

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Sperm Motility Analysis from Microscopic Videos

Rita do Carmo Pereira Marques



Master in Bioengineering - Biomedical Engineering

Supervisor: Prof. Luís Filipe Teixeira, PhD

Proponent: David Carrageta, PhD

October 28, 2025

Sperm Motility Analysis from Microscopic Videos

Rita do Carmo Pereira Marques

Master in Bioengineering - Biomedical Engineering

October 28, 2025

Resumo

O panorama moderno da medicina reprodutiva é marcado por duas tendências convergentes: uma crescente prevalência de infertilidade, particularmente de etiologia masculina, e o aumento da dependência das tecnologias de reprodução assistida. Ambos estes fatores dependem da avaliação precisa motilidade dos espermatozoides, um processo que continua a ser baseado em análises manuais subjetivas, demoradas e propensas a erro. Embora tenham sido desenvolvidos sistemas de análise de sémen assistida por computador (CASA) para facilitar uma avaliação mais objetiva, a sua adoção generalizada é limitada pelos elevados custos, complexidade técnica e adaptabilidade insuficiente a amostras humanas. Avanços recentes em inteligência artificial abrem perspectivas promissoras para superar estas limitações, abrangendo desde modelos de deep learning end-to-end capazes de prever distribuições de motilidade diretamente a partir de vídeos, até pipelines baseados em tracking que integram detecção de espermatozoides, reconstrução de trajetórias e classificação da motilidade.

Esta dissertação explora estes dois paradigmas concorrentes para análise automatizada da motilidade dos espermatozoides em vídeos de microscopia de campo claro. Em primeiro lugar, implementámos um método end-to-end de estado-da-arte, no qual uma rede neuronal convolucional (CNN) 3D foi treinada com sequências de fluxo óptico para capturar características de movimento e prever a distribuição de motilidade, enquanto uma ResNet18 foi treinada em frames brutos para extrair características espaciais e prever a concentração. Estes modelos pré-treinados foram então combinados numa rede multimodal que aproveitou tanto a informação de movimento como a espacial para melhorar a previsão de motilidade. Em segundo lugar, a principal contribuição deste trabalho foi o desenvolvimento de um pipeline inovador baseado em tracking, começando pela detecção de espermatozoides. Foram avaliadas múltiplas arquiteturas baseadas em YOLO para esta tarefa, selecionando-se o modelo de melhor desempenho para a implementação de um algoritmo de estimativa de concentração e posterior análise de trajetórias. O tracking foi realizado utilizando BoT-SORT-ReID com hiperparâmetros ajustados, integrando embeddings de características profundas extraídas do modelo de detecção para uma associação de identidade robusta. Adicionalmente, foram implementados métodos inovadores para calcular parâmetros cinemáticos precisos a partir das trajetórias obtidas, independentemente da taxa de frames do vídeo, garantindo uma classificação rigorosa de cada espermatozoide nas categorias de motilidade padrão recomendadas pela Organização Mundial da Saúde.

Embora a CNN 3D unimodal tenha capturado efetivamente padrões de movimento e apresentado desempenho razoável na previsão de motilidade com MAE global de 10.71 e R^2 de 0.78, a rede multimodal apresentou métricas superiores em todas as classes de motilidade, alcançando MAE de 8.49 e R^2 de 0.86 e demonstrando o benefício da integração de sinais de movimento e espaciais. Por sua vez, o algoritmo proposto baseado em tracking obteve o melhor desempenho, com MAE global de 8.46 e R^2 de 0.87, oferecendo ainda reconstrução de trajetórias celulares e estimativa robusta de parâmetros cinemáticos.

Para suportar aplicação prática, o algoritmo desenvolvido foi integrado numa interface gráfica

intuitiva, permitindo a estimativa de concentração e análise de motilidade a partir de vídeos de amostras de sémen, incluindo a avaliação de distribuições de motilidade e parâmetros cinemáticos. Em conclusão, o framework proposto alcança um equilíbrio entre precisão e interpretabilidade, destacando o seu potencial como uma ferramenta transparente e acessível para a análise de espermatozoides.

Abstract

The modern landscape of reproductive medicine is defined by two converging trends: a rising prevalence of infertility, particularly of male etiology, and a growing reliance on assisted reproductive technologies. Both of these factors critically hinge on the precise assessment of sperm cell motility, a process still reliant on subjective, time-consuming and error-prone manual analysis. Although computer-assisted semen analysis (CASA) systems have been developed to facilitate more objective evaluation, their widespread adoption is limited by high costs, technical complexity, and insufficient adaptability to human samples. Recent advances in artificial intelligence offer promising avenues to overcome these limitations, with approaches ranging from end-to-end deep learning models that predict motility distributions directly from video data, to tracking-based pipelines that combine sperm detection, trajectory reconstruction, and per-cell motility classification.

This dissertation explores these two competing paradigms for automated sperm motility analysis in brightfield microscopy videos. Firstly, we implemented a state-of-the-art end-to-end method in which a 3D convolutional neural network (CNN) was trained on optical flow sequences to capture motion features and predict motility distribution, while a ResNet18 was trained on raw frames to extract spatial features and predict concentration. These pretrained models were then combined into a multimodal network that leveraged both motion and spatial information to improve motility prediction. Secondly, the main contribution of this work was the development of a novel tracking-based pipeline, beginning with sperm detection. Multiple YOLO-based architectures were evaluated for this task with the best-performing model selected for the establishment of a concentration estimation algorithm and subsequent trajectory analysis. Tracking was carried out using BoT-SORT-ReID with tuned hyperparameters while integrating deep feature embeddings extracted from the detection model for robust identity association. Furthermore, innovative methods were implemented to compute precise kinematic parameters from the obtained tracks, regardless of video frame rate, ensuring accurate classification of each sperm cell into standard motility categories recommended by the World Health Organization.

While the unimodal 3D CNN effectively captured motion patterns and provided reasonable performance in motility prediction with overall MAE of 10.71 and R^2 of 0.78, the multimodal framework had improved metrics over all motility classes reaching MAE of 8.49 and R^2 of 0.86, demonstrating the benefit of integrating motion and spatial cues. Meanwhile, the proposed tracking-based algorithm achieved the best performance, with overall MAE of 8.46 and R^2 of 0.87, while additionally offering per-cell trajectory reconstruction and robust estimation of kinematic parameters.

To support practical application, the developed algorithm was ultimately integrated into an intuitive graphical interface, enabling sperm concentration estimation and motility analysis from microscopic semen sample videos, including assessment of motility distributions and kinematic parameters. In conclusion, the proposed framework achieves a balance of accuracy and interpretability, highlighting its potential as a transparent and accessible tool for sperm analysis.

UN Sustainable Development Goals

In any research project, it is valuable to consider not only the immediate technical outcomes but also the wider implications the work may have on society. The United Nations Sustainable Development Goals (SDG) provide a useful framework for reflecting on these broader contributions, encouraging researchers to align their efforts with global priorities such as health, equality, and innovation.

This dissertation contributes to a selection of SDG by developing a practical and accessible computational tool for sperm motility analysis, directly supporting research in reproductive health. The specific SDGs and specific targets aligned with this project include:

SDG 3 Ensure healthy lives and promote well-being for all at all ages.

- **Target 3.7:** Ensure universal access to sexual and reproductive health-care services.

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.

- **Target 9.5:** Enhance scientific research and upgrade technological capabilities.

SDG	Target	Contribution	Performance Indicators and Metrics
3	3.7	The development of an accessible system for semen analysis improves male infertility diagnosis in resource-limited settings. Additionally, it enables early detection of reproductive health issues, supporting timely interventions for family planning.	3.7.1 Proportion of women of reproductive age who have access to reproductive health services. Project metrics: Number of resource-limited laboratories adopting the sperm analysis tool.
9	9.5	Promotes scientific research directly, specifically in resource-limited settings, by developing an accessible computational tool that strengthens research infrastructure and fosters innovation in male reproductive health.	9.5.1 Research and development expenditure as a proportion of GDP. Project metrics: Number of research projects, publications or presentations utilizing the developed tool.

Acknowledgements

Em primeiro lugar, gostaria de agradecer ao meu orientador, Professor Luís Filipe Teixeira, pelo acompanhamento, orientação e apoio ao longo de todo este trabalho.

Agradeço também ao David Carrageta e a todos os membros do laboratório UMIB por disponibilizarem o dataset e por estarem sempre prontos a ajudar. Espero que este trabalho vos seja genuinamente útil e que possa contribuir para os vossos futuros projetos.

Obrigada ainda a todos os amigos que fiz neste percurso, de Bioquímica a Bioengenharia, pelas memórias que tornaram estes anos inesquecíveis. À minha família, pelo carinho e incentivo constantes, e por me ouvirem sempre, mesmo sem perceberem muito bem do que eu estava a falar. Por último, mas de forma alguma menos importante, ao Lucas, obrigada por cada momento de motivação, compreensão e paciência. Seria impossível sem todos vocês.

Rita Marques

“All things are difficult before they are easy.”

Thomas Fuller

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivation	2
1.3	Objectives	2
1.4	Main Contributions	2
1.5	Dissertation Structure	3
2	Fundamentals and Clinical Relevance of Sperm Motility Analysis	4
2.1	Semen and Spermatozoa	4
2.1.1	Semen: Function and Composition	4
2.1.2	Spermatozoa: Structure and Morphology	5
2.2	Human Sperm Motility	6
2.2.1	Definition and Biological Role	7
2.2.2	Motility Patterns and Heterogeneity	7
2.2.3	Determinants of Sperm Motility	7
2.3	Semen and Sperm Motility Analysis	8
2.3.1	Standard Semen Analysis	8
2.3.2	Clinical Significance of Sperm Motility Analysis	10
2.4	Summary	12
3	Technical Advances in Sperm Motility Analysis	13
3.1	Automated Semen Analysis	13
3.1.1	Historical Development of CASA	13
3.1.2	Current CASA Market Landscape	14
3.2	Computer Vision and AI in Sperm Motility Analysis	17
3.2.1	Traditional Approach	17
3.2.2	Artificial Intelligence in Sperm Motility Analysis	20
3.3	Summary	37
4	Methodology	38
4.1	Dataset	38
4.1.1	Source and Ethical Compliance	38
4.1.2	Sample Collection and Preparation	38
4.1.3	Video Acquisition and Dataset Overview	39
4.2	Computational Resources	41
4.3	End-to-End Approach	41
4.3.1	Input Modality Generation	41
4.3.2	Ground Truth Annotation and Scaling	43

4.3.3	Network Architectures	44
4.3.4	Training Setup	46
4.3.5	Training and Performance Evaluation	47
4.4	Proposed Algorithm	48
4.4.1	Data Preparation	48
4.4.2	Sperm Detection	52
4.4.3	Concentration Estimation	55
4.4.4	BoT-SORT-ReID Tracking	56
4.4.5	Kinematic Parameter Calculation	59
4.4.6	Motility Classification	62
4.4.7	Graphical Interface and Integration	63
5	Results and Discussion	66
5.1	End-to-end Approach Results	66
5.1.1	Training Monitoring	66
5.1.2	Performance Evaluation	68
5.2	Proposed Method Results	71
5.2.1	Detection	71
5.2.2	Tracking	73
5.2.3	Kinematic Parameters	77
5.2.4	Motility Classification	80
5.2.5	GUI Functionality	81
5.3	Discussion	85
6	Conclusions and Future Work	88
6.1	Conclusions	88
6.2	Future Work	89
6.3	Final Remarks	89
	References	90
A	Architectural Details of YOLOv9, v10, and v11	100
A.1	YOLOv9	100
A.2	YOLOv10	102
A.3	YOLOv11	102
B	BoT-SORT-ReID Workflow	104
C	Detection Training Results	107

List of Figures

2.1	Human semen sample under brightfield (A) and phase-contrast (B, adapted from [18]) microscopy. Examples of spermatozoa (red arrows) and non-sperm cells or debris (blue arrowheads) are indicated.	5
2.2	Human sperm cell structure. A) Schematic representation of a sperm cell, created with BioRender.com [21]. B) Scanning electron microscopy image of a real human sperm cell, adapted from: [22].	6
2.3	Morphological diversity of spermatozoa across species. A) Examples from various mammalian species, including hamster, human, bull, rat, echidna, guinea pig, dog, cat, and bandicoot, adapted from [24]. B) Examples from other vertebrate species, including fish (toadfish, elasmobranch), amphibians (toad, frog, salamander), reptiles (lizard), and birds (fringilla, domestic fowl), adapted from [25]. . .	6
3.1	Illustration of the most commonly implemented CASA systems in andrology laboratories. MES, Medical Electronic Systems. Source: [81].	15
3.2	Illustration of CASA trajectory paths and key kinematic parameters. VCL, curvilinear velocity. VSL, straight-line velocity. VAP, average path velocity. ALH, Amplitude of lateral head displacement. BCF, Beat-cross frequency.	18
3.3	Examples of challenges and limitations in traditional CASA systems. A) Bright-field microscopy image showing sperm with overlapping paths; collisions are marked with red circles. Source: [94]. B) Effect of frame rate on trajectory accuracy: dots indicate sperm positions per frame; the grey line shows the full path a high temporal resolution. Adapted from: [92]. C) True curvilinear path (red) vs. straight-line path (grey) over increasing time intervals, highlighting misrepresentation of non-linear motion. Adapted from: [92].	19
3.4	Conceptual hierarchy of artificial intelligence, machine learning, neural networks and deep learning. Source: [96].	20
3.5	Architecture of artificial neural networks. A) Schematic of an artificial neuron showing inputs (x), weights (w), bias (b), summation, activation function, and output. B) Neural network architecture composed of input, hidden, and output layers. Adapted from [99].	21
3.6	Overview diagram of the two main computational paradigms for state-of-the-art sperm motility assessment: Tracking-based and End-to-end approaches.	22
4.1	Example of variability between videos. First frame from four different samples recorded under varying imaging conditions.	39
4.2	Video dataset overview. A) Histogram of video duration distribution, in seconds. B) Histogram of video frame rate distribution, in FPS.	40

4.3	Box plots of sperm motility percentages (immotile, non-progressive, progressive) across the 90-video manually annotated dataset.	40
4.4	Generation of the D1 modality using dense optical flow with specified stride (X) and depth (Z), followed by stacking into video-based input. Source: [129].	42
4.5	Extraction of D2 input from the source frames used in D1 optical flow computation. Source: [129].	43
4.6	Example of a D1 sample (A, B) and its corresponding D2 sample (C) from the same video segment, used in our implementation.	43
4.7	Architectural diagrams of the unimodal models. A) 3D CNN for D1-based motility classification: left, single convolutional block; right, full 3D CNN. Our implementation differs from the original by using 3D max pooling with a reduced kernel size of $1 \times 2 \times 2$ and applying average pooling over spatial dimensions instead of global pooling, resulting in a fully connected layer input size of 256 instead of 512. B) Modified ResNet18 architecture for D2-based concentration regression. Source: [129].	45
4.8	Architectural diagram of the multimodal single-output network for motility classification. Our implementation differs from the original in that the 3D CNN output is 256, making the fully connected layer input 1256. Source: [129].	46
4.9	Overview of the proposed sperm motility analysis algorithm.	48
4.10	Examples of sperm head annotations in two sample images (A and B).	49
4.11	Examples of two frames (A and B) and their augmented versions as included in the final training set.	52
4.12	COCO mAP vs. inference latency (ms/image) for various YOLO architectures on T4 GPU with TensorRT (FP16). Source: [139].	52
4.13	IDF1, MOTA and HOTA comparisons of commonly used trackers with BoT-SORT and BoT-SORT-ReID on the MOT17 and MOT20 test sets. Source: [141].	57
4.14	Effect of fixed-window moving average on sperm trajectories at varying frame rates. The gray line represents the curvilinear path, while each colored path corresponds to a different frame rate using an 11-point moving average. Adapted from: [92].	60
4.15	GUI workflow for the sperm analysis algorithm. The diagram illustrates the sequence of user interactions and system responses, from loading input videos, selecting analysis options and parameter configuration to result generation. The border style and color of the boxes represent different output behaviors, which vary depending on whether single or batch video analysis is performed and the combination of analysis tasks selected by the user.	65
5.1	Training and validation loss across epochs for different cross-validation folds of the unimodal 3D CNN model.	66
5.2	Training and validation loss across epochs for different cross folds of unimodal 2D CNN model.	67
5.3	Training and validation loss across epochs for different cross folds of multimodal model for motility prediction.	68
5.4	F1-Confidence curve of the selected YOLOv11n model, where F1 score is plotted against the confidence score. The optimal confidence level of 0.591 corresponds to a maximum F1 score of 0.93.	73
5.5	Example YOLOv11n detection results on the test set using a confidence threshold of 0.591, showing true positives (green), false positives (red), and false negatives (blue).	73

5.6	Visual comparison of tracking output between (A, C) default and (B, D) tuned BoT-SORT-ReID parameters in two different videos. Frames shown correspond to the final frame of each video. Colored lines represent individual sperm tracks, with different colors indicating different track IDs.	74
5.7	Distribution of track lengths (frames) for the default (blue) and tuned (orange) configurations of the BoT-SORT-ReID, aggregated over the 90-video dataset. . .	75
5.8	Examples of occlusion handling by the tracker. (A–C) show successful identity preservation across an occlusion event. (D–F) depict a failure case, where a crossing leads to an ID switch. Frames are sampled before, during, and after the occlusion.	75
5.9	Tracking failure due to erratic cell movement, where different colors indicate separate track IDs incorrectly assigned to the same sperm cell. Frames are sampled in alternating sequence.	76
5.10	A spermatozoon with an abnormally large midpiece is incorrectly segmented into multiple tracks, each assigned a different ID, represented in distinct colors. . . .	76
5.11	Example of track overlays on the final frame of two different videos, showcasing valid, complete tracks (in blue) and excluded short tracks (in red).	76
5.12	Comparison of smoothing methods for a slower ($VCL = 13.25\mu\text{m/s}$) and a faster ($VCL = 21.45\mu\text{m/s}$) progressive sperm. Black shows the raw curvilinear path, red dashed the traditional moving average with a fixed temporal window (1.5s in A–B, 0.5s in C–D), and blue the proposed adaptive smoothing.	77
5.13	Limitations of the traditional riser-based method for ALH and BCF estimation. (A, B) Riser signal plots for a slower sperm track ($VCL = 13.25\mu\text{m/s}$) and a faster sperm track ($VCL = 21.45\mu\text{m/s}$), respectively, with peaks and troughs highlighted. (C, D) Corresponding zoomed-in views of the sperm tracks showing the curvilinear path (black) and the average path (blue). ALH peaks and BCF crossing points are derived from the riser signal.	78
5.14	Performance of the Proposed Geometric Method for ALH and BCF Estimation. (A) Track of a slower sperm ($VCL = 13.25\mu\text{m/s}$) with curvilinear path (black) and smoothed average path (blue). Obtained metrics: Mean ALH = $0.51\mu\text{m}$, Max ALH = $0.67\mu\text{m}$, BCF = 1.59 crossings/s. (B) Track of a faster sperm ($VCL = 21.45\mu\text{m/s}$) with the same representation. Obtained metrics: Mean ALH = $0.07\mu\text{m}$, Max ALH = $0.24\mu\text{m}$, BCF = 8.28 crossing/s.	79
5.15	Four examples of sperm tracks, including curvilinear (dark blue), straight (red) and average paths (light blue), and their calculated kinematic parameters obtained using the developed algorithm.	80
5.16	Predicted versus ground-truth motility distributions across 90 videos.	81
5.17	Sperm analysis tool main window.	82
5.18	Different tabs of sperm analysis tool results window: Counting results (A), Motility Results (B), and Kinematic parameters (C).	84
A.1	Overview of the YOLO main architecture, including backbone, neck and head. Source: [147]	101
C.1	YOLOv9t loss curves and validation metrics across 200 epochs.	107
C.2	YOLOv9s loss curves and validation metrics across 200 epochs.	108
C.3	YOLOv9m loss curves and validation metrics across 200 epochs.	108
C.4	YOLOv9c loss curves and validation metrics across 200 epochs.	109
C.5	YOLOv9e loss curves and validation metrics across 200 epochs.	109
C.6	YOLOv10n loss curves and validation metrics across 200 epochs.	110

C.7	YOLOv10s loss curves and validation metrics across 200 epochs.	110
C.8	YOLOv10m loss curves and validation metrics across 200 epochs.	111
C.9	YOLOv10b loss curves and validation metrics across 200 epochs.	111
C.10	YOLOv10l loss curves and validation metrics across 200 epochs.	112
C.11	YOLOv10x loss curves and validation metrics across 200 epochs.	112
C.12	YOLOv11n loss curves and validation metrics across 200 epochs.	113
C.13	YOLOv11s loss curves and validation metrics across 200 epochs.	113
C.14	YOLOv11m loss curves and validation metrics across 200 epochs.	114
C.15	YOLOv11l loss curves and validation metrics across 200 epochs.	114

List of Tables

2.1	Reference values for macroscopic semen parameters. Adapted from [6], as reported in [17].	8
2.2	Lower reference limits (5th percentiles) for microscopic semen parameters relevant to fertility prediction, adapted from [6].	11
3.1	Overview of CASA systems currently on the market: applications and sperm analysis device type.	15
3.2	Summary of state-of-the-art sperm detection.	25
3.3	Summary of state-of-the-art sperm tracking.	29
3.4	Summary of State-of-the-art Sperm Motility Assessment Studies.	34
4.1	Differences between the dataset used in the original study (VISEM) and our dataset.	41
4.2	Comparison of D1 input configuration: original article vs. our adaptation.	42
4.3	Training configurations for each model architecture in end-to-end multimodal approach, in accordance with [129].	47
4.4	Distribution of sperm detection dataset across train, validation, and test sets.	50
4.5	Applied data augmentations using Albumentations, with corresponding parameters and probabilities.	51
4.6	YOLO model scales selected for sperm detection, with corresponding number of parameters, number of layers, and GFLOPs (Giga Floating Point Operations Per Second), a measure of computational complexity.	53
4.7	BoT-SORT-ReID default and selected parameters in our implementation	58
4.8	Sperm motility classification criteria of proposed algorithm.	63
5.1	Unimodal 3D CNN performance metrics on the test set.	69
5.2	Original article’s unimodal 3D CNN performance metrics [129].	69
5.3	Unimodal 2D CNN performance metrics on the test set.	69
5.4	Multimodal model performance metrics on the test set.	70
5.5	Original article’s multimodal model performance metrics [129].	70
5.6	Performance comparison of YOLO detection models on validation and test sets. Metrics include precision, recall, mAP50, mAP50-95, and inference time per image. Selected model is shown in bold.	72
5.7	Evaluation metrics for predicted versus ground-truth motility distributions across 90 videos.	80
5.8	Summary of average performance metrics for different approaches tested for sperm motility analysis.	86

List of Acronyms

AI	Artificial Intelligence
ALH	Amplitude of Lateral Head displacement
ART	Assisted Reproductive Technologies
BCF	Beat-cross Frequency
CAGR	Compound Annual Growth Rate
CASA	Computer-assisted Semen analysis
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Networks
CSR-DCF	Channel and Spatial Reliability Discriminative Correlation Filter
DL	Deep Learning
ESHRE	European Society of Human Reproduction and Embryology
FPS	Frames Per Second
GELAN	Generalized Efficient Layer Aggregation Network
GFLOP	Giga Floating Point Operations Per Second
GMC	Global Motion Compensation
GUI	Graphic User Interface
HOTA	Higher Order Tracking Accuracy
ICBAS	Abel Salazar Institute of Biomedical Sciences
ICSI	Intracytoplasmic Sperm Injection
IoU	Intersection over Union
ID	Identity
IUI	Intrauterine Insemination
IVF	In Vitro Fertilization
JPDA	Joint Probabilistic Data Association
LIN	Linearity
LR	Learning Rate
MAE	Mean Absolute Error
ML	Machine Learning
MOTA	Multiple Object Tracking Accuracy
MS-CDMCM	Mean Shift–Collision Detection and Modified Covariance Matrix
MSE	Mean Squared Error
NMS	Non-maximum Suppression
NN	Nearest Neighbor
OSPA	Optimal Subpattern Assignment

PGI	Programmable Gradient Information
PSA	Partial Self-attention
RMSE	Root Mean Squared Error
SDG	United Nations Sustainable Development Goals
STR	Straightness
UMIB	Unit for Multidisciplinary Research in Biomedicine
VAP	Velocity of Average Path
VCL	Velocity of Curved Line
VSL	Velocity of Straight Line
WHO	World Health Organization
WOB	Wobble

Chapter 1

Introduction

1.1 Context

Male infertility, defined as the inability to conceive after one year of regular, unprotected sexual intercourse due to factors attributable to the male partner, has become an increasingly pressing global health issue [1]. Currently contributing to nearly 50% of all infertility cases worldwide [2], recent studies indicate that the age-standardized prevalence of male infertility is rising at a steady rate of 0.3% annually [3]. In parallel, the global use of assisted reproductive technologies (ART) such as in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI) has increased significantly, resulting in the birth of over 10 million children worldwide to date [4], with the ART market projected to grow at an annual rate of 8.3% through 2028 [5].

The ability of sperm cells to move efficiently, known as sperm motility, is widely recognized as one of the most informative biomarkers of male reproductive potential, as it directly affects the likelihood of successful fertilization [6]. As such, its assessment is essential for not only the diagnosis of male infertility but also guiding fertility treatment decisions, including the selection of ART. These are often based on the distribution of established sperm motility categories defined by the World Health Organization (WHO) [6], which reflect the quality and functional competence of sperm movement.

Despite its clinical importance, current motility assessment relies predominantly on manual microscopic evaluation, a method plagued by subjectivity, labor-intensiveness, and inter-observer variability. Computer-Semen Analysis (CASA) systems emerged in the 1970's to address these shortcomings, offering automated alternatives aimed at improving accuracy and consistency [7]. However, these systems are not without limitations. Persistent barriers include high costs, steep learning curves, insufficient robustness for routine use, and a focus on veterinary applications that perform poorly in humans due to interspecies differences in sperm morphology and the greater complexity and variability of human semen samples [6, 8, 9].

These limitations hinder widespread adoption in human clinical settings and scalable research, creating an urgent need for solutions that reconcile accessibility, clinical accuracy and operational efficiency. In response, Artificial intelligence (AI) and specially deep learning (DL) have emerged

as a promising approach [10]. These data-driven methods offer the potential for automated, high-precision motility analysis, with applications ranging from fully DL-based frameworks to hybrid strategies that couple DL-driven detection with traditional tracking algorithms.

1.2 Motivation

This dissertation is motivated by a specific need identified by the Unit for Multidisciplinary Research in Biomedicine (UMIB) research unit at Abel Salazar Institute of Biomedical Sciences (ICBAS) for a more practical and reliable solution for sperm motility assessment. This laboratory conducts experiments with human semen across a variety of studies and currently faces challenges in evaluating motility due to the limitations of manual methods and the lack of accessible automated tools suited to human samples. These issues hinder both the efficiency and reproducibility of their research workflows. By addressing this gap, the project also aligns with broader efforts to improve the quality of male fertility research and to strengthen the technical approaches available for sperm analysis. The motivation, therefore, lies not only in contributing to the scientific field but also in delivering tangible improvements to an active research environment that stands to directly benefit from such a tool.

1.3 Objectives

The primary objective of this dissertation is to develop an automated computational tool for sperm motility analysis using standard brightfield microscopic videos. This tool aims to directly address the challenges faced by the UMIB research unit and enhance their current sperm analysis practices. The specific objectives are as follows:

- Accurately detect and count sperm cells in video samples to enable reliable estimation of sperm concentration.
- Quantify the distribution of sperm motility categories - progressive, non-progressive and immotile - in each sample, following WHO guidelines.
- Reduce the time required for sperm motility assessment compared to manual evaluation, improving research workflow efficiency.

1.4 Main Contributions

The main contributions of this dissertation are:

- Benchmarking of state-of-the-art end-to-end DL models for sperm motility distribution prediction in brightfield and low-frame-rate microscopy videos.
- Development of a tracking-based pipeline for sperm motility analysis, which combines detection, tracking, and kinematic assessment. This approach comprised:

- Creation of a manually annotated dataset for sperm head detection, enabling training and evaluation of detection models.
- Systematic evaluation of YOLOv9, YOLOv10, and YOLOv11 architectures for sperm head detection, enabling informed selection of detection backbones for motility studies.
- Customization of BoT-SORT-ReID hyperparameters to improve the robustness and accuracy of sperm track reconstruction.
- Design of algorithms for improved kinematic parameter estimation, enabling reliable analysis of sperm tracks even at low frame rates.
- Integration of the full pipeline into a graphical interface, allowing practical use of the methods in laboratory settings and facilitating reproducible analysis.

1.5 Dissertation Structure

Following the introduction, this dissertation is organized into five additional chapters.

Chapter 2 covers the biological fundamentals of semen and sperm, and the clinical relevance of sperm motility. Chapter 3 reviews the development of CASA systems, traditional methods, and state-of-the-art advances in sperm motility analysis. Chapter 4 describes the dataset and methodology, covering both the implementation of a state-of-the-art end-to-end class distribution predictor and the proposed algorithm that includes sperm detection, concentration estimation, tracking, kinematic parameter computation, and classification. Chapter 5 presents the results from both approaches and discusses their implications. Chapter 6 evaluates the achievement of the set goals, synthesizes the main findings, and suggests directions for future work.

Chapter 2

Fundamentals and Clinical Relevance of Sperm Motility Analysis

This chapter introduces the biological and clinical foundations of sperm motility analysis, providing essential context for understanding its role in reproductive health and fertility assessment. It begins with an overview of semen composition and spermatozoa structure, then explores the mechanisms and patterns of human sperm motility, along with the factors that influence it. Building on this foundation, the chapter examines standardized approaches to semen analysis, highlighting the clinical significance of motility parameters in diagnosing male infertility and guiding assisted reproductive technologies.

2.1 Semen and Spermatozoa

2.1.1 Semen: Function and Composition

Semen is a complex biological fluid produced by the male reproductive system of mammals and other vertebrates. It acts as a transport medium for the male gametes, sperm cells (also known as spermatozoa), ensuring their nourishment and protection during transit through the female reproductive tract to achieve fertilization [11].

In humans, the production of this non-homogeneous fluid involves coordinated processes across the fully developed male reproductive organs. Spermatozoa originate in the testes through spermatogenesis, then mature and acquire motility in the epididymis. Upon ejaculation, sperm are propelled through the vas deferens and combined with secretions from the seminal vesicles (contributing 50-65% of semen volume), prostate gland (20-30%), and bulbourethral glands (<5%) [12, 13]. This mixture is expelled through the urethra as the final ejaculate.

Human semen is thus primarily composed of sperm cells – typically around 80 million per milliliter in healthy individuals, accounting for approximately 5% of the total volume – suspended in the combined glandular secretions, known as seminal plasma [12]. This organic medium contains a variety of proteins, enzymes, lipids, and other biochemical components that support sperm

survival, facilitate motility, and help neutralize the acidic environment of the female reproductive tract [14, 15]. Residual cellular components, including immune cells (e.g., leukocytes), epithelial cells shed from the male reproductive system, immature germ cells, and occasionally aggregated spermatozoa can also be found in healthy semen [16]. As illustrated in Figure 2.1, microscopy reveals the heterogeneous nature of semen, where spermatozoa can be distinguished from non-sperm cells or debris.

It is important to note that the composition of semen is dynamic, varying not only between individuals but also throughout a man's lifetime depending on factors such as age, health status and hormonal levels [17].

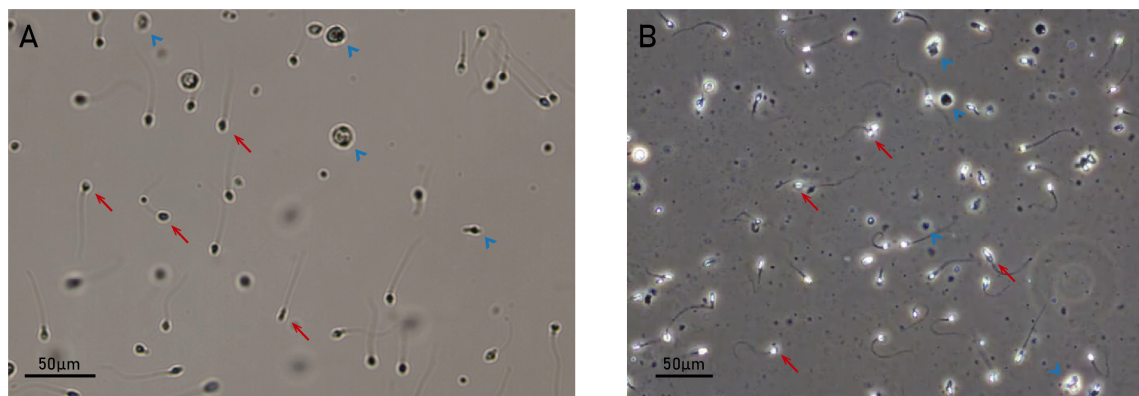


Figure 2.1: Human semen sample under brightfield (A) and phase-contrast (B, adapted from [18]) microscopy. Examples of spermatozoa (red arrows) and non-sperm cells or debris (blue arrowheads) are indicated.

2.1.2 Spermatozoa: Structure and Morphology

Spermatozoa are small, haploid, polarized and highly specialized cells, evolved to deliver the male's genetic material to the female egg during fertilization, with a structure optimized for this role [19].

As shown in Figure 2.2, a sperm cell consists of three primary regions – head, midpiece, and tail – with each part contributing to the cell's overall function. The head (in humans, 3–5 μm length and 2–3 μm width) holds the genetic material and is covered by the acrosome, which contains hydrolytic enzymes that aid egg penetration [20]. The midpiece (in humans, 7–8 μm), as the name entails, stands in between the head and the tail, and is packed with mitochondria to provide energy, in the form of ATP, that powers the movement of the tail [20]. The tail or flagellum (in humans, 45 μm) contains the axoneme, the motile apparatus responsible for movement, which drives the sperm forward [20].

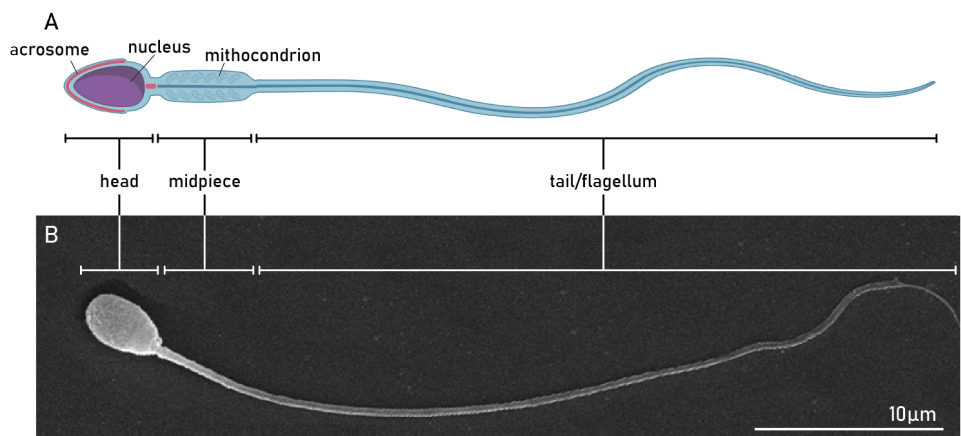


Figure 2.2: Human sperm cell structure. A) Schematic representation of a sperm cell, created with BioRender.com [21]. B) Scanning electron microscopy image of a real human sperm cell, adapted from: [22].

While the fundamental structure and function of spermatozoa are broadly conserved across sexually reproducing animals, notable variations exist between species [23]. Particularly among mammals, differences can be observed in head shape, midpiece size, tail length and overall sperm dimensions (Figure 2.3A). Outside of mammals, sperm morphology can differ even more drastically, reflecting a wide range of reproductive strategies across the animal kingdom (Figure 2.3B).

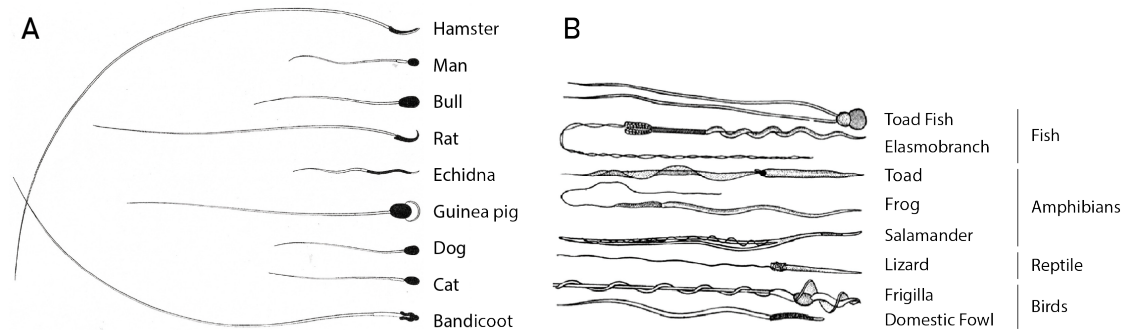


Figure 2.3: Morphological diversity of spermatozoa across species. A) Examples from various mammalian species, including hamster, human, bull, rat, echidna, guinea pig, dog, cat, and bandicoot, adapted from [24]. B) Examples from other vertebrate species, including fish (toadfish, elasmobranch), amphibians (toad, frog, salamander), reptiles (lizard), and birds (fringilla, domestic fowl), adapted from [25].

2.2 Human Sperm Motility

Just as sperm morphology varies between species, motility patterns and mechanisms can differ considerably [26, 27]. In this dissertation, the focus is placed on human sperm motility due to its direct relevance in clinical settings.

2.2.1 Definition and Biological Role

Sperm motility refers to the self-propelled movement of spermatozoa through fluid environments, enabled by the rhythmic flagellar beating. This movement originates in the axoneme, where the mitochondrial ATP powers specific proteins that slide adjacent microtubules, generating whip-like bends along the flagellum [28]. These bends create shear forces against viscous media, propelling sperm forward.

This function is biologically indispensable for natural human fertilization with two key purposes. First, it enables sperm to traverse the complex topography of the female reproductive tract, covering an average distance of approximately 19 cm, through viscous cervical mucus, epithelial crypts and uterine contractions [29, 30]. Second, it allows the penetration of the protective layer, zona pellucida, that surrounds the oocyte [29].

2.2.2 Motility Patterns and Heterogeneity

Spermatozoa within a single healthy ejaculate exhibit intrinsic heterogeneity in motility potential, manifesting as a spectrum of velocities and swimming trajectories. This biological variation includes sperm displaying rapid linear paths, slower wide circular movements, highly confined tight circling patterns, and varying forms of immobility, ranging from complete stillness to cases where the head remains stationary while the tail continues to beat without producing effective movement.

Among the diverse motility patterns observed in spermatozoa is hyperactivation, a specialized high-energy state that plays a critical role in the final stages of fertilization [31, 32]. This physiological transformation is triggered during capacitation, a biochemical maturation process initiated upon exposure to specific factors within the female reproductive tract [33]. Hyperactivation manifests as high-amplitude, asymmetrical flagellar bends, generating powerful, non-linear thrust patterns essential for penetrating the zona pellucida [31].

2.2.3 Determinants of Sperm Motility

Spermatozoa are highly sensitive cells, and their behavior depends on biochemical, environmental, and physiological conditions. Reflecting this sensitivity, sperm motility can be influenced by a complex interplay of endogenous and exogenous factors.

In terms of the immediate environment surrounding sperm cells, motility is strongly influenced by the composition of the surrounding medium, as it directly affects intracellular conditions. Research has indicated that concentrations of specific ions, such as calcium and bicarbonate, along with reactive oxygen species and pH levels, must be precisely regulated as imbalances in any of these factors can negatively affect sperm motility and overall viability [34]. Furthermore, temperature also plays a significant role: while lower temperatures reduce motility – particularly by decreasing their velocity – they help preserve cell vitality within certain thresholds and are therefore commonly used in sperm storage [35].

Beyond these immediate regulators, sperm motility is affected by broader physiological and lifestyle factors [36]. Unhealthy habits such as smoking, excessive alcohol consumption, and

chronic stress have all been associated with reduced motility [37–39]. Prolonged exposure to heat or radiation, including through frequent use of electronic devices near the groin, may also impair sperm function [40–42]. Certain medications, including anticancer and psychotropic drugs, have been linked to motility loss [43, 44]. Infections – whether bacterial, viral, or fungal – can disrupt sperm function through inflammation, oxidative stress, or mitochondrial damage [45, 46]. Additionally, conditions like varicocele [47], genetic disorders affecting flagellar structure [48], and the presence of antisperm antibodies [49] further contribute to impaired motility in some men.

Understanding the many factors that influence sperm motility is essential not only for interpreting biological variation but also for designing accurate and reliable methods for semen and motility analysis in clinical and research settings.

2.3 Semen and Sperm Motility Analysis

2.3.1 Standard Semen Analysis

Semen analysis constitutes the cornerstone of male fertility assessment, providing quantitative and qualitative evaluation of ejaculate parameters essential for diagnosing reproductive health and guiding treatment decisions. Its clinical utility depends on rigorous standardization, which is especially important given the high sensitivity of sperm parameters to analytical conditions [50]. The WHO Laboratory Manual for Human Semen Examination, first published in 1980 and updated through five editions (with the most recent in 2021), provides the globally recognized framework for standardization [6]. This manual establishes evidence-based protocols to ensure reproducibility, procedural uniformity across laboratories and comparability of results worldwide.

Semen analysis involves two complementary levels of evaluation: macroscopic and microscopic. Macroscopic assessment evaluates physical and chemical characteristics - including volume, pH, viscosity, appearance, and liquefaction time - measured shortly after sample collection, with typical reference values provided in Table 2.1. Microscopic analysis focuses on the cellular and subcellular features of the sample, enabling the assessment of sperm concentration, motility, morphology, and additional functional parameters.

Table 2.1: Reference values for macroscopic semen parameters. Adapted from [6], as reported in [17].

Semen parameter	Reference value
Volume (mL)	1.4–6.3
Viscosity	Forms small discrete drops (thread < 2 cm long)
Liquefaction	Within 60 min
pH	7.2–7.8
Appearance	Gray-opalescent

Microscopic analysis involves examination using brightfield or phase-contrast optics. To prevent cellular overlap while maintaining a representative distribution, samples are typically diluted prior to observation. Evaluation at $\times 200$ to $\times 400$ magnification enables precise sperm counting, motility classification, and identification of non-sperm cellular elements such as round cells, including leukocytes and immature germ cells. Using a Neubauer hemocytometer, a gridded chamber with standardized depth and dimensions, precise counting of sperm cells is enabled, thereby allowing calculation of both concentration and total sperm number. For motility assessment, the WHO guidelines advise temperature control at approximately 37°C and the use of an ocular lens reticle with a grid to precisely track sperm movement paths and distances. In turn, morphological assessment requires high-magnification analysis ($\geq \times 1000$), with stained smears examined under oil immersion to detect abnormalities in the sperm head, midpiece, and tail in accordance with the WHO criteria. Additional advanced assessments may include evaluations of sperm vitality, DNA integrity, and acrosome status using specialized laboratory techniques.

2.3.1.1 Standard Sperm Motility Analysis

Sperm motility is considered one of the primary indicators of a semen sample's fertility potential. Motility assessment involves evaluating individual sperm cells under the microscope and classifying their movement into defined categories. The proportion of sperm falling into each category is then calculated and used to assess whether the sample meets established reference thresholds predictive of fertility. Contemporary sperm motility analysis, classifies motility into four distinct categories based on velocity and progression patterns [6]:

- **Rapidly progressive (or grade a):** spermatozoa demonstrate active linear or large circular motion with velocities of at least $25\text{ }\mu\text{m/s}$, traversing distances equivalent to half the tail length within one second.
- **Slowly progressive (or grade b):** spermatozoa demonstrate active linear or large circular motion at velocities between $5\text{ }\mu\text{m/s}$ and $25\text{ }\mu\text{m/s}$, covering distances of at least one head length but less than half the tail length within one second.
- **Non-progressive (or grade c):** spermatozoa exhibit active flagellar motion without effective progression with velocities less than $5\text{ }\mu\text{m/s}$, sometimes characterized by patterns such as tight circling where head movement remains confined to less than one head length from the origin.
- **Immotile (or grade d):** spermatozoa show a complete absence of tail movement.

This current four-category classification system is presented in the 6th edition of the WHO Laboratory Manual for Human Semen Examination (2021) and represents a refinement of earlier frameworks [6]. While the 4th edition (1999) [51] similarly employed these four categories, the 5th edition (2010) [52] consolidated motility assessment into a three-tier system: progressive (combining rapidly and slowly progressive), non-progressive, and immotile, a change motivated

by the difficulty of reliably distinguishing between rapidly and slowly progressive sperm under manual observation [53]. This simplified classification drove technological development and dominated research literature during the subsequent decade, with most studies adopting the three-tier classification during this period. The 6th edition reinstated the four-category system based on research evidence demonstrating that the ability to distinguish rapidly progressive motility possesses specific clinical importance [53, 54].

Nonetheless, clinical reference thresholds continue to rely on aggregated motility categories - particularly total and progressive motility (further detailed in Section 2.3.2.1) - as these remain the most widely validated predictors of fertility potential in large population studies.

2.3.2 Clinical Significance of Sperm Motility Analysis

2.3.2.1 Male Infertility Diagnosis

The WHO estimates that approximately 17.5% of the adult population, or about 1 in 6 individuals globally, experience infertility [55], defined as a couple's inability to conceive after one year of regular unprotected sexual intercourse [1]. Males are solely responsible for approximately 20% of infertility cases, while they are also a contributing factor in an additional 30% to 40% of cases overall [56]. In total, male factors significantly contribute to around 50% of all infertility cases [2]. The prevalence of male infertility has been on the rise in recent years. In 2019, it was estimated that global male infertility increased by 76.9% since 1990 [57]. Furthermore, the age-standardized prevalence of male infertility is reportedly rising by 0.3% each year [3]. This trend is believed to be attributed to a variety of factors including environmental toxins, genetic disorders, lifestyle choices, urogenital abnormalities, and certain medications [58]. Reflecting this growing concern, the market size for male infertility was valued at USD 4.0 billion in 2022, with projections indicating a continued expansion at an expected compound annual growth rate (CAGR) of 4.9% from 2023 to 2030 [59].

It is important to note that the rise in male factor infertility rates has not been uniform across regions, varying significantly from 20% to 70% [3]. Notably, infertility is particularly prevalent in developing countries, where access to healthcare and fertility treatments may be limited [60].

Clinical Diagnosis Using Standardized Semen Parameters

Semen analysis stands as the gold standard for assessing male fertility [61, 62]. This assessment relies on standardized evidence-based limits established by the WHO 6th edition manual [6]. These limits, summarized in Table 2.2, are derived from a population of fertile men (whose partners conceived within 12 months) and represent the lower 5th percentiles of measured values for key microscopic parameters. Values below these thresholds are linked to a significantly reduced likelihood of natural conception and thus serve as thresholds for clinical diagnosis of male infertility.

Among the parameters evaluated, sperm motility is critically important. One of the most clinically relevant diagnoses based on these thresholds is asthenozoospermia, defined by the WHO as

Table 2.2: Lower reference limits (5th percentiles) for microscopic semen parameters relevant to fertility prediction, adapted from [6].

Parameter	Reference Limit (Lower 5th Percentile)
Sperm concentration ($10^6/\text{mL}$)	16
Total sperm number (10^6)	39
Progressive motility (Rapidly Progressive + Slowly Progressive, %)	30
Total motility (Progressive + Non-progressive, %)	42
Sperm morphology (normal forms, %)	4
Vitality (%)	54

the presence of abnormally low sperm motility [6]. Specifically, a sample is classified as asthenozoospermic when progressive motility (grades a + b) falls below 30% or total motility (grades a + b + c) falls below 42%. This condition is among the most common forms of male infertility and underscores the clinical importance of motility assessment.

2.3.2.2 Role in ART and Fertility Treatments

Multiple fertility treatments can address both male and female infertility issues. According to the Centres for Disease Control and Prevention, Assisted Reproductive Technology (ART) is defined as any fertility-related treatment involving the manipulation of eggs or embryos, aimed at helping individuals who have difficulty conceiving naturally, achieve pregnancy [63]. This includes in vitro fertilization (IVF), where eggs are retrieved and fertilized with sperm in a lab through natural interaction, and intracytoplasmic sperm injection (ICSI), where a single sperm is directly injected into an egg to assist fertilization, particularly in cases of male infertility [64]. In both IVF and ICSI, fertilization occurs outside the uterus, and the resulting embryo is then transferred into the uterus to establish a pregnancy. While intrauterine insemination (IUI) is not classified as ART, it remains an important less invasive procedure that involves placing prepared sperm directly into the uterus to increase the chances of conception [65].

As of 2018, over 10 million children worldwide were born through ART, representing up to 7.9% of births in Europe and 5.1% in the United States [4, 66]. At this time, the ART market was already expanding, with a steady annual growth rate of 5.6% which has driven its value to \$37.6 billion in 2023. Projections indicate that the market could reach \$55.91 billion by 2028, growing at a rate of 8.3% per year [5].

The previously mentioned rise in infertility rates is a significant factor driving demand for ART and other fertility treatments such as IUI. Additionally, the trend of delayed parenthood, often supported by options like egg freezing, has led more individuals to turn to these medical interventions to support family planning at later stages [67]. Meanwhile, advancements in ART techniques in the past few years have improved both success rates and safety, reducing complications and making these options more accessible and appealing [68]. Together, these factors contribute to the growing reliance on fertility treatments to achieve pregnancy.

Motility Parameters Guide ART Treatment Selection

To determine the most effective fertility treatment, specialists rely on several semen parameters obtained through semen analysis of the male partner, with motility being a key factor in selecting the best therapeutic approach [36]. If at least 5 million motile sperm can be obtained, IUI is recommended [69, 70]. When the motile sperm count falls between 2 and 5 million, IVF is usually preferred, while ICSI, a more targeted intervention, is advised if fewer than 2 million motile spermatozoa are present in the sample [71].

This highlights the crucial importance of analyzing sperm motility, as the increasing demand for ART and fertility treatments like IUI drives the need for precise diagnostic parameters to effectively triage patients and optimize therapeutic pathways.

2.4 Summary

Semen is a heterogeneous fluid composed of sperm cells suspended in seminal plasma, with its composition and quality influenced by a range of physiological and environmental factors. Spermatozoa are highly specialized cells whose structure – particularly the flagellum – enables motility, a key determinant of fertilizing capability. Human sperm motility is characterized by intrinsic heterogeneity, with distinct patterns ranging from rapidly progressive to immotile. These patterns are shaped by both internal cellular mechanisms and external environmental conditions. Standardized semen analysis, as outlined by the WHO, remains the clinical cornerstone of male fertility evaluation. Motility assessment, performed under controlled laboratory conditions, plays a central role in diagnosing male infertility and informing the selection of appropriate ART. Given the global rise in infertility and the increasing reliance on ART, precise sperm motility evaluation has become more clinically relevant than ever.

Chapter 3

Technical Advances in Sperm Motility Analysis

This chapter explores the technical advances in sperm motility assessment, with a focus on tools developed to standardize and objectify analysis while addressing the limitations of manual methods. It reviews the development, evolution, and commercialization of automated systems, and examines the recent application of artificial intelligence, highlighted in contemporary scientific literature, as a means of overcoming the inherent challenges of this complex task.

3.1 Automated Semen Analysis

3.1.1 Historical Development of CASA

Antonie van Leeuwenhoek's 17th-century single-lens microscopes, with magnifications up to $\times 270$, allowed the first detailed observations of human sperm [72, 73]. Observations quickly highlighted its role in reproduction, emphasizing the significance of understanding sperm characteristics, particularly how they affect fertility [74]. The gradual evolution of microscopy techniques that followed – from basic compound lenses to phase-contrast optics before World War II – set the stage for more sophisticated sperm studies.

Between 1940 and 1970, researchers focused on sperm function started developing methods to quantify sperm motion [75]. The limitations of manual methods became evident: they remain slow, labour-intensive, offer only basic metrics, and are prone to subjective errors, requiring significant technician training for consistency. A few scientists recognized the need for objective, bias-free measurements of sperm motility, aiming to establish reproducible standards, especially for animal genetics and sperm function research [7].

In the 1970s, the emergence of programmable digital computers combined with governmental interest in tracking moving objects, such as missiles, sparked biologists' interest in using these technologies to quantify microscopic movements, including sperm [7]. This culminated in the development of a primitive CASA system, initiated in 1974 and completed in 1975 [76].

Building on this proof-of-concept, the first fully automated CASA system emerged in 1985. This system adapted the pre-existing Expert Vision motion analysis platform, originally designed for non-biological applications, to the task of sperm motility assessment [77]. This period also saw the first publications on automated sperm morphology evaluation using CASA technology [78], expanding its diagnostic scope.

Commercialization rapidly followed, targeting both human andrology labs and veterinary applications, particularly livestock sperm storage optimization. The CellSoft system, launched in 1985, became the first dedicated commercial CASA system for motility assessment, leading to several studies on human semen and other species [75]. Both CellSoft and the adapted Expert Vision platform (later CellTrak) used a phase-contrast microscope with a video camera and computer for real-time video capture and analysis [7].

In 1986, the HTM-2000 was introduced by Hamilton Thorne as a compact system for quantifying stallion sperm during dilution and storage, showcasing multiple technical advancements. Its success paved the way for Hamilton Thorne's subsequent IVOS and CEROS systems, developed in 1992, which established themselves as dominant platforms through continuous updates [72]. Throughout the 1990s, additional CASA systems, including the SM-CMA and Hobson Sperm Tracker, were introduced to meet specific needs in both human and animal applications. Although no longer widely marketed, some of these systems remain in use [7].

3.1.2 Current CASA Market Landscape

Since its inception, the CASA market has expanded significantly fueled by a rising demand for automated semen analysis among laboratory technicians and active promotion by commercial CASA system manufacturers. In 2022, it was valued at USD 48.40 million and is projected to grow to USD 598.80 million by 2030, reflecting a CAGR of 7.80% [79]. This growth is driven by increasing adoption in clinical laboratories worldwide. In China, for example, over 50% of andrology laboratories have been equipped with CASA systems [80].

Nowadays, modern CASA systems offer a wide array of products, encompassing the hardware and software, designed to evaluate not only sperm motility but also other key parameters of standard semen analysis, including concentration, morphology, vitality, DNA integrity, leukocyte presence, and the acrosome reaction. These systems cater to distinct application domains: human clinical diagnostics, where accurate classification and precise measurement are essential for guiding infertility treatments; human research, which focuses on understanding reproductive biology and health; and veterinary applications, focused on optimizing breeding programs for livestock and companion animals. Reflecting diverse user needs, hardware configurations vary significantly between conventional microscope-based setups and specialized standalone devices that may rely on non-microscopy technologies. Illustrating this landscape, examples of the most commonly used systems are shown in Figure 3.1, while Table 3.1 lists a broader selection, with details on manufactures, target applications, and hardware configurations.



Figure 3.1: Illustration of the most commonly implemented CASA systems in andrology laboratories. MES, Medical Electronic Systems. Source: [81].

Table 3.1: Overview of CASA systems currently on the market: applications and sperm analysis device type.

Company	CASA System	Application			Sperm Analysis Device			Specialized Device
		Veterinary	Human		Conventional Microscope		Fluorescence	
			Research	Clinical*	Brightfield	Phase Contrast		
Hamilton Thorne	IVOS II	✓	✓	✓	-	-	-	✓
	CEROS II	✓	✓	✓	✓	✓	-	-
Microptic	SCA	✓	✓	✓	✓	✓	✓	-
	SCA Scope	-	✓	✓	-	-	-	✓
LensHooke	X3 PRO	-	✓	✓	-	-	-	✓
	X12 PRO	-	✓	✓	-	-	-	✓
Medical	SQA-Vision	-	✓	✓	-	-	-	✓
Electronic	SQA-V Gold	-	✓	✓	-	-	-	✓
Systems	SQA-iO	-	✓	✓	-	-	-	✓
ARQUIMERA	ISAS	✓	-	-	-	✓	✓	-
	ISAS PRO	✓	-	-	-	-	-	✓
	SQS 2	✓	-	-	-	-	-	✓
Minitube	Sperm Vision	✓	-	-	-	✓	-	-
	AndroScope	✓	-	-	-	-	-	✓
Magapor	Magavision	✓	-	-	-	✓	-	-
	Magavision One	✓	-	-	-	✓	-	-

Note: Although devices are described by manufacturers as suitable for clinical use, their regulatory approval varies by country, and the WHO does not currently recommend using them as a substitute for manual semen analysis.

Industry leading systems include those from Hamilton Thorne, IVOS II, and CEROS II, which employ various microscopy techniques designed for applications in human clinical diagnostics, research, and veterinary use. Microptic similarly provides advanced CASA systems, such as the microscope-based SCA and the specialized standalone SCA Scope, catering to both human and veterinary markets. LensHooke provides more recent systems like the X3 PRO and X12 PRO, they produce the analysis devices tailored specifically for human sperm analysis. Medical Electronic Systems provides additional specialized devices, which do not rely on traditional microscopy, for human use, including the SQA-Vision, SQA-V Gold and the more recent SQA-iO. In the veterinary field, several systems are available exclusively for animal applications, including those from ARQIMEA, Minitube, and Magapor, among many others.

3.1.2.1 Market Needs and Limitations

While the CASA market has expanded, most available systems remain primarily designed for veterinary applications. Manufacturers often focus on this sector due to less stringent regulatory requirements and easier market access [82]. Moreover, semen from domesticated animals typically present fewer challenges for automated analysis than human samples: they tend to contain less debris and fewer non-sperm cells, have lower rates of dead or aggregated sperm, and display more uniform morphology within a species [8]. Due to the significant sperm morphological differences across species, these systems are unsuitable for human use, exposing a gap in effective CASA solutions for human semen analysis.

Furthermore, CASA systems present a high cost barrier, as most require specialized equipment, whether microscope-based technologies or not, along with tailored consumables such as custom slides and reagents. These tools are necessary to ensure functionality but add significant expense, limiting the practical deployment of CASA technology in many laboratories.

Moreover, it has been reported that this type of equipment is not fully automated, requires constant calibration by trained personnel, is difficult to learn, and is not sufficiently robust for widespread use [9]. Currently, the WHO does not recommend CASA systems as a replacement for manual semen analysis in routine clinical practice, due to variability between systems, lack of universal internal quality control, and limited reproducibility across laboratories [6, 83, 84].

As a result, manual and visual sperm analysis continues to be the predominant approach in clinical settings, emphasizing the pressing need for CASA solutions that are not only specifically adapted for human applications, standardized and reliable but also but also cost-effective to increase accessibility. Such advancements would allow for broader adoption in clinics, improving access to objective, standardized and more efficient sperm analysis.

3.2 Computer Vision and AI in Sperm Motility Analysis

3.2.1 Traditional Approach

Traditional CASA systems track sperm motility using microscopic imaging - most commonly with phase contrast microscopy, as it enhances the visibility of sperm heads compared to other imaging techniques [85]. These systems capture visual data of sperm cells and convert it into a digital format. A grey-level threshold is typically applied to each video frame, allowing the software to detect edges or diffraction patterns that identify individual sperm based on head shape or the middle piece, distinguishing them from the background [72, 86]. Each sperm head is assigned a "centroid" to track its movement across consecutive frames, using a circular measurement validation region (or "gate") centred on the centroid to identify the corresponding position in the next frame, thereby creating a continuous trajectory map [86]. Nearly all commercial CASA instruments employ a nearest-neighbor (NN) tracking scheme, in which the nearest measurement to a given track is assumed to correspond to the same sperm cell [86].

Key motility kinematic parameters are then derived from these tracks, as illustrated on Figure 3.2. According to the WHO [6], these parameters include:

- **VCL** (velocity of curved line, $\mu\text{m/s}$): Time-averaged velocity of the sperm head along its actual curvilinear path; also known as curvilinear velocity.
- **VSL** (velocity of straight-line, $\mu\text{m/s}$): Velocity calculated along a straight line between the first and last points of the path.
- **VAP** (velocity of average path, $\mu\text{m/s}$): Time-averaged velocity along a smoothed average path.
- **ALH** (amplitude of lateral head displacement, μm): Magnitude of lateral head displacement about the average path.
- **LIN** (linearity): Linearity of the curvilinear path (VSL/VCL ratio).
- **WOB** (wobble): Oscillation of the curvilinear path about the average path (VAP/VCL ratio).
- **STR** (straightness): Linearity of the average path (VSL/VAP ratio).
- **BCF** (beat-cross frequency, Hz): Frequency at which the curvilinear path crosses the average path.

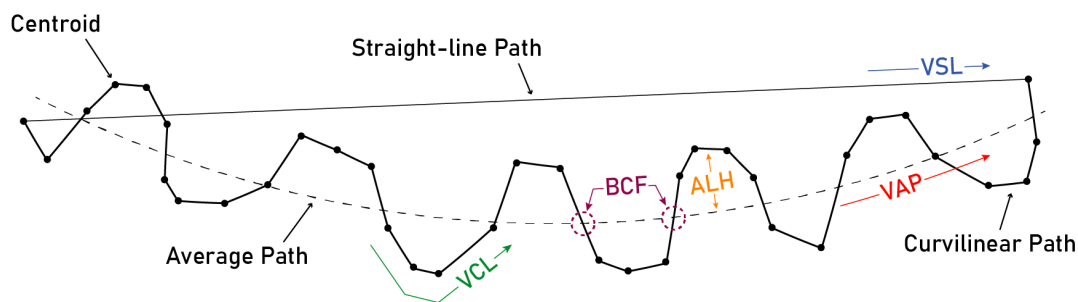


Figure 3.2: Illustration of CASA trajectory paths and key kinematic parameters. VCL, curvilinear velocity. VSL, straight-line velocity. VAP, average path velocity. ALH, Amplitude of lateral head displacement. BCF, Beat-cross frequency.

The average path is obtained by applying a smoothing technique to the curvilinear trajectory, which varies depending on the specific algorithm implemented in each CASA system [87]. There is no standardized approach for this calculation, and methodologies may differ between manufacturers. Following computation of the kinematic parameters, the resulting values are typically used to classify individual spermatozoa into motility categories according to the WHO framework. However, universally accepted threshold values for CASA-derived parameters are also lacking, with each system adopting its own classification rules based on different criteria [8, 87, 88]. Once motility classes are assigned to sperm cells, the relative proportion of each motility group in the sample is computed and interpreted based on the context of use. In addition to classification, CASA systems compute average kinematic values for the entire sample and for specific motility subgroups. This enables the extraction of quantitative parameters that are not easily observable by the human eye, offering a more nuanced and objective assessment of sperm movement patterns and behavior.

3.2.1.1 Challenges and Limitations

The biological characteristics of human semen samples present notable challenges for this type of traditional automated analysis [8]. These include high levels of debris and round cells, low sperm motility and the occurrence of dead spermatozoa often clumped together [8]. Distinguishing non-progressive and immotile sperm from debris or dead sperm remains difficult [89], and the clumping of live sperm with debris or non-viable cells can hinder the tracking process, leading to interruptions and missed detections [8, 90]. These issues are particularly pronounced in samples from infertile individuals – which are common in clinical settings – further complicating accurate analysis in this context.

In addition, CASA performance is highly dependent on microscopy and visualization conditions [91]. Techniques such as brightfield microscopy, where sperm heads can be more difficult to distinguish from the background than in phase contrast, may pose additional challenges for accurate detection. This sensitivity extends to variations in video quality, such as brightness, contrast,

and focus, which can significantly affect detection outcomes. Fixed grey-level thresholds, commonly used in these systems, may not adapt well to such changes, leading to inaccurate detection and, consequently, unreliable tracking. Furthermore, sperm concentration also impacts tracking reliability [83]. When visualizing more concentrated samples, the increased likelihood of collisions and occlusions (as visible in Figure 3.3A) can distort trajectories or cause tracking failures, particularly when using simple distance-based algorithms, such as NN.

Moreover, the accuracy of kinematic parameters extracted by CASA systems is highly dependent on technical factors such as frame rate, duration of capture and algorithm choice [83, 92]. A lower frame rate results in simplified trajectories that fail to capture finer movement nuances, as illustrated in Figure 3.3B, significantly affecting parameters such as VCL, LIN, ALH and BCF [92, 93]. Likewise, the duration of the captured sequence directly impacts the calculated motion descriptors – for instance, a sperm cell moving in tight circles may appear to have a very low VSL over a longer time window, while appearing more progressive in a shorter one, as seen in Figure 3.3C [92]. Furthermore, the algorithm employed to compute the average path is crucial, as it directly affects VAP, ALH, STR, and BCF, all of which are derived from this smoothed trajectory. Collectively, these factors demonstrate that kinematic outputs from different CASA systems cannot be considered equivalent unless standardized methods and validation protocols are applied.

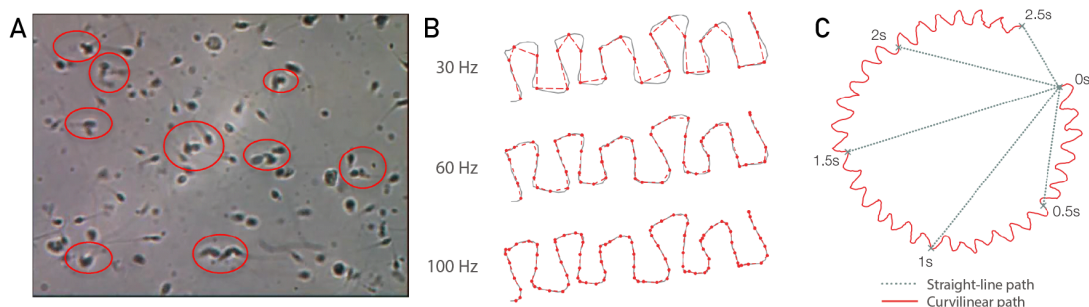


Figure 3.3: Examples of challenges and limitations in traditional CASA systems. A) Brightfield microscopy image showing sperm with overlapping paths; collisions are marked with red circles. Source: [94]. B) Effect of frame rate on trajectory accuracy: dots indicate sperm positions per frame; the grey line shows the full path at high temporal resolution. Adapted from: [92]. C) True curvilinear path (red) vs. straight-line path (grey) over increasing time intervals, highlighting misrepresentation of non-linear motion. Adapted from: [92].

Another critical limitation of traditional CASA systems lies in distinguishing non-progressive motile sperm from truly immotile ones. Human observation defines motility based on both the spatial displacement of the sperm cell and the pattern of flagellar movement. However, CASA systems track only the movement of the sperm head and do not directly analyse flagellar beating. As a result, spermatozoa that exhibit tail movement without head displacement – classified manually as non-progressive – may be incorrectly labeled as immotile by CASA [8]. This leads to an overestimation of immotile sperm in automated assessments compared to manual analysis.

3.2.2 Artificial Intelligence in Sperm Motility Analysis

3.2.2.1 AI and DL Applications

The limitations of the traditional approach for sperm motility assessment, coupled with the recent surge in artificial intelligence (AI), have generated increased interest in developing more advanced solutions.

AI is a broad discipline that refers to the development of computational systems capable of performing tasks that typically require human intelligence, such as reasoning, perception, learning, and decision-making [95]. Its conceptual scope is represented in Figure 3.4. Machine learning (ML) is one of the subfields of AI that centers on enabling systems to learn from data and create predictions without being explicitly programmed. Depending on the nature of the task, learning can be supervised, where the model is trained on input-output pairs to infer a mapping from inputs to known targets, or unsupervised, where the goal is to identify underlying patterns or structures within data that lacks predefined labels [95].

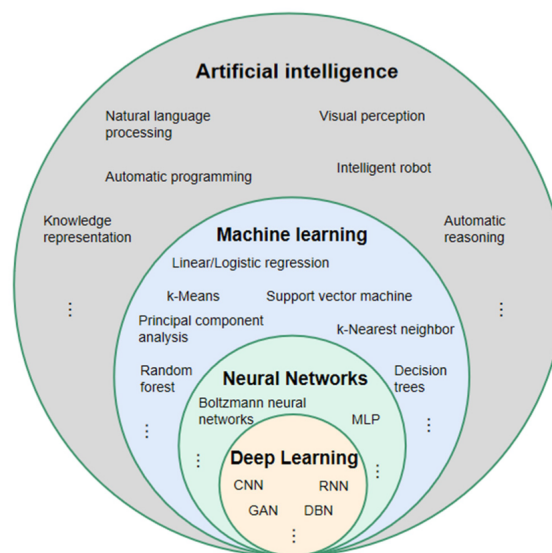


Figure 3.4: Conceptual hierarchy of artificial intelligence, machine learning, neural networks and deep learning. Source: [96].

Neural networks are a type of ML inspired by the structure and function of the human brain, designed to process information through interconnected units known as artificial neurons [97]. Each neuron computes a weighted sum of its inputs, adds a bias term, and applies a nonlinear activation function to produce an output, as illustrated in Figure 3.5A [98]. Typically, these models are organized in layers, where a basic architecture comprises an input layer, one or more hidden layers that perform intermediate computations, and an output layer that generates the final prediction. This structure is represented in Figure 3.5B. Training involves iteratively adjusting connection weights using algorithms like backpropagation with gradient descent, minimizing the difference between predictions and targets. This layered architecture enables hierarchical feature learning, where early layers capture simple features and deeper layers build complex abstractions.

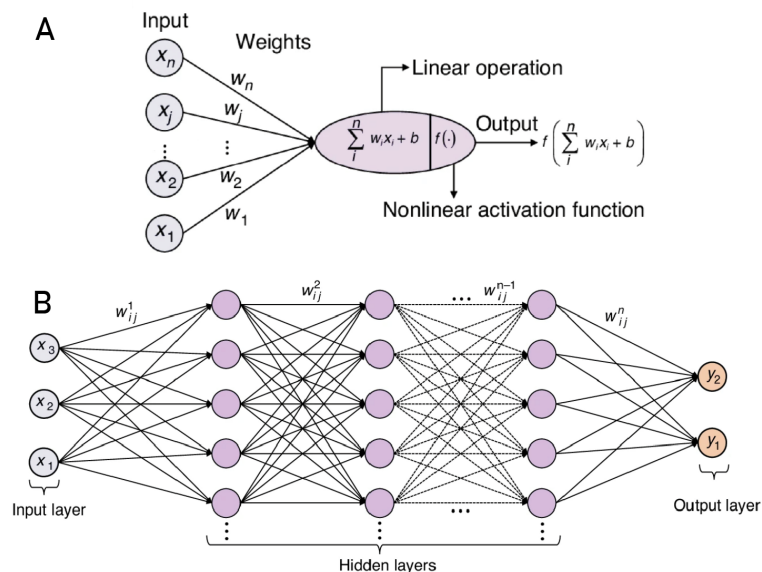


Figure 3.5: Architecture of artificial neural networks. A) Schematic of an artificial neuron showing inputs (x), weights (w), bias (b), summation, activation function, and output. B) Neural network architecture composed of input, hidden, and output layers. Adapted from [99].

Deep learning (DL) refers to neural networks with a high number of hidden layers. The increased depth allows these models to automatically learn multiple levels of feature abstraction directly from raw data [98]. This contrasts with traditional ML approaches, which often require handcrafted feature extraction as a separate step. The depth and representational capacity of DL architectures make them particularly effective for solving complex tasks such as those in natural language processing and computer vision [100]. A prominent example is the Convolutional Neural Network (CNN), a DL architecture distinguished by its ability to automatically learn spatial features from images through localized connections and weight sharing, enabling efficient detection of patterns and structures in visual data.

The use of AI is becoming increasingly prevalent in medical fields due to its capability to analyse complex data, enhance accuracy, and enable automation [101, 102]. These attributes are particularly important for sperm motility analysis, given the complexity of the data and the existing limitations of CASA systems being related to accuracy and lack of complete automation [103]. As a result, sperm motility has emerged as a central focus within this domain of AI research [10]. ML can be employed to classify sperm motility types and predict motility scores based on extracted features. DL, in particular, advances this process by learning spatial and temporal patterns directly from raw microscopy frames or videos supporting more robust detection of sperm cells and even enabling direct prediction of motility characteristics [100].

3.2.2.2 State-of-the-art

The identification of relevant literature was conducted through systematic queries of PubMed, IEEE Xplore, Scopus, and Google Scholar databases. Primary search terms included: *automated*

sperm analysis, deep learning sperm motility, computer-aided sperm analysis (CASA), sperm detection, sperm tracking, sperm motility classification and related keywords, published since 2020. Critically, citation chaining from papers and scrutiny of reference lists supplemented initial results.

Through this research, two distinct methodological paradigms for sperm motility analysis have been identified, as illustrated in Figure 3.6. The first, **tracking-based approaches** typically consists of three principal stages: sperm detection (generally DL-based), tracking algorithms to reconstruct individual sperm trajectories, and classification based on extracted motion parameters. These per-cell classifications are subsequently aggregated to estimate sample-level distributions. The second paradigm is **end-to-end approaches**, where DL models bypass explicit tracking and individual classification, learning instead to predict motility distributions directly from spatiotemporal video data. While end-to-end models offer a streamlined solution, tracking-based pipelines provide greater interpretability, as their classifications are grounded in measurable motion features and explicit cell-level analysis.

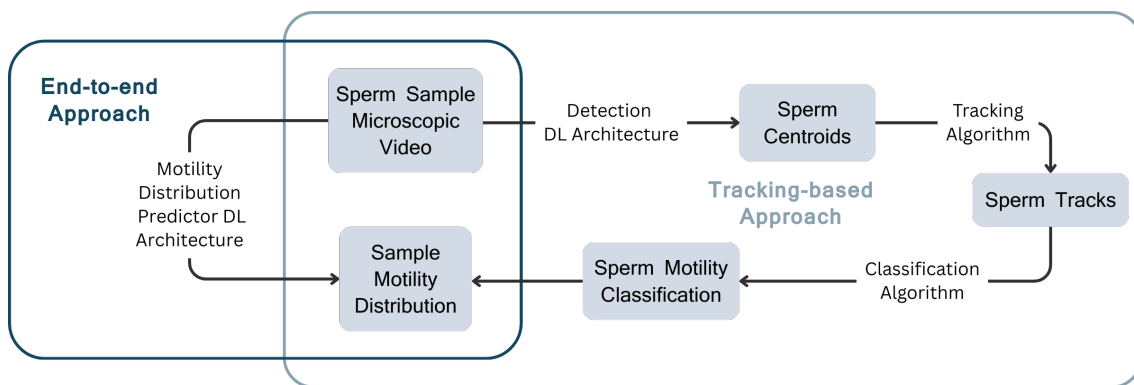


Figure 3.6: Overview diagram of the two main computational paradigms for state-of-the-art sperm motility assessment: Tracking-based and End-to-end approaches.

These diverse computational strategies contribute to considerable variation in methodological approaches within sperm motility analysis, reflecting differing levels of task specialization, integration, and analytical scope. Accordingly, this review adopts a task-oriented structure by dividing the analysis pipeline into three core components: **detection**, **tracking**, and **motility assessment**. While numerous studies encompass multiple stages of this pipeline, the present review isolates and discusses each contribution within its respective functional category.

3.2.2.2.1 Detection

Nissen et al. [104] pioneered DL for human sperm detection, comparing custom CNN architectures of different depths. Their approach used full-image training with class re-weighting and upsampling, achieving 93.9% precision and 91.9% recall with a deep CNN – significantly outperforming classical computer vision methods and establishing DL as the new paradigm.

Subsequent research explored alternative CNN-based frameworks. Mohammadi et al. [105] introduced a modified RetinaNet architecture incorporating a multi-frame input strategy (concatenating sequences of 3 frames), enabling the network to extract temporal as well as spatial features. This approach proved especially effective in dense samples and achieved an impressive 99.1% average precision. Around the same time, Valiūškaitė et al. [106] employed Faster R-CNN with an InceptionV2 backbone, trained on the VISEM dataset, reporting a detection accuracy of 91.77%, which reinforced the suitability of region proposal-based methods for sperm head localization.

In the following years, the YOLO family of object detectors began to dominate the domain due to their speed and robustness across various applications [107]. Hidayatullah et al. [108] validated this shift with their YOLO-based DeepSperm framework including innovations like a dedicated small-object detection layer, increased input resolution (640×640), and targeted augmentation, yielding 94.11% mAP50 and real-time processing at 51.9 FPS – making real-time analysis feasible for the first time. Comparative studies further solidified YOLO’s position. Yuzkat et al. rigorously evaluated single-stage (YOLOv5, SSD, EfficientDet, Detectron2) and two-stage (Mask R-CNN) detectors on their novel SeSViD dataset. YOLOv5 consistently outperformed others in patient-oriented testing with a mAP50 of 0.880, demonstrating superior efficiency and accuracy for within-patient detection, though fusion strategies helped marginally on challenging unseen cases in independent testing.

Progress in the field was further driven by organized challenges, notably the MediaEval Medico task *Transparent Sperm Tracking*. This challenge encouraged contributions using the VISEM dataset and involved complex subtasks, including multi-category detection (normal, small/pinhead, clusters), real-time tracking, and patient-level prediction. In 2022, Kosela et al. [109] employed YOLOv5 with a carefully designed temporal train-validation split (first 24s/last 6s per video) and confidence threshold optimization, achieving 91% mAP50. Concurrently, Huynh et al. [110] developed a tail-aware model based on YOLOv7. By doubling input resolution and leveraging edge detection to filter non-tail candidates, they significantly reduced false positives in crowded scenes. In 2023, the MediaEval challenge recurrence retained the same core theme and subtasks introducing a stronger emphasis on computational efficiency, intensifying architectural refinements under identical evaluation protocols. Nguyen et al. [111] benchmarked YOLOv8s with deliberate recall-oriented tuning, lowering confidence thresholds to 0.25 to prioritize sensitivity (62.8% recall) at the expense of precision (50%). Concurrently, Nguyen-Mau et al. [112] tackled severe class imbalance (e.g., 93.3% sperm vs. 3% minority classes) through a targeted augmentation strategy that synthetically increased underrepresented categories like small/pinhead sperm and clusters by cropping and pasting under-represented objects into non-overlapping image regions. While their approach improved detection rates for rare morphologies, it did not yield high overall performance metrics. Osaid et al. [113] reported 96.5% mAP using the highly efficient YOLOv8n, demonstrating strong performance from a lightweight model.

Significant effort focused on optimizing YOLO architectures for efficiency and specific sperm detection challenges. Zhang et al. [114] created Sperm-YOLOv5t, a lightweight variant by removing the large-object head and reducing kernels by 50%, achieving 95.7% mAP50 with a tiny

2.8MB footprint. Zhu et al. [115] improved YOLOv5s by integrating depthwise separable convolutions and a Shuffle Attention mechanism, which enhances feature selection by allowing the network to better capture both spatial and channel-wise information. These modifications increased average precision by 6.9% (reaching 87.5%) while maintaining 55.8 FPS, with notable improvements in detecting occluded sperm.

As YOLO-based approaches continued to evolve, Nawae et al. [116] conducted a direct comparison of YOLOv5, v6, v7, and v8s for the detection of both sperm and debris under small-dataset conditions. While YOLOv5 performed well on the original data, targeted augmentation, such as shear and blur, enabled YOLOv8s to achieve exceptional performance (precision: 99.4%, recall: 96.5%, mAP: 99.3%) on the augmented dataset, demonstrating the importance of augmentation for robustness with limited data.

Further architectural refinements pushed performance boundaries by explicitly addressing sperm-specific challenges like small size, transparency, and dense clusters. Jiao et al. [117] presented a meticulously optimized YOLOv5, incorporating key strategies: mosaic augmentation to simulate crowded scenes, a focus mechanism to preserve fine details during downsampling, and advanced feature fusion to effectively integrate multi-scale information. This configuration achieved an impressive 97.9% detection accuracy at 0.027s/frame, significantly outperforming traditional methods in complex environments. More recently, Zhang et al. [118] introduced major enhancements to YOLOv8 specifically tailored for detecting small, densely packed sperm. Their model included a dedicated small-object detection layer and integrated attention mechanisms to prioritize sperm-specific visual patterns and contextual cues within the feature maps. They also replaced standard downsampling convolutions with specific modules designed to minimize information loss, adopted more efficient downscaling and feature refinement modules, and optimized the loss function to enhance bounding box accuracy for microscopic targets. This comprehensive redesign resulted in a 1.3% increase in precision, a 1.4% gain in recall, and a 2.0% improvement in mAP@0.5:0.95 over baseline YOLOv8 on the VISEM-Tracking dataset.

Overall, these advancements highlight a clear trend toward YOLO-based models as the leading approach for sperm detection, thanks to their balance of speed, accuracy and adaptability through ongoing optimization. Refinements have effectively addressed major challenges such as small cell size and dense samples by combining architectural improvements, data augmentation strategies, and efficiency enhancements. This progress has resulted in highly reliable detection performance even in challenging conditions, providing a reliable basis for further analysis stages such as tracking and classification.

Table 3.2: Summary of state-of-the-art sperm detection.

Ref.	Dataset	Task	Architecture(s)	Best Performance
[104]	Non-public (765 images, 38,708 annotations)	Sperm Head Detection	Custom CNN	Precision: 93.87% Recall: 91.89%
[105]	Non-public (900 images, 1,628 sperm instances)	Sperm Head Detection	Modified RetinaNet	mAP50: 99.1% Accuracy: 96.3% Recall: 98.7% F1 score: 98.1%
[106]	WISEM (650 images, ~19,500 annotations)	Sperm Head Detection	Faster R-CNN with Inception_V2 backbone	Accuracy: 91.11–92.43%
[108]	Non-public (120 frames, 44,048 annotations)	Sperm Head Detection	DeepSperm (YOLO-based)	mAP50: 94.11% F1 score: 93%
[109]	WISEM-Tracking (29,196 images, 179,802 annotations)	Sperm Head (Normal or Small) and Cluster Detection	YOLOv5	mAP50: ~91% mAP50-95: ~66%
[110]	WISEM-Tracking (29,196 images, 179,802 annotations)	Sperm Head (Normal or Small) and Cluster Detection	YOLOv7 + Tail-aware sperm detection algorithm	mAP50: 80.8% Precision: 84.3% Recall: 95.9% F1 Score: 98.7%
[114]	WISEM (200 images, 1,072 sperm instances)	Sperm Cell Detection	Modified YOLOv5s	mAP50: 95.7% mAP50-95: 56.8%
[119]	SeSViD (1,275 frames, 68,544 sperm + 23,785 debris annotations)	Sperm Head and Debris Detection	YOLOv5 , Mask R-CNN, Detectron2, SSD, EfficientDet	mAP50: 88.0% mAP50 (Sperm): 83.9% mAP50 (Debris): 92.3%

Continued on next page

Ref.	Dataset	Task	Architecture(s)	Best Performance
[10]	WISEM (382 frames, 3,500 annotations)	Sperm Head Detection	YOLOv5 (n, s, m, l, x)	mAP50: 72.1% Precision: 88.6% Recall: 52.6%
[115]	WISEM (600 frames, 167,115 annotations)	Sperm Head Detection	Modified YOLOv5s	mAP50: 87.5% Precision: 83.1% Recall: 88.1% F1 score: 85.6%
[120]	Non-public (540 images, 5,624 sperm instances)	Sperm Cell Detection	YOLOv8s	Precision: 81.71–83.45 % Accuracy: 91.14–92.76% Recall: 75.08%
[116]	Non-public (256 images, 9,088 annotations)	Sperm Head and Debris Detection	YOLOv5, YOLOv6, YOLOv7, YOLOv8s	mAP50: 99.3% Accuracy: 99.4% Recall: 96.5% F1 score: 97.9%
[117]	Non-public (4,000 images, 3,800 annotations)	Sperm Head Detection	YOLOv5	mAP50: 98.5% Accuracy: 97.9% Recall: 98.0%
[111]	WISEM-Tracking (29,196 images, 179,802 annotations)	Sperm Head (Normal or Small) and Cluster Detection	YOLOv8s	mAP50: 50.6% mAP50-95: 19.1% Precision: 50% Recall: 62.8%
[112]	WISEM-Tracking (29,196 images, 179,802 annotations)	Sperm Head (Normal or Small) and Cluster Detection	YOLOv8 (n, s, m, l, x)	mAP50: 28% mAP50-95: 10.8% Precision: 31.7% Recall: 27.2%
[113]	WISEM-Tracking (29,196 images, 179,802 annotations)	Sperm Head (Normal or Small) and Cluster Detection	YOLOv8n	Accuracy: 96.5%
[118]	WISEM-Tracking (29,196 images, 179,802 annotations)	Sperm Head (Normal or Small) and Cluster Detection	Modified YOLOv8	mAP50: 95.1% mAP50-95: 82.7% Precision: 98.2% Recall: 96.2%

3.2.2.2.2 Tracking

While not necessarily AI-based, tracking plays a central role within the sperm motility analysis framework by bridging detection and classification through assignment of consistent identities (IDs) to individual sperm, enabling monitoring individual sperm trajectories across frames. Although many studies have primarily focused on cell detection, the tracking of spermatozoa remains comparatively underexplored [4]. Sperm tracking introduces unique challenges beyond initial localization, such as managing collisions between cells, occlusions, and the inherently variable nature of sperm trajectories. An overview of state-of-the-art approaches is provided below, and selected works are summarized in Table 3.3.

Early research adapted multi-target tracking algorithms from radar/sonar domains to sperm motility analysis. Notably, the Joint Probabilistic Data Association (JPDA) method improves over NN traditional tracking by using probabilistic associations instead of hard one-to-one assignments to reduce ID switches and handle ambiguities in crowded scenes. Urbano et al. [86] pioneered the use of this method in sperm motility analysis with a customized JPDA algorithm integrated with Kalman filters to predict and update sperm positions. Their approach employed a Continuous White Noise Acceleration motion model and adaptive noise covariance to handle abrupt motility changes, achieving robust tracking in dense samples while minimizing ID swaps during near-collisions ($\geq 10 \mu\text{m}$ separation). Quantitative evaluation via the Optimal Subpattern Assignment (OSPA) metric, a standard measure for multi-target tracking accuracy, on simulated sperm videos demonstrated superiority over NN methods, though persistent occlusions (>5 frames) remained challenging. Wang et al. [121] subsequently addressed track coalescence – a failure mode where converging trajectories merge irreversibly – through a JPDA variant that computed only the optimal joint association event per frame. Despite increased computational demands, this variant significantly outperformed the others in dense environments, maintaining accurate IDs in both synthetic and real sperm datasets. Years later, Jiao et al. [117] achieved notably high performance, with a Multiple Object Tracking Accuracy (MOTA) of 91%, using JPDA with Kalman filters post-YOLOv5 detection. Their probabilistic association handling proved effective for occlusion management in dense environments.

In parallel, Mohammadi et al. [105] introduced a filter-based approach using a modified Channel and Spatial Reliability Discriminative Correlation Filter (CSR-DCF) tracker – an object tracking algorithm that learns to follow targets by modeling both spatial reliability and feature channel (e.g., color, texture) importance. Initialized on sperm detections from a RetinaNet, individual CSR-DCF trackers were assigned in the first frame and updated via distance-based matching (≤ 15 pixels). To manage common tracking challenges like occlusions and re-entries, the method included a "Missing Tracks Joiner", which heuristically reconnected interrupted tracks based on motion consistency, spatial proximity (≤ 10 pixels), and expected sperm speed (≈ 5 px/frame). Tested on 9 phase-contrast microscopy videos with 411 spermatozoa, the method achieved 96.61% tracking F1 (i.e., ratio of correctly matched tracks to the total number of predicted and annotated tracks) with imperfect detections and reached 100% F1 under perfect detections.

Addressing occlusion challenges, Tan et al. [94] developed the Mean Shift–Collision Detection and Modified Covariance Matrix (MS-CDMCM) method integrated with their sperm head segmentation method. Their approach identifies collisions based on deviations in pixel weights from the ideal value of 1, isolating non-collided regions for more accurate tracking. These regions are then used to update the covariance matrix, effectively redefining the search area. In cases of sustained collisions, the algorithm reduces the search radius and reverts to previous positions to maintain track consistency. This strategy led to lower trajectory errors compared to standard benchmark methods, particularly in scenarios with frequent sperm interactions.

Building on their detection approaches, the teams participating in the aforementioned MediaEval Medico *Transparent Sperm Tracking* challenge in 2022 also integrated tracking methods. For instance, Huynh et al. [110] after YOLOv7 and tail-aware sperm detection, employs the open-source SORT algorithm. SORT uses Kalman filters to predict object positions and then matches these predictions to new detections based on distance, enabling fast, real-time tracking using motion cues only. To mitigate ID errors arising from visually similar sperm cells during abrupt microscope field-of-view changes, they incorporated a scene change detection module. It calculates frame-to-frame peak signal-to-noise ratio values and validates changes by checking for abrupt shifts in sperm count, resetting tracks upon scene transitions. This feature specifically addresses ID persistence errors across discontinuous scenes, as reflected by improved IDF1 metrics, which quantify the accuracy of correctly tracking individual sperm over time. Concurrently, Kosela et al. [109] combined their YOLOv5-based detection pipeline with StrongSORT, a tracker that improves upon the original SORT by adding appearance-based re-identification. While SORT relies solely on motion, StrongSORT incorporates deep appearance embeddings extracted using OS-Net, a lightweight CNN architecture trained to capture fine-grained and scale-invariant features. The accuracy of this integrated pipeline was evaluated using the Higher Order Tracking Accuracy (HOTA) metric, achieving a score of 0.283 for normal sperm tracking.

In the 2023 edition of the same challenge, Nguyen et al. [111] reported modest tracking performance (HOTA: 41.5%) but did not provide methodological details as their focus was the detection task. Osaid et al. [113] employed ByteTrack, a tracking-by-detection method that extends SORT by handling low-confidence detections to reduce missed targets and ID switches, though they focused on classification metrics without reporting tracking-specific results. Lastly, Nguyen-Mau et al. [112] implemented BoT-SORT, a more advanced tracker that uses a modified Kalman filter and global motion compensation to account for camera movements. They used default parameters but did not include quantitative tracking results in their submission.

Zhang et al. [118] introduced SpermTrack-EVD, an enhanced version of DeepOCSORT tracking algorithm tailored for sperm tracking. DeepOCSORT is a tracking algorithm that, similarly to StrongSORT, extends SORT by combining motion cues with appearance features, but it is specifically designed to maintain object IDs in challenging scenarios such as occlusions or crowded regions. SpermTrack-EVD introduces several enhancements: a stronger re-identification backbone (OSNet trained on VISEM-Tracking), dynamic weighting of motion and appearance cues, and camera motion compensation. These improvements reduce ID switches in clustered or occluded

regions while maintaining real-time processing (≈ 10.55 FPS). Evaluated on VISEM-Tracking, it achieved 74.303% HOTA and 71.167% MOTA, outperforming baseline DeepOCSORT with significantly fewer ID switches.

In summary, sperm tracking has largely relied on adapting trackers originally developed for general object tracking tasks, subsequently modified to address the specific challenges of sperm motility analysis. Over time, sperm tracking has progressed considerably, with state-of-the-art methods demonstrating strong performance in handling occlusions, ID switches, and crowded scenes. Recently, approaches that incorporate DL techniques have emerged, such as appearance-based re-identification, to further enhance tracking accuracy and robustness in these complex sperm motility scenarios.

Table 3.3: Summary of state-of-the-art sperm tracking.

Ref.	Dataset	Input Method	Tracker/ Association Method	Performance
[86]	Non-public (Synthetic sperm videos, human sperm videos)	Traditional detection	JPDA	Superior OSPA scores vs. traditional methods
[121]	Non-public (Synthetic sperm videos, human sperm videos, tadpole videos)	—	JPDA with track-coalescence avoidance	Maintained 100% track ID through collisions, OSPA superiority vs. traditional methods
[105]	Non-public (sperm and sperm track annotation)	RetinaNet detection (Section 3.2.2.2.1)	Modified CSR-DCF	Recall: 97.32% Precision: 95.2% Accuracy: 93.45% F1 score: 96.91%
[94]	Non-public (annotated sperm but not tracks)	Sperm head segmentation	MS-CDMCM	MAE: 13.32% MSE: 50.16% RMSE: 7.08%
[109]	VISEM-Tracking (sperm and sperm track annotation)	YOLO detection (Section 3.2.2.2.1)	StrongSORT and OSNet	HOTA: 28.3%
[110]	VISEM-Tracking (sperm and sperm track annotation)	YOLO detection (Section 3.2.2.2.1)	SORT	HOTA: 42.21% IDF1: 44.58%
Continued on next page				

Ref.	Dataset	Input Method	Tracker/ Association Method	Performance
[117]	Non-public (sperm and sperm track annotation)	YOLO detection (Section 3.2.2.2.1)	JPDA	MOTA: 91%
[111]	VISEM-Tracking (sperm and sperm track annotation)	YOLO detection (Section 3.2.2.2.1)	-	HOTA: 41.5% OWTA: 46.0%
[112]	VISEM-Tracking (sperm and sperm track annotation)	YOLO detection (Section 3.2.2.2.1)	BoT-SORT	-
[113]	VISEM-Tracking (sperm and sperm track annotation)	YOLO detection (Section 3.2.2.1)	ByteTrack	-
[118]	VISEM-Tracking (sperm and sperm track annotation)	YOLO detection (Section 3.2.2.2.1)	Modified DeepOCSORT-ReID	HOTA: 74.30% MOTA: 71.17%

3.2.2.2.3 Motility Assessment

Sperm motility assessment in computational analysis involves identifying clinically relevant motility categories from microscopy video data. This section focuses on this task specifically, examining how sperm motility is quantified or categorized and the effectiveness of the approach proposed – whether through tracking-based pipelines or end-to-end models. For tracking-based methods, the emphasis is placed on motility classification derived from reconstructed trajectories, as the detection and tracking components were addressed in the previous section. End-to-end approaches are considered holistically, as they integrate the full analysis pipeline into a single predictive model. A summary of representative studies is provided in Table 3.4.

Tracking-based Approaches

Following Faster R-CNN-based detection and conventional tracking, Valiushkaitė et al. [106] applied a simple threshold-based approach, classifying sperm as motile when their movement speed – computed via Euclidean distance between consecutive bounding box centers – exceeded 34.5 $\mu\text{m/s}$. The motility percentage was then calculated as the proportion of motile cells in the sample, achieving strong agreement with manual annotations (MAE = 2.92%, Pearson r = 0.969).

Mohammadi et al. [105] also relied on trajectory-based motion analysis, this time extracting five kinematic parameters – VSL, VCL, VAP, STR, and LIN – from tracked sperm. Motility classification was performed using predefined threshold rules: VAP values segmented speed categories

(e.g., rapid $> 50 \mu\text{m/s}$; slow $< 30 \mu\text{m/s}$), while $\text{STR} \geq 70\%$ indicated progressive motion. This rule-based system divided sperm into six motility classes. However, no classification accuracy metrics were provided as the approach depended critically on the reliability of the underlying detection and tracking.

In a more learning-driven framework, Jiao et al. [117] mapped kinematic features to motility classes using a three-layer backpropagation neural network. The model used three input parameters (VCL, VSL, VAP), a hidden layer with 10 nodes, and an output layer representing the current four WHO categories. Trained on 1000 samples extracted from 30 annotated videos, the network did not report direct classification performance.

Ottl et al. [90] developed MotilitAI to demonstrate that unsupervised feature learning and quantization can effectively predict sperm motility, even with noisy tracks. They applied simple tracking algorithms (Lucas-Kanade and Crocker-Grier) to the VISEM dataset to extract per-sperm features such as custom movement statistics and mean squared displacement. These features were aggregated per video using either Bag-of-words or ensemble averaging, and fed into various ML regressors to estimate percentages of progressive, non-progressive, and immotile sperm. The best performance (MAE: 7.31) was achieved using Crocker-Grier tracking, individual mean squared displacement features, Bag-of-words aggregation, and a Linear Support Vector Regressor. Progressive motility was predicted with particularly high accuracy ($R^2 = 0.74$), reinforcing the value of unsupervised, quantized features.

From the *Transparent Sperm Tracking MediaEval Challenge*, discussed in previous sections, both Kosela et al. [109] and Osaïd et al. [113] proposed motion-based strategies for motility classification following the detection and tracking of individual sperm cells. Kosela et al. [109] introduced a velocity-based classification method that compared average velocities over short ($T_{\text{short}} = 5/50 \text{ s}$) and long ($T_{\text{long}} = 20/50 \text{ s}$) intervals. Motility states were defined relative to the maximum observed velocity (v_{max}): sperm were labeled immotile if $v_{\text{short}} \leq 0.003 v_{\text{max}}$, non-progressive if $v_{\text{short}} > 0.003 v_{\text{max}}$ but $v_{\text{long}} \leq 0.01 v_{\text{max}}$ and progressive if both exceeded their respective thresholds. This interval-comparison approach exploited distinct velocity decay patterns among motility classes and yielded a mean absolute error of 15.6% for progressive, 8.6% for non-progressive, and 22.2% for immotile sperm estimation. In contrast, Osaïd et al. [113] applied a so-called "Greedy Shape Geometry Technique", that analyses the shape of sperm trajectories. Linear trajectories indicated progressive motility, circular paths denoted non-progressive movement, and near-static behavior implied immotility. Their method achieved the lowest MAE for progressive motility (3.096%), though performance was less robust for immotile distribution (59.987%).

End-to-end Deep Learning Models

The MediaEval *Multimedia for Medicine* task in 2019 catalyzed early DL for motility classification, using the VISEM dataset with expert manual analysis annotations. It defined a mandatory subtask to predict sample-level percentages of progressive, non-progressive, and immotile spermatozoa, alongside a separate morphology prediction task. Hicks et al. [122] adopted an InceptionV3 model, a deep CNN, with transfer learning, feeding it short grayscale video clips where frames were stacked as channels. They experimented with varying clip lengths and frame strides, finding

that stride 10 clips of three frames best captured movement patterns while maintaining pre-trained feature effectiveness with a MAE of 8.96. Concurrently, Rosenblad et al. [123] employed 3D CNNs, a type of CNN where convolutions are applied across both spatial and temporal dimensions, enabling the model to learn motion patterns directly from video sequences. They designed two architectures using 15-frame sequences: a shallow model with two convolutional layers and a deeper one with three convolutional blocks and pooling. Using data augmentation (random crops, flips, noise), the deeper model significantly outperformed the shallow one, achieving lower MAE and RMSE of 8.83 and 11.39, respectively. Thambawita et al. [124] presented two distinct approaches. In one approach, they used an autoencoder to compress temporal dynamics from video clips into spatial feature-images, which were then processed by a ResNet34 for regression. By comparing four input formats – from single frames to stacks of 18 frames – they demonstrated improved performance with longer temporal context, achieving a MAE of 9.43 using 18-frame stacks. In a second study, the authors modified ResNet34 to handle multi-modal inputs combining grayscale frames with dense optical flow (stride 1 and 10) and introduced a two-stage architecture with a Multi-Layer Perceptron to mitigate overfitting. This fusion of appearance and explicit motion representations yielded the lowest error among MediaEval entries (MAE: 8.825), underscoring optical flow's efficacy in capturing kinematics.

In the same year, Hicks et al. [125] published a foundational study in *Nature Scientific Reports* demonstrating the use of CNNs to predict motility proportions in progressive, non-progressive, and immotile categories based on video data. They evaluated several CNN architectures tailored to the different tested input types: ResNet-50 for single RGB frames and optical flow images, a modified ResNet-50 trained from scratch for 30-channel stacked grayscale frames, ResNet-18 for vertical frame matrices, and a custom two-stream network integrating raw frames with optical flow. The best results (MAE = 8.8) were achieved using either grayscale stacking of 30-frame sequences or dense optical flow, showing that temporal and motion-based cues significantly enhance prediction. In a later follow-up published in the same journal [126], the research group extended their approach by training a ResNet-50 deep CNN on sparse optical flow images. Their models predicted motility distribution using both a three-category (progressive, non-progressive, immotile) and a four-category (rapid progressive, slow progressive, non-progressive, immotile) classification scheme. Trained on 65 videos from a quality assessment dataset and evaluated with tenfold cross-validation, the three-category model achieved an MAE of 5 with high correlation to manual assessments ($r = 0.88\text{--}0.89$), while the four-category model yielded an MAE of 7, with strongest agreement for progressive motility.

Building on this line of research, Priyansi et al. [127] explored a 3D CNN-based classification pipeline, also using the VISEM dataset, but focusing on direct three-class classification (progressive, non-progressive, immotile) of sperm videos. They evaluated three models: a baseline 3D ResNet18, and two variants combining video with 19 tabular patient data features using 3D ResNet18 and ResNet34. The baseline model achieved perfect accuracy (100%) on validation and test sets, while the addition of tabular data reduced performance (88.9% and 77.8%, respectively). Despite high accuracy, the method assigns a single motility class per sample, lacking proportional

estimates needed for clinical assessments.

Fujii et al. [128] proposed a transformer-based model for sperm motility assessment from videos, named RoSTFine. Unlike traditional CNN-based architectures that rely on fixed receptive fields, transformers use self-attention mechanisms to dynamically capture long-range dependencies in both space and time. RoSTFine uses a Patch Selection Module that automatically identifies and focuses on the most informative sperm regions – such as the head, neck, and tail – by assigning attention scores to video patches. Separate branches then extract spatial features (morphology) and temporal features (motion), which are combined to predict a 5-grade motility score. Notably, this study used an uncommon annotation approach: multiple experts independently assigned a single motility grade to each sample, and the final ground truth was defined as the distribution of these expert-assigned labels. Thanks to this targeted attention mechanism, RoSTFine outperformed baseline video models such as TimeSformer and 3D CNNs, achieving a MSE of 1.104×10^{-2} in matching expert grade distributions and a balanced accuracy of 36.37% across all motility grades.

Goh et al. [129] introduced a multimodal DL model that simultaneously predicts sperm motility and concentration. The model combines temporal motion features extracted by a 3D CNN from stacked dense optical flow inputs with spatial morphological features captured by a ResNet18 from RGB frames. These two streams are fused at a late stage with batch normalization, enabling the network to simultaneously estimate motility proportions (progressive, non-progressive and immotile) and regress sperm concentration values. The authors evaluated various optical flow strides and temporal depths, identifying an optimal configuration with stride 1 and depth 8. This combined approach attains state-of-the-art motility prediction achieving a MAE of 8.048, while providing reliable concentration estimates comparable to commercial systems, with a strong Pearson's correlation coefficient of 0.853.

In summary, sperm motility classification has progressed through tracking-based methods that use explicit motion features and end-to-end DL models that learn directly from video data. While tracking approaches provide interpretable results, their accuracy depends on reliable detection and tracking as well as selected classification rules. DL methods, especially those using 3D CNNs, have shown improved accuracy by capturing complex spatiotemporal patterns. However, varying classification outputs and the lack of manual annotations in some studies limit consistent validation. Despite these challenges, both paradigms hold significant potential for future clinical use.

Table 3.4: Summary of State-of-the-art Sperm Motility Assessment Studies.

Ref.	Dataset	General Approach	Motility Assessment Methodology	Input	Output*	Best Performance
[125]	VISEM (85 videos w/ manual annotation)	End-to-end architecture	ResNet18, Modified ResNet50, ResNet50, Two-stream Network	Raw frames; grayscale-stacked sequences; optical flow; tabular data	3 motility classes distribution	MAE: 8.786
[122]	VISEM (85 videos w/ manual annotation)	End-to-end architecture	InceptionV3	Frame sequences with various strides and lengths, with/without binarization	3 motility classes distribution	MAE: 8.962 RMSE: 12.046
[123]	VISEM (85 videos w/ manual annotation)	End-to-end architecture	Custom 3D CNN	15 consecutive frames	3 motility classes distribution	MAE:8.83 RMSE: 11.39
[130]	VISEM (85 videos w/ manual annotation)	End-to-end architecture	Autoencoder + ResNet34	9 and 18 stacked consecutive frames	3 motility classes distribution	MAE: 9.427
[124]	VISEM (85 videos w/ manual annotation)	End-to-end architecture	ResNet34, ResNet34 + MLP	Original frame; dense optical flows (strides 1 & 10)	3 motility classes distribution	MAE: 8.825
[105]	Non-public (Not annotated)	Tracking-based	RetinaNet (Section 3.2.2.2.1) + Modified CSR-DCF (Section 3.2.2.2.2) + Velocity threshold-based motility classification	Raw videos	6 motility classes distribution	-
[106]	VISEM (85 videos w/ manual annotation)	Tracking-based	Faster R-CNN (Section 3.2.2.2.1) + Traditional Tracking + Threshold-based motility classification	Raw videos	2 motility classes distribution	MAE: 2.92 Pearson correlation: 0.969

Continued on next page

Ref.	Dataset	General Approach	Motility Assessment Methodology	Input	Output*	Best Performance
[127]	VISEM (85 videos w/ manual annotation)	End-to-end architecture	3DResNet18, 3DResNet34	50 consecutive frames	Single-label classification out of 3 motility classes	Accuracy: 100%
	YOLO					
[109]	VISEM (85 videos w/ manual annotation)	Tracking-based	(Section 3.2.2.2.1) + StrongSORT (Section 3.2.2.2.2) + Dual-interval velocity thresholds	Raw videos	3 motility classes distribution	MAE: 15.5 RMSE: 16.6
	YOLO					
[90]	VISEM (85 videos w/ manual annotation)	Tracking-based	Traditional tracking + Extraction of custom features + Bag-of-Words + ML algorithms	Raw videos	3 motility classes distribution	MAE: 7.31
	YOLO					
[113]	VISEM-Tracking (20 videos w/ manual annotation)	Tracking-based	(Section 3.2.2.2.1) + ByteTrack (Section 3.2.2.2.2) + Trajectory shape analysis	Raw videos	3 motility classes distribution	MAE: 25.251
	YOLO					
[126]	VISEM-QC (65 videos w/ manual annotation)	End-to-end architecture	ResNet-50	Sparse optical flow images	3 and 4 motility classes distribution	MAE (3 classes): 5 MAE (4 classes): 7
	YOLO					
[117]	Non-public (30 videos w/ manual annotation)	Tracking-based	(Section 3.2.2.2.1) + JPDA (Section 3.2.2.2.2) + BP Neural Network trained on VCL, VSL, VAP	Raw videos	4 motility classes distribution	-
	YOLO					
[128]	Non-public (615 videos w/ manual annotation)	Tracking-based	Traditional Detection and Tracking + Custom RoSTFine	16 consecutive frames	5 motility classes distribution	MSE: 1.104×10^{-2} Balanced Accuracy: 36.37%
	YOLO					

Continued on next page

Ref.	Dataset	General Approach	Motility Assessment Methodology	Input	Output*	Best Performance
[129]	VISEM (85 videos w/ manual annotation)	End-to-end architecture	Custom Multimodal Architecture	Dense optical flow images with different stride and depth and single raw frames	3 motility classes distribution	MAE: 8.048

Note: These studies report motility classifications ranging from 2 to 6 classes. Example categories include: motile/immotile (2 classes), progressive/non-progressive/immotile (3 classes), rapidly/slowly progressive/non-progressive/immotile (4 classes), and additional granular classes in some tracking-based studies.

3.2.2.3 Code Availability

Publicly available code for automated sperm motility assessment remains relatively rare, which limits reproducibility and wider adoption of methods. While many studies describe their algorithms and present promising results, only a few provide open-source implementations. Notable exceptions include Ottl et al. [90], Mohammadi et al. [105], and Priyansi et al. [127], who have made their codebases accessible, contributing to transparency and facilitating further research in this field.

3.2.2.4 Datasets

Although a wide range of studies have investigated sperm motility, there is a limited number of publicly accessible comprehensive datasets. Phase-contrast microscopy is by far the most common imaging modality used in both public and private datasets, as it enhances the contrast of spermatozoa, making them more distinguishable from the background.

Across various studies, the VISEM dataset stands out as the most widely used for motility assessment [18]. Developed by Oslo Metropolitan University and Simula Research Laboratory, this multi-modal dataset includes live phase-contrast microscopy videos of spermatozoa from 85 participants, collected per WHO standards, along with semen analysis results, participant details (age, abstinence time, body mass index), and metrics like sex hormone and fatty acid levels. The videos (640 × 480 resolution, 50 frames per second) last 2–7 minutes. Building on this foundation, the VISEM-Tracking dataset was introduced in 2023 [67]. It extends VISEM by adding manual frame-level annotations with bounding boxes and unique tracking IDs for individual sperm, including normal sperm, sperm clusters, and small/pinhead sperm, in 20 of the videos.

The VISEM-QC dataset, also developed by Simula Research Laboratory, contains videos of 65 wet semen samples recorded at 400× magnification, collected between 2006 and 2018 through the External Quality Assessment Programme by ESHRE [131]. Recorded at the Andrology Unit, Karolinska University Hospital and Institutet, each sample includes multiple 5–10 second clips

capturing at least 200 spermatozoa. Videos were independently analysed by 4 to 10 laboratories, enabling inter-laboratory comparison and quality control. Motility was manually classified using both the current WHO 4-tier system and the earlier 3-tier classification, with percentage annotations provided for each motility category to support benchmarking.

The publicly available SeSViD dataset includes 12 brightfield semen videos recorded at 100x magnification [119]. However, it currently lacks manual annotations for sperm detection or motility classification, which limits its utility for supervised learning. Additionally, its very low frame rate of 5 fps further restricts its usefulness for detailed motion analysis.

To address the scarcity of diverse annotated datasets and improve model generalizability, efforts also include creating synthetic datasets [67] and developing semi-automatic annotation methods [68]. Beyond motility, several datasets are available specifically for morphology assessment [132–134], reflecting the distinct challenges and techniques involved in that domain.

3.3 Summary

Traditional CASA technologies have provided a foundation for objective and reproducible sperm assessment, yet their limitations – such as high cost, lack of full automation, and reduced accuracy with heterogeneous human samples – have hindered widespread clinical adoption. Recent advancements in AI, especially DL, have markedly enhanced sperm analysis capabilities, giving rise to two predominant methodological paradigms: one based on sequential detection, tracking, and classification, and another focusing on direct prediction of distributions in an end-to-end architecture. Both approaches have shown promising results. YOLO-based architectures have emerged as the most commonly used and most effective for sperm head detection. Progress in tracking has benefited from probabilistic models and appearance-based re-identification methods, improving ID consistency even in dense or occluded scenes. Overall, the integration of AI is pushing the field toward more accurate, scalable, and clinically relevant tools for sperm motility assessment. This, in turn, highlights the growing need for larger, well-annotated datasets to further improve model performance.

Chapter 4

Methodology

This chapter presents the methodological framework adopted to achieve the objectives of this thesis. In addition to describing the dataset and computational resources, the work is structured around two main computational approaches. The first involves the implementation and evaluation of a state-of-the-art end-to-end DL architecture for direct prediction of sperm motility categories. The second, which constitutes the main contribution of this work, focuses on the development and validation of a novel, comprehensive tracking-based algorithm, further integrated into a user-friendly graphical interface designed for practical laboratory use.

4.1 Dataset

4.1.1 Source and Ethical Compliance

The dataset used in this study comprises brightfield microscopy videos of human spermatozoa, provided by the UMIB research unit at ICBAS. The seminal samples were collected at the Centre for Reproductive Genetics A. Barros (Porto, Portugal) from male patients undergoing fertility assessment or from healthy donors, using only surplus freshly ejaculated semen samples. All procedures complied with ethical standards for research involving human biological material. Written informed consent was obtained from all participants and approved by the Joint Ethics Committee ICBAS/ULSSA. The Centre for Reproductive Genetics A. Barros operates under the National Medically Assisted Procreation Act (2006), supervised by the National Council for Medically Assisted Procreation. According to these regulations, both clinical data and biological material can be used for research purposes, provided that donor anonymity is preserved and consent is properly documented.

4.1.2 Sample Collection and Preparation

Samples were obtained by masturbation after a sexual abstinence period of three to five days. Upon collection, the semen was kept at 37°C until complete liquefaction (up to 2 hours) and then processed following the WHO guidelines [6]. Spermatozoa were isolated through centrifugation

at $300 \times g$ for 5 minutes at room temperature, followed by two washing steps using isotonic phosphate-buffered saline. The videos may include additional treatments or post-processing steps, resulting in variable protocols depending on the experimental conditions, with samples sometimes further diluted or prepared in different media as needed.

4.1.3 Video Acquisition and Dataset Overview

The prepared samples were then observed and recorded using a Nikon Eclipse E400 brightfield microscope (200 $\times$ magnification), equipped with a DS-Fi3 microscope camera (Nikon, Tokyo, Japan). Image acquisition was performed with the Nikon NIS-Elements Br 4.60.00 software.

The full dataset comprises a total of 463 brightfield microscopy videos of human spermatozoa of 2880 $\times$ 2048 resolution. As the recordings originate from semen samples obtained from different individuals, including patients undergoing fertility assessments, some videos contain a high concentration of round cells and debris, adding visual complexity and posing additional challenges for computational analysis. The samples also differ in sperm concentration and dilution, reflecting the natural variability in laboratory preparation protocols. In addition, video acquisition was carried out by multiple researchers, resulting in variations in imaging conditions such as lighting, contrast, and focus. This variability is illustrated in Figure 4.1.

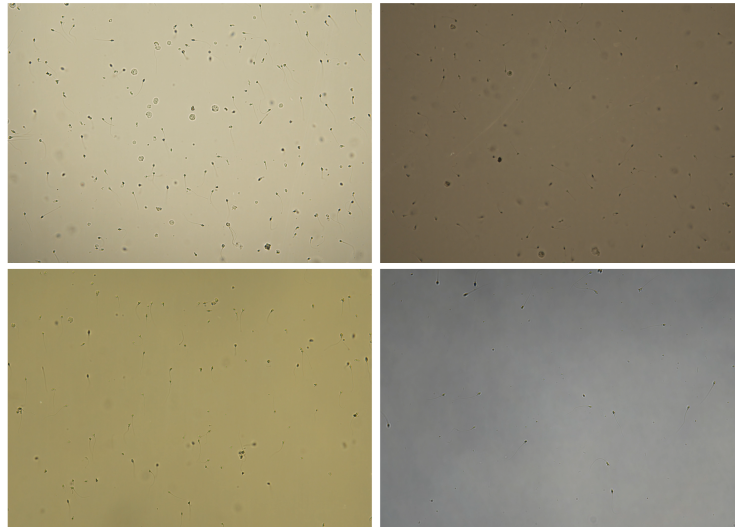


Figure 4.1: Example of variability between videos. First frame from four different samples recorded under varying imaging conditions.

Out of the full dataset, 90 videos include manual annotations of the three-category sperm motility distributions, performed by researchers at the UMIB laboratory. These annotations include the total number of sperm considered for the motility assessment (not the overall sperm count, but a representative subset sufficient to estimate the sample’s motility distribution); the number of sperm classified as progressive, non-progressive, and immotile; and the corresponding percentage of each motility class. The remaining 373 videos are unannotated, and only a single frame from each was used to train a detection model, as further described in Section 4.4.2.

The annotated videos have durations ranging from 9.06 to 10.02 seconds, with frame rates between 9.76 and 16.23 frames per second (FPS). These variations stem from differences in acquisition settings influenced by, not only operator choices, but also hardware performance. The distribution of video durations and frame rates is shown in Figure 4.2.

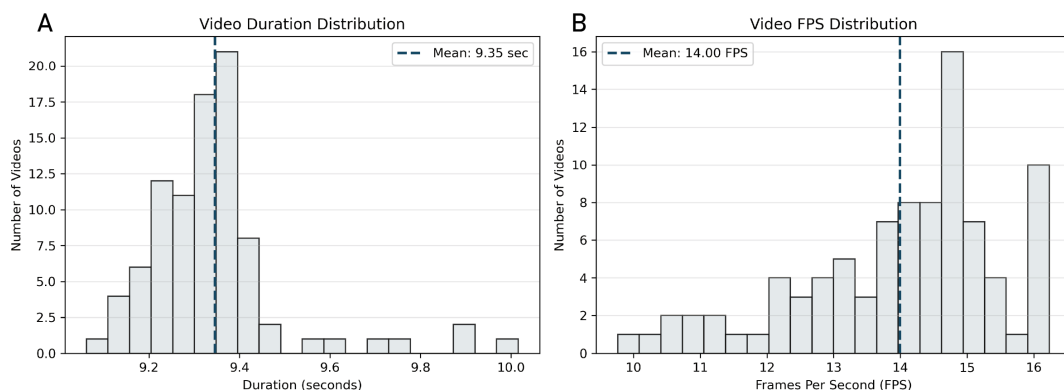


Figure 4.2: Video dataset overview. A) Histogram of video duration distribution, in seconds. B) Histogram of video frame rate distribution, in FPS.

The dataset annotations reveal an uneven distribution across sperm motility classes, as can be verified in the box plots in Figure 4.3. Immotile sperm dominate the dataset, with most samples showing percentages above 50%. While the distribution of immotile sperm is somewhat spread, values generally remain high across samples. Progressive sperm appear at moderate levels, with percentages ranging roughly from 0% to 60%. Their median lies below that of immotile sperm but above non-progressive sperm. Non-progressive sperm are the least represented, with percentages mostly under 20% and a very tight distribution around these low values, reflecting minimal variability.

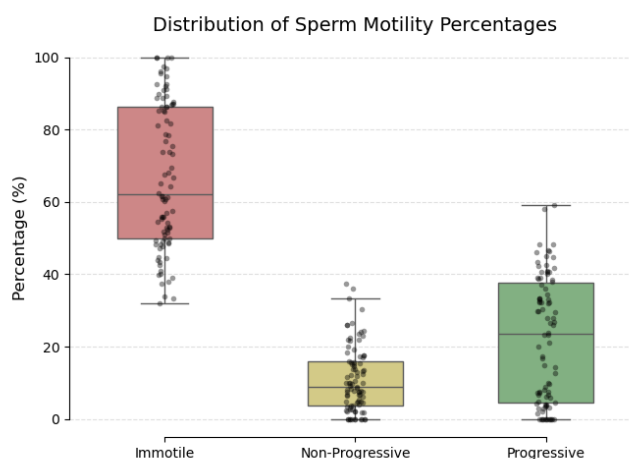


Figure 4.3: Box plots of sperm motility percentages (immotile, non-progressive, progressive) across the 90-video manually annotated dataset.

4.2 Computational Resources

All model training was conducted on a remote server equipped with an NVIDIA GeForce RTX 2080 Ti GPU (11 GB VRAM) running CUDA version 12.6. Subsequent prediction tasks and further algorithm development were carried out on a personal computer running Windows 11, featuring an AMD Ryzen 5 4600H processor with 12 cores operating at approximately 3.0 GHz and 8 GB of RAM. An NVIDIA GeForce RTX 2060 GPU with 6 GB of VRAM was additionally used on this personal computer to accelerate prediction computations.

4.3 End-to-End Approach

To explore the feasibility of an end-to-end DL approach for sperm motility classification, we implemented the method proposed in [129]. This architecture was selected as a representative example of recent state-of-the-art models capable of directly predicting motility categories from video data. The decision was supported by its strong reported performance and the availability of sufficient implementation details in the original publication.

Since the original source code was not publicly available, the entire pipeline was reimplemented based on the methodological descriptions provided in the publication. This included generating the input modalities, designing the network architectures, and reproducing the training setup as closely as possible. Some adaptations were required due to differences between the dataset used in the original study and the one employed in this work, as detailed in Table 4.1, in order to ensure compatibility and reliable performance in the context of this study.

Table 4.1: Differences between the dataset used in the original study (VISEM) and our dataset.

	VISEM	Our dataset
Number of annotated videos	85	90
Frame rate	50 FPS	~14 FPS (mean)
Magnification	$\times 400$	$\times 200$
Resolution	640×480	2880×2048

4.3.1 Input Modality Generation

As done in the original methodology, two input modalities were generated: a video-based modality (D1) for temporal analysis using a 3D CNN, and an image-based modality (D2) for spatial analysis using a conventional 2D CNN, ResNet18. As in the original work, the D1 modality consisted of sequences of stacked dense optical flow frames, while the D2 modality comprised single RGB frames extracted from each video sequence.

D1 Input - Video-based Modality

The D1 modality was generated by computing dense optical flow between consecutive frames using Gunnar Farneback’s algorithm implemented in OpenCV, capturing the motion of sperm

cells. The generation process for this modality is illustrated in Figure 4.4. In the original study, the best-performing configuration consisted of D1 samples composed of 8 dense optical flow frames (Z), computed with a stride (X) of 1 (i.e., between every pair of consecutive frames), resized to 144×144 pixels and stacked into a tensor of shape 3×144×144×8.

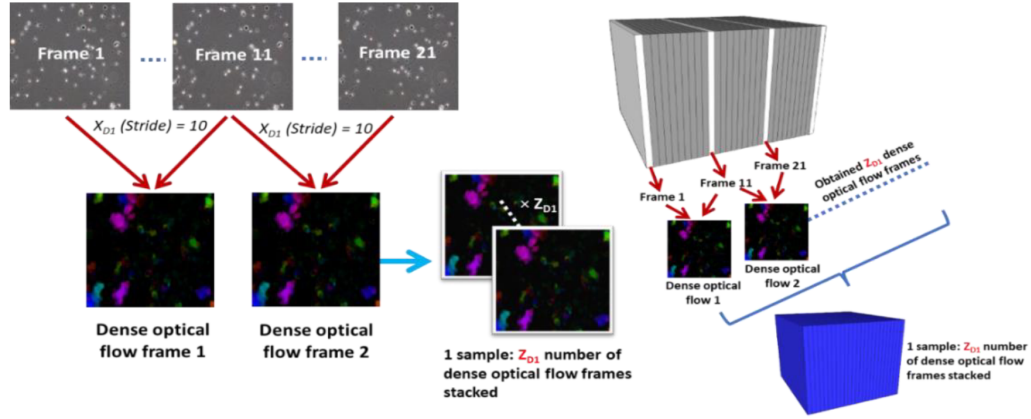


Figure 4.4: Generation of the D1 modality using dense optical flow with specified stride (X) and depth (Z), followed by stacking into video-based input. Source: [129].

In our implementation, it was necessary to adjust the input configuration to account for the relatively lower frame rate of our videos. While the same stride ($X=1$) was retained – as it represents the minimum possible temporal gap between frames – the depth of the input stack was reduced to $Z=2$, in order to proportionally reflect the reduced temporal resolution, ensuring that the optical flow captures a comparable motion span relative to the original study. To align with the magnification difference, each frame was first center-cropped to approximately simulate the original field of view, before optical flow computation and resizing to 144×144 pixels. To obtain a similar number of total samples, we extracted 47 frames per video, generating 24 samples per video. This approach maintained consistency in dataset size for both training and evaluation purposes. Table 4.2 summarizes the main differences between the original configuration (i.e., Set A from their experimentation) and the adapted one used in this study.

Table 4.2: Comparison of D1 input configuration: original article vs. our adaptation.

D1 Input	Original Article	Our Adaptation
Stride (X)	1	1
Depth (Z)	8	2
Total Samples Generated (T)	2890	2880
Video Number	85	90
Total Frames Per Video (N)	272	64
Samples Obtained Per Video	34	32
Preprocessing Steps	Optical flow computation, Resize	Center-crop, Optical flow computation, Resize
Dimension	$3 \times 144 \times 144 \times 8$	$3 \times 144 \times 144 \times 2$

D2 Input – Image-based Modality

For the generation of D2 inputs (image-based modality), each sample corresponds to the first raw frame used in the optical flow computation for the associated D1 sample, as depicted in Figure 4.5. In the original study, these frames were resized from their native resolution of 640×480 pixels to 512×384, using a downscaling ratio of 0.8, in order to reduce computational cost while preserving relevant spatial information.

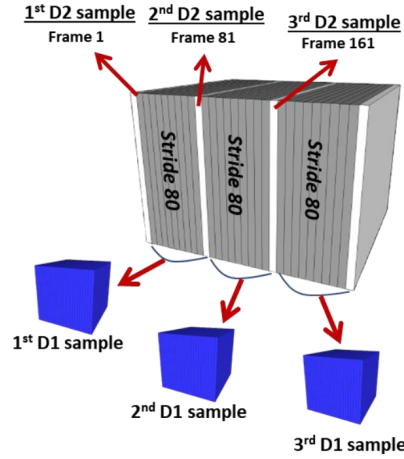


Figure 4.5: Extraction of D2 input from the source frames used in D1 optical flow computation. Source: [129].

In our implementation, a corresponding preprocessing pipeline was applied. To address the magnification difference between the datasets, a center-crop was first performed on each frame to approximate the field of view and spatial scale of the original study. Following cropping, frames were resized to 540×384 pixels, maintaining the same height and slightly adjusting the width to match the aspect ratio of our videos. An example of a D1 sample (A, B) and its corresponding D2 sample (C) from the same video segment is shown in Figure 4.6.

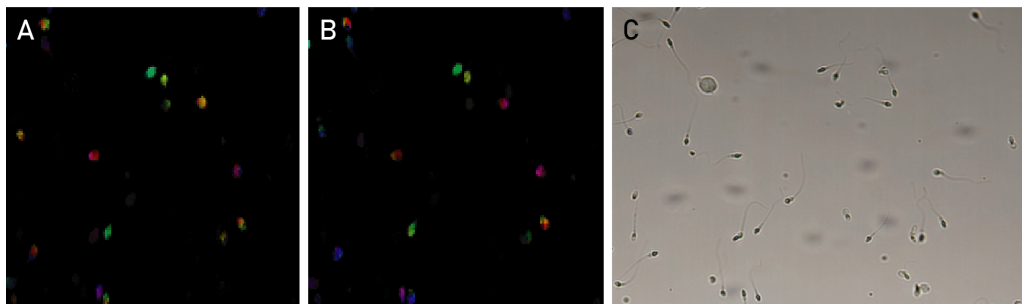


Figure 4.6: Example of a D1 sample (A, B) and its corresponding D2 sample (C) from the same video segment, used in our implementation.

4.3.2 Ground Truth Annotation and Scaling

For motility prediction, the output target was represented as the proportion of spermatozoa in each motility category (i.e., progressive, non-progressive and immotile), in line with the original study.

These percentages, constrained to sum to 100%, were normalized by dividing by 100, allowing compatibility with the Softmax-based output layer of the model. For concentration prediction, unlike the original study, which used annotated ground truth values, our dataset lacked direct annotations for this parameter. To address this, we estimated concentration values using a custom algorithm described in Section 4.4.3, based on the detection outputs of our YOLO-based pipeline. Consistent with the original approach, the concentration, in 10^6 sperm/mL, was logarithmically transformed prior to training to reduce skewness in the distribution and improve the performance of the regression model.

4.3.3 Network Architectures

Following the original methodology, our implementation relied on a combination of unimodal and multimodal DL architectures to perform sperm motility prediction. The overall strategy began with pretraining two separate models – a 3D CNN for the D1 modality and a ResNet18 for the D2 modality. These were then combined into a multimodal network designed to improve motility predictions. While the original study culminated in a dual-output network for simultaneous motility and concentration prediction, in this work concentration information from the YOLO-based detection pipeline is used solely as an auxiliary feature to enhance motility classification. Therefore, the dual-output network was not implemented.

The first unimodal model is a 3D CNN designed to capture temporal motion features from stacked dense optical flow frames (D1). The network consists of three convolutional blocks, each containing a 3D convolution layer (with kernel size $3 \times 3 \times 3$, stride 1 and padding 1), followed by ReLU activation, a 3D max pooling layer (with kernel size $1 \times 2 \times 2$ and stride $1 \times 2 \times 2$), and batch normalization. While the original methodology employs a max pooling kernel of $2 \times 2 \times 2$, this configuration was intentionally modified to preserve temporal resolution, given the limited number of input frames per sample ($Z = 2$), while still reducing spatial dimensions progressively. The convolutional layers have filter counts of 64, 64, and 128 respectively, with a stride of 1 applied in convolution layers. After the final convolutional block, a pooling operation is required to reduce the feature map to a vector suitable for classification. Considering the input depth of only two frames, we opted for average pooling over the spatial dimensions while preserving the temporal axis. The resulting feature vector is then flattened and passed through a fully connected layer to produce the three motility class scores, followed by a Softmax activation to generate class probabilities. For ease of integration in the multimodal setup, the model is conceptually split into two parts: Part 1 includes all layers up to average pooling, and Part 2 comprises the fully connected and Softmax layers. The architecture design proposed in the original article is illustrated in Figure 4.7A, serving as the basis for our implementation with the aforementioned adaptations.

The second unimodal model processes video image frames (D2) via a modified ResNet18 architecture. ResNet18 is a deep convolutional neural network composed of 18 layers, featuring skip (or residual) connections [135]. These connections enable the network to pass the input of a layer directly to a deeper layer, effectively ‘skipping’ some intermediate layers. This helps avoid

the vanishing gradient problem – a common issue where gradients become too small to effectively train very deep networks – thus allowing the network to learn better and train more efficiently. This implementation maintains the standard structure of ResNet18, including multiple blocks of 3×3 convolutions and periodic downsampling. The modification lies in the network’s output head, where the default classification layer was replaced with a custom regression head specifically designed for concentration prediction. The network is divided into three conceptual parts: Part 1 extends from input up to the last convolutional layer, Part 2 includes average pooling and a fully connected layer producing 1000 features, and Part 3 contains a ReLU activation, dropout, and final fully connected regression layer outputting the predicted concentration. This architecture, exactly as implemented in our approach, is illustrated in 4.7B.

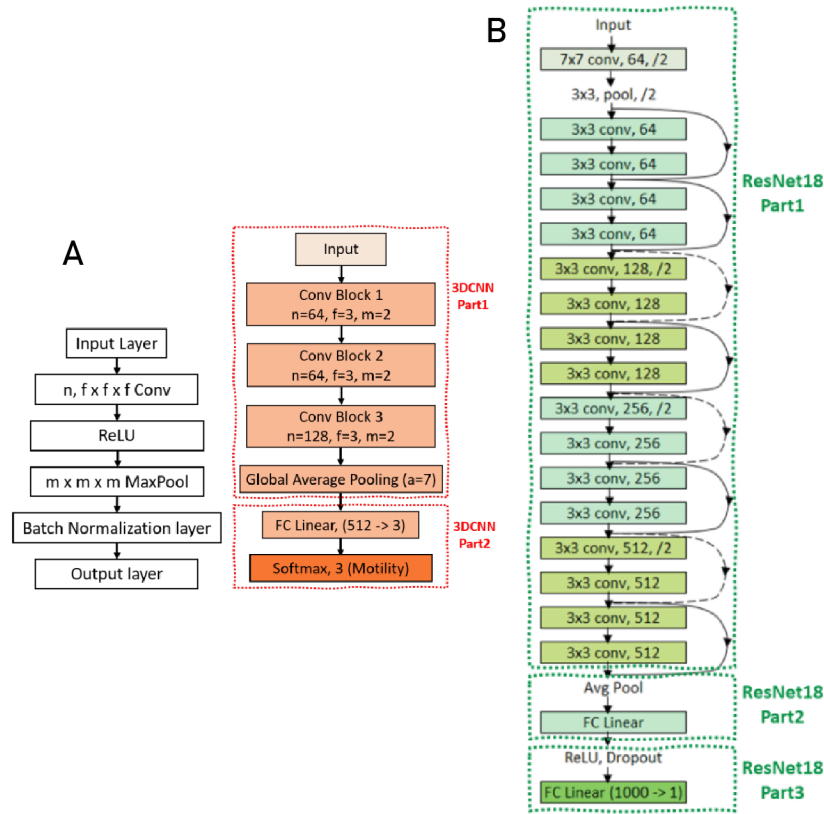


Figure 4.7: Architectural diagrams of the unimodal models. A) 3D CNN for D1-based motility classification: left, single convolutional block; right, full 3D CNN. Our implementation differs from the original by using 3D max pooling with a reduced kernel size of $1 \times 2 \times 2$ and applying average pooling over spatial dimensions instead of global pooling, resulting in a fully connected layer input size of 256 instead of 512. B) Modified ResNet18 architecture for D2-based concentration regression. Source: [129].

To leverage complementary information from both modalities, pretrained 3D CNN and ResNet18 components are combined into a multimodal network to predict motility. This architecture’s design is illustrated in Figure 4.8. The D1 input passes through the initial layers of the 3D CNN, followed by batch normalization. Simultaneously, the D2 input passes through the early layers

of ResNet18, with frozen weights to retain previously learned visual features. It then continues through deeper ResNet layers and batch normalization. Batch normalization between concatenation and linear layers helps stabilize training by normalizing feature distributions from different modalities. The outputs from both streams (i.e. 256 features from the 3D CNN and 1000 from ResNet18, in our implementation) are concatenated into a 1256-dimensional feature vector. This vector is passed through a ReLU activation, followed by a fully connected layer that reduces the dimensionality to 1000, another ReLU, and a final fully connected layer that outputs the 3-class motility prediction, followed by a Softmax activation.

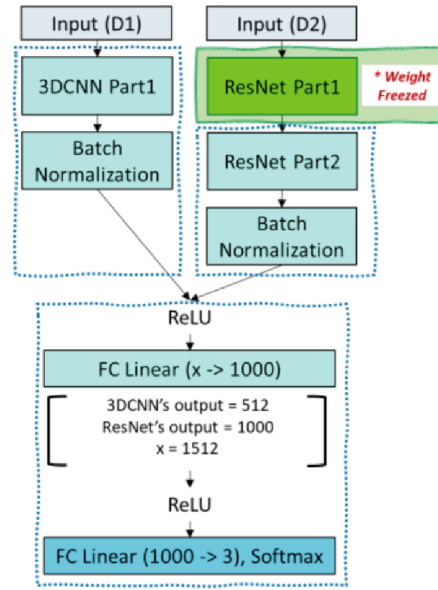


Figure 4.8: Architectural diagram of the multimodal single-output network for motility classification. Our implementation differs from the original in that the 3D CNN output is 256, making the fully connected layer input 1256. Source: [129].

4.3.4 Training Setup

Overall, this method involves training three different model architectures. All trained models used the Adam optimizer combined with a learning rate (LR) scheduler, StepLR, decreasing the LR by 5% every 10 training epochs. A batch size of 30 was applied consistently across these models. Training was performed using 3-fold cross-validation to ensure robustness and generalization. Details on initial LRs, training epochs, and loss functions are summarized in Table 4.3.

Table 4.3: Training configurations for each model architecture in end-to-end multimodal approach, in accordance with [129].

Model	Optimizer	LR Scheduler	Batch Size	Initial LR	Epochs	Loss Function
Unimodal 3D CNN (Motility)	Adam	StepLR ($\gamma=0.95$, step=10)	30	1×10^{-6}	Early stop (patience=10)	MSE
Unimodal ResNet18 (Concentration)	Adam	StepLR ($\gamma=0.95$, step=10)	30	3×10^{-4}	100	RMSE
Multimodal Single-Output (Motility)	Adam	StepLR ($\gamma=0.95$, step=10)	30	3×10^{-5}	40	MSE

Note: γ is the multiplicative factor used in StepLR to decay the LR at each step.

4.3.5 Training and Performance Evaluation

During training, we tracked the loss over both training and validation sets to assess convergence and generalization. This allowed us to evaluate how well each architecture learned its respective task. In particular, MSE and RMSE were monitored in the models where they were the designated loss function:

- **Mean Squared Error (MSE):** Measures the average of the squared differences between predicted values and ground truth, penalizing large errors heavily:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (4.1)$$

where y_i is the true value, $\hat{y}_i$ is the predicted value and N is the number of samples.

- **Root Mean Squared Error (RMSE):** The square root of MSE, providing error in the same units as the target variable:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (4.2)$$

After training, each trained network was evaluated on the test set. Two complementary regression metrics were used to assess performance:

- **Mean Absolute Error (MAE):** The average absolute difference between predicted and true values, giving a straightforward measure of average error magnitude:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (4.3)$$

- **Coefficient of Determination (R^2):** Measures the proportion of variance in the ground truth explained by the model predictions. Values closer to 1 indicate better explanatory power:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (4.4)$$

where $\bar{y}$ is the mean of true values.

For motility prediction, represented as a continuous distribution across three categories (progressive, non-progressive, and immotile), these metrics were computed both per category and on the overall output vector, with values expressed as percentages. For concentration prediction, the same metrics were applied directly to the predicted and true concentration values, expressed in 10^6 sperm/mL.

4.4 Proposed Algorithm

Our main contribution in this thesis is the development of a tracking-based algorithm that integrates multiple stages of sperm motility analysis. We propose a pipeline that includes sperm detection, concentration estimation, tracking of detected cells, extraction of kinematic parameters, and motility classification. A user-friendly graphical user interface (GUI) was also developed to integrate and facilitate the use of these functionalities. Figure 4.9 illustrates the overall pipeline of the proposed approach.

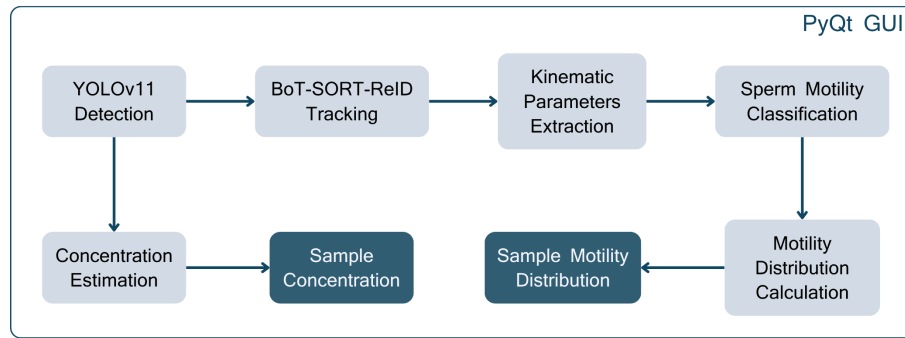


Figure 4.9: Overview of the proposed sperm motility analysis algorithm.

4.4.1 Data Preparation

4.4.1.1 Dataset Annotation

Accurate ground truth annotations of sperm are a prerequisite for training a reliable detection model. The annotation process was conducted using Roboflow, a web-based platform designed to streamline dataset annotation and also offers additional capabilities such as data augmentation and model training for computer vision applications [136]. Roboflow provides a convenient interface for labeling, supports various annotation formats, and integrates tools for automated labeling,

version control, and export, making it particularly suitable for iterative and large-scale annotation workflows.

As previously discussed, the original videos consisted of large, high-resolution frames (2880×2048 pixels) and with varying image characteristics due to differences in microscope configurations such as brightness, contrast, and focus. To ensure representative coverage of these conditions, the first frame of each video was selected for annotation. For more practical annotation, each frame was divided into four smaller tiles (1440×1024 pixels), allowing for a more efficient use of Roboflow’s annotation features.

A bounding box was fit around each sperm head, including varying morphologies, provided they were visually distinguishable and sufficiently in focus, even in cases of partial occlusion or when partially outside the field of view, as seen in Figure 4.10. An initial set of 525 images was manually annotated and then used to train a YOLOv11s model, which generated preliminary annotations for the remaining frames. Despite the use of automation, all predicted annotations were manually reviewed and corrected to ensure quality and consistency.

In total, 373 different video frames were processed into 1,492 annotated image tiles, resulting in 21,586 labeled sperm bounding boxes.

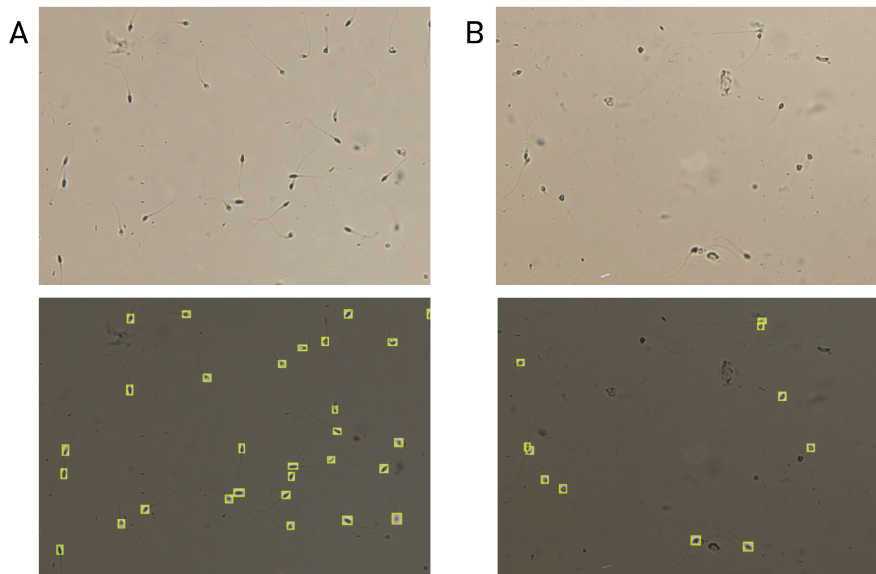


Figure 4.10: Examples of sperm head annotations in two sample images (A and B).

4.4.1.2 Data Split and Preprocessing

The annotated dataset was randomly divided into training, validation, and test sets following a 70%–15%–15% split. This resulted in 1,051 images for training, 221 for validation, and 220 for testing, as can be seen in Table 4.4.

To reduce input dimensionality and mitigate the impact of color variations introduced by imaging conditions, all images were converted to grayscale, allowing the model to focus on relevant morphological features.

Table 4.4: Distribution of sperm detection dataset across train, validation, and test sets.

Dataset Split	Percentage	Number of Images
Train Set	70%	1,051
Validation Set	15%	221
Test Set	15%	220

4.4.1.3 Data Augmentation

Data augmentation is a widely adopted technique that artificially expands the size and variability of the training dataset, mitigating overfitting and enhancing the model’s generalization performance on previously unseen data [137]. In light of these benefits, data augmentation was systematically applied to the training images to enhance the robustness of the detection model against variability inherent in microscopy imaging conditions.

Augmentations were implemented using the Albumentations library, which offers a fast and flexible framework for applying a wide range of image transformations while preserving and adapting bounding box annotations [138].

Data augmentation encompassed two distinct categories of transformations applied to the training images: pixel-level modifications and geometric alterations. The pixel-level augmentations were designed to replicate microscope parameters variability:

- **Random brightness** adjusts image overall brightness within a controlled range to simulate variability in this microscopy setting.
- **Random contrast** adjusts image overall contrast within a controlled range to simulate variability in this microscopy setting.
- **Contrast Limited Adaptive Histogram Equalization (CLAHE)** enhances contrast within small sections of the image independently, simulating varying illumination conditions.
- **Multiplicative noise** randomly scales pixel intensities, introducing subtle noise that mimics real-world imaging imperfections.
- **Motion blur** applies directional blur, simulating the effect of slight movement during image capture.

In parallel, geometric transformations were applied to simulate spatial variability in the orientation and positioning of sperm cells within the field of view. This pipeline included **horizontal and vertical flipping** and **random rotations** to introduce invariance to cell orientation.

These augmentations were applied probabilistically and according to a selective strategy, as detailed in Table 4.5. For pixel-level modifications, exactly one transformation from brightness adjustment, contrast adjustment, or CLAHE was applied per image, preventing excessive or unrealistic alterations. For geometric transformations, at least one augmentation was always applied from flipping and rotation operations, ensuring consistent spatial variability across samples.

Table 4.5: Applied data augmentations using Albumentations, with corresponding parameters and probabilities.

Augmentation Group	Augmentation Type	Parameters
Pixel-level (One of)	Random Brightness	Range: $\pm 20\%$ Probability: 0.4
	Random Contrast	Range: $\pm 20\%$ Probability: 0.4
	CLAHE	Clip limit: 2.0 Tile grid: (8,8) Probability: 0.2
Pixel-level (Applied individually)	Motion Blur	Kernel size: 5–9 px Probability: 0.3
	Multiplicative Noise	Multiplier: 0.95–1.05 Probability: 0.3
Geometric (At least one)	Horizontal Flip	Probability: 0.33
	Vertical Flip	Probability: 0.33
	Rotation	Range: $\pm 20^\circ$ Black border padding Probability: 0.33

Data augmentation was applied exclusively to the training subset of the data. Each original image was augmented twice using transformation parameters sampled randomly according to pre-defined probabilities, resulting in a threefold expansion of the training set. All transformations were applied while ensuring the corresponding bounding boxes were accurately adjusted, preserving spatial coherence between annotations and image content. This augmentation strategy significantly increased the diversity of training examples, thereby enabling the detection model to learn robust representations of sperm heads across a wide spectrum of imaging conditions.

Finally, after augmentation, all images were resized to 1440×1440 pixels by adding black padding where necessary to preserve the original aspect ratio, ensuring standardized input dimensions for model training. Examples of original frames and their augmented versions included in the final training set are shown in Figure 4.11.

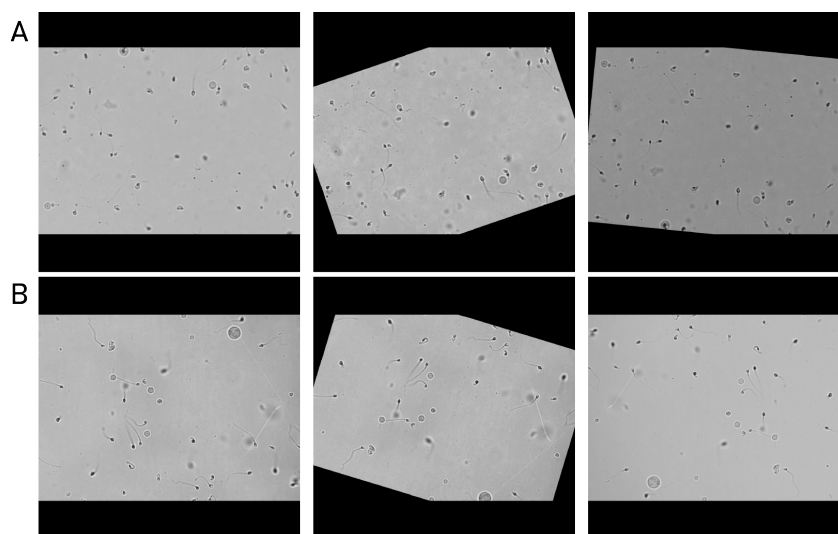


Figure 4.11: Examples of two frames (A and B) and their augmented versions as included in the final training set.

4.4.2 Sperm Detection

As mentioned in Section 3.2.2.2, YOLO has been widely adopted in sperm detection owing to its balance of speed and detection accuracy. Accordingly, this project follows the current trend by employing YOLO-based models for the task of sperm head detection.

New versions of YOLO are released regularly, each aiming to improve detection accuracy while reducing inference time. Benchmark results from the Ultralytics repository demonstrate consistent enhancements in mAP in the COCO dataset and latency, as can be seen in Figure 4.12. We selected YOLOv9, YOLOv10, and YOLOv11 for evaluation, as they are among the most recent versions available. To the best of our knowledge, these models have not yet been applied to sperm detection in published research. A detailed explanation of YOLO and the architectures of the specific models used is provided in Appendix A.

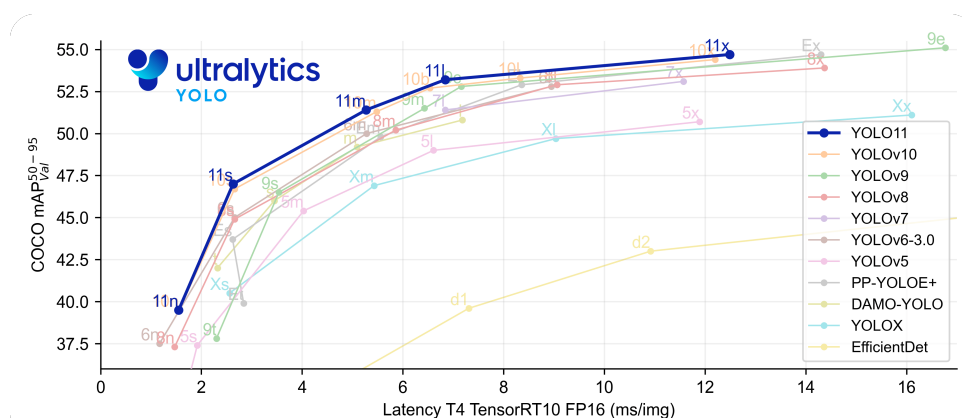


Figure 4.12: COCO mAP vs. inference latency (ms/image) for various YOLO architectures on T4 GPU with TensorRT (FP16). Source: [139].

Each YOLO version is typically released in multiple model scales, differing mainly in depth (i.e., number of layers) and width (i.e., number of channels per layer), allowing users to choose the appropriate balance between speed and accuracy for their specific computational resources and requirements. In this work, we tested all available scales for each version to evaluate their performance on our sperm detection task, each characterized by distinct model size and computational complexity, as detailed in Table 4.6.

Table 4.6: YOLO model scales selected for sperm detection, with corresponding number of parameters, number of layers, and GFLOPs (Giga Floating Point Operations Per Second), a measure of computational complexity.

Model	Parameters (M)	Layers	GFLOPs
YOLOv9t	2.01	544	7.8
YOLOv9s	7.29	544	27.4
YOLOv9m	20.16	348	77.5
YOLOv9c	25.53	358	103.7
YOLOv9e	58.15	721	192.7
YOLOv10n	2.71	223	8.4
YOLOv10s	8.07	234	24.8
YOLOv10m	16.49	288	64.0
YOLOv10b	20.45	300	98.7
YOLOv10l	25.77	364	127.2
YOLOv10x	31.66	400	171.0
YOLOv11n	2.59	181	6.4
YOLOv11s	9.43	181	21.5
YOLOv11m	20.05	231	68.2
YOLOv11l	25.31	357	87.3

4.4.2.1 Training Setup

All YOLO models were trained using the Ultralytics framework, which provides a flexible and reproducible implementation of various YOLO architectures [140]. The annotated dataset, formatted according to the YOLO standard, was used as input. The image size was fixed at 1440×1440 pixels, ensuring that the models could leverage high-resolution features, which is particularly beneficial in detecting small objects such as sperm heads.

Each model was initialized with pretrained weights derived from the COCO dataset – a widely used benchmark for general-purpose object detection – in accordance with its specific architecture. Training was conducted for 200 epochs using the framework’s default LR scheduling and optimization strategies. The AdamW optimizer was automatically selected based on the total number of training iterations, with an initial LR of 0.002, scaled to accommodate the characteristics of

the single-class dataset, and a momentum value set to 0.9. The optimizer divides model parameters into groups with different weight decay settings for better training and regularization control. The YOLO loss function combined three key components: box loss, classification loss, and distribution focal loss, with respective weights of 7.5, 0.5, and 1.5. These components are described in more detail in the following section.

The batch size was adapted for each model scale to make the most of the available GPU memory. Specifically, lighter models (e.g., n or s) were trained with larger batch sizes, while heavier models (e.g., l or x) used smaller batches to avoid out-of-memory errors. Despite these variations, a constant nominal batch size of 64 was maintained through gradient accumulation. This approach ensured that weight updates were based on a consistent number of samples, allowing for a fair and stable comparison of convergence dynamics between models of differing sizes.

4.4.2.2 Training and Performance Evaluation

During training, we monitored several key metrics over epochs to evaluate learning progress and model behavior. These included the following loss functions computed on both the training and validation sets of the data:

- **Bounding Box Regression Loss:** Measures the discrepancy between predicted and ground truth bounding box coordinates, indicating how well the model's predicted boxes align with the objects.
- **Classification Loss:** Quantifies the error in class prediction for each detected object. In our case, as only one class is used (i.e., sperm head), the model's confidence in correctly distinguishing sperm from background.
- **Distribution Focal Loss:** A specialized loss function used in YOLO to improve bounding box regression by modeling the possible positions of box edges as a probability distribution, enhancing localization precision.

Alongside these, we assessed the model's performance on unseen data using several validation measures throughout training:

- **Precision:** The proportion of correctly identified sperm heads among all predictions made:

$$Precision = \frac{TP}{TP + FP} \quad (4.5)$$

where TP is the number of true positives and FP is the number of false positives.

- **Recall:** The proportion of correctly identified sperm heads among all actual sperm heads present in the ground truth:

$$Recall = \frac{TP}{TP + FN} \quad (4.6)$$

where FN is the number of false negatives.

- **mAP50:** The average precision computed at an Intersection over Union (IoU) threshold of 0.5, offering a standard measure of object detection performance. IoU measures the overlap between predicted and ground truth boxes and is defined by:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|} \quad (4.7)$$

where B_p is the predicted bounding box and B_{gt} is the ground truth bounding box.

- **mAP50-95:** The average precision computed over multiple IoU thresholds (from 0.50 to 0.95 in steps of 0.05), providing a more detailed view of both localization and classification accuracy.

For each YOLO architecture, the best-performing checkpoint, defined as the epoch with the highest validation performance, was selected based on a weighted fitness function that combined the aforementioned measures. Upon completion of training, these selected models were evaluated on both the validation and test sets of the data to assess detection accuracy and practical applicability. At the optimal confidence threshold, we report the previously described performance measures - precision, recall, mAP50, and mAP50-95 - along with:

- **Inference Time:** The average time taken by the model to process a single image, used to assess its suitability for real-time or time-sensitive applications.

Based on these indicators, the best-performing model was selected for subsequent stages of the algorithm. To thoroughly evaluate detection performance across different confidence thresholds on the chosen model, we generated an F1 curve, which represents the F1 score as a function of the model's confidence in its detections. This analysis allows for the identification of the confidence threshold that provides the optimal balance between precision and recall:

- **F1 Score:** The harmonic mean of precision and recall, providing a balanced metric that penalizes both false positives and false negatives:

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (4.8)$$

4.4.3 Concentration Estimation

Traditionally, the Neubauer chamber, a specialized counting slide used in microscopy, provides a known observation volume that enables manual estimation of sperm concentration by counting the number of cells within a defined grid area. Building on the same principle, we estimated sperm concentration from the video data by analysing the number of detected spermatozoa per frame and converting it into a concentration value using the physical dimensions of the imaged volume. All detections considered for concentration estimation were obtained using a confidence threshold of 0.591, which corresponds to the point of maximum F1 score achieved by the best detection model, as will be further detailed in Section 5.2.1.

After detecting spermatozoa in each frame, we calculated the average number of detections per frame ($\bar{N}$):

$$\bar{N} = \frac{1}{F} \sum_{i=1}^F N_i \quad (4.9)$$

where F is the total number of frames and N_i is the number of detections in frame i .

To convert this count into a concentration value expressed in sperm per milliliter (sperm/mL), the physical dimensions of the field of view must be determined based on the pixel-to-micrometer ratio provided by the microscope's acquisition software. Let r be the pixel resolution in micrometers per pixel ($\mu\text{m}/\text{px}$). Given a video resolution of $W \times H$ pixels, the physical area of the field of view (A) is given by:

$$A = (r \times W) \times (r \times H) \quad (4.10)$$

$$= r^2 \times W \times H \quad (4.11)$$

This yields the area in square micrometers (μm^2), which is then converted to square millimeters (mm^2) by a factor of 10^{-6} . Assuming imaging is performed using a standard Neubauer chamber with a known depth d of typically 0.1 mm, the volume V in milliliters is calculated as:

$$V = A \times d \times 10^{-6} \quad (4.12)$$

Finally, the sperm concentration C in sperm/mL is obtained by dividing the average number of detections per frame by the imaged volume:

$$C = \frac{\bar{N}}{V} \quad (4.13)$$

If the sample was diluted prior to imaging, the resulting value can be multiplied by the dilution factor to obtain the final concentration (C_{final}) in sperm/mL:

$$C_{\text{final}} = C \times \text{Dilution Factor} \quad (4.14)$$

This approach allows for an automated estimation of sperm concentration based on video analysis and calibrated imaging parameters, eliminating the need for manual cell counting.

4.4.4 BoT-SORT-ReID Tracking

BoT-SORT-ReID is a state-of-the-art multi-object tracking algorithm developed to overcome limitations present in earlier ‘‘SORT-like’’ trackers such as SORT, DeepSORT, and ByteTrack [141]. It enhances ByteTrack by incorporating an improved Kalman filter, global motion compensation (GMC), and a re-identification fusion. The reidentification extension strengthens ID preservation during occlusions and complex object interactions by integrating appearance-based reidentification with advanced motion modeling. This results in fewer ID switches and more reliable tracking.

As shown in Figure 4.13, BoT-SORT-ReID outperforms widely used trackers on the MOT17 and MOT20 benchmarks, which are standard datasets commonly used to evaluate multi-object tracking algorithms in video sequences. The tracker achieves higher IDF1, MOTA, and HOTA scores, reflecting improvements in tracking accuracy, ID preservation, and overall performance.

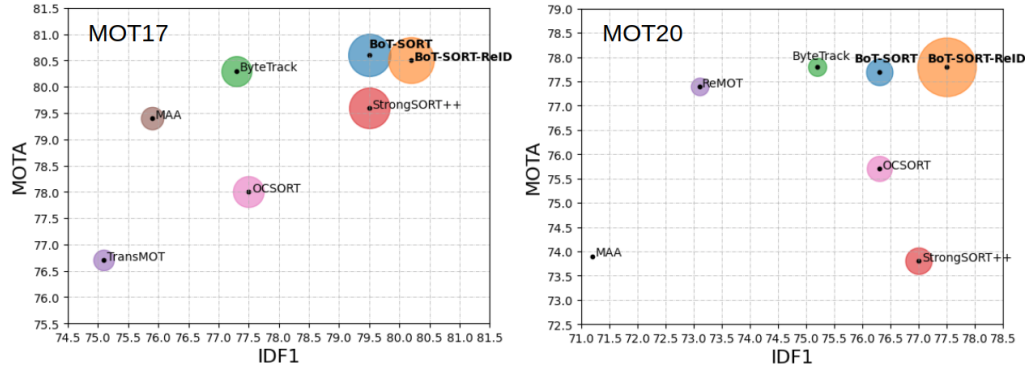


Figure 4.13: IDF1, MOTA and HOTA comparisons of commonly used trackers with BoT-SORT and BoT-SORT-ReID on the MOT17 and MOT20 test sets. Source: [141].

This algorithm was selected for the present study not only for these performance advantages but also due to its open-source availability, recent direct integration of YOLO outputs and intermediate features within the Ultralytics framework, and the flexibility it offers through customizable parameters, allowing for fine-tuning to the specific characteristics of our problem. A detailed description of the BoT-SORT-ReID workflow, which provides context for the parameter settings discussed below, is presented in Appendix B.

4.4.4.1 Parameter Configuration

BoT-SORT-ReID includes several tunable parameters that influence how object tracks are initiated, updated, and removed over time. This flexibility was essential in adapting the tracker to the specific dynamics of sperm motion. As ground-truth tracking annotations were not available, we conducted a qualitative visual analysis by manually inspecting tracking outputs across different videos while adjusting the parameters. Through this iterative process, we identified a configuration that consistently yielded more stable and accurate tracking results. Table 4.7 lists both the default parameter values of BoT-SORT-ReID and the specific configuration selected for our implementation.

Table 4.7: BoT-SORT-ReID default and selected parameters in our implementation

Parameter	Description	Default	Our Implementation
track_high_thresh (τ_{high})	Confidence threshold for associating detections with existing tracks (stage 1 association)	0.25	0.20
track_low_thresh (τ_{low})	Confidence threshold for associating unmatched tracks with detections (stage 2 association)	0.10	0.05
new_track_thresh (η)	Minimum score for initiating a new track from an unmatched detection	0.25	0.20
track_buffer (Δ_{buffer})	Maximum number of frames to keep an unmatched track before removing it	30	15
match_thresh (θ_{match})	Maximum allowed cost for matching a detection to a predicted track in the first association step	0.80	0.95
proximity_thresh (θ_{IoU})	Minimum IoU required to validate a ReID-based match	0.50	0.40
appearance_thresh (θ_{emb})	Minimum similarity score for appearance-based matching	0.80	0.85

These parameter adjustments were made to better accommodate the unique challenges posed by sperm tracking in microscopy videos. Specifically, the reduced confidence thresholds allow the tracker to retain detections even under conditions of partial occlusion with other cells or debris where usually the confidence level lowers. Decreasing the track buffer provides a more responsive handling of tracks, ensuring that tracks are removed sooner when detections are consistently missing (e.g., when leaving the frame), which can help prevent false continuation of tracks. A more tolerant matching cost threshold helps to reduce track fragmentation, allowing the tracker to maintain consistent IDs even when detections momentarily deviate in position or appearance. Furthermore, the proximity threshold was lowered to account for more erratic movement patterns and the fact that sperm may abruptly change direction or speed. Lastly, a more permissive appearance threshold was set to account for slight visual variations sometimes caused by occlusions or imaging noise.

4.4.4.2 Performance Evaluation

As mentioned, no tracking ground-truth annotations were available for the dataset, which means standard quantitative metrics such as IDF1, MOTA, or HOTA could not be computed. Instead, we conducted a qualitative evaluation on a representative subset of the videos to assess the tracker's

performance. This included visual inspection of selected sequences to verify ID consistency, track continuity, and the algorithm's robustness to occlusions and erratic motion.

In addition, we performed an indirect quantitative analysis over the 90-video dataset using two surrogate metrics:

- **Mean track length:** Measures the average duration (in frames) that the tracker is able to maintain a consistent ID for a single object.
- **Mean number of unique track IDs:** Quantifies the total number of distinct track IDs assigned per video, serving as an indicator of track fragmentation and potential false positives.

Longer average track lengths and fewer IDs per video are indicative of more stable tracking and reduced ID switches or track restarts, reflecting consistent and accurate tracking performance.

4.4.5 Kinematic Parameter Calculation

One major benefit of using a tracking-based approach is that it enables the calculation of kinematic parameters. By tracking individual sperm across frames, it becomes possible to quantify specific motion patterns and characteristics. These offer a more detailed understanding of movement of each sperm cell, going beyond simple motility classification into predefined categories.

The kinematic parameters previously described in Section 3.2.1 were computed from the tracked (x, y) coordinates in pixel space for each individual track. All distances were converted from pixels to micrometers using the known spatial calibration factor ($\mu\text{m}/\text{px}$), ensuring that the resulting measurements were expressed in biologically meaningful units and remained valid regardless of the specific pixel-to-micrometer ratio. For the average path, ALH and BCF, in this section, we present both the traditional calculation methods and our improved versions, which are later compared in Section 5.2.3.

4.4.5.1 Basic Velocity Metrics: VSL and VCL

VSL and VCL are directly computed from the tracked coordinates, following their standard definitions. VSL was computed by dividing the Euclidean distance between the first and last point of the track by the track time duration:

$$\text{VSL} = \frac{\|\mathbf{p}_{\text{end}} - \mathbf{p}_{\text{start}}\|}{T} \quad (4.15)$$

where $\mathbf{p}_{\text{start}}$ and $\mathbf{p}_{\text{end}}$ are the coordinates of the first and last frames of the track, respectively, and the time duration T is calculated as the number of frames the track persisted divided by the video frame rate.

VCL was obtained by summing the distances between all consecutive points along the trajectory and dividing by the same time duration:

$$\text{VCL} = \frac{\sum_{i=1}^{N-1} \|\mathbf{p}_{i+1} - \mathbf{p}_i\|}{T} \quad (4.16)$$

where $\mathbf{p}_i$ is the position of the object in frame i , for a track present in N frames.

4.4.5.2 Average Path Estimation and VAP

An essential aspect of calculating kinematic parameters is the estimation of the average path, which smooths the raw trajectory to reduce noise and better represent the general movement trend.

The most common method employs a moving average with a fixed window size defined by a number of frames [92]. However, this approach has a significant limitation: since the window size is based on frame count, it produces inconsistent results across videos with different frame rates. This effect is demonstrated in Figure 4.14. For this reason, the fixed-frame moving average was not used in our analysis.

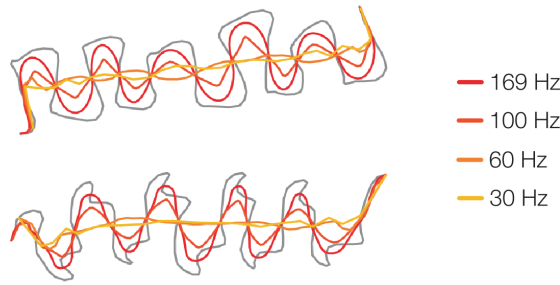


Figure 4.14: Effect of fixed-window moving average on sperm trajectories at varying frame rates. The gray line represents the curvilinear path, while each colored path corresponds to a different frame rate using an 11-point moving average. Adapted from: [92].

A natural improvement would be to normalize the window size by time rather than frame count, ensuring temporal consistency. Yet we found that this still results in variability dependent on sperm speed. For instance, a small window suitable for fast and more progressive sperm may fail to adequately smooth the trajectories of slower sperm, leaving residual noise. Conversely, a larger window tuned for slower and less progressive sperm may oversmooth fast sperm tracks, potentially masking important directional changes.

To address these issues, we implemented an adaptive smoothing approach in which the window size is dynamically adjusted based on the sperm's VCL. Instead of relying on a fixed duration or number of frames, the window is defined so that it corresponds to a consistent distance traveled by the sperm. This ensures that the smoothing process is both temporally and biologically consistent, regardless of the sperm's speed or the video frame rate.

This distance, denoted by K and expressed in micrometers, is a tunable constant used to compute the number of frames in the smoothing window according to the following relation:

$$\text{window size (in frames)} = \frac{K}{\text{VCL}} \times \text{FPS} \quad (4.17)$$

where VCL is expressed in micrometers per second and FPS is the frame rate of the video in frames per second. The value of K was set to 20, based on empirical testing across multiple

videos with varying sperm motility profiles, aiming to find a balance between noise reduction and preservation of trajectory fidelity across different conditions.

Additionally, our algorithm addresses the issue of trajectory shortening at the start and end of the sequence when applying a moving average. To handle these boundary conditions, we use symmetric edge-padding, which allows the smoothing window to be centered even at the edges of the trajectory. The smoothed average path is thus computed by applying the moving average to this padded sequence, preserving consistent treatment across all frames.

From the calculated average path, several additional kinematic parameters can be derived, including VAP. This is computed as the total length of the average path obtained divided by the elapsed time:

$$\text{VAP} = \frac{\sum_{i=1}^{N-1} \|\mathbf{a}_{i+1} - \mathbf{a}_i\|}{T} \quad (4.18)$$

where $\mathbf{a}_i$ represents the position of the sperm head at point i along the smoothed average path, present in N frames.

4.4.5.3 Beat-related Parameters: ALH and BCF

ALH and BCF rely on the identification of peak deviations and crossing points between the curvilinear and average paths. BCF is then quantified as the number of such crossings ($N_{\text{crossings}}$) divided by the elapsed time. Formally:

$$\text{BCF} = \frac{N_{\text{crossings}}}{T} \quad (4.19)$$

ALH is computed based on the maximum perpendicular distance between the curvilinear path and the average path. It is typically reported as both the mean and the maximum amplitude over the full track. For a track with P beat cycles, where each cycle corresponds to a complete lateral bending and recovery motion between two successive crossing points, let d_k be the maximum perpendicular distance in the k -th cycle. We then define:

$$\text{Mean ALH} = \frac{1}{P} \sum_{k=1}^P d_k \quad (4.20)$$

$$\text{Max ALH} = \max_{1 \leq k \leq P} d_k \quad (4.21)$$

Traditionally, ALH and BCF are estimated based on the analysis of ‘risers’, which correspond to the Euclidean distances between corresponding points on the original (curvilinear) and smoothed (average) paths [88]. The ALH peaks occur at the maxima of the riser signal, while the BCF crossing points correspond to the minima (i.e., the points where the riser signal approaches zero). In our implementation of this method, we employed the `find_peaks` function from the SciPy signal processing library to accurately identify both riser peaks and troughs. However, important limitations arise when applying this approach to low frame rate videos. Particularly for faster-moving sperm, the reduced temporal resolution means that each beat is represented by too

few data points. This sparsity in the riser signal undermines the reliability of both ALH and BCF estimates, as it becomes difficult to accurately identify true peaks and crossing points.

To address these limitations, we propose a more geometrically consistent approach that models both the curvilinear and smoothed trajectories as sequences of line segments rather than isolated points. Crossings between the two paths are identified by detecting geometric intersections between line segments of both trajectories. To make this process computationally efficient, a KD-Tree is employed to restrict intersection checks to segments that are spatially close. These intersections define the crossing points, which are then used to compute BCF.

For ALH computation, each beat cycle is defined as the interval between two successive crossings of the curvilinear trajectory with the smoothed path. Within each cycle, the deviation of the curvilinear trajectory from the smoothed path is quantified by computing the shortest distance from each trajectory point to the nearest segment of the smoothed path. To accelerate this calculation, a KD-Tree built from the midpoints of the smoothed segments is employed to identify the closest candidate segments for each trajectory point, and the exact point-to-segment distance is then computed for these candidates. The smallest of these distances is taken as the true lateral deviation of the point. The maximum deviation observed within each cycle is recorded as the ALH peak, and the mean and maximum values across all cycles is reported as the mean ALH and max ALH, respectively.

4.4.5.4 Kinematic Ratios: LIN, WOB and STR

Using the previously calculated velocities, the remaining kinematic ratios are derived as follows:

$$\text{LIN} = \frac{VSL}{VCL} \quad (4.22)$$

$$\text{WOB} = \frac{VAP}{VCL} \quad (4.23)$$

$$\text{STR} = \frac{VSL}{VAP} \quad (4.24)$$

4.4.6 Motility Classification

We classified sperm motility in each video sample using a four-class scheme. Although our dataset's manual annotations categorize sperm into only three classes, we followed the most recent WHO guidelines to allow a more granular classification for future applications of the algorithm. While descriptive criteria for classifying individual sperm motility are provided, strict thresholds for computational classification are not defined. To translate these definitions into a quantitative algorithm, we used the velocity values implied by the established guidelines, as presented in Section 2.3.1.1 (i.e., $\geq 25 \mu\text{m/s}$ for rapidly progressive and $\geq 5 \mu\text{m/s}$ for slowly progressive) and defined a set of simple rules based primarily on the VSL, which directly reflects net forward progression.

Therefore, sperm with $VSL < 5 \mu\text{m/s}$ fall below the threshold for forward progression and are either classified as non-progressive or immotile. Although distinguishing non-progressive from

immotile sperm ideally requires analysis of flagellar motion, our approach does not include explicit tail tracking. Instead, we approximated this distinction using VCL as a proxy for movement activity. Sperm with $VCL \geq 5 \mu\text{m/s}$ were classified as non-progressive, as this metric captures unproductive but active motion - such as tight circular swimming or oscillatory displacement - that VSL alone may underestimate due to its reliance on net forward distance. Conversely, sperm with $VCL < 5 \mu\text{m/s}$ were classified as immotile, reflecting minimal overall displacement, which may include small passive vibrations within the fluid or due to collisions rather than active movement.

The classification rules implemented in our algorithm are summarized in Table 4.8.

Table 4.8: Sperm motility classification criteria of proposed algorithm.

Sperm Motility Category	Criteria
Rapidly Progressive	$VSL \geq 25 \mu\text{m/s}$
Slowly Progressive	$5 \mu\text{m/s} \leq VSL < 25 \mu\text{m/s}$
Non-Progressive	$VSL < 5 \mu\text{m/s}$ and $VCL \geq 5 \mu\text{m/s}$
Immotile	$VSL < 5 \mu\text{m/s}$ and $VCL < 5 \mu\text{m/s}$

The motility composition of each video sample is then calculated by determining the percentage of tracked sperm that fall into each class, following the same general approach as manual WHO-based evaluation.

4.4.6.1 Performance Evaluation

To assess the accuracy and consistency of the algorithm's predictions, the same regression-based metrics described in Section 4.3.5 were employed: MAE and R^2 .

Considering the four-category continuous distribution predictions, evaluation metrics were computed for each individual class as well as on aggregated categories. These include total progressive motility (i.e., the sum of rapidly progressive and slowly progressive classes) and the overall output vector representing the complete distribution across all four categories. This approach allows for detailed insight into the model's ability to predict both specific motility classes and broader motility patterns, enabling effective comparison with the alternative tested approach.

4.4.7 Graphical Interface and Integration

To facilitate analysis of sperm motility from microscopy videos in the proponent laboratory, we developed a software tool that integrates core computational procedures with a user-friendly graphical interface for non-technical users. These functions, encapsulated in a dedicated backend module, include loading the trained detection model, performing video analysis (e.g., detection and tracking), estimating sperm concentration, classifying sperm motility, generating visual outputs such as plots and annotated videos, and exporting results to Excel files. This modular structure separates backend processing from the frontend interface, enhancing maintainability.

The GUI was developed using PyQt5, a Python library for building cross-platform desktop applications [142] and designed to align with the specific workflow needs of the UMIB laboratory.

The primary objective was to create an intuitive and flexible tool that supports a range of analysis tasks. The interface should allow users to load one or multiple videos for assessment and to select the desired type of analysis: sperm counting and concentration estimation and/or motility classification into the four standard motility classes. Kinematic parameter estimation should also be available as a separate analysis option, providing both individual sperm parameters and summary statistics for the entire sample, including subgroup-specific values for rapidly progressive, slowly progressive, total motile, total progressive and non-progressive spermatozoa. To support visual inspection and result validation, the system should include an option to generate tracking videos with customizable display settings - either showing sperm tracks with motility class labels or displaying tracks without classification overlays.

Based on these requirements, the GUI workflow was implemented as illustrated in Figure 4.15. The interface guides the user through input loading, analysis selection, and parameter configuration. The software then generates results according to the chosen options. Users can perform either single or batch analyses: in single analysis mode, results are displayed in a dedicated window, whereas batch analysis automatically saves results to Excel files. When kinematic parameter estimation is selected, individual sperm parameters are additionally saved in a separate Excel file. Users can also optionally generate videos with overlaid sperm tracks, which are saved in the same folder as the original video, with display settings determined by the selected video options. Importantly, the software executes only the analyses explicitly selected by the user, ensuring that only the relevant tabs are displayed and only the chosen outputs are generated, maintaining flexibility and efficiency in processing.

To enhance the visual presentation, a custom theme was applied to give the interface a more modern and cohesive look, aligned with current user interface design standards.

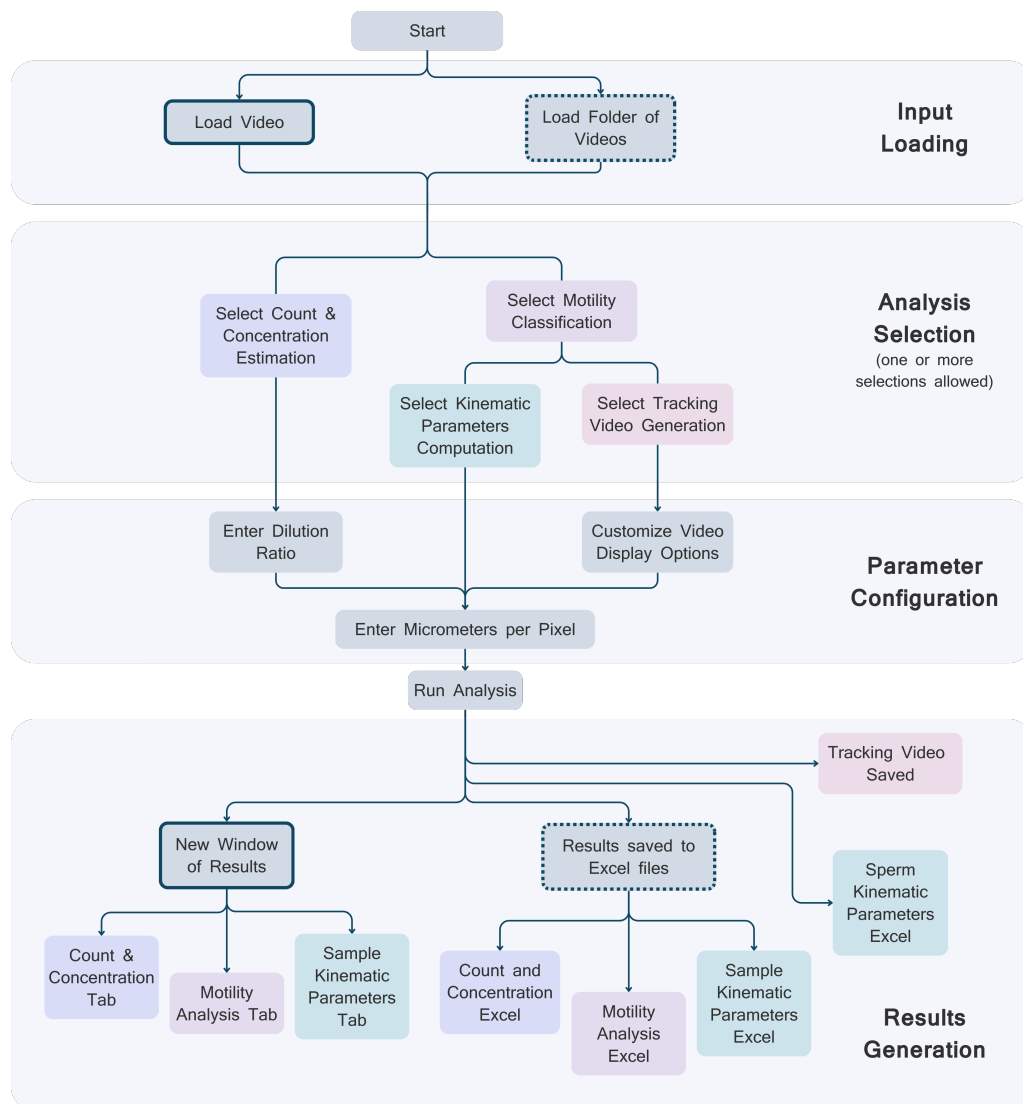


Figure 4.15: GUI workflow for the sperm analysis algorithm. The diagram illustrates the sequence of user interactions and system responses, from loading input videos, selecting analysis options and parameter configuration to result generation. The border style and color of the boxes represent different output behaviors, which vary depending on whether single or batch video analysis is performed and the combination of analysis tasks selected by the user.

Chapter 5

Results and Discussion

5.1 End-to-end Approach Results

5.1.1 Training Monitoring

As detailed in Section 4.3.4, the training process of this state-of-the-art implementation was carefully monitored across all models by analyzing the evolution of training and validation loss during cross-validation.

For the unimodal 3D CNN trained on motility prediction, the loss decreased smoothly across folds, indicating stable convergence and consistent learning, as illustrated in Figure 5.1. This indicates that the chosen training parameters provided a solid setup for the task. The narrow gap between training and validation loss suggests that overfitting was limited. While the model consistently converged relatively early in all folds, and training could have been slightly more efficient by reducing the patience parameter, the small gap indicates that the results are reliable and the training process was satisfactory.

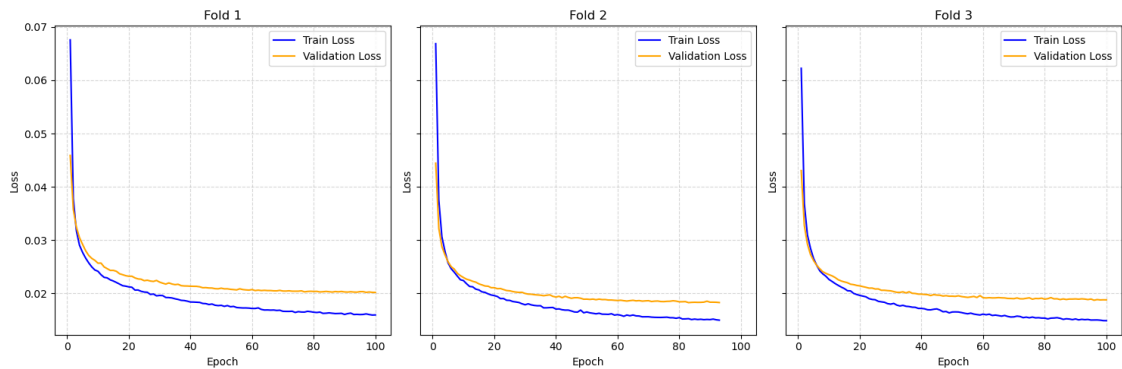


Figure 5.1: Training and validation loss across epochs for different cross-validation folds of the unimodal 3D CNN model.

The training dynamics for the unimodal 2D CNN trained on concentration prediction, displayed in Figure 5.2, exhibited a different behavior. The validation curves were significantly noisier, particularly in the early stages of training, with large fluctuations that gradually decreased as

the model learned. This instability is consistent with the nature of the task: unlike motility prediction, which is modeled as a constrained regression problem where outputs represent proportions that sum to one, concentration estimation is an unconstrained regression task, making it more sensitive to fluctuations. In addition, as previously reported, the dataset presented considerable image variability due to differences in microscope conditions, which may have further complicated this task. The overall trend still indicated progressive stabilization across epochs, with the training loss eventually plateauing and maintaining only a small gap relative to the validation loss, indicating effective learning.

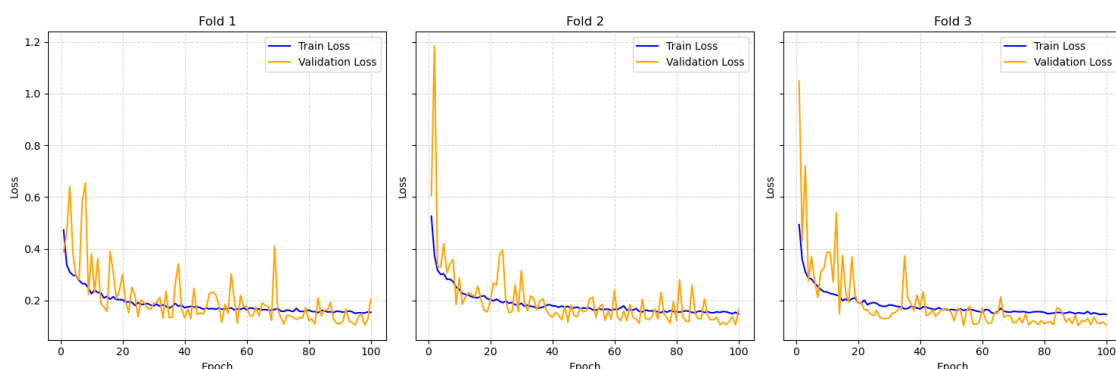


Figure 5.2: Training and validation loss across epochs for different cross folds of unimodal 2D CNN model.

Combining the two unimodal architectures, the multimodal model was trained for motility prediction, with the training and loss curves shown in Figure 5.3. The curves indicate successful convergence, suggesting that the model effectively leveraged information from both modalities. Compared to the unimodal motility predictor, the multimodal loss converged much faster, as expected, due to the use of pretrained, partially frozen feature extractors, which reduce the need to learn features from scratch, and the complementary information provided by the two modalities, which enhances learning efficiency. Moreover, the model exhibited a notably smaller gap between training and validation curves, likely reflecting the inclusion of spatially informative features extracted by the ResNet18 backbone. Seemingly, these features provided strong, discriminative representations, reducing confusion from non-sperm motion and improving generalization.

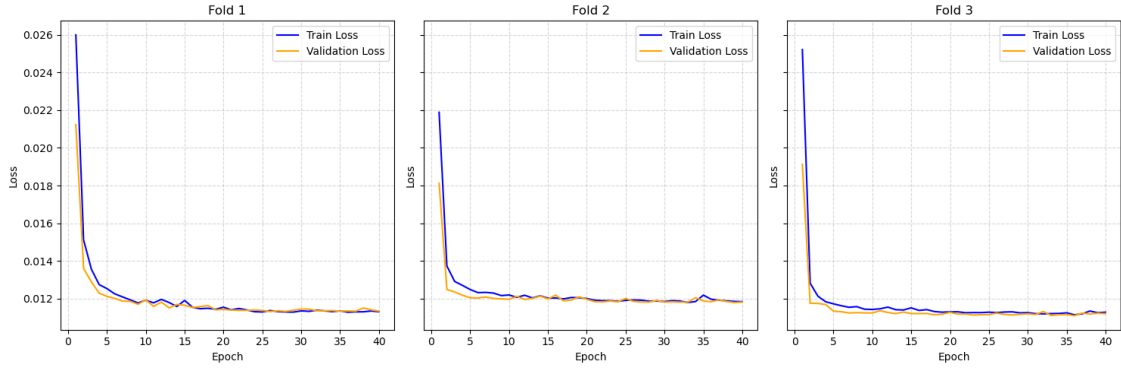


Figure 5.3: Training and validation loss across epochs for different cross folds of multimodal model for motility prediction.

5.1.2 Performance Evaluation

To assess the effectiveness of the implemented end-to-end DL approach, each model was systematically evaluated on the test set across three cross-validation folds.

For motility prediction with the unimodal 3D CNN, performance was relatively stable across folds, with average MAE values of 12.27 (immotile), 9.14 (non-progressive), and 10.72 (progressive). These results indicate a reasonable predictive ability. The higher MAE for immotile sperm is likely due to the complete reliance on optical flow in the D1 modality: immotile sperm produce minimal motion, with very weak or non-existent flow signals, which makes accurate prediction more challenging. In terms of correlation, R^2 scores reveal notable differences between classes: immotile and progressive sperm achieved moderate values (0.38 and 0.32, respectively), while non-progressive sperm showed a strongly negative score (-0.66). This discrepancy is likely related to the low variability of the non-progressive annotations; since R^2 depends on the variance of the ground truth, even small absolute errors can produce very low or negative values when the target distribution is narrow. In contrast, the greater variability observed in immotile and progressive sperm allowed the model to capture meaningful trends, resulting in more interpretable R^2 scores. Importantly, overall metrics suggest that, despite these class-specific differences, the model captures the general motility distribution with acceptable accuracy.

Table 5.1: Unimodal 3D CNN performance metrics on the test set.

Metric	Fold	Immotile	Non-progressive	Progressive	Overall
MAE ↓	Fold 1	11.94	9.10	10.69	10.58
	Fold 2	12.31	9.29	10.59	10.73
	Fold 3	12.56	9.03	10.88	10.82
	Average	12.27	9.14	10.72	10.71
R² ↑	Fold 1	0.41	-0.63	0.33	0.78
	Fold 2	0.37	-0.68	0.31	0.77
	Fold 3	0.35	-0.67	0.32	0.77
	Average	0.38	-0.66	0.32	0.78

Comparing our results with those reported in the original article, shown in Table 5.2, our MAE values are higher across all classes. This may be attributed to differences in the dataset such as the imaging modality. While the original work used phase-contrast microscopy, our dataset consists of brightfield videos, where reduced contrast and increased background noise can introduce small errors in flow computation, impacting the model’s predictive accuracy.

Table 5.2: Original article’s unimodal 3D CNN performance metrics [129].

Metric	Immotile	Non-progressive	Progressive	Overall
MAE ↓	8.92	7.41	10.16	8.83

The next step in the evaluation involved assessing the concentration prediction model, designed to capture spatial information from individual frames. Table 5.3 shows stable performance across folds, with an average MAE of 0.23 (in units of 10^6 sperm/mL), corresponding to an absolute error of approximately 230,000 sperm/mL. R^2 values were consistently very high (0.95 on average), indicating a strong correlation between predicted and ground truth concentrations. Overall, these findings confirm that the ResNet18 model reliably extracts spatial features that allow accurate differentiation of sperm cells in our dataset.

Table 5.3: Unimodal 2D CNN performance metrics on the test set.

Fold	MAE ↓	R ² ↑
Fold 1	0.21	0.97
Fold 2	0.21	0.95
Fold 3	0.25	0.93
Average	0.23	0.95

Finally, the two unimodal models were combined to form the multimodal architecture, which predicts sperm motility with improved performance, as shown in Table 5.4. Once again, metrics were stable across folds, demonstrating consistent performance on different subsets of data. Compared to the unimodal 3D CNN, the multimodal model shows improved MAE and R^2 values across all motility classes as well as overall. This improvement confirms successful combination of motion information from the 3D CNN’s D1 input with spatial features extracted from the ResNet18 D2 input, which effectively distinguishes sperm from the background. By integrating these complementary features, the multimodal model can better identify and classify sperm cells, even in cases where motion signals are weak, such as immotile sperm. This results in more accurate and reliable motility predictions.

Table 5.4: Multimodal model performance metrics on the test set.

Metric	Fold	Immotile	Non-progressive	Progressive	Overall
MAE ↓	Fold 1	9.53	7.92	8.40	8.62
	Fold 2	9.75	7.67	7.74	8.39
	Fold 3	9.39	8.54	7.44	8.46
	Average	9.56	8.04	7.86	8.49
R^2 ↑	Fold 1	0.63	-0.16	0.59	0.86
	Fold 2	0.61	-0.14	0.63	0.86
	Fold 3	0.65	-0.39	0.67	0.86
	Average	0.63	-0.23	0.63	0.86

Comparing the final multimodal motility predictions with the metrics reported in the original article shown in Table 5.5, our results are very similar, though slightly lower. This difference is likely due to the dataset variations previously discussed, such as the use of brightfield rather than phase-contrast imaging. Importantly, the metrics remain robust even though sperm concentration was not manually annotated but instead derived from our YOLO detection model. These findings indicate that the state-of-the-art architecture achieves strong performance on brightfield data and constitutes a reliable framework for motility prediction.

Table 5.5: Original article’s multimodal model performance metrics [129].

Metric	Immotile	Non-progressive	Progressive	Overall
MAE ↓	9.37	7.04	7.74	8.05

5.2 Proposed Method Results

5.2.1 Detection

All 15 YOLO models were trained under the same experimental conditions described in Section 4.4.2.1, with consistent input resolution, number of epochs, optimizer settings, and learning schedule. Training loss and performance metrics were monitored throughout the process, and the corresponding plots are presented in Appendix C.

Across most models, the training loss followed a consistent pattern: it began at a high value and decreased steadily throughout training. Some models exhibited plateaus at various points during training, particularly in the classification loss, with no significant instability observed. In contrast, the validation loss exhibited greater variability across model architectures and scale variants. In several models – notably YOLOv9m, YOLOv9c, YOLOv10n, YOLOv10s, YOLOv10m, YOLOv10l, YOLOv10x, YOLOv11s, and YOLOv11l – the validation loss began increasing around epoch 150, while in some cases this occurred as early as epoch 100, while training loss continued to decline. This divergence indicates overfitting, where the model becomes overly specialized to the training set, leading to reduced performance on unseen data. Another indication of overfitting was the decreasing validation metrics in the final epochs. These phenomenon were more pronounced in the larger models, whose increased number of parameters allows them to capture fine-grained patterns in the training data, but also makes them more prone to memorization. These training dynamics suggest that the majority of models would benefit from early stopping or more aggressive regularization strategies (e.g. increasing dropout, or weight decay) to maintain generalization during extended training. Interestingly, YOLOv9e, the largest model trained, exhibited unusual behavior: from around epoch 120, the validation metrics remained constant. We attribute this to a possible bug or instability in the training run.

Following training, the checkpoint achieving the highest validation performance was selected. By prioritizing the epoch with optimal validation metrics, the impact of overfitting is minimized, as this typically corresponds to the stage before overfitting occurs. Comparative evaluation of the selected models on the test and validation sets was performed, with results summarized in Table 5.6, including key detection metrics: precision, recall, mAP50, mAP50-95, and inference time per image.

As shown, across all 15 YOLO models evaluated, performance on the sperm head detection task was generally high. The majority of models achieved precision and recall values above 0.91 on both the validation and test sets, with varying trade-offs between these two metrics. The mAP50 also remained consistently high, frequently exceeding 0.97, which reflects robust overall localization and detection performance. By contrast, mAP50-95 values exhibited greater variability across models, highlighting differences in the ability of each architecture to produce tightly fitting bounding boxes under more stringent IoU thresholds. Overall, most models demonstrated reasonably consistent generalization from validation to test data, with only minor fluctuations in performance metrics. Nevertheless, models that showed signs of overfitting during training typ-

Table 5.6: Performance comparison of YOLO detection models on validation and test sets. Metrics include precision, recall, mAP50, mAP50-95, and inference time per image. Selected model is shown in bold.

Model	Precision $\uparrow$		Recall $\uparrow$		mAP50 $\uparrow$		mAP50-95 $\uparrow$		Inference Time (ms) $\downarrow$	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
YOLOv9t	0.946	0.912	0.918	0.941	0.980	0.973	0.771	0.767	6.4	22.4
YOLOv9s	0.955	0.936	0.908	0.917	0.980	0.970	0.774	0.762	12.1	50.0
YOLOv9m	0.938	0.942	0.921	0.915	0.978	0.970	0.768	0.759	26.1	110.0
YOLOv9c	0.938	0.924	0.935	0.928	0.979	0.962	0.767	0.759	30.4	139.0
YOLOv9e	0.948	0.938	0.917	0.912	0.978	0.971	0.759	0.756	73.7	122.7
YOLOv10n	0.935	0.919	0.925	0.933	0.978	0.975	0.779	0.768	10.0	23.4
YOLOv10s	0.943	0.928	0.920	0.937	0.980	0.975	0.782	0.767	12.2	46.3
YOLOv10m	0.941	0.937	0.917	0.919	0.978	0.972	0.769	0.759	25.0	103.3
YOLOv10b	0.939	0.930	0.925	0.916	0.979	0.971	0.768	0.752	29.1	136.4
YOLOv10l	0.934	0.937	0.924	0.921	0.977	0.971	0.767	0.756	35.5	162.3
YOLOv10x	0.948	0.921	0.914	0.927	0.975	0.966	0.769	0.755	59.3	256.5
YOLOv11n	0.940	0.934	0.926	0.923	0.981	0.975	0.769	0.766	4.6	17.8
YOLOv11s	0.928	0.932	0.937	0.921	0.978	0.972	0.769	0.763	10.5	42.1
YOLOv11m	0.932	0.915	0.933	0.948	0.979	0.973	0.771	0.761	23.6	103.0
YOLOv11l	0.933	0.940	0.929	0.928	0.981	0.973	0.757	0.750	32.1	132.3

ically experienced a slight decline in test performance relative to validation, suggesting a somewhat reduced ability to generalize. Notably, smaller-scale models delivered substantially faster inference times while maintaining competitive detection accuracy, highlighting their potential for efficient deployment.

Among all the evaluated models, YOLOv11n was selected for deployment in the final pipeline. Although other models offered marginally higher scores on individual metrics such as precision, recall or mAP50-95, YOLOv11n achieved a strong overall balance between performance and efficiency. It reached the highest mAP50 on both the validation (0.981) and test (0.975) sets, ranked among the top models for mAP50-95 (0.769 validation, 0.766 test), and maintained a balanced precision-recall profile, ensuring accurate detections with minimal missed targets. Crucially, it presented the lowest inference time of all tested models of 4.6ms on validation and 17.8ms on test, making it well-suited for scenarios that demand rapid analysis such as video processing. Additionally, its smaller size results in lower computational demands, allowing efficient deployment even on moderately equipped hardware.

The F1 score curve of the selected model, YOLOv11n, was then analyzed to determine the optimal confidence threshold for prediction, as shown in Figure 5.4. The optimal threshold was identified at a confidence level of 0.591, corresponding to a maximum F1 score of 0.93. This threshold represents the best trade-off between precision and recall, which is particularly important

in this task to simultaneously minimize false positives – such as debris incorrectly identified as sperm – and false negatives - where true sperm detections might be missed.

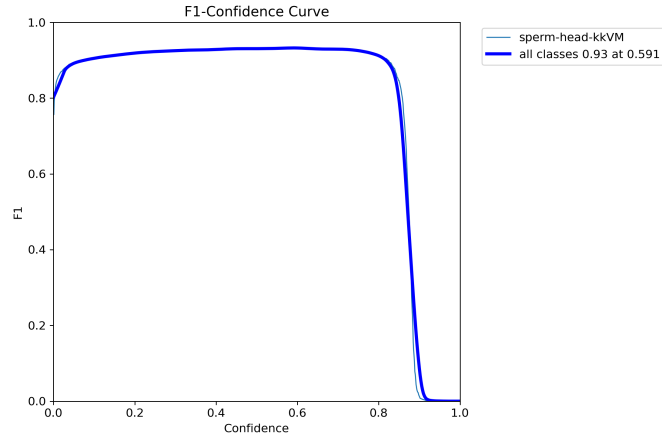


Figure 5.4: F1-Confidence curve of the selected YOLOv11n model, where F1 score is plotted against the confidence score. The optimal confidence level of 0.591 corresponds to a maximum F1 score of 0.93.

Upon visual analysis, it becomes evident that the reduced precision is primarily due to debris being incorrectly identified as sperm heads, while the reduced recall is largely associated with sperm heads that are less visually distinct by exhibiting low contrast, are more out of focus or have a different morphology. This is illustrated in Figure 5.5, which highlights correctly detected sperm heads, false positives, and missed detections in the test set.

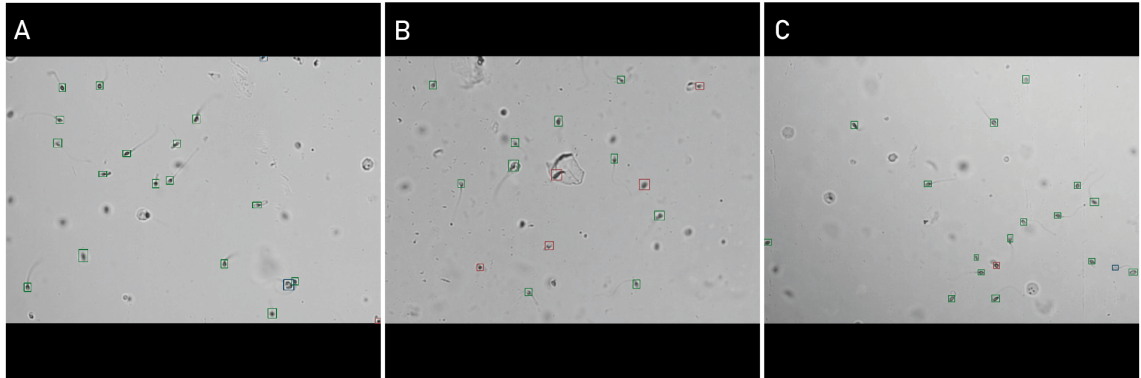


Figure 5.5: Example YOLOv11n detection results on the test set using a confidence threshold of 0.591, showing true positives (green), false positives (red), and false negatives (blue).

5.2.2 Tracking

To evaluate the impact of the parameter adjustments made to the BoT-SORT-ReID algorithm, we compared tracking performance between the default configuration and our customized parameter set. Due to the absence of ground-truth annotations, our evaluation relied primarily on visual

inspection, track length distribution, and the number of unique IDs generated per video, which serve as proxies for identity preservation and tracking stability.

Figure 5.6 illustrates representative tracking outputs using both the default and tuned BoT-SORT-ReID configurations. In the default setup (A and C), frequent ID switches are evident, even when no interactions or path crossings occur between spermatozoa. In contrast, the tuned configuration (B and D) demonstrates substantially improved track continuity and a clear reduction in identity switches.

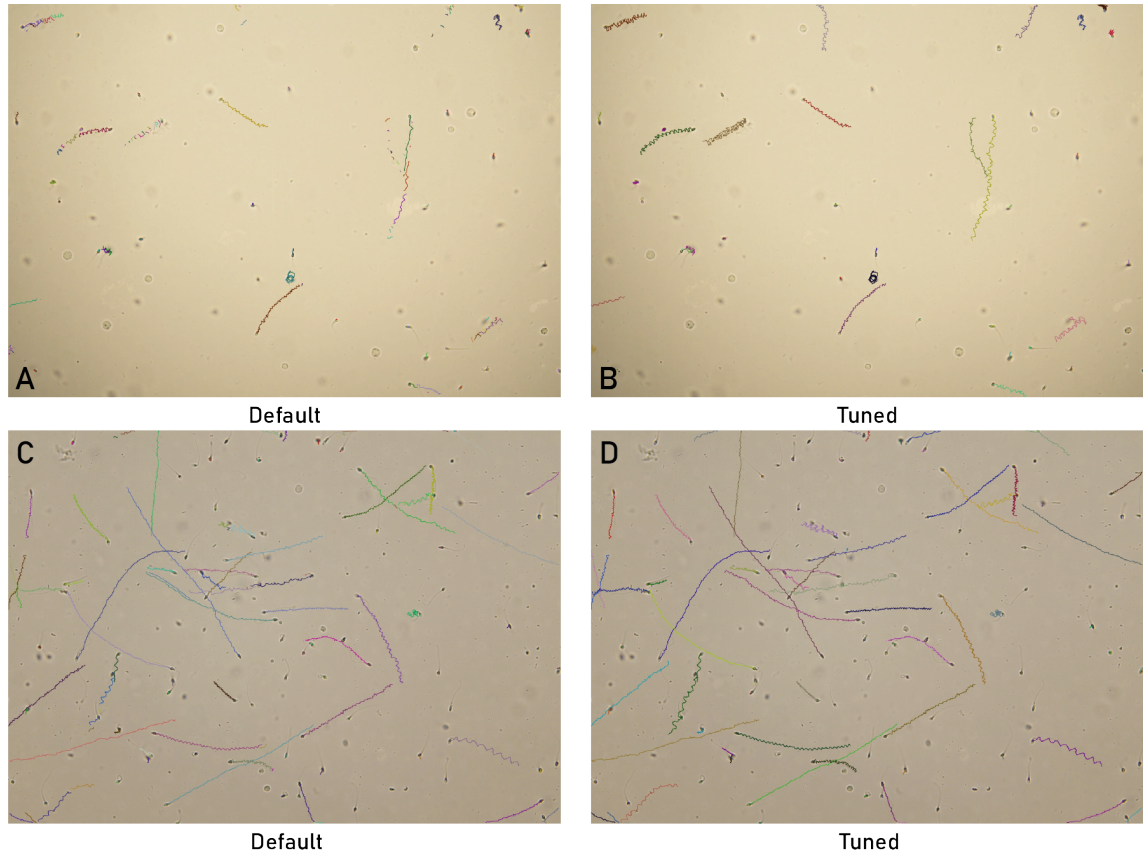


Figure 5.6: Visual comparison of tracking output between (A, C) default and (B, D) tuned BoT-SORT-ReID parameters in two different videos. Frames shown correspond to the final frame of each video. Colored lines represent individual sperm tracks, with different colors indicating different track IDs.

To quantify tracking stability and identity preservation, we analyzed the track length distribution and the number of unique IDs generated per video. As shown in Figure 5.7, the default configuration resulted in a large number of short-lived tracks, indicating that spermatozoa were more frequently lost during tracking. In contrast, the tuned parameter setup yielded significantly fewer fragmented trajectories. However, it is worth noting that even with the tuned configuration, a substantial number of short tracks still persisted, with their count remaining higher to that of the longer tracks. This highlights the potential for further enhancement. Moreover, we assessed the number of unique IDs generated by each setup. On average, the default configuration produced

approximately **235** unique IDs per video, while the tuned setup reduced this to around **134** for the same set of 90 videos. This substantial difference further indicates that the default tracker often interrupted tracks and reassigned IDs, resulting in an inflated count of unique objects.

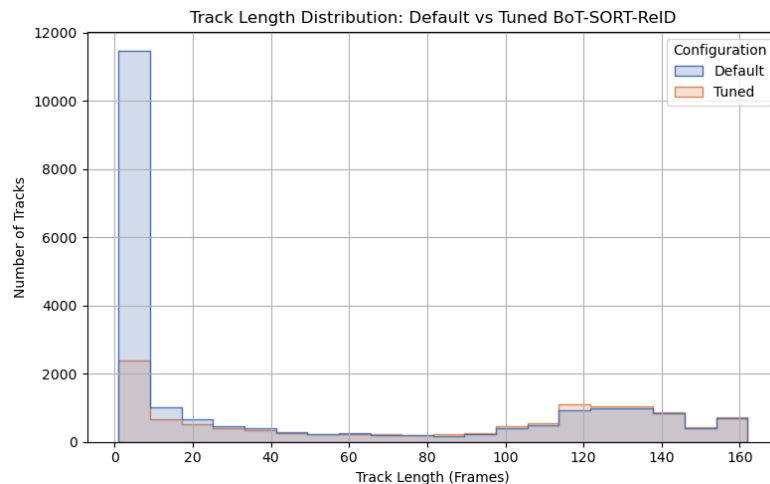


Figure 5.7: Distribution of track lengths (frames) for the default (blue) and tuned (orange) configurations of the BoT-SORT-ReID, aggregated over the 90-video dataset.

Despite overall improvements, the tuned algorithm exhibited limitations in some cases. While it successfully maintained identification through certain occlusions, it encountered difficulties in others with ID switching between crossing sperm, as illustrated in Figure 5.8. Consequently, in scenarios with frequent occlusions and overlapping trajectories, such as high-density samples, switching occurred more often. Additionally, track fragmentation tended to arise when tracking spermatozoa with erratic or abrupt motion. This issue was exacerbated by the low frame rate of the videos in the dataset, as sudden substantial changes in an unexpected direction between frames made it difficult for the tracker to maintain consistent trajectories. This effect, demonstrated in Figure 5.9, is likely due to the Kalman filter’s limited ability to accurately predict spermatozoa location under these conditions.

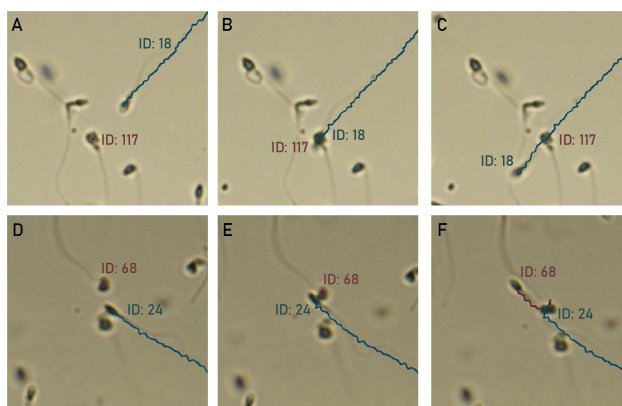


Figure 5.8: Examples of occlusion handling by the tracker. (A–C) show successful identity preservation across an occlusion event. (D–F) depict a failure case, where a crossing leads to an ID switch. Frames are sampled before, during, and after the occlusion.

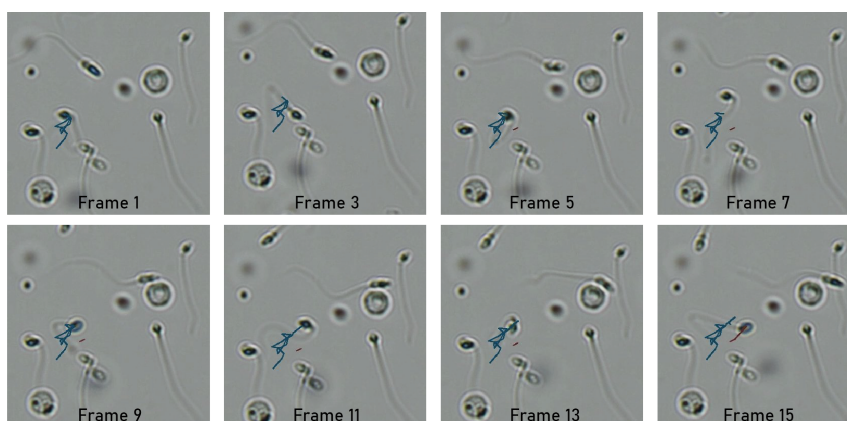


Figure 5.9: Tracking failure due to erratic cell movement, where different colors indicate separate track IDs incorrectly assigned to the same sperm cell. Frames are sampled in alternating sequence.

Rare errors also occurred due to atypical sperm morphologies – for example, spermatozoa with an unusually large midpiece were occasionally misdetected as two separate objects, leading to false positives and tracking inconsistencies, as illustrated in Figure 5.10.

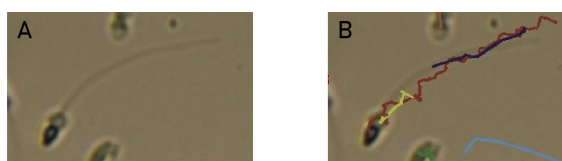


Figure 5.10: A spermatozoon with an abnormally large midpiece is incorrectly segmented into multiple tracks, each assigned a different ID, represented in distinct colors.

In light of these challenges, the final algorithm was designed to include only tracks comprising more than 10 points. This threshold helped reduce the impact of short, fragmented trajectories on the motility distribution analysis, as illustrated in Figure 5.11.

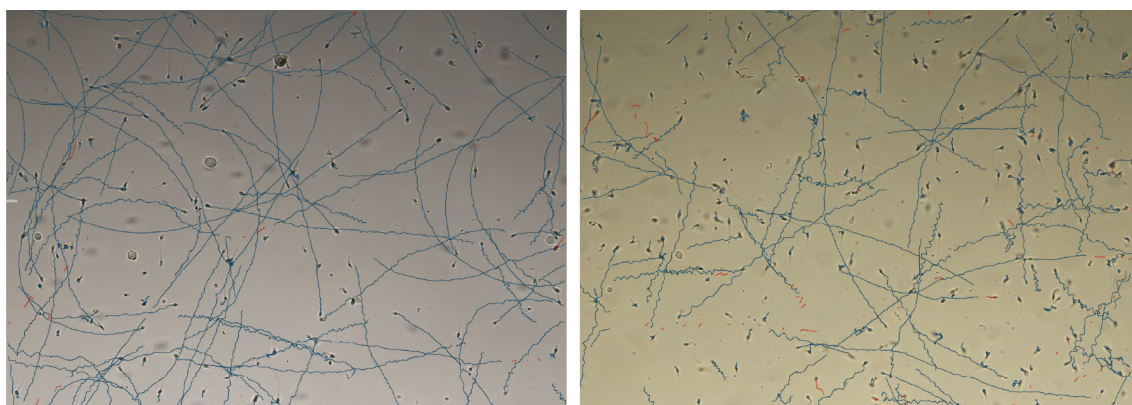


Figure 5.11: Example of track overlays on the final frame of two different videos, showcasing valid, complete tracks (in blue) and excluded short tracks (in red).

5.2.3 Kinematic Parameters

The core methodological innovations in our kinematic parameter analysis lie in two key components: the adaptive estimation of the average path and a robust geometric method for detecting crossing points and lateral peaks, enabling improved calculation of VAP, BCF and ALH.

To evaluate the performance of our adaptive smoothing approach, we first compared it with the traditional moving average method temporal windows. As illustrated in Figure 5.12, we observed that the effectiveness of a fixed window strongly depends on the sperm's velocity. For instance, a window that adequately smooths the trajectory of a faster sperm often fails to reduce noise in a slower one 5.12(C,D). Conversely, a longer window suitable for slower sperm tends to oversmooth the trajectory of faster sperm, masking meaningful directional changes 5.12(A,B). In contrast, our adaptive method adjusts the smoothing window based on the distance traveled by the sperm, providing more consistent results across different motility profiles.

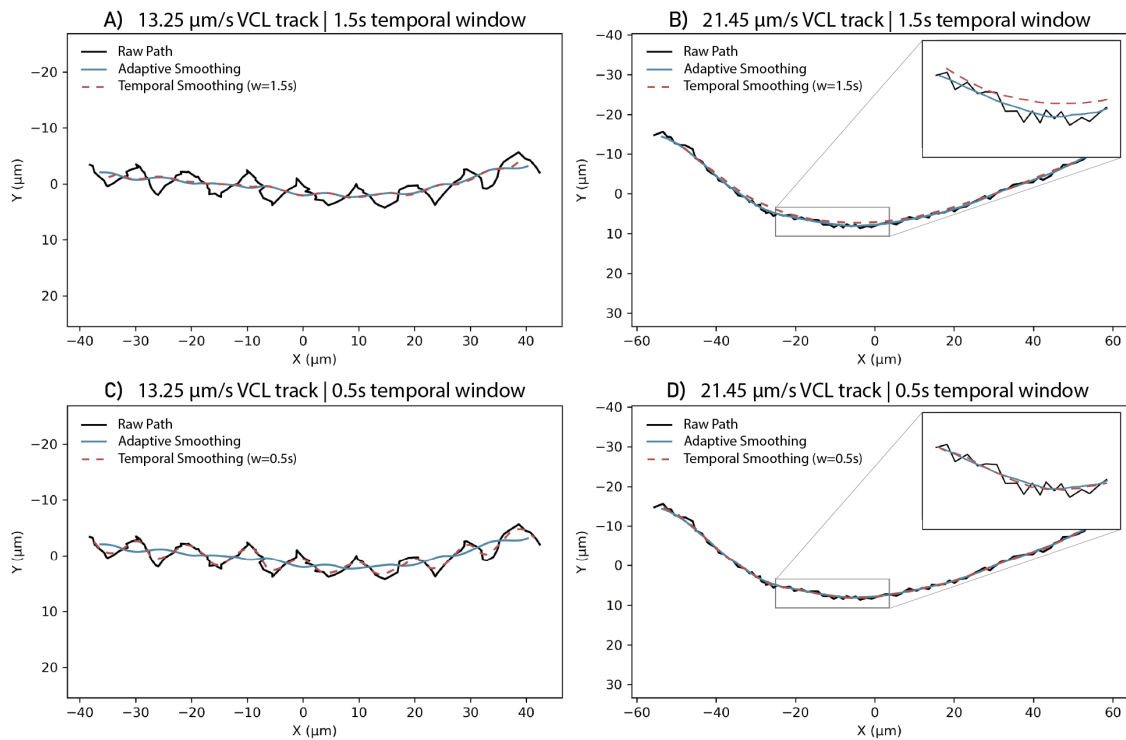


Figure 5.12: Comparison of smoothing methods for a slower (VCL = 13.25 $\mu\text{m/s}$) and a faster (VCL = 21.45 $\mu\text{m/s}$) progressive sperm. Black shows the raw curvilinear path, red dashed the traditional moving average with a fixed temporal window (1.5s in A–B, 0.5s in C–D), and blue the proposed adaptive smoothing.

We also examined the limitations of the traditional method for estimating beat-related parameters, which relies on detecting peaks and zero-crossings in the riser signal - the frame-by-frame distances between the curvilinear and average paths. As illustrated in Figure 5.13, this approach can perform reasonably well when applied to sperm tracks with sufficient temporal resolution but shows its limitations outside of these conditions. For example, in the slower sperm track, with

lower VCL, the high number of data points per beat cycle provides a dense riser signal where peaks and troughs accurately correspond to ALH peaks and crossing points between the curvilinear and average paths (Figure 5.13A,C). However, for faster sperm, especially in low frame rate videos, this traditional method becomes unreliable. As shown in Figure 5.13B,D, the reduced temporal resolution leads to sparse data points per beat cycle, and the riser signal no longer consistently represents the perpendicular distance from the curvilinear path to the average path. Instead, it appears diagonal due to temporal undersampling, causing missed or inaccurately placed peaks and crossing points along the track. Therefore, it's safe to conclude that ALH and BCF estimates derived from this method become unreliable under these conditions.

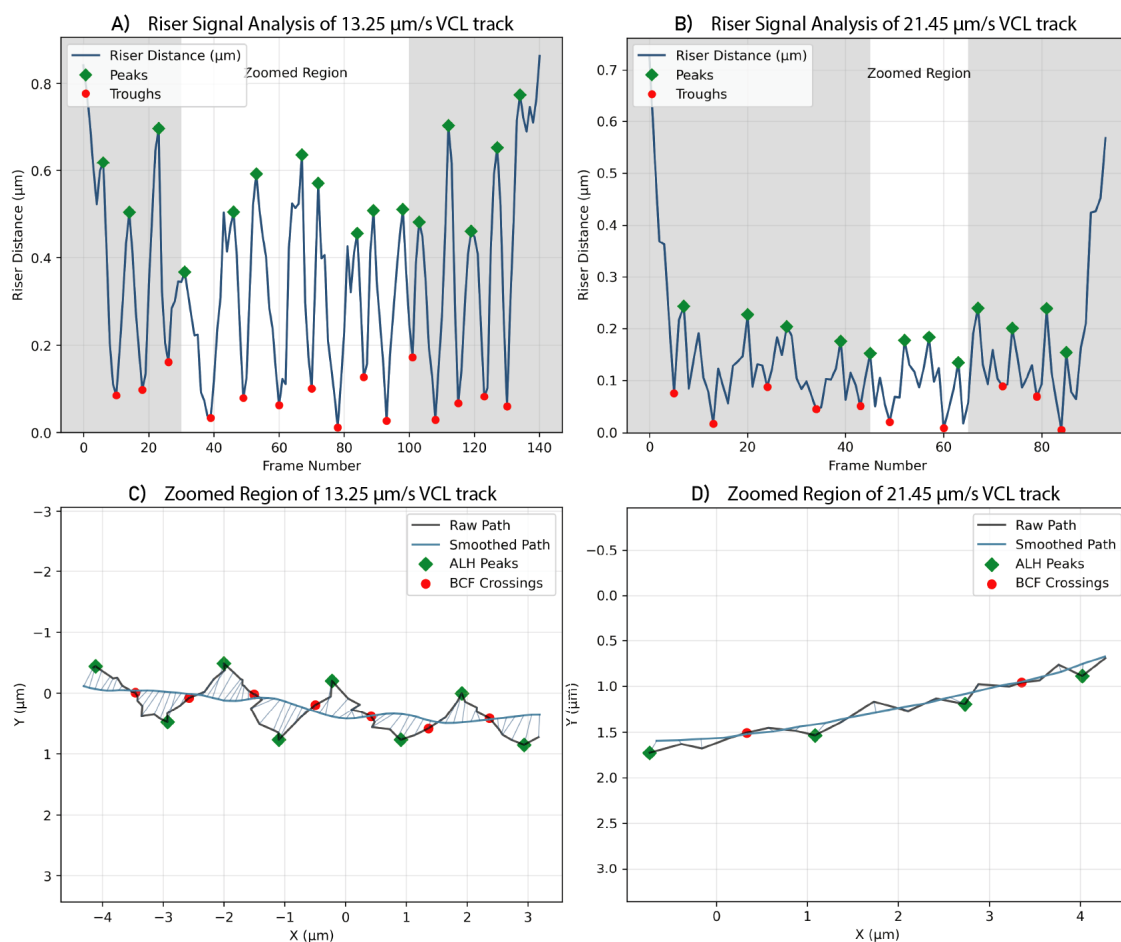


Figure 5.13: Limitations of the traditional riser-based method for ALH and BCF estimation. (A, B) Riser signal plots for a slower sperm track (VCL = 13.25 $\mu\text{m/s}$) and a faster sperm track (VCL = 21.45 $\mu\text{m/s}$), respectively, with peaks and troughs highlighted. (C, D) Corresponding zoomed-in views of the sperm tracks showing the curvilinear path (black) and the average path (blue). ALH peaks and BCF crossing points are derived from the riser signal.

To address these limitations, we applied our previously described geometric method. As shown in Figure 5.14, it demonstrated robustness across both slow and fast sperm tracks. By treating trajectories as sequences of line segments and identifying actual intersections between the curvilinear

and average paths, the method defines beat cycles more accurately and calculates ALH using true perpendicular distances. This approach remains effective even at a relatively low frame rate and with high sperm velocities.

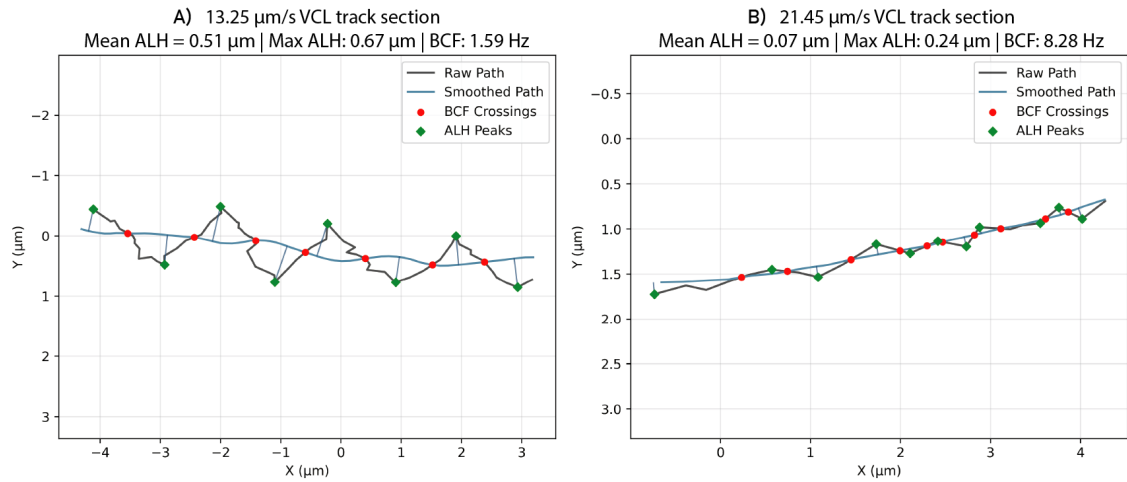


Figure 5.14: Performance of the Proposed Geometric Method for ALH and BCF Estimation. (A) Track of a slower sperm (VCL = 13.25 $\mu\text{m/s}$) with curvilinear path (black) and smoothed average path (blue). Obtained metrics: Mean ALH = 0.51 μm , Max ALH = 0.67 μm , BCF = 1.59 crossings/s. (B) Track of a faster sperm (VCL = 21.45 $\mu\text{m/s}$) with the same representation. Obtained metrics: Mean ALH = 0.07 μm , Max ALH = 0.24 μm , BCF = 8.28 crossing/s.

It is still worth noting, however, that higher frame rates would improve accuracy, particularly for fast-moving sperm. In these cases, the sparsity of points of our low frame rate dataset, may be limiting the fidelity of the reconstructed path. Nonetheless, our method mitigates these limitations and enables the extraction of meaningful information from these tracks.

Overall, the calculated kinematic parameters with the developed approach offered valuable insights into sperm motility patterns, allowing us to distinguish between different track types. Figure 5.15 illustrates this point with four examples of tracks and their corresponding parameters.

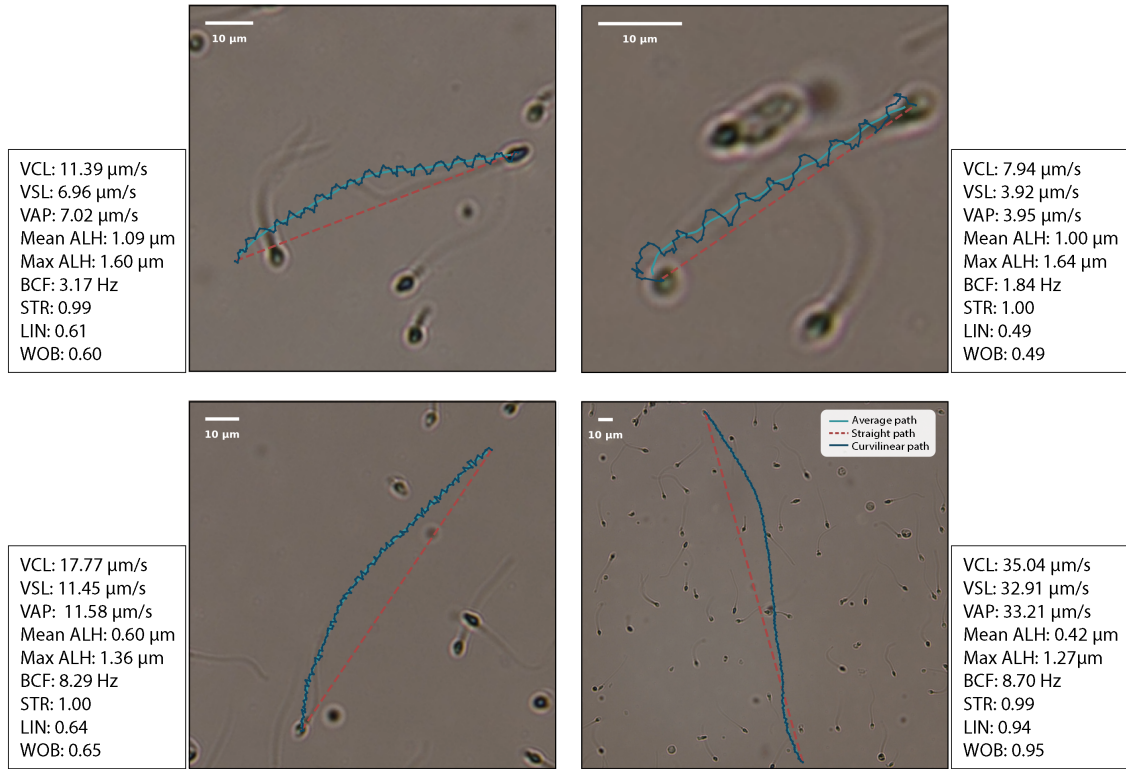


Figure 5.15: Four examples of sperm tracks, including curvilinear (dark blue), straight (red) and average paths (light blue), and their calculated kinematic parameters obtained using the developed algorithm.

5.2.4 Motility Classification

Once kinematic parameters were extracted, each sperm was classified into one of the motility categories, and the proportions of sperm in each class were computed. These predicted distributions were then compared with the annotated ground truth across the 90-video dataset. Table 5.7 summarizes the results for each class and overall.

Table 5.7: Evaluation metrics for predicted versus ground-truth motility distributions across 90 videos.

Metric	Immotile	Non-progressive	Progressive	Overall
MAE ↓	11.71	6.46	7.21	8.46
R ² ↑	0.39	0.02	0.60	0.87

The MAE varied across classes, being highest for immotile sperm (11.71), intermediate for progressive sperm (7.21), and lowest for non-progressive sperm (6.46). This indicates that, in absolute terms, predictions for immotile sperm deviate more from the ground truth than for the other classes, while predictions for non-progressive sperm are relatively closer to the true values. Similarly, R² values, which reflect how well the predicted distributions follow the trends in the ground

truth, also varied across classes. Progressive sperm exhibited the highest correlation ($R^2=0.60$), immotile sperm were moderately correlated ($R^2=0.39$), and non-progressive sperm showed very low correlation ($R^2=0.02$). These results suggest that the tracking is able to capture the movement of progressively motile sperm relatively well, while estimation of the distribution of immotile and non-progressive sperm is less consistent. This discrepancy can be explained by the method's reliance on head trajectory features: cells with minimal net displacement but clear tail movement, which belong to the non-progressive category, may be misclassified as immotile. Moreover, for the non-progressive class, this especially low correlation is also consistent with the previously noted sensitivity of R^2 for classes with narrow distributions. Despite some variations across classes, the proposed method captures the overall motility distribution effectively. Despite some differences between classes, the method effectively captures the overall sperm motility patterns with very strong performance, achieving an overall MAE of 8.46 and R^2 of 0.87.

For a more detailed analysis, Figure 5.16 presents the predicted versus ground-truth motility distributions. A consistent trend is observed in which the algorithm overestimates the proportion of immotile sperm while underestimating non-progressive sperm, consistent with the expected misclassification pattern arising from reliance on sperm head tracking. Despite these challenges, the regression plot indicates that most predictions cluster close to the line of perfect agreement, with a few outliers.

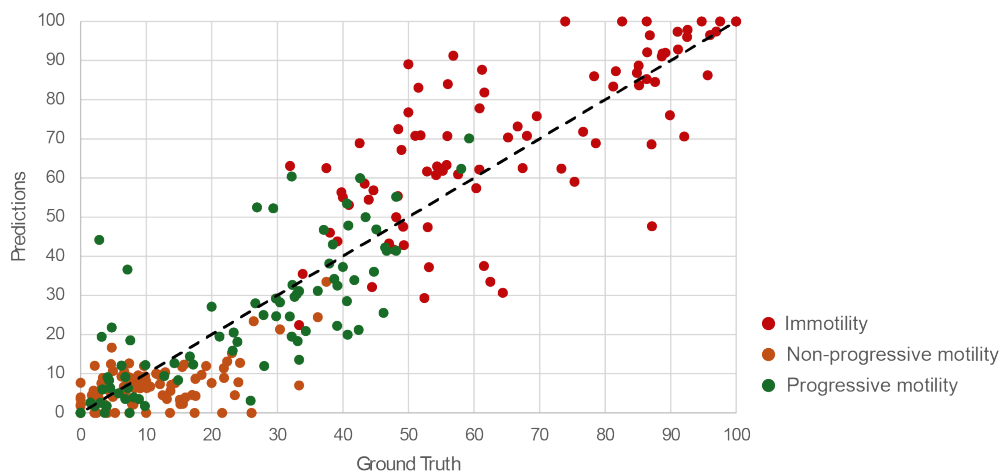


Figure 5.16: Predicted versus ground-truth motility distributions across 90 videos.

5.2.5 GUI Functionality

The developed GUI enables a simple and organized integration of all the proposed algorithm features, providing an accessible platform for laboratory users to perform sperm video analysis without requiring programming expertise. This section presents an overview of the interface design and its main functionalities, illustrating how users can interact with the system to carry out sperm counting and motility evaluation.

Upon launching the program, the main window, shown in Figure 5.17, is displayed, enabling users to select between single video or batch analysis. Users can load either an individual video file or a folder containing multiple videos, with the selected input visibly indicated in the interface. Subsequently, users choose the desired analysis type: sperm counting and concentration estimation, motility classification, or both. If motility analysis is selected, additional options become available for kinematic parameter estimation and generation of a tracking video with overlaying sperm trajectories. The tracking video includes customizable display settings, allowing users to toggle the visibility of sperm identification numbers, motility class labels adjacent to each spermatozoon, and color-coded tracks corresponding to motility types. Next, user can input calibration parameters directly within the interface, including the micrometers-per-pixel ratio, which should correspond to the microscope camera's settings. When counting and concentration estimation is enabled, the dilution factor must also be specified to ensure accurate concentration calculations. Analysis is initiated by clicking the “Run Analysis” button, upon which a progress bar indicates execution status.

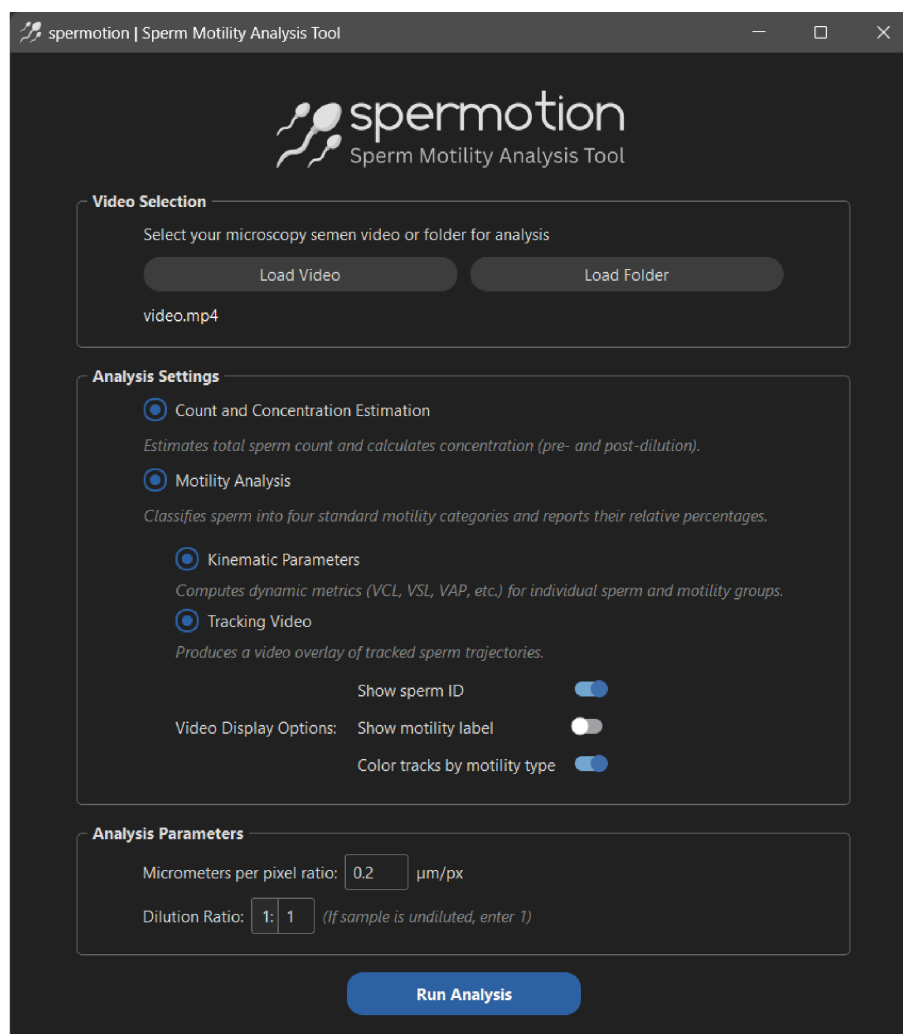


Figure 5.17: Sperm analysis tool main window.

Once a single video analysis is completed, results are presented in a new window, within dedicated tabs corresponding to the selected analysis types. The ‘Counting Results’ tab (Figure 5.18) displays statistics such as the total number of frames analyzed, average sperm count per frame, and the distribution of sperm counts across frames. A histogram visualizes the frequency distribution of sperm counts per frame. Concentration calculations are summarized, including both undiluted and even diluted concentrations when applicable.

The ‘Motility Results’ tab (Figure 5.18B) provides detailed motility statistics, including the number and concentration of spermatozoa within each of the four standard motility classes, as well as aggregated totals for total motile, total progressive, and overall sperm counts. If tracking video generation was requested, the directory path for the saved annotated video is displayed, typically bearing a filename based on the original video name, modified to indicate the tracking results.

The ‘Kinematic Parameters’ tab (Figure 5.18C) presents sample-level kinematic metrics. Users can filter results by motility groups – including motile, progressive, rapidly progressive, slowly progressive, and non-progressive – with parameter descriptions dynamically updated based on the selected subgroup. This tab also features a button to open an Excel file which was saved alongside the original video upon analysis, containing detailed per-sperm kinematic data.

In batch analysis mode, results tabs are not displayed within the interface. Instead, a pop-up notification appears once the processing is complete and output files are automatically saved as multiple Excel spreadsheets for each performed analysis, including counting and concentration data, motility results, and kinematic parameters. Additionally, for each video, individual kinematic parameter Excel files are generated, consistent with those produced in single video analyses.

Overall, the developed GUI provides an intuitive, user-friendly interface that adapts to different laboratory workflows. The interface is designed to be flexible and informative, offering clear visualization of results, customizable tracking displays, and direct access to detailed data outputs, making it a practical tool for routine laboratory use. In addition, the developed tool demonstrates efficient computational performance. With GPU acceleration, full analysis of a single 9-10 second video of the dataset was done in around 1 minute, whereas on a CPU it requires around 2 minutes and 30 seconds to 3 minutes.

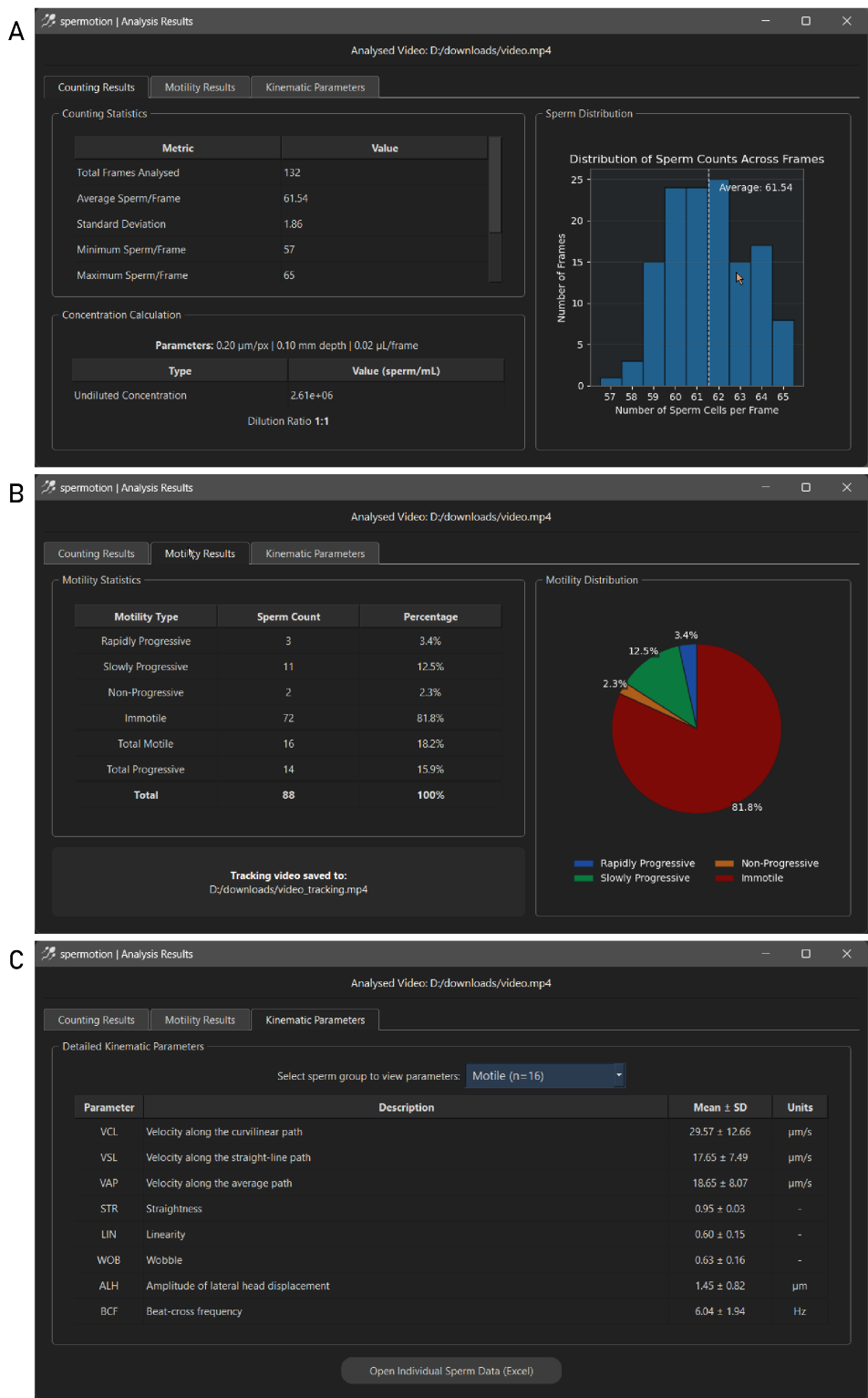


Figure 5.18: Different tabs of sperm analysis tool results window: Counting results (A), Motility Results (B), and Kinematic parameters (C).

5.3 Discussion

In this work, two different approaches to sperm motility analysis were explored: an end-to-end DL method and a tracking-based pipeline. The end-to-end implementation, adapted from a state-of-the-art technique to work in our low-frame-rate dataset, was successfully trained, as confirmed by analysis of the training dynamics, which showed stable learning and convergence across epochs. It successfully reproduced similar performance to what is reported in the literature and served as a reliable baseline. The relatively simple 3D CNN provided reasonable results on its own, but performance improved significantly when combined with spatial features extracted from a pretrained concentration predictor. The adaptations we implemented to account for differences in the datasets were proven successful, ensuring reliable training and motility predictions across all three model architectures. This approach confirmed the feasibility of applying such models to our dataset, while also highlighting their limited interpretability and dependence on large, annotated datasets. Notably, the sperm concentration used for training the 2D CNN was estimated using the developed YOLO detection model; ideally, experimentally measured annotations would be used to further improve accuracy.

Building on this baseline, we developed a tracking-based pipeline that expanded the analysis toward detailed sperm-level kinematics. Its first component was a YOLO-based sperm head detector. Several architectures were tested, most of which yielded strong results while using default hyperparameters. Some larger models showed signs of overfitting during training, but selecting the checkpoint with the highest validation performance mitigated this effect on the final models. YOLOv11n was chosen for the final pipeline due to its exceptionally fast inference and lightweight architecture, which enables efficient operation on less powerful hardware and improves accessibility, all while maintaining high detection performance. Overall, the obtained results underscore the strong potential of YOLO architectures for integration into automated sperm analysis workflows. It is important to note, however, that the detection annotations were not performed by an expert, as is common in other state-of-the-art datasets. This may have introduced some errors and could have slightly affected detection performance, therefore, refinement with expert-annotated data could further enhance the results.

BoT-SORT-ReID was adopted as the subsequent tracking component, with a key advantage being the use of appearance features extracted directly from intermediate layers of the trained YOLO detector. This strengthens ID association without requiring a separate reidentification model, making it a particularly effective and efficient approach with strong potential for integration into automated sperm analysis pipelines. Importantly, our customized hyperparameter configuration for the sperm tracking task resulted in clear improvements over the default setup. The tuned settings produced longer and more stable tracks while reducing the number of fragmented trajectories and redundant IDs, thereby enhancing the reliability of motility analysis. Nevertheless, some limitations remained in high-density samples with frequent occlusions and in cases of erratic sperm motion. As such, progress could be further supported by an annotated tracking dataset, enabling objective evaluation of the algorithm under different configurations and further refinement of its

parameters to enhance tracking performance.

Leveraging the improved tracking results, we developed novel methods for calculating kinematic parameters that remain robust across varying frame rates and sperm velocities. These approaches overcome the limitations of traditional methods, providing accurate estimates of parameters such as ALH and BCF, even in videos with lower temporal resolution or faster-moving sperm. Notwithstanding, using higher frame rate data is always advantageous and could further refine both tracking accuracy and the resulting kinematic parameter estimates, improving precision and, in turn, biological relevance.

Individual sperm were classified into four categories according to rules designed to approximate current WHO recommendations for motility assessment, while performance was evaluated using the three-tier categorization of our manually annotated dataset, yielding strong results. Table 5.8 summarizes the average performance metrics of the end-to-end baseline models and our approach. Our tracking-based pipeline exhibited improved performance compared to the simple unimodal 3D CNN across every single metric. When compared to the multimodal CNN, differences can be observed at the class-specific level: the multimodal model achieved slightly better performance for immotile sperm (MAE 9.56 vs 11.71, R^2 0.63 vs 0.39), whereas our method outperformed the multimodal approach for non-progressive (MAE 6.46 vs 8.04, R^2 0.02 vs -0.23) and progressive sperm (MAE 7.21 vs 7.86, R^2 0.60 vs 0.63). Notably, our method achieved the best overall metrics, with MAE 8.46 and R^2 0.87 demonstrating state-of-the-art performance. Some discrepancies between predicted and ground-truth distributions were observed, particularly for immotile and non-progressive sperm, due to approximating motility based solely on head movement rather than full flagellar activity, a common challenge of CASA systems [8]. Despite these limitations, our developed approach reliably captures overall motility patterns and provides indirect validation of the effectiveness of the performance of the tracking algorithm.

Table 5.8: Summary of average performance metrics for different approaches tested for sperm motility analysis.

Model		Immotile	Non-progressive	Progressive	Overall
<i>End-to-end Approach</i>					
Unimodal	MAE ↓	12.27	9.14	10.72	10.71
3D CNN	R^2 ↑	0.38	-0.66	0.32	0.78
Multimodal	MAE ↓	9.56	8.04	7.86	8.49
CNN	R^2 ↑	0.63	-0.23	0.63	0.86
<i>Tracking-based Approach</i>	MAE ↓	11.71	6.46	7.21	8.46
	R^2 ↑	0.39	0.02	0.60	0.87

Ultimately, both tested approaches in this work achieved similar overall metrics; however, the tracking-based pipeline offers clear advantages that led us to adopt it in the final tool. In both

clinical and research contexts, tracking-based methods provide significant benefits over end-to-end models. One key advantage is interpretability: by analyzing individual sperm trajectories, the results can be clearly justified, which is particularly important in clinical practice where explainability is a major concern for AI adoption [143]. This transparency allows clinicians to evaluate the reasonableness of the outcomes, increasing confidence in the system. Moreover, the calculation of kinematic parameters provides valuable insights beyond simple motility classification. Detailed information on sperm movement patterns is particularly useful in research settings, as it can reveal subtle effects of experimental conditions that categorical classifications alone might overlook. Another important consideration is robustness across diverse experimental conditions. Sperm motility can be affected differently by various factors, and analyses must remain unbiased. End-to-end DL approaches are more susceptible to such biases due to their reliance on overall patterns learned from a specific dataset. In contrast, tracking-based algorithms classify each spermatozoon individually, producing a granular and condition-independent assessment of motility distribution, which reduces the risk of biased results.

The final developed algorithm provides a practical and efficient tool for sperm analysis, demonstrating reliable performance across diverse brightfield microscopy conditions as well as different magnifications through flexible calibration with the $\mu\text{m}/\text{px}$ ratio. Feedback from the proponent laboratory highlighted the intuitive GUI and the comprehensive information it provides, which they expect will enrich their analyses in future work, for example by enabling evaluation of kinematic parameters. They especially valued the tracking video output, which not only allows for verification of results but also facilitates clear visualization of motility differences for research presentations. Another significant advantage highlighted was the considerable time saved as their manual evaluation typically requires 10-15 minutes per sample video and is prone to fatigue-induced errors and restarts. This translates to $\approx 3\times$ faster processing on CPU and $\approx 10\times$ faster on GPU when using the developed tool, and allows for researchers to focus on other activities in parallel. Beyond efficiency, the laboratory noted that the outputs deliver more detailed insights than their previous manual process. In particular, they emphasized that manual evaluation of sperm velocity is highly challenging, as distinguishing between rapidly and slowly progressive sperm is error-prone and often inconsistent. The algorithm overcomes this issue by providing objective, velocity-based classification, enabling a more refined assessment within the progressive category and ensuring a level of detail and consistency that is difficult to achieve manually. The only limitation reported aligns with the previously mentioned challenge: the difficulty in differentiating non-progressive sperm that exhibit tail movement without noticeable head displacement, as the tracking relies on head trajectories. As such, while it is not yet intended to fully replace human assessment, the developed tool serves as a reliable complementary resource for enhancing efficiency and supporting more detailed and systematic assessments.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this dissertation we successfully developed an automated tool for sperm motility analysis from standard brightfield microscopy videos, addressing the primary objective of enhancing the workflow at the UMIB research unit. The tool enables accurate detection and counting of sperm cells, allowing reliable estimation of sperm concentration. Sperm motility was classified not only into the initially targeted categories of progressive, non-progressive, and immotile, but also into rapidly and slowly progressive sperm according to recent WHO guidelines, providing a more detailed and biologically meaningful analysis. Compared to traditional manual assessment, the tool accelerates processing up to threefold on a CPU and tenfold on a GPU, supporting fast and scalable sperm evaluation. It further provides detailed kinematic metrics, improving the precision and interpretability of motility characterization, alongside an integrated user interface that ensures accessible and efficient interaction with the algorithm’s functionalities.

Additionally, we implemented a state-of-the-art DL approach, including both a unimodal 3D CNN and a multimodal CNN, that directly predict motility distribution. This demonstrated the feasibility of applying modern CNN architectures of this end-to-end sperm motility analysis paradigm to our brightfield, low-frame-rate dataset and allowed establishment of performance baselines. Our tracking-based pipeline was evaluated and shown to reliably classify sperm motility into progressive, non-progressive, and immotile categories. It outperformed the unimodal 3D CNN across all metrics, and while the multimodal CNN performed slightly better for immotile sperm, our method achieved superior results for non-progressive and progressive sperm. Overall, the developed pipeline attained the best metrics (MAE 8.46, R^2 0.87), demonstrating strong performance despite some discrepancies for immotile and non-progressive cells. Importantly, the developed framework provided not only high predictive performance but also interpretable outputs grounded in cell trajectories, making it a robust and practical solution for automated sperm motility assessment.

6.2 Future Work

Several directions could further enhance the developed sperm analysis pipeline. Future improvements could include developing a dedicated tracking dataset with expert annotations, enabling objective evaluation and additional fine-tuning of tracker parameters. Such a dataset could also support the creation of a sperm-specific tracking algorithm, designed to better handle challenges such as occlusions, erratic motion, and high-density samples, common in real experimental conditions.

An important enhancement to improve the usability of the developed tool would be extending the GUI to allow users to manually correct sperm tracks and adjust motility classifications. Providing an interactive interface where users can review and correct these tracks would not only improve overall accuracy but also increase user confidence in the final motility distribution results. By combining automated efficiency with expert oversight, this feature would make the analysis framework more robust and adaptable to a wider range of experimental conditions.

Incorporating flagellar movement tracking could further improve motility classification, particularly in distinguishing non-progressive from immotile sperm, a critical limitation in the majority of CASA systems. Beyond more precise classification, it could enable the detection of hyperactivation, a critical functional parameter that is currently missed by head-based tracking alone. This extension would make the analysis pipeline more biologically informative for research applications focused on sperm functionality under different experimental conditions.

Finally, the pipeline could be expanded to include sperm morphology assessment. Similar to motility analysis, morphology evaluation is time-intensive and highly subjective when performed manually. Automating this step would provide a more complete assessment of sperm quality, complementing concentration and kinematic measurements, offering a more comprehensive understanding of sperm health and behavior.

6.3 Final Remarks

In conclusion, this dissertation demonstrates that integrating AI-driven analysis with careful feature design and expert oversight marks a significant step forward in semen analysis. The tool developed herein shows the feasibility of creating accessible, low-cost solutions without compromising biological relevance. By simplifying workflows and eliminating the need for specialized microscopy and complex equipment, it helps bridge financial and infrastructural gaps, particularly in resource-limited settings. Beyond directly promoting more standardized and objective research practices, this work establishes a foundation for future clinical translation. This is especially vital given the high global prevalence of infertility and the growing demand for ART, where accurate and affordable automated tools could deliver the greatest impact.

References

- [1] World Health Organization, *Infertility*, May 2024. visited on 24/10/2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/infertility>.
- [2] P. N. Schlegel et al., “Diagnosis and Treatment of Infertility in Men: AUA/ASRM Guideline Part I,” *The Journal of Urology*, vol. 205, no. 1, pp. 36–43, Jan. 2021.
- [3] H. Sun, T.-T. Gong, Y.-T. Jiang, S. Zhang, Y.-H. Zhao, and Q.-J. Wu, “Global, regional, and national prevalence and disability-adjusted life-years for infertility in 195 countries and territories, 1990-2017: Results from a global burden of disease study, 2017,” *Aging*, vol. 11, no. 23, pp. 10 952–10 991, Dec. 2019.
- [4] A. Pinborg, U.-B. Wennerholm, and C. Bergh, “Long-term outcomes for children conceived by assisted reproductive technology,” *Fertility and Sterility*, vol. 120, pp. 449–456, Apr. 2023.
- [5] Research and Markets, “Assisted Reproductive Technology Global Market Opportunities and Strategies to 2033,” Tech. Rep. visited on 27/10/2024. [Online]. Available: <https://www.researchandmarkets.com/reports/5982227/assisted-reproductive-technology-global-market>.
- [6] World Health Organization, *WHO Laboratory Manual for the Examination and Processing of Human Semen*, 6th ed. Geneva, 2021.
- [7] R. P. Amann and D. F. Katz, “Andrology Lab Corner: Reflections on CASA After 25 Years,” *Journal of Andrology*, vol. 25, no. 3, pp. 317–325, May 2004.
- [8] S. T. Mortimer, G. van der Horst, and D. Mortimer, “The future of computer-aided sperm analysis,” *Asian Journal of Andrology*, vol. 17, no. 4, pp. 545–553, 2015.
- [9] S. C. Sikka and W. J. G. Hellstrom, “Current updates on laboratory techniques for the diagnosis of male reproductive failure,” *Asian Journal of Andrology*, vol. 18, no. 3, pp. 392–401, 2016.
- [10] M. Dobrovolny, J. Benes, J. Langer, O. Krejcar, and A. Selamat, “Study on Sperm-Cell Detection Using YOLOv5 Architecture with Labaled Dataset,” *Genes*, vol. 14, no. 2, p. 451, Feb. 2023.
- [11] J. Auger, “Spermatozoa and Sperm Structure,” in *Encyclopedia of Reproduction (Second Edition)*, M. K. Skinner, Ed., vol. 1, Oxford: Academic Press, 2018, pp. 62–67.
- [12] N. Lawrentschuk, G. Ptaszniak, and S. Ong, “Benign Prostate Disorders,” in *Endotext*, K. R. Feingold et al., Eds., South Dartmouth (MA): MDText.com, Inc., 2000.
- [13] L. H. J. Ricardo, “Male Accessory Glands and Sperm Function,” in *Spermatozoa*, R. Maccariello and R. Chianese, Eds., Rijeka: IntechOpen, 2018.

- [14] B. R. Hopkins, I. Sepil, and S. Wigby, "Seminal fluid," *Current Biology*, vol. 27, no. 11, R404–R405, 2017.
- [15] J. C. Perry, L. Sirot, and S. Wigby, "The seminal symphony: How to compose an ejaculate," *Trends in Ecology & Evolution*, vol. 28, no. 7, pp. 414–422, 2013.
- [16] A. S. Kumurya, D. E. Bitet, and A. Garba, "Semen Parameters: A Review," *UMYU Journal of Microbiology Research (UJMR)*, vol. 3, no. 2, pp. 72–81, Dec. 2018.
- [17] M. M. Mason, K. Schuppe, A. Weber, A. Gurayah, A. Muthigi, and R. Ramasamy, "Ejaculation: The Process and Characteristics From Start to Finish," *Current sexual health reports*, vol. 15, no. 1, pp. 1–9, Mar. 2023.
- [18] T. B. Haugen et al., "VISEM: A multimodal video dataset of human spermatozoa," in *Proceedings of the 10th ACM Multimedia Systems Conference*, ser. MMSys '19, New York, NY, USA: Association for Computing Machinery, Jun. 2019, pp. 261–266.
- [19] P. Kannan, "Ultrastructure of Human Spermatozoa," in *Male Infertility: A Clinical Approach*, K. Gunasekaran and N. Pandiyan, Eds., New Delhi: Springer India, 2017, pp. 27–39.
- [20] P. Sunanda, B. Panda, C. Dash, R. N. Padhy, and P. Routray, "An illustration of human sperm morphology and their functional ability among different group of subfertile males," *Andrology*, vol. 6, no. 5, pp. 680–689, Sep. 2018.
- [21] *Biorender - scientific Image and Illustration Software*. visited on 16/06/2025. [Online]. Available: <https://www.biorender.com/>.
- [22] W.-J. Zhu, "Preparation and observation methods can produce misleading artefacts in human sperm ultrastructural morphology," *Andrologia*, vol. 50, no. 7, e13043, Sep. 2018.
- [23] M. Ramón et al., "Understanding Sperm Heterogeneity: Biological and Practical Implications," *Reproduction in Domestic Animals*, vol. 49, no. s4, pp. 30–36, 2014.
- [24] "The Testes and Spermatozoa," in *Mammalian Sexuality: The Act of Mating and the Evolution of Reproduction*, A. F. Dixon, Ed., Cambridge: Cambridge University Press, 2021, pp. 155–186.
- [25] P. D. Temple-Smith, A. Ravichandran, and F. E. H. Nunez, "Sperm: Comparative vertebrate," in *Encyclopedia of Reproduction: Volume 6: Comparative Reproduction*, Elsevier, Jan. 2018, pp. 210–220.
- [26] N.-H. Gu, W.-L. Zhao, G.-S. Wang, and F. Sun, "Comparative analysis of mammalian sperm ultrastructure reveals relationships between sperm morphology, mitochondrial functions and motility," *Reproductive Biology and Endocrinology*, vol. 17, no. 1, p. 66, Aug. 2019.
- [27] K. Inaba and K. Shiba, "Microscopic analysis of sperm movement: Links to mechanisms and protein components," *Microscopy (Oxford, England)*, vol. 67, no. 3, pp. 144–155, Jun. 2018.
- [28] G. A. Cornwall, "Epididymis: Sperm Maturation and Motility," in *Encyclopedia of Reproduction (Second Edition)*, M. K. Skinner, Ed., Oxford: Academic Press, Jan. 2018, pp. 292–297.
- [29] J. P. Rickard, K. R. Pool, X. Druart, and S. P. de Graaf, "The fate of spermatozoa in the female reproductive tract: A comparative review," *Theriogenology*, Proceedings of the IX International Conference of Boar Semen Preservation, vol. 137, pp. 104–112, Oct. 2019.

- [30] E. A. Gaffney, H. Gad  lha, D. J. Smith, J. R. Blake, and J. C. Kirkman-Brown, "Mammalian Sperm Motility: Observation and Theory," *Annual Review of Fluid Mechanics*, vol. 43, no. Volume 43, 2011, pp. 501–528, Jan. 2011.
- [31] S. Suarez and H.-C. Ho, "Hyperactivated Motility in Sperm," *Reproduction in Domestic Animals*, vol. 38, no. 2, pp. 119–124, 2003.
- [32] "Cooperation, Conflict and Cryptic Female Choice," in *Mammalian Sexuality: The Act of Mating and the Evolution of Reproduction*, A. F. Dixson, Ed., Cambridge: Cambridge University Press, 2021, pp. 205–247.
- [33] S. S. Suarez and H. C. Ho, "Hyperactivation of mammalian sperm," *Cellular and Molecular Biology (Noisy-Le-Grand, France)*, vol. 49, no. 3, pp. 351–356, May 2003.
- [34] S. Chakraborty and S. Saha, "Understanding sperm motility mechanisms and the implication of sperm surface molecules in promoting motility," *Middle East Fertility Society Journal*, vol. 27, no. 1, p. 4, Feb. 2022.
- [35] R. A. Appell and P. R. Evans, "The effect of temperature on sperm motility and viability," *Fertility and Sterility*, vol. 28, no. 12, pp. 1329–1332, Dec. 1977.
- [36] R. Dcunha et al., "Current Insights and Latest Updates in Sperm Motility and Associated Applications in Assisted Reproduction," *Reproductive Sciences*, vol. 29, no. 1, pp. 7–25, Dec. 2020.
- [37] R. M. Mostafa, Y. S. Nasrallah, M. M. Hassan, A. F. Farrag, A. Majzoub, and A. Agarwal, "The effect of cigarette smoking on human seminal parameters, sperm chromatin structure and condensation," *Andrologia*, vol. 50, no. 3, e12910, 2018.
- [38] E. Ricci et al., "Semen quality and alcohol intake: A systematic review and meta-analysis," *Reproductive BioMedicine Online*, vol. 34, no. 1, pp. 38–47, Jan. 2017.
- [39] R. N. Clarke, S. C. Klock, A. Geoghegan, and D. E. Travassos, "Relationship between psychological stress and semen quality among in-vitro fertilization patients," *Human Reproduction*, vol. 14, no. 3, pp. 753–758, Mar. 1999.
- [40] P. Thonneau, L. Bujan, L. Multigner, and R. Mieusset, "Occupational heat exposure and male fertility: A review.," *Human Reproduction*, vol. 13, no. 8, pp. 2122–2125, Aug. 1998.
- [41] S. V. Andreychenko, A. V. Klepko, L. V. Gorban, O. A. Motryna, L. V. Grubska, and O. V. Trofimenko, "Post-Chornobyl Remote Radiation Effects On Human Sperm And Seminal Plasma Characteristics," *Experimental Oncology*, vol. 38, no. 4, pp. 245–251, Dec. 2016.
- [42] I. Gorpinchenko, O. Nikitin, O. Banyra, and A. Shulyak, "The influence of direct mobile phone radiation on sperm quality," *Central European Journal of Urology*, vol. 67, no. 01, 2014.
- [43] D. H. Williams, E. Karpman, J. C. Sander, P. E. Spiess, L. L. Pisters, and L. I. Lipshultz, "Pretreatment semen parameters in men with cancer," *Journal of Urology*, vol. 181, pp. 736–740, Feb. 2009. visited on 19/05/2020.
- [44] J. Ding et al., "Fda-approved medications that impair human spermatogenesis," *Oncotarget*, vol. 8, no. 6, pp. 10 714–10 725, 2017.
- [45] F. Gimenes et al., "Male infertility: A public health issue caused by sexually transmitted pathogens," *Nature Reviews Urology*, vol. 11, no. 12, pp. 672–687, Dec. 2014.
- [46] Y. Shang, C. Liu, D. Cui, G. Han, and S. Yi, "The Effect of Chronic Bacterial Prostatitis on Semen Quality in Adult Men: A Meta-Analysis of Case-control Studies," *Scientific Reports*, vol. 4, no. 1, p. 7233, Nov. 2014.

- [47] A. Agarwal, R. Sharma, A. Harlev, and S. C. Esteves, "Effect of varicocele on semen characteristics according to the new 2010 World Health Organization criteria: A systematic review and meta-analysis," *Asian Journal of Andrology*, vol. 18, no. 2, pp. 163–170, 2016.
- [48] M. W. Leigh et al., "Clinical and genetic aspects of primary ciliary dyskinesia/Kartagener syndrome," *Genetics in Medicine*, vol. 11, no. 7, pp. 473–487, Jul. 2009.
- [49] D. Cui et al., "Antisperm antibodies in infertile men and their effect on semen parameters: A systematic review and meta-analysis," *Clinica Chimica Acta*, vol. 444, pp. 29–36, Apr. 2015.
- [50] A. Agarwal et al., "Standardized Laboratory Procedures, Quality Control and Quality Assurance Are Key Requirements for Accurate Semen Analysis in the Evaluation of Infertile Male," *The World Journal of Men's Health*, vol. 40, no. 1, pp. 52–65, Jan. 2022.
- [51] World Health Organization, Ed., *WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction*, 4. ed. Cambridge, UK New York, NY, 1999.
- [52] World Health Organization, *WHO laboratory manual for the examination and processing of human semen*, 5th ed. Geneva, 2010.
- [53] C. Barratt, L. Björndahl, R. Menkveld, and D. Mortimer, "ESHRE special interest group for andrology basic semen analysis course: A continued focus on accuracy, quality, efficiency and clinical relevance†," *Human Reproduction*, vol. 26, no. 12, pp. 3207–3212, Dec. 2011.
- [54] L. Björndahl, "The usefulness and significance of assessing rapidly progressive spermatozoa," *Asian Journal of Andrology*, vol. 12, no. 1, pp. 33–35, Jan. 2010.
- [55] World Health Organization, *1 in 6 people globally affected by infertility*, Apr. 2023. visited on 26/10/2024. [Online]. Available: <https://www.who.int/news/item/04-04-2023-1-in-6-people-globally-affected-by-infertility>.
- [56] M. G. Hull et al., "Population study of causes, treatment, and outcome of infertility," *British Medical Journal (Clinical research ed.)*, vol. 291, no. 6510, pp. 1693–1697, Dec. 1985.
- [57] B. Huang, Z. Wang, Y. Kong, M. Jin, and L. Ma, "Global, regional and national burden of male infertility in 204 countries and territories between 1990 and 2019: An analysis of global burden of disease study," *BMC Public Health*, vol. 23, no. 1, p. 2195, Nov. 2023.
- [58] S. W. Leslie, T. L. Soon-Sutton, and M. A. Khan, "Male Infertility," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025.
- [59] *Male Infertility Market Size, Share & Trends Analysis Report By Test (DNA Fragmentation Technique, Oxidative Stress Analysis, Microscopic Examination), By Treatment, By Region, And Segment Forecasts, 2023 - 2030*. visited on 25/10/2024. [Online]. Available: <https://www.grandviewresearch.com/industry-analysis/male-infertility-market>.
- [60] M. C. Inhorn and P. Patrizio, "Infertility around the globe: New thinking on gender, reproductive technologies and global movements in the 21st century," *Human Reproduction Update*, vol. 21, no. 4, pp. 411–426, 2015.
- [61] S. Baskaran, R. Finelli, A. Agarwal, and R. Henkel, "Diagnostic value of routine semen analysis in clinical andrology," *Andrologia*, vol. 53, no. 2, e13614, Mar. 2021.
- [62] M. Sunder and S. W. Leslie, "Semen Analysis," in *StatPearls*, Treasure Island (FL), 2025.

- [63] Centers for Disease Control and Prevention (CDC), *Assisted Reproductive Technology (ART)*, Oct. 2019. visited on 27/10/2024. [Online]. Available: <https://www.cdc.gov/art/whatis.html>.
- [64] M. Jain and M. Singh, “Assisted Reproductive Technology (ART) Techniques,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025.
- [65] G. N. Allahbadia, “Intrauterine Insemination: Fundamentals Revisited,” *Journal of Obstetrics and Gynaecology of India*, vol. 67, no. 6, pp. 385–392, Dec. 2017.
- [66] f. t. E. S. o. H. R. a. E. (European IVF Monitoring Consortium (EIM) et al., “ART in Europe, 2018: Results generated from European registries by ESHRE†,” *Human Reproduction Open*, vol. 2022, no. 3, hoac022, Jan. 2022.
- [67] S. Zabak, A. Varma, S. Bansod, and M. R. Pohane, “Exploring the Complex Landscape of Delayed Childbearing: Factors, History, and Long-Term Implications,” *Cureus*, vol. 15, no. 9, e46291,
- [68] G. D. Adamson, F. Zegers-Hochschild, and S. Dyer, “Global fertility care with assisted reproductive technology,” *Fertility and Sterility*, vol. 120, no. 3 Pt 1, pp. 473–482, Sep. 2023.
- [69] ESHRE Capri Workshop Group, “Intrauterine insemination,” *Human Reproduction Update*, vol. 15, no. 3, pp. 265–277, 2009.
- [70] A. Starosta, C. E. Gordon, and M. D. Hornstein, “Predictive factors for intrauterine insemination outcomes: A review,” *Fertility Research and Practice*, vol. 6, no. 1, p. 23, Dec. 2020.
- [71] H. W. Michelmann, “Minimal criteria of sperm quality for insemination and IVF therapy,” *International Journal of Andrology*, vol. 18 Suppl 2, pp. 81–87, Dec. 1995.
- [72] R. P. Amann and D. Waberski, “Computer-assisted sperm analysis (CASA): Capabilities and potential developments,” *Theriogenology*, vol. 81, no. 1, 5–17.e3, Jan. 2014.
- [73] E. G. Ruestow, “Images and ideas: Leeuwenhoek’s perception of the spermatozoa,” *Journal of the History of Biology*, vol. 16, no. 2, pp. 185–224, Jun. 1983.
- [74] J. Puerta Suárez, S. S. du Plessis, and W. D. Cardona Maya, “Spermatozoa: A Historical Perspective,” *International Journal of Fertility & Sterility*, vol. 12, no. 3, pp. 182–190, 2018.
- [75] S. Boyers, R. Davis, and D. F. Katz, “Automated semen analysis,” *Current Problems in Obstetrics Gynecology and Fertility*, vol. 12, no. 5, pp. 165–200, Jan. 1989.
- [76] Y. T. Liu and P. K. Warne, “Computerized evaluation of sperm cell motility,” *Computers and Biomedical Research*, vol. 10, no. 2, pp. 127–138, Apr. 1977.
- [77] D. F. Katz, R. O. Davis, B. A. Delandmeter, and J. W. Overstreet, “Real-time analysis of sperm motion using automatic video image digitization,” *Computer Methods and Programs in Biomedicine*, vol. 21, no. 3, pp. 173–182, Dec. 1985.
- [78] J. F. Moruzzi, A. J. Wyrobek, B. H. Mayall, and B. L. Gledhill, “Quantification and classification of human sperm morphology by computer-assisted image analysis,” *Fertility and Sterility*, vol. 50, no. 1, pp. 142–152, Jul. 1988.
- [79] Data Bridge Market Research, *Global Computer Assisted Semen Analysis Market – Industry Trends and Forecast to 2030*, Jul. 2023. visited on 25/10/2024. [Online]. Available: <https://www.databridgemarketresearch.com/reports/global-computer-assisted-semen-analysis-market>.

- [80] J.-C. Lu, H.-Y. Zhang, Y.-A. Hu, Y.-F. Huang, and N.-Q. Lü, "A survey on the status of semen analysis in 118 laboratories in China," *Asian Journal of Andrology*, vol. 12, no. 1, pp. 104–110, Jan. 2010.
- [81] R. Finelli, K. Leisegang, S. Tumallapalli, R. Henkel, and A. Agarwal, "The validity and reliability of computer-aided semen analyzers in performing semen analysis: A systematic review," *Translational Andrology and Urology*, vol. 10, no. 7, pp. 3069–3079, Jul. 2021.
- [82] M. Yeste, S. Bonet, J. E. Rodríguez-Gil, and M. M. Rivera, "Evaluation of sperm motility with casa-mot: Which factors may influence our measurements?" *Reproduction, Fertility and Development*, vol. 30, pp. 789–789, Jan. 2018. visited on 30/03/2024.
- [83] J. C. Lu, Y. F. Huang, and N. Q. Lü, "Computer-aided sperm analysis: Past, present and future," *Andrologia*, vol. 46, no. 4, pp. 329–338, 2014.
- [84] "Guidelines on the application of CASA technology in the analysis of spermatozoa. ESHRE Andrology Special Interest Group. European Society for Human Reproduction and Embryology.," *Human Reproduction*, vol. 13, no. 1, pp. 142–145, Jan. 1998.
- [85] M. U. Daloglu and A. Ozcan, "Computational imaging of sperm locomotion," *Biology of Reproduction*, vol. 97, no. 2, pp. 182–188, Jul. 2017.
- [86] L. F. Urbano, P. Masson, M. VerMilyea, and M. Kam, "Automatic Tracking and Motility Analysis of Human Sperm in Time-Lapse Images," *IEEE Transactions on Medical Imaging*, vol. 36, no. 3, pp. 792–801, Mar. 2017.
- [87] L. Björndahl, "The 13th International Symposium on Spermatology," in *XIIIth International Symposium on Spermatology*, L. Björndahl, J. Flanagan, R. Holmberg, and U. Kvist, Eds., Cham: Springer International Publishing, 2021, pp. 3–8.
- [88] D. Mortimer and S. Mortimer, "Routine Application of CASA in Human Clinical Andrology and ART Laboratories," in Jul. 2021, pp. 183–197.
- [89] P. Hidayatullah and M. Zuhdi, "Automatic sperms counting using adaptive local threshold and ellipse detection," in *2014 International Conference on Information Technology Systems and Innovation (ICITSI)*, Nov. 2014, pp. 56–61.
- [90] S. Ottl, S. Amiriparian, M. Gerczuk, and B. W. Schuller, "motilitAI: A machine learning framework for automatic prediction of human sperm motility," *iScience*, vol. 25, no. 8, p. 104 644, Aug. 2022.
- [91] S. Boryshpolets, R. K. Kowalski, G. J. Dietrich, B. Dzyuba, and A. Ciereszko, "Different computer-assisted sperm analysis (CASA) systems highly influence sperm motility parameters," *Theriogenology*, vol. 80, no. 7, pp. 758–765, Oct. 2013.
- [92] G. Cupples, M. T. Gallagher, D. J. Smith, and J. C. Kirkman-Brown, "Heads and Tails: Requirements for Informative and Robust Computational Measures of Sperm Motility," in *XIIIth International Symposium on Spermatology*, L. Björndahl, J. Flanagan, R. Holmberg, and U. Kvist, Eds., Cham: Springer International Publishing, 2021, pp. 135–150.
- [93] D. Bompert et al., "CASA-Mot technology: How results are affected by the frame rate and counting chamber," *Reproduction, Fertility, and Development*, vol. 30, no. 6, pp. 810–819, Jun. 2018.
- [94] W. C. Tan, N. A. Mat Isa, and M. Mohamed, "Automated human sperm tracking using mean shift - collision detection and modified covariance matrix method," *Multimedia Tools and Applications*, vol. 79, no. 39, pp. 28 551–28 585, Oct. 2020.

- [95] S. A. Bini, “Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?” *The Journal of Arthroplasty*, vol. 33, no. 8, pp. 2358–2361, Aug. 2018.
- [96] S. Li, Y.-Q. Deng, Z.-L. Zhu, H.-L. Hua, and Z.-Z. Tao, “A comprehensive review on radiomics and deep learning for nasopharyngeal carcinoma imaging,” *Diagnostics*, vol. 11, p. 1523, Aug. 2021.
- [97] M. G. M. Abdolrasol et al., “Artificial Neural Networks Based Optimization Techniques: A Review,” *Electronics*, vol. 10, no. 21, p. 2689, Jan. 2021.
- [98] R. Indrakumari, T. Poongodi, and K. Singh, “Introduction to Deep Learning,” in *Advanced Deep Learning for Engineers and Scientists: A Practical Approach*, K. B. Prakash, R. Kannan, S. Alexander, and G. R. Kanagachidambaresan, Eds., Cham: Springer International Publishing, 2021, pp. 1–22.
- [99] T. Fu et al., “Optical neural networks: Progress and challenges,” *Light: Science & Applications*, vol. 13, no. 1, p. 263, Sep. 2024.
- [100] J. Ahmad, H. Farman, and Z. Jan, “Deep Learning Methods and Applications,” in *Deep Learning: Convergence to Big Data Analytics*, M. Khan, B. Jan, and H. Farman, Eds., Singapore: Springer, 2019, pp. 31–42.
- [101] J. L. Girela, D. Gil, M. Johnsson, M. J. Gomez-Torres, and J. De Juan, “Semen parameters can be predicted from environmental factors and lifestyle using artificial intelligence methods,” *Biology of Reproduction*, vol. 88, no. 4, p. 99, Apr. 2013.
- [102] R. A. Ghayda et al., “Artificial Intelligence in Andrology: From Semen Analysis to Image Diagnostics,” *The World Journal of Men’s Health*, vol. 42, no. 1, pp. 39–61, Jan. 2024.
- [103] M. K. Panner Selvam, A. K. Moharana, S. Baskaran, R. Finelli, M. C. Hudnall, and S. C. Sikka, “Current Updates on Involvement of Artificial Intelligence and Machine Learning in Semen Analysis,” *Medicina*, vol. 60, no. 2, p. 279, Feb. 2024.
- [104] M. S. Nissen, O. Krause, K. Almstrup, S. Kjærulff, T. T. Nielsen, and M. Nielsen, *Convolutional neural networks for segmentation and object detection of human semen*, Apr. 2017.
- [105] M. r. Mohammadi, M. Rahimzadeh, and A. Attar, *Sperm Detection and Tracking in Phase-Contrast Microscopy Image Sequences using Deep Learning and Modified CSR-DCF*, Apr. 2020.
- [106] V. Valiuškaitė, V. Raudonis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, “Deep Learning Based Evaluation of Spermatozoid Motility for Artificial Insemination,” *Sensors (Basel, Switzerland)*, vol. 21, no. 1, p. 72, Dec. 2020.
- [107] L. T. Ramos and A. D. Sappa, *A decade of you only look once (yolo) for object detection: A review*, 2025.
- [108] P. Hidayatullah et al., “DeepSperm: A robust and real-time bull sperm-cell detection in densely populated semen videos,” *Computer Methods and Programs in Biomedicine*, vol. 209, p. 106302, Sep. 2021.
- [109] M. Kosela, J. Aszyk, M. Jarek, J. Klimek, and T. Prokop, “Tracking of Spermatozoa by YOLOv5 detection and StrongSORT with OSNet tracker,” *MediaEval’22: Multimedia Evaluation Workshop*, Jan. 2023.
- [110] T.-L. Huynh et al., “Tail-Aware Sperm Analysis for Transparent Tracking of Spermatozoa,” *MediaEval’22: Multimedia Evaluation Workshop*, Jan. 2023.

- [111] B.-T. Nguyen, V.-L. Nguyen, and M.-T. Tran, “Transparent Tracking of Spermatozoa with YOLOv8,” *MediaEval’23: Multimedia Evaluation Workshop*, Feb. 2024.
- [112] T.-H. Nguyen-Mau et al., “Optimizing Sperm Detection and Tracking in Fluids with Equalize Class Representation Augmentation,” *MediaEval’23: Multimedia Evaluation Workshop*, Feb. 2024.
- [113] M. Osaïd, A. Samad, O. Qureshi, M. A. Tahir, and M. N. Durrani, “Tracking and Prediction Of Human Spermatozoa Motility Using YOLOv8n with Greedy Shape Geometry Technique,” *MediaEval’23: Multimedia Evaluation Workshop*, Feb. 2024.
- [114] Z. Zhang, B. Qi, S. Ou, and C. Shi, “Real-Time Sperm Detection Using Lightweight YOLOv5,” Dec. 2022, pp. 1829–1834.
- [115] R. Zhu et al., “YOLOv5s-SA: Light-Weighted and Improved YOLOv5s for Sperm Detection,” *Diagnostics*, vol. 13, no. 6, p. 1100, Jan. 2023.
- [116] M. Nawae, P. Maneelert, C. Choksuchat, T. Phairatana, and J. Jaruenpunyasak, “A Comparative Study of YOLO Models for Sperm and Impurity Detection Based on Proposed Augmentation in Small Dataset,” in *2023 15th International Conference on Information Technology and Electrical Engineering (ICITEE)*, ISSN: 2766-0419, Oct. 2023, pp. 305–310.
- [117] W. Jiao, Y. Xie, Y. Li, and J. Qi, “Sperm Recognition and Viability Classification Based on Artificial Intelligence,” in *Cognitive Systems and Information Processing*, F. Sun, Q. Meng, Z. Fu, and B. Fang, Eds., Singapore: Springer Nature, 2024, pp. 404–418.
- [118] C. Zhang, Y. Zhang, Z. Chang, and C. Li, “Sperm YOLOv8E-TrackEVD: A Novel Approach for Sperm Detection and Tracking,” *Sensors*, vol. 24, no. 11, p. 3493, Jan. 2024.
- [119] M. Yuzkat, H. O. İlhan, and N. Aydin, “Detection of sperm cells by single-stage and two-stage deep object detectors,” *Biomedical Signal Processing and Control*, vol. 83, p. 104 630, May 2023.
- [120] D. Goss et al., *AI facilitated sperm detection in azoospermic samples for use in ICSI*, Oct. 2023.
- [121] C. Wang, J.-w. Choi, L. F. Urbano, P. Masson, M. VerMilyea, and M. Kam, “Tracking of Human Sperm in Time-Lapse Images,” in *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*, Jul. 2018, pp. 197–201.
- [122] S. Hicks, T. B. Haugen, P. Halvorsen, and M. Riegler, “Using Deep Learning to Predict Motility and Morphology of Human Sperm,” *MediaEval’19*, Oct. 2019.
- [123] J.-M. Rosenblad, S. Hicks, H. K. Stensland, T. B. Haugen, P. Halvorsen, and M. Riegler, “Using 2D and 3D Convolutional Neural Networks to Predict Semen Quality,” *MediaEval’19*, Oct. 2019.
- [124] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, and T. B. Haugen, *Stacked dense optical flows and dropout layers to predict sperm motility and morphology*, Nov. 2019.
- [125] S. A. Hicks et al., “Machine Learning-Based Analysis of Sperm Videos and Participant Data for Male Fertility Prediction,” *Scientific Reports*, vol. 9, no. 1, p. 16 770, Nov. 2019.
- [126] T. B. Haugen, O. Witczak, S. A. Hicks, L. Björndahl, J. M. Andersen, and M. A. Riegler, “Sperm motility assessed by deep convolutional neural networks into WHO categories,” *Scientific Reports*, vol. 13, no. 1, p. 14 777, Sep. 2023, Publisher: Nature Publishing Group.
- [127] Priyansi, B. Bhattacharjee, and J. Rahim, *Predicting Semen Motility using three-dimensional Convolutional Neural Networks*, Jan. 2021.

- [128] T. Fujii, H. Nakagawa, T. Takeshima, Y. Yumura, and T. Hamagami, “Automated Sperm Assessment Framework and Neural Network Specialized for Sperm Video Recognition,” in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2024, pp. 7660–7669.
- [129] V. H. Goh, M. A. Mansor, M. A. As’ari, and L. H. Ismail, “Multimodal Convolutional Neural Networks for Sperm Motility and Concentration Predictions,” *Malaysian Journal of Fundamental and Applied Sciences*, vol. 20, no. 2, pp. 347–359, Apr. 2024.
- [130] V. Thambawita, P. Halvorsen, H. Hammer, M. Riegler, and T. B. Haugen, *Extracting temporal features into a spatial domain using autoencoders for sperm video analysis*, Nov. 2019.
- [131] *Simula Datasets - VISEM QC*. visited on 18/06/2025. [Online]. Available: <https://datasets.simula.no/visem-qc/>.
- [132] *Sperm Morphology Image Data Set (SMIDS)*. visited on 19/06/2025. [Online]. Available: <https://www.kaggle.com/datasets/orvile/sperm-morphology-image-data-set-smids>.
- [133] *HuSHeM (Sperm Morphology Analysis Dataset)*. visited on 19/06/2025. [Online]. Available: <https://www.kaggle.com/datasets/orvile/hushem-dataset>.
- [134] S. Javadi, *Soroushj/mhsma-dataset*, Jul. 2025. visited on 19/06/2025. [Online]. Available: <https://github.com/soroushj/mhsma-dataset>.
- [135] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015.
- [136] B. Dwyer, J. Nelson, T. Hansen, et al., *Roboflow (version 1.0) [software]*, <https://roboflow.com>, Computer vision, 2025.
- [137] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *Journal of Big Data*, vol. 6, no. 1, p. 60, Jul. 2019.
- [138] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and Flexible Image Augmentations,” *Information*, vol. 11, no. 2, p. 125, Feb. 2020.
- [139] Ultralytics, *Yolo11*, 2024. visited on 21/07/2025. [Online]. Available: <https://docs.ultralytics.com/pt/models/yolo11/>.
- [140] G. Jocher, J. Qiu, and A. Chaurasia, *Ultralytics YOLO*, Jan. 2023. visited on 19/06/2025. [Online]. Available: <https://github.com/ultralytics/ultralytics>.
- [141] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, *BoT-SORT: Robust Associations Multi-Pedestrian Tracking*, Jul. 2022.
- [142] R. C. Limited, *Pyqt5: Python bindings for the qt cross platform application toolkit*, PyPI, Jul. 2024. [Online]. Available: <https://pypi.org/project/PyQt5/>.
- [143] E. G. Poon, C. H. Lemak, J. C. Rojas, J. Guptill, and D. Classen, “Adoption of artificial intelligence in healthcare: Survey of health system priorities, successes, and challenges,” *Journal of the American Medical Informatics Association*, vol. 32, May 2025.
- [144] Y. Sun, Z. Sun, and W. Chen, “The evolution of object detection methods,” *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108 458, 2024.
- [145] S. A. Alhashmi and A. Al-azawi, “A review of the single-stage vs. two-stage detectors algorithm:comprehensive insights into object detection,” *International Journal of Environmental Sciences*, vol. 11, pp. 775–787, 2025.

- [146] A. Vijayakumar and S. Vairavasundaram, “Yolo-based object detection models: A review and its applications,” *Multimedia Tools and Applications*, vol. 83, no. 35, pp. 83 535–83 574, Oct. 2024.
- [147] A. Scius-Bertrand, M. Jungo, B. Wolf, A. Fischer, and M. Bui, “Transcription Alignment of Historical Vietnamese Manuscripts without Human-Annotated Learning Samples,” *Applied Sciences*, vol. 11, p. 4894, May 2021.
- [148] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, *YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information*, Feb. 2024.
- [149] A. Wang et al., *YOLOv10: Real-Time End-to-End Object Detection*, Oct. 2024.
- [150] R. Khanam and M. Hussain, *YOLOv11: An Overview of the Key Architectural Enhancements*, Oct. 2024.

Appendix A

Architectural Details of YOLOv9, v10, and v11

Traditional object detectors often use multi-stage pipelines that first propose regions of interest and then classify them [144]. In contrast, YOLO models unify these tasks by dividing the input image into a grid and predicting bounding boxes, along with their confidence scores and class probabilities, directly for each grid cell [145]. This single-stage detection approach allows for fast inference, which is very advantageous for applications like sperm detection where many objects must be located in every frame. YOLO models share a common architecture, structured in three main components: Backbone, Neck, and Dense Prediction Head, as shown in Figure A.1.

The Backbone is a deep CNN responsible for extracting hierarchical features from the input image. It consists of modular blocks, which are sequences of convolutional layers and operations, designed to progressively capture essential visual patterns such as edges, textures, shapes, and semantic regions [146]. Also composed of convolutional blocks, the Neck serves as an intermediate feature aggregator, combining information from different scales and enhancing spatial and semantic representations. It often includes specific modules to improve the model’s ability to detect objects of varying sizes [146]. Finally, the Dense Prediction Head, built on convolutional layers as well, transforms these refined features into bounding boxes, objectness scores, and class probabilities, operating on the processed features to localize and classify multiple objects in a single pass [146].

This modular structure allows targeted improvements to each component, as seen in different YOLO versions, each addressing specific detection challenges. The key innovations in the selected models are highlighted below, focusing on their impact on detection performance.

A.1 YOLOv9

YOLOv9, introduced in early 2024, builds upon prior iterations of the YOLO family addressing one of the key limitations of deep CNNs: the progressive loss of information as input data passes through multiple layers [148]. To mitigate this, YOLOv9 integrates two primary components:

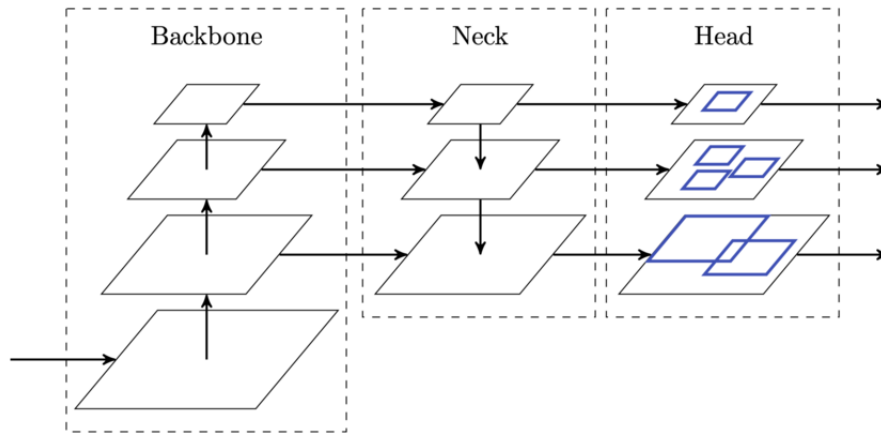


Figure A.1: Overview of the YOLO main architecture, including backbone, neck and head. Source: [147]

Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN).

PGI is a mechanism introduced at training time to enhance the reliability of gradient updates, particularly for small object detection. It includes an auxiliary, reversible branch that runs in parallel to the main network backbone. This branch consists of a small number of reversible blocks, which apply invertible transformations, thereby enabling the full reconstruction of feature maps. During backpropagation, this auxiliary path computes gradients based on raw, uncompressed features, which are then merged with those from the main branch through channel-wise attention gates. These gates assign dynamic weights to different feature channels based on their relevance. Importantly, the reversible branch is discarded at inference time, incurring no additional computational cost during deployment.

The second innovation, GELAN, is a reimagined backbone architecture that generalizes and improves upon prior designs. GELAN adopts a hierarchical fusion strategy that enhances feature aggregation across layers while reducing parameter redundancy. By employing partial transitions and multi-path aggregation, GELAN captures both shallow and deep semantic information efficiently. Moreover, the architecture incorporates adaptive gradient paths – learned skip connections that directly link early high-resolution layers with deeper layers – allowing fine-grained spatial features to propagate through the network. A stochastic depth mechanism is also used during training, whereby a subset of layers (typically 10–15%) is randomly bypassed, improving generalization and regularization [148].

Together, these enhancements allow YOLOv9 to preserve over 95% of the original input information even in very deep architectures, extending up to 200 layers [148]. These design choices enable this model to support both object detection and instance segmentation tasks across various model scales.

A.2 YOLOv10

Released in mid-2024, YOLOv10 represents a significant evolution in the YOLO series, primarily addressing the challenges of inference latency and detection accuracy in densely populated scenes [149]. One of its hallmark innovations is the removal of Non-Maximum Suppression (NMS) during inference, a post-processing step traditionally used to eliminate duplicate detections but known to incur both computational overhead and failure cases when objects overlap densely.

The core enabler of this NMS-free design is a dual assignment training strategy, which employs two parallel prediction heads: The One-to-Many head generates multiple predictions per object during training, providing rich supervisory signals that help the model learn more generalized and discriminative features. The One-to-One head produces a single, top-ranked prediction per object, replacing the need for traditional Hungarian matching with a novel unified scoring metric that combines spatial priors, classification confidence, and localization quality. By aligning the scoring parameters between these heads, YOLOv10 ensures that the best candidate detection is consistently selected without requiring NMS, leading to approximately 46% faster inference speeds compared to YOLOv9 [149].

In addition to the reimagined detection pipeline, YOLOv10 incorporates several architectural improvements aimed at optimizing efficiency and accuracy. A lightweight classification head using depthwise separable convolutions reduces computational cost by 64% with minimal impact on accuracy [149]. A novel Spatial-Channel Decoupled Downsampling module improves feature retention of fine details during resolution reduction. A rank-guided block design replaces redundant convolutional blocks in deeper layers with more compact modules, saving parameters without accuracy loss. For small models (YOLOv10n and YOLOv10s), large-kernel convolutions with structural reparameterization improve detection of small objects by expanding receptive fields, while larger models (YOLOv10b and above) incorporate Partial Self-Attention (PSA) modules to capture global context efficiently, enhancing detection in cluttered backgrounds.

A.3 YOLOv11

Lastly, YOLOv11 was released in 2025 and continues the trend of improving both accuracy and efficiency in object detection models, with several innovations tailored for high-resolution, real-time analysis tasks [150]. This version introduces several architectural improvements designed to reduce computational complexity, increase feature discrimination, and improve detection performance.

One of the key modifications in YOLOv11 is its refined backbone and neck structure with the replacement of the earlier C2f block, used since YOLOv8, with a newly designed C3k2 module. This block is optimized to maintain a wide receptive field while reducing the number of parameters and inference latency. It achieves this efficiency by employing a sequence of smaller convolutional kernels instead of larger, more costly ones. It further enhances feature diversity

through channel shuffling, ensuring that information flows effectively across feature maps. This design results in a 15% reduction in latency compared to YOLOv10, while maintaining detailed edge information [150].

Another major enhancement is the integration of the PSA mechanism within a new block known as C2PSA. It maintains spatial information during feature weighting by generating attention maps from both average and maximum pooling. These maps guide the weighting of feature regions, highlighting object clusters and suppressing irrelevant background artifacts.

To improve multi-scale context awareness, YOLOv11 further incorporates a revised spatial pooling module known as Spatial Pyramid Pooling - Fast. Instead of pooling features in parallel, as done in previous versions, SPPF applies sequential pooling with increasing receptive fields. This helps the model better capture both localized and global information, which is particularly beneficial when objects vary in size and shape. Additionally, to mitigate overfitting, feature dimensionality is reduced before pooling, and the resulting features are refined using attention mechanisms. [150].

Appendix B

BoT-SORT-ReID Workflow

The BoT-SORT-ReID algorithm follows a frame-by-frame pipeline consisting of four main stages: detection processing, motion estimation, data association, and track management. It operates on an input video using detections and appearance features extracted for re-identification, tracking objects consistently throughout the sequence.

Initially, for each frame in the video f_k , detections D_k are obtained from the detector, each with a bounding box and confidence score. The detection results are divided into two sets based on confidence scores and two predefined thresholds: a high-confidence threshold τ_{high} and a low-confidence threshold $\tau_{\text{low}} < \tau_{\text{high}}$. Detections with confidence scores greater than τ_{high} are assigned to the high-confidence set D_{high} , while those with scores between τ_{low} and τ_{high} are placed in the low-confidence set D_{low} . Detections with scores below τ_{low} are discarded. Formally:

$$D_{\text{high}} = \{d \in D_k \mid \text{score}(d) > \tau_{\text{high}}\} \quad (\text{B.1})$$

$$D_{\text{low}} = \{d \in D_k \mid \tau_{\text{low}} < \text{score}(d) < \tau_{\text{high}}\} \quad (\text{B.2})$$

For all high-confidence detections, appearance features are extracted for re-identification using a feature encoder, producing a set of embeddings. In our implementation, the features are directly sourced from intermediate layers within the trained YOLO architecture - specifically the Feature Pyramid Network layers and final output layer of the detection head. This strategy takes advantage of Ultralytics' implementation, allowing the reuse of rich spatial and semantic information already computed during detection, without requiring an additional re-identification model.

Each existing track in the current track set T undergoes a state prediction using the modified Kalman filter. This filter estimates the object's future position and size by modeling its motion over time - including direction and velocity - while accounting for uncertainties in both how the object moves (process noise) and how it is observed (measurement noise). The balance between the predicted state and the actual measurement is determined by the Kalman gain, which adjusts the influence of each source according to their respective uncertainties. This implementation uses a modified Kalman filter with an eight-dimensional state vector for predicting and updating each

object's position, size, and motion over time:

$$x_k = \begin{bmatrix} x_c(k), y_c(k), w(k), h(k), \dot{x}_c(k), \dot{y}_c(k), \dot{w}(k), \dot{h}(k) \end{bmatrix}^T, \quad (\text{B.3})$$

where x_c and y_c denote the object's center coordinates, w and h its width and height, and the dotted terms represent their respective velocities. Unlike earlier approaches that encoded scale using area or aspect ratio in a seven-dimensional vector, this formulation improves robustness to changes in object size by tracking width and height independently. The process noise and measurement noise are dynamically scaled based on the current size of the object, allowing greater variance for larger objects and tighter constraints for smaller ones.

To refine predicted states, GMC is applied to estimate the transformation between consecutive frames f_{k-1} and f_k using a warp matrix A_k^{k-1} , modeled as an affine transformation.

Data association is performed in two stages. In the first stage, the predicted track locations (i) are matched with the high-confidence detections (j) by minimizing a combined cost matrix C_{high} which incorporates both motion and appearance cues. The IoU distance $d_{i,j}^{\text{IoU}}$ captures spatial proximity, while the cosine distance $d_{i,j}^{\text{cos}}$ measures appearance dissimilarity between track and detection embeddings. To penalize weak matches and give preference to pairs that are both close in position and look similar, a filtering rule is applied:

$$\hat{d}_{i,j}^{\text{cos}} = \begin{cases} 0.5 \times d_{i,j}^{\text{cos}}, & \text{if } d_{i,j}^{\text{cos}} < \theta_{\text{emb}} \text{ and } d_{i,j}^{\text{IoU}} < \theta_{\text{IoU}} \\ 1, & \text{otherwise} \end{cases} \quad (\text{B.4})$$

Here, the appearance threshold θ_{emb} controls whether embeddings are considered similar enough to reinforce a match. The final cost matrix C_{high} is then computed as:

$$C_{i,j}^{\text{high}} = \min(d_{i,j}^{\text{IoU}}, \hat{d}_{i,j}^{\text{cos}}) \quad (\text{B.5})$$

The resulting cost matrix is then solved optimally using the Hungarian algorithm. Only matches with a total cost less than or equal to the threshold θ_{match} are accepted, while pairs exceeding this threshold are discarded as unmatched. Detections and tracks that remain unmatched are forwarded to the second association stage.

In the second stage, unmatched tracks (i') are matched to low-confidence detections (j') solely based on IoU distance:

$$C_{i',j'}^{\text{low}} = d_{i',j'}^{\text{IoU}} \quad (\text{B.6})$$

again by employing the Hungarian algorithm. Tracks still unmatched after this step are considered lost and are removed from the active track set if they have been unmatched beyond a predefined buffer period, Δ_{buffer} .

Following data association, matched tracks are updated with the corresponding new detections, including refinement of their state estimates through the Kalman filter and updating of their

appearance embeddings. New tracks are initialized from unmatched high-confidence detections that surpass a minimum score threshold η .

Appendix C

Detection Training Results

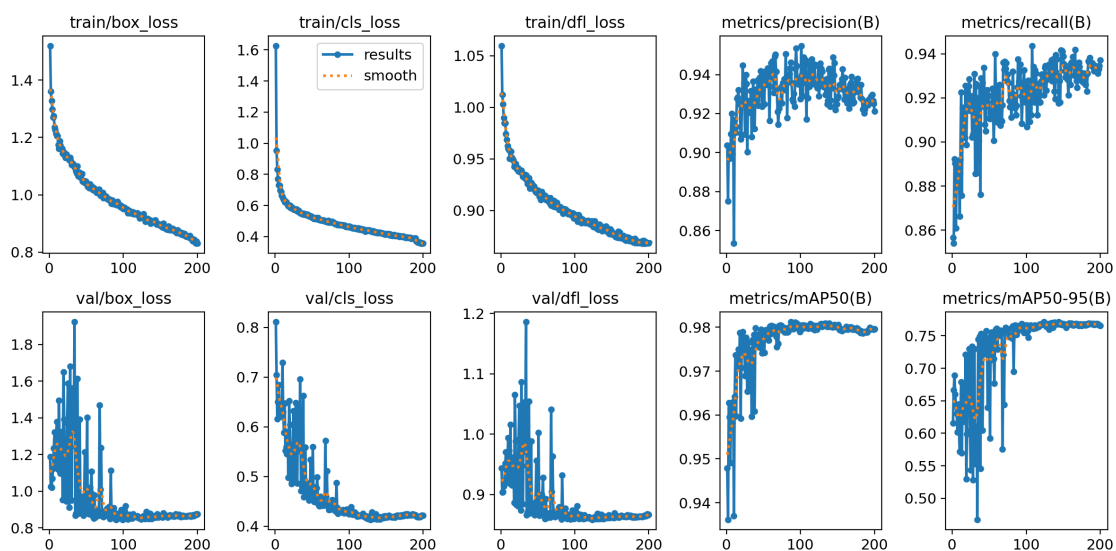


Figure C.1: YOLOv9t loss curves and validation metrics across 200 epochs.

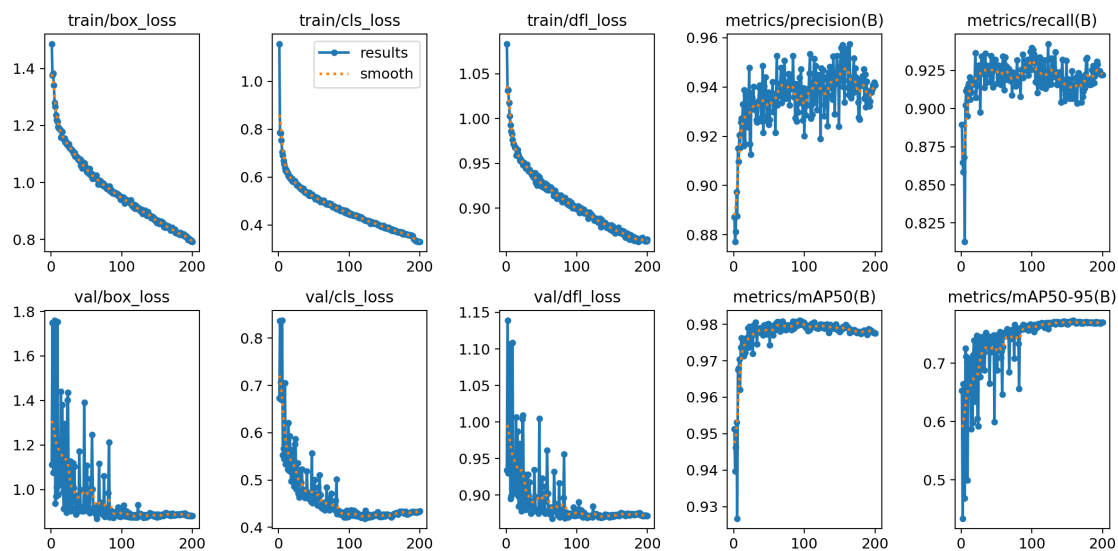


Figure C.2: YOLOv9s loss curves and validation metrics across 200 epochs.

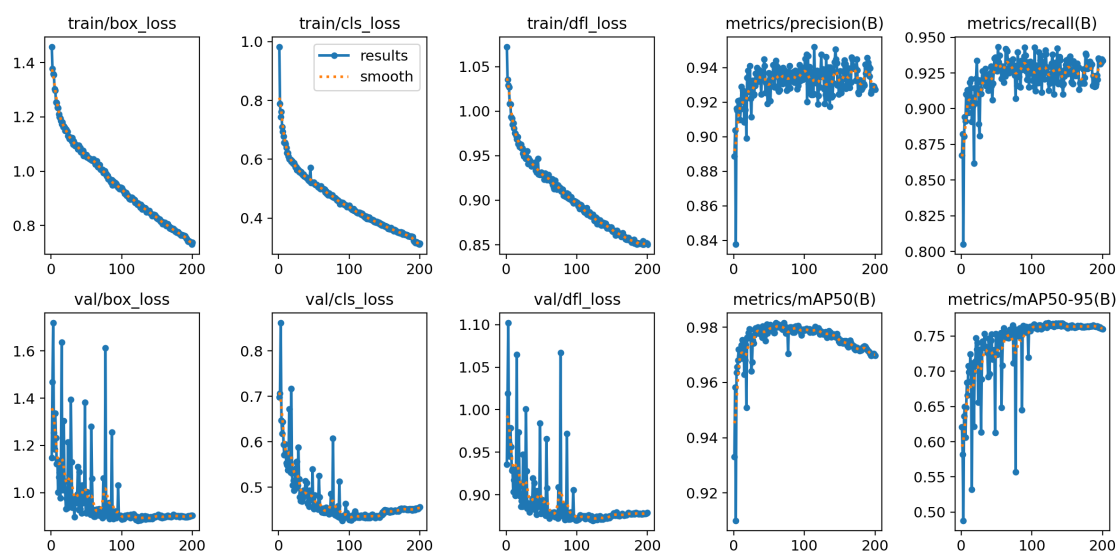


Figure C.3: YOLOv9m loss curves and validation metrics across 200 epochs.

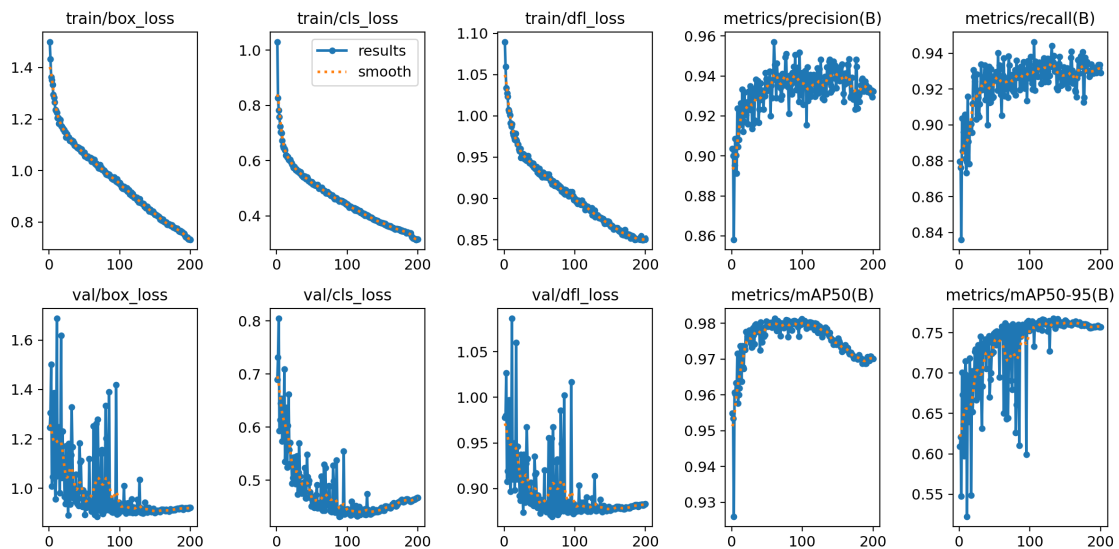


Figure C.4: YOLOv9c loss curves and validation metrics across 200 epochs.

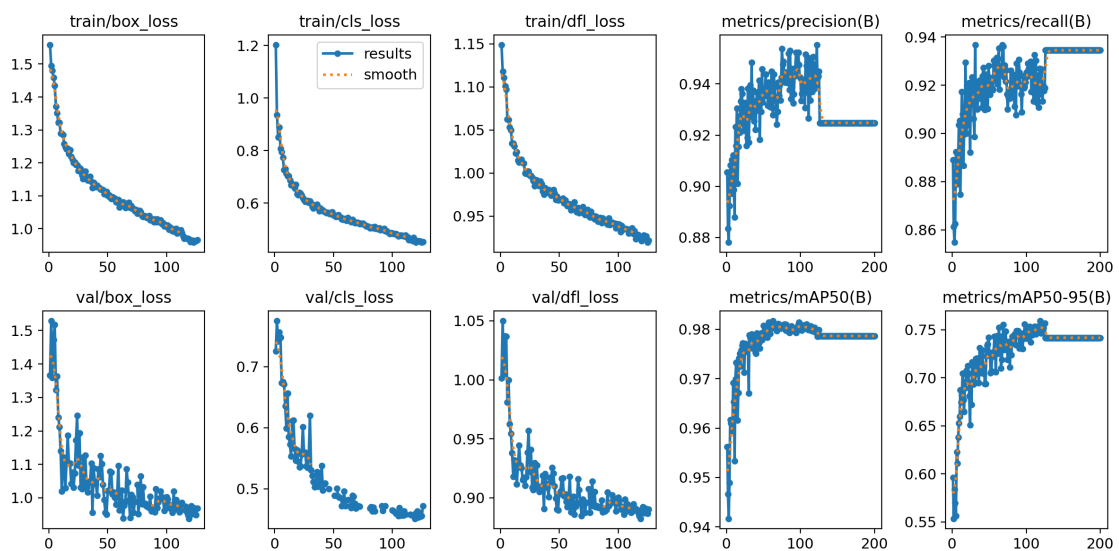


Figure C.5: YOLOv9e loss curves and validation metrics across 200 epochs.

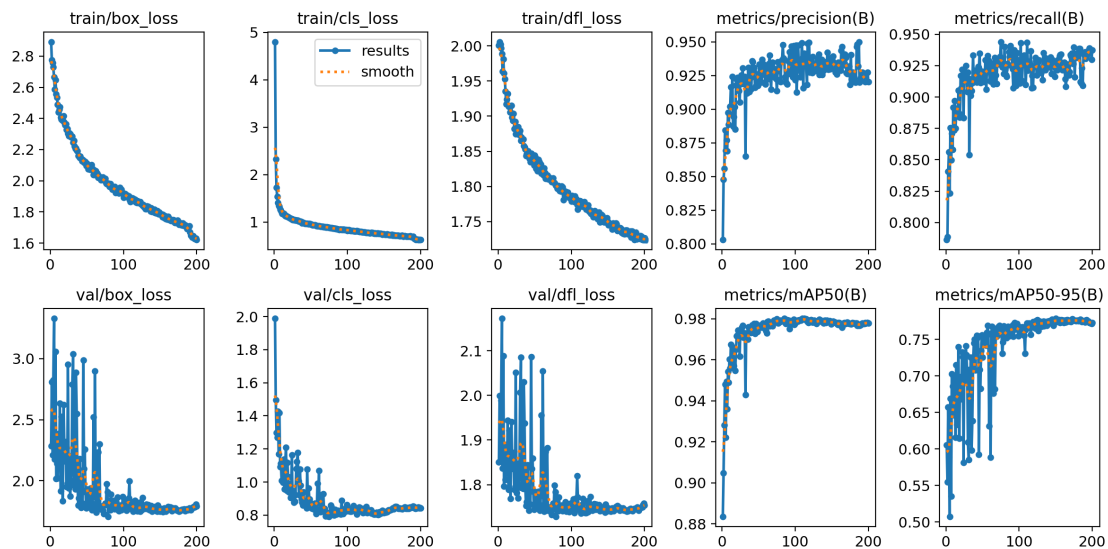


Figure C.6: YOLOv10n loss curves and validation metrics across 200 epochs.

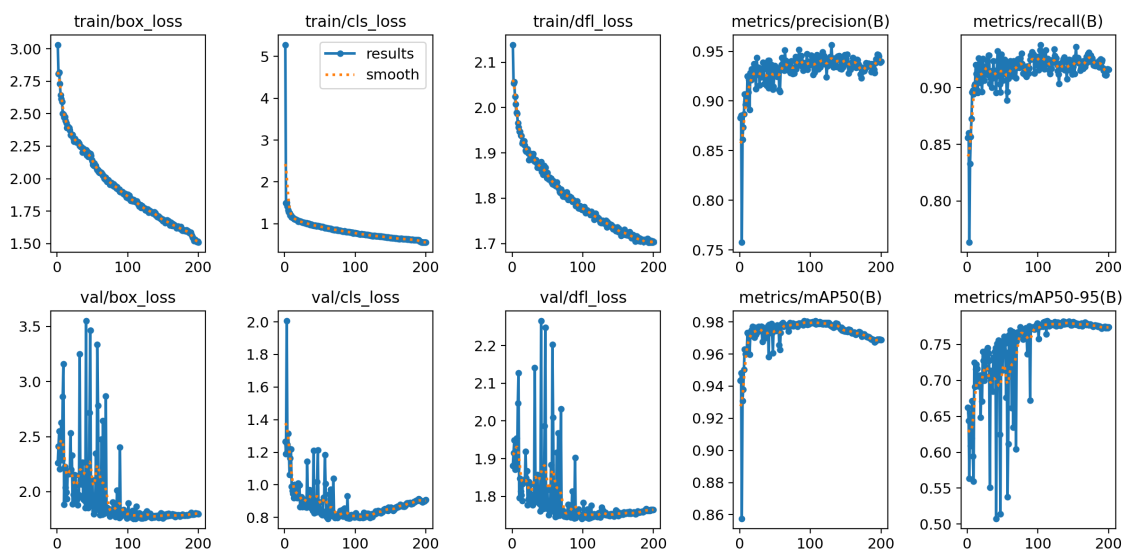


Figure C.7: YOLOv10s loss curves and validation metrics across 200 epochs.

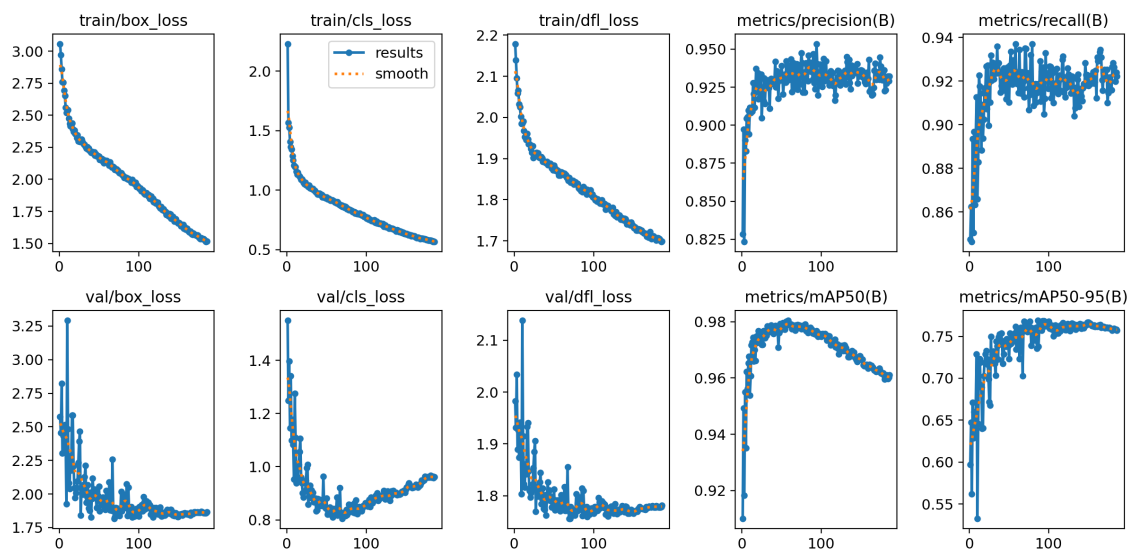


Figure C.8: YOLOv10m loss curves and validation metrics across 200 epochs.

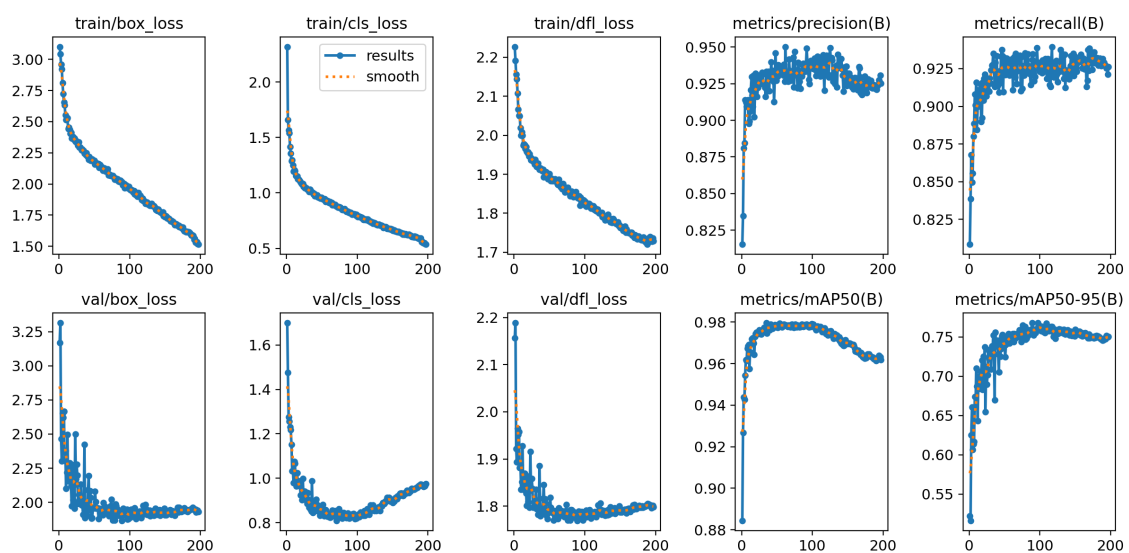


Figure C.9: YOLOv10b loss curves and validation metrics across 200 epochs.

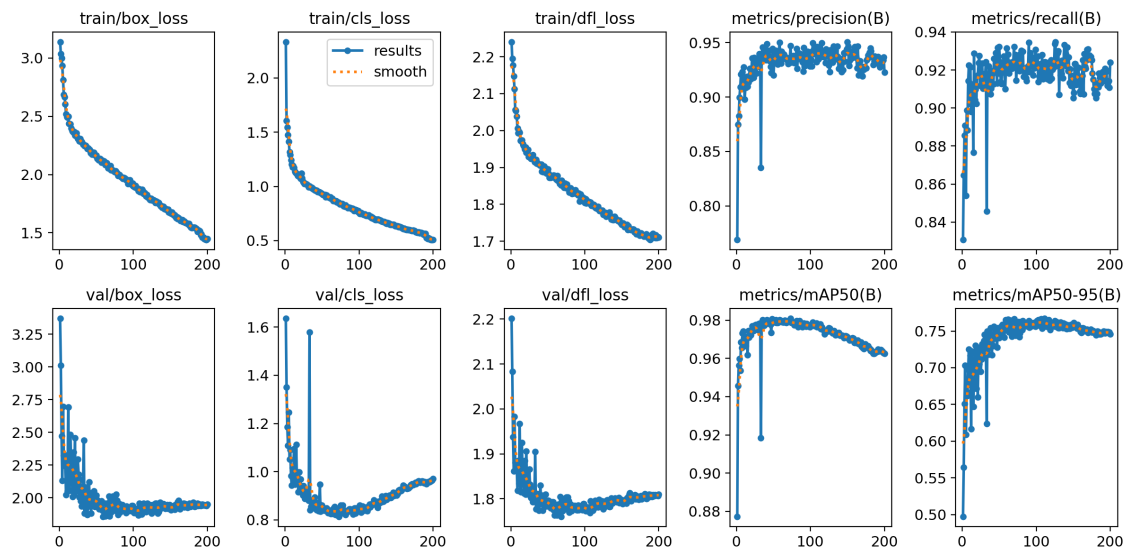


Figure C.10: YOLOv10l loss curves and validation metrics across 200 epochs.

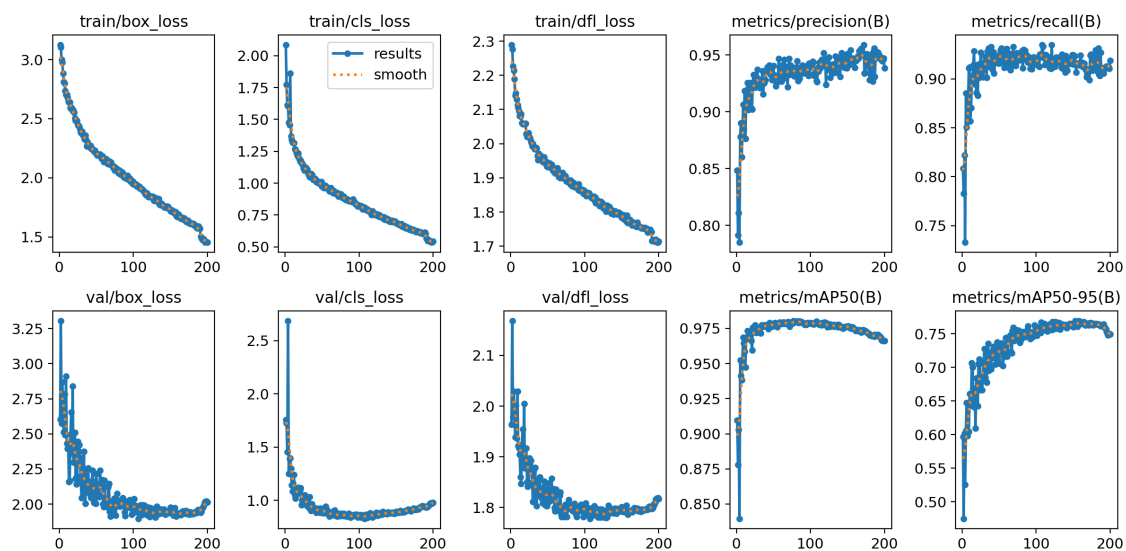


Figure C.11: YOLOv10x loss curves and validation metrics across 200 epochs.

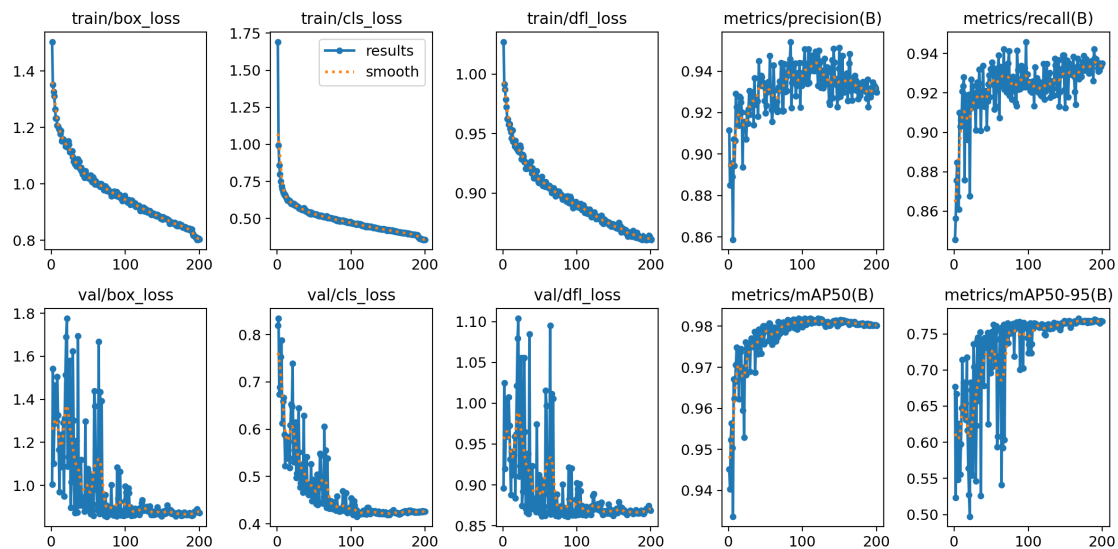


Figure C.12: YOLOv11n loss curves and validation metrics across 200 epochs.

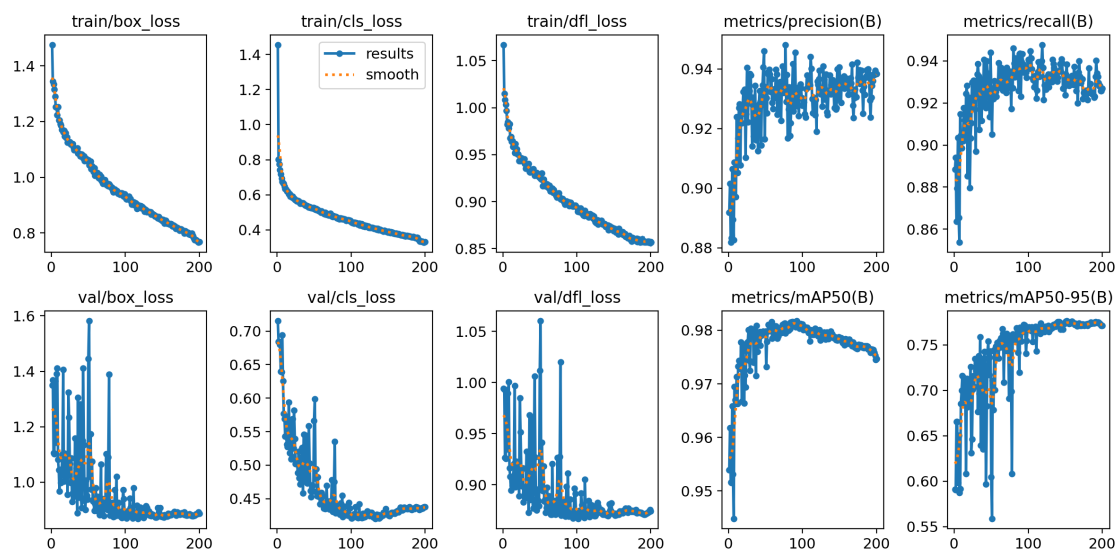


Figure C.13: YOLOv11s loss curves and validation metrics across 200 epochs.

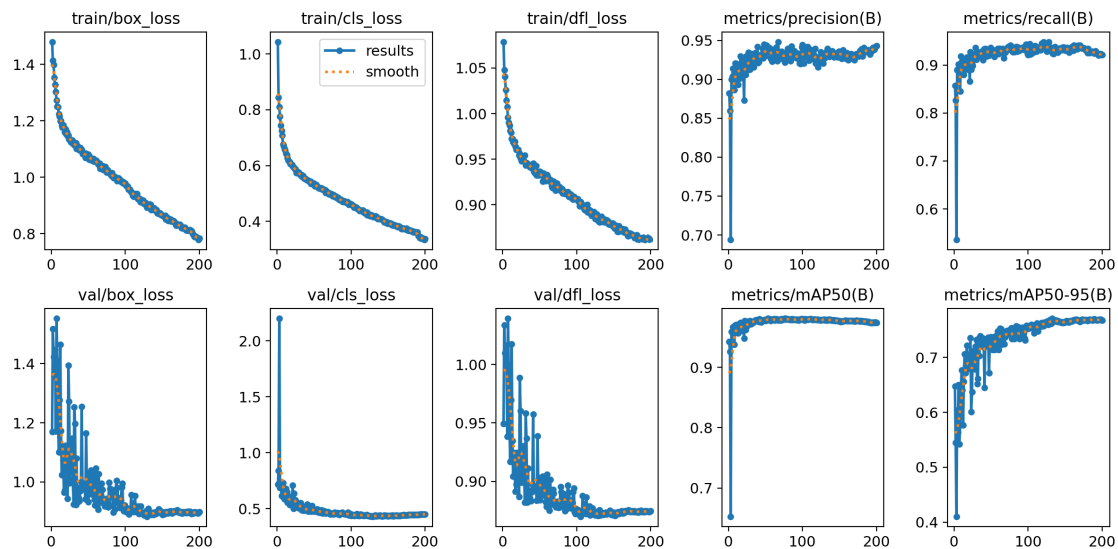


Figure C.14: YOLOv11m loss curves and validation metrics across 200 epochs.

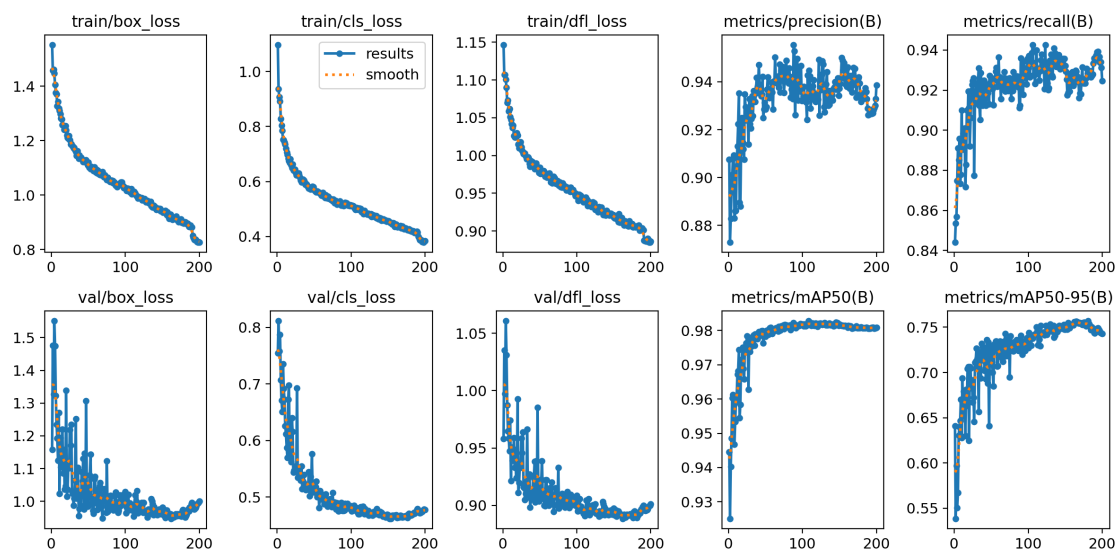


Figure C.15: YOLOv11l loss curves and validation metrics across 200 epochs.