and provide a more realistic scenario, a different data partitioning method was applied in Scenario II. This approach aims to enable more comprehensive evaluations for both detection and tracking problems.

In Scenario II, in order to calculate tracking metrics in addition to detection metrics, the videos in the dataset were divided such that the last 15 s were allocated to the test set, while the remaining part was assigned to the training set. This approach allowed for obtaining a video sequence in the test set, unlike the commonly used random splitting method in the literature. The video sequences enabled the calculation of tracking metrics for the trained model in addition to detection metrics. In this scenario, using the YOLOv8n architecture, detection metrics were achieved as follows: 92.81% Precision, 84.19% Recall, 87.63% F1-score, 83.88% mAP50, and 63.98% mAP50:95. For tracking problems, the following metrics were calculated: 72.27% HOTA, 77.88% MOTA, and 86.31% IDF1. These results demonstrate that the model can be evaluated comprehensively, not only for detection but also for tracking problems.

The performance values obtained in Scenarios I and II demonstrate highly promising results when compared to studies on sperm detection and tracking listed in Table 2. In a study utilizing a small subset of the VISEM dataset videos, Valiuskaite et al. applied the Scenario I data partitioning method and achieved an accuracy of 91.77%.[20] Similarly, another study employing the same data partitioning method[43] reported an mAP50 value of 87.5% using the YOLOv5s model. Additionally, Yuzkat et al. achieved an mAP50 of 85.2% on their self-constructed SeSVid dataset by fusing three different deep learning-based detection algorithms.[21] These comparisons demonstrate that the Hi-LabSpermTracking dataset and the methods employed in Scenarios I and II are competitive with, and in some cases outperform, existing studies in the literature. Zhang et al. utilized a modified YOLOv8s model along with their proposed tracking algorithm on the VISEM-Tracking dataset.[22] In their study, a total of 29 196 frames from the dataset were partitioned randomly, and data augmentation techniques were applied to both the training and test sets. The authors reported 74.303% HOTA and 71.767% MOTA metrics, although they did not explicitly specify which videos from the dataset were used during the testing phase. However, applying data augmentation to both training and test data implies that the results may be biased and have limitations in representing real-world problems. This raises a notable concern about the generalizability of the model.

When the results obtained in Scenario I and Scenario II are compared with the studies in the literature, although more consistent results are obtained, the performance metrics of the models trained with the implemented data partitioning methods do not reflect the actual performance values for real world problems. The main reason for this is that the training and test sets contain frames belonging to the same video, even if they are different. In the real world, new sperm samples that the model has never

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

encountered before are most likely to be encountered. For this reason, in Scenario III, a data partition method is adopted in which the training and test videos are completely separated from each other. In this scenario, sperms were detected using the proposed ensemble learning methods and sperm tracking was performed with the DeepSORT method. This approach is designed to increase the real-world applicability of the model and to obtain more reliable results.

The individual models trained in Scenario III were combined with three different ensemble methods and their sperm detection and tracking performances were evaluated. As a result of the experiments, the proposed mean ensemble method achieved the most successful detection performance with 87.87% F1-score and 86.55% mAP50 values on Hi-LabSpermTracking dataset. Similarly, in MOT problems, the mean ensemble method showed the best tracking performance with 66.60% HOTA, 76.42% MOTA, and 85.79% IDF1 metrics. The same data partitioning method used in Scenario III was used by Thambawita et al. on the VISEM-Tracking dataset.[19] In this study, variants of the YOLOv5 model were used to demonstrate the suitability of the dataset for sperm detection problems. In the experiment with the YOLOv5m model, the most successful results were obtained with 57.12% P, 22.79% R and 22.16% mAP50 metrics. When the results of the study are analyzed, it is observed that the VISEM-Tracking dataset exhibits very low performances in sperm detection problems.[19] This indicates that the dataset used does not provide sufficient diversity and richness for sperm detection.

Our experimental findings demonstrate that deep-learning-based detectors, such as YOLOv8n and RT-DETR-L, can achieve strong performance for sperm detection and tracking on the Hi-LabSpermTracking dataset. However, interpreting the results requires distinguishing between different aspects of tracking performance, 1) detection (how well each sperm cell is found and localized) and 2) association (how consistently each sperm cell is tracked across frames). The HOTA metric offers a unifying framework for dissecting these components and understanding trade-offs in tracking algorithms' performance. One of the central insights of the HOTA metric is that tracking performance should weigh both detection and association errors equally. In the context of our study: Scenario II revealed high detection accuracy (precision, recall, mAP) across many videos but showed that, despite strong detection, the overall tracking metrics (HOTA-like measures such as MOTA and IDF1) could still suffer if identity consistency (i.e., low ID switches) was not maintained. This aligns with the HOTA philosophy that excellent detection alone is insufficient for robust tracking; strong association is equally critical. Scenario III, which splits training and test sets by entirely different videos (mimicking real-world usage), underscores this balance further. Both YOLOv8n and RT-DETR-L individually displayed varying degrees of detection success—some videos achieved higher recall but moderate precision, or vice versa. However, as in HOTA, the overall tracking success also depends on maintaining identity consistency across frames. In our results, YOLOv8n generally produced fewer missed detections (higher recall) and fewer total ID switches than RT-DETR-L. By contrast, RT-DETR-L had more ID switches overall (182 vs 149) and also missed more sperm in challenging frames, contributing to its lower average recall. This interplay echoes HOTA's motivation to avoid metrics that focus narrowly on detection (e.g., MOTA) or association (e.g., IDF1). Similarly, our data show that emphasizing detection or association alone can misrepresent true tracking quality.

HOTA highlights the importance of looking beyond short-term identity switches and examining long-term consistency in tracking. Our results show that certain videos—especially those with high sperm density—exhibit increased IDSW after collisions or occlusions. In simpler videos (e.g., those with fewer sperm cells and minimal overlap), most of our tested models showed near-ideal detection and stable tracking—once a sperm cell was identified, the ID rarely changed. In more crowded videos (e.g., Vid-36), identity switches grew more frequent due to overlapping trajectories. Nevertheless, YOLOv8n, which excelled at detecting most sperm (higher recall), also ended up with fewer ID switches overall (149) than RT-DETR-L (182). Although RT-DETR-L was more conservative in detection (lower recall), it did not reduce ID switches as one might expect. This outcome underscores HOTA's core premise that sustained, long-term association consistency is just as critical as strong detection—incremental gains in detection alone (or attempts to be conservative) do not necessarily translate to fewer identity errors over a long video sequence.

Our proposed ensemble strategies (best, mean, weighted) reflect an important principle: combining detection outputs can reduce certain types of errors but must be done carefully to avoid inflating false positives or missing the credit for correct associations. Much as HOTA balances detection and association, our ensemble approach aims to ensure that overly aggressive bounding-box predictions from one model do not overshadow the precise bounding boxes of the other—or vice versa. The best-performing ensemble (mean) consistently lowered identity switches and increased mAP50:95. This synergy reaffirms the HOTA viewpoint that reducing both false detections (precision) and missed detections (recall) while maintaining consistent tracking identities leads to a stronger overall performance.

6. Conclusion

Spermiogram testing is performed using traditional visual assessment methods and computer-aided sperm analysis systems. Although visual assessment is considered easy and cost-effective to implement, training qualified experts in the field is a lengthy and costly process. Additionally, the evaluation results obtained by this method vary significantly depending on the experts' experience. CASA systems, as another evaluation method, are not widely used due to their high costs and complex structures. In this context, the development of inexpensive, easily accessible, and reliable systems as alternatives to existing methods has become an important global need.

In this study, the labeled Hi-LabSpermTracking dataset required for the development of deep learning-based systems enabling automatic detection and tracking of spermatozoa is introduced, and baseline results obtained with deep learning-based sperm detection and tracking algorithms are presented. The Hi-LabSpermTracking dataset, consisting of videos acquired through a mobile phone-based imaging system integrated with a microscope, allows the developed models to be used in many laboratories worldwide with a traditional microscope and mobile phone, without the need for any additional advanced imaging systems.

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Furthermore, the YOLOv8n and RT-DETR-L models used in the study were combined with three different ensemble learning methods, and detections were tracked using the DeepSORT algorithm. As a result, the proposed mean ensemble fusion approach achieved more successful detection and tracking results compared to individual models.

7. Limitations and Future Work

In future studies, it is aimed to automate the labeling processes, which are labor-intensive and costly in terms of time, by using the models trained within this study through a pseudolabeling method, and to increase the number of labeled videos in the Hi-LabSpermTracking dataset. This method will enable more effective and systematic labeling of other unlabeled videos. Consequently, it is expected to obtain models with better generalization capabilities. Results obtained by training different sperm detection and tracking algorithms on the newly created dataset will be presented and compared. Similar to the fusion methods for sperm detection performed in this study, it is planned to fuse sperm trajectories obtained using algorithms such as ByteTrack, BotSort, and DeepOC-Sort.[55–57] Additionally, future work aims to utilize trajectories obtained through deep learning-based models to extract key motility metrics, including VCL, VAP, and VSL. These metrics will then be used to classify sperm motility into Grade A, B, C, and D categories, as defined by the WHO.[1] This approach leverages wide-field video recordings to ensure a more comprehensive and accurate analysis, thereby enhancing the reliability of sperm motility classification and broadening the applicability of the proposed system in clinical and research contexts.

Acknowledgements

This work was supported by TUSEB (Turkey Health Institutes Presidency) under the project Mobile Phone-Based Sperm Motility Analysis and Reporting System, Project Number: 27698. The authors express their sincere gratitude for the financial and scientific support provided by TUSEB, which contributed significantly to the successful realization of this study.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

dataset benchmark, deep learning, infertility, sperm detection and tracking

Received: January 26, 2025
Revised: June 11, 2025
Published online:

[1] World Health Organization, *Infertility Prevalence Estimates, 1990–2021*, World Health Organization, Geneva, Switzerland **2023**.
[2] World Health Organization, *WHO Laboratory Manual for the Examination and Processing of Human Semen*, World Health Organization, Geneva, Switzerland **2021**.
[3] J. Karibe, S. Kuroda, T. Saito, Y. Ishibashi, K. Usui, T. Takeshima, M. Komeya, Y. Yumura, *Andrologia* **2022**, *54*, e14266.
[4] R. Cannarella, M. Gül, A. Rambhatla, A. Agarwal, *Endocrine* **2023**, *79*, 1.
[5] A. Sharma, S. Minhas, W. S. Dhillo, C. N. Jayasena, *J. Clin. Endocrinol. Metab.* **2021**, *106*, e442.
[6] A. Agarwal, S. Baskaran, N. Parekh, C. L. Cho, R. Henkel, S. Vij, M. Arafa, M. K. P. Selvam, R. Shah, *The Lancet* **2021**, *397*, 319.
[7] F. Gholami Jourabi, S. Yari, P. Amiri, A. Heidarianpour, H. Hashemi, *Andrologia* **2021**, *53*, e13850.
[8] D. Durairajanayagam, D. Singh, A. Agarwal, R. Henkel, *Andrologia* **2021**, *53*, e13666.
[9] J. S. Su, N. J. Farber, S. C. Vij, *Andrologia* **2021**, *53*, e13576.
[10] H. O. Ilhan, N. Aydin, *Biomed. Signal Process. Control* **2018**, *41*, 129.
[11] A. Aktas, G. Serbes, H. O. Ilhan, in *2023 8th Inter. Conf. on Computer Science and Engineering (UBMK)*, IEEE, Piscataway, NJ **2023**, pp. 330–335.
[12] A. Chen, C. Li, M. M. Rahaman, Y. Yao, H. Chen, H. Yang, P. Zhao, W. Hu, W. Liu, S. Zou, N. Xu, M. Grzegorzek, *Intell. Med.* **2023**, *4*, 114.
[13] H. O. Ilhan, G. Serbes, *Biomed. Signal Process. Control* **2022**, *71*, 103246.
[14] H. O. Ilhan, G. Serbes, N. Aydin, *Comput. Biol. Med.* **2020**, *122*, 103845.
[15] L. Spencer, J. Fernando, F. Akbaridoust, K. Ackermann, R. Nosrati, *Adv. Intell. Syst.* **2022**, *4*, 2200111.
[16] M. Yüzkat, H. O. Ilhan, N. Aydin, *Comput. Biol. Med.* **2021**, *137*, 104790.
[17] S. Adinugroho, A. Nakazawa, *Inform. Med. Unlocked* **2024**, *45*, 101459.
[18] H. O. Ilhan, M. Yuzkat, N. Aydin, *Expert Syst. Appl.* **2021**, *186*, 115774.
[19] V. Thambawita, S. A. Hicks, A. M. Storås, T. Nguyen, J. M. Andersen, O. Witczak, T. B. Haugen, H. L. Hammer, P. Halvorsen, M. A. Riegler, *Sci. Data* **2023**, *10*, 260.
[20] V. Valiuškaitė, V. Raudonis, R. Maskeliūnas, R. Damaševičius, T. Krilavičius, *Sensors* **2020**, *21*, 72.
[21] M. Yuzkat, H.O. Ilhan, N. Aydin, *Biomed. Signal Process. Control* **2023**, *83*, 104630.
[22] C. Zhang, Y. Zhang, Z. Chang, C. Li, *Sensors* **2024**, *24*, 3493.
[23] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, IEEE, Piscataway, NJ **2024**, pp. 16965–16974.
[24] W. Yi, K. Park, J. Paick, in *Proceedings of the 20th Annual Inter. Conf. of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286)*, IEEE, Piscataway, NJ **1998**, Vol. *2*, pp. 974–977.
[25] L. Ramos, J. C. Hendriks, P. Peelen, D. D. Braat, A. M. Wetzels, *J. Androl.* **2002**, *23*, 882.
[26] V. Abbiramy, A. Tamilarasi, in *Inter. Conf. on Digital Image Processing and Information Technology*, Springer, Berlin **2011**, pp. 220–228.
[27] F. Ghasemian, S. A. Mirroshandel, S. Monji-Azad, M. Azarnia, Z. Zahiri, *Comput. Methods Prog. Biomed.* **2015**, *122*, 409.
[28] F. Shaker, S. A. Monadjemi, J. Alirezaie, A. R. Naghsh-Nilchi, *Comput. Biol. Med.* **2017**, *91*, 181.
[29] J. Riordon, C. McCallum, D. Sinton, *Comput. Biol. Med.* **2019**, *111*, 103342.
[30] V. Chang, A. Garcia, N. Hitschfeld, S. Härtel, *Comput. Biol. Med.* **2017**, *83*, 143.
[31] S. Javadi, S. A. Mirroshandel, *Comput. Biol. Med.* **2019**, *109*, 182.
[32] K. K. Mohammed, H. M. Afify, F. Fouda, A. E. Hassanien, S. Bhattacharyya, S. Vaclav, in *Inter. Conf. on Innovative Computing*

 © 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

and Communications: Proceedings of ICICC 2019, Springer, Berlin **2020**, Vol. *1*, pp. 857–871.
[33] H. O. Ilhan, I. O. Sigirci, G. Serbes, N. Aydin, *Med. Biol. Eng. Comput.* **2020**, *58*, 1047.
[34] M. I. Mahali, J. S. Leu, J. T. Darmawan, C. Avian, N. Bachroin, S. W. Prakosa, M. Faisal, N. A. S. Putro, *Sensors* **2023**, *23*, 6613.
[35] A. Chen, C. Li, S. Zou, M. M. Rahaman, Y. Yao, H. Chen, H. Yang, P. Zhao, W. Hu, W. Liu, M. Grzegorzek, *Biocybern. Biomed. Eng.* **2022**, *42*, 204.
[36] T. B. Haugen, S. A. Hicks, J. M. Andersen, O. Witczak, H. L. Hammer, R. Borgli, P. Halvorsen, M. Riegler, in *Proc. of the 10th ACM Multimedia Systems Conf.*, Amherst, MA, USA, June **2019**, pp. 261–266.
[37] A. Aktas, G. Serbes, M. Huner Yigit, N. Aydin, H. Uzun, H. Osman Ilhan, *IEEE Access* **2024**, *12*, 196070.
[38] K. Dewan, T. Rai Dastidar, M. Ahmad, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, IEEE, Piscataway, NJ **2018**, pp. 2299–2306.
[39] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, T. B. Haugen, Extracting temporal features into a spatial domain using autoencoders for sperm video analysis. *arXiv preprint arXiv:1911.03100* **2019**.
[40] S. A. Hicks, J. M. Andersen, O. Witczak, V. Thambawita, P. Halvorsen, H. L. Hammer, T. B. Haugen, M. A. Riegler, *Sci. Rep.* **2019**, *9*, 16770.
[41] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, T. B. Haugen, Stacked dense optical flows and dropout layers to predict sperm motility and morphology. *arXiv preprint arXiv:1911.03086* **2029**.
[42] S. Ottl, S. Amiriparian, M. Gerczuk, B. W. Schuller, *iscience* **2022**, *25*, 104644.
[43] R. Zhu, Y. Cui, J. Huang, E. Hou, J. Zhao, Z. Zhou, H. Li, *Diagnostics* **2023**, *13*, 1100.
[44] B. Li, Q. Meng, X. Li, Z. Wang, X. Liu, S. Kong, *Electronics* **2024**, *13*, 4411.
[45] V. Eddye, M. J. Iorpenda, V. Willert, in *2024 22nd Inter. Conf. on Research and Education in Mechatronics (REM)*, IEEE, Piscataway, NJ **2024**, pp. 262–267.
[46] N. Wojke, A. Bewley, D. Paulus, in *2017 IEEE Inter. Conf. on Image Processing (ICIP)*, IEEE, Piscataway, NJ **2017**, pp. 3645–3649.
[47] R. E. Kalman, A new approach to linear filtering and prediction problems **1960**.
[48] H. W. Kuhn, *Naval Res. Logist. Q.* **1955**, *2*, 83.
[49] L. St, S. Wold, *Chemom. Intell. Lab. Syst.* **1989**, *6*, 259.
[50] D. C. Montgomery, *Design and Analysis of Experiments*, John Wiley & Sons, Hoboken, NJ **2017**.
[51] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, B. Leibe, *Int. J. Comput. Vis.* **2021**, *129*, 548.
[52] W. F. Hendria, Q. T. Phan, F. Adzaka, C. Jeong, *ICT Express* **2023**, *9*, 258.
[53] W. Lu, C. Lan, C. Niu, W. Liu, L. Lyu, Q. Shi, *IEEE J. Selected Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1211.
[54] Ł. Witkowski, *Biocybern. Biomed. Eng.* **2013**, *33*, 179.
[55] N. Aharon, R. Orfaig, B. Z. Bobrovsky, Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651* **2022**.
[56] G. Maggiolino, A. Ahmad, J. Cao, K. Kitani, in *2023 IEEE Inter. Conf. on Image Processing (ICIP)*, IEEE, Piscataway, NJ **2023**, pp. 3025–3029.
[57] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, *European Conf. on Computer Vision*, Springer, Berlin **2022**, pp. 1–21.

© 2025 The Author(s). Advanced Intelligent Systems published by Wiley-VCH GmbH

26404567, 0, Downloaded from https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202500115 by Recep Tayyip Erdoan Üniversitesi, Wiley Online Library on [04/08/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License