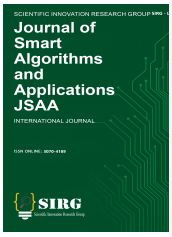

Journal of Smart Algorithms and Applications
JSAA
ISSN: 3070-4189/
Journal Homepage
https://pub.scientificirg.com/index.php/JSAA

Artificial Intelligence for Automated Semen Analysis from Microscopy to Clinical Translation

Shimaa B. Abdallah [a,1] **, Mostafa Ahmed Hamed** [b] **, and Ahmed A. Elngar** [c]

[a] *Management Technology and Health Informatics Department, Faculty of Applied Health Sciences Technology, Beni-Suef University, Beni-Suef City, 62511, Egypt; E-mail: Shimaa.Bahy@fahst.bsu.edu.eg*

[b] *Department of Andrology, Sexology and STDs, Faculty of Medicine, Beni-Suef University, Beni-Suef City, 62511, Egypt; E-mail: elgebalymostafa@gmail.com*

[c] *Department of Computer Science, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt; E-mail: elngar_7@yahoo.co.uk*

ABSTRACT

Male-factor infertility is commonly evaluated using semen analysis; however, manual assessment of sperm concentration, motility, and morphology remains susceptible to observer variability and laboratory heterogeneity. This review synthesizes artificial intelligence methods for automated semen analysis, covering sperm detection, segmentation, morphology classification, motility assessment, tracking, and hybrid computational pipelines. A literature search was conducted across PubMed/MEDLINE, IEEE Xplore, Scopus, Web of Science Core Collection, and Google Scholar. Eligible studies were peer-reviewed, English-language investigations that applied machine-learning or deep-learning methods to at least one semen-analysis task and reported quantitative performance measures. The reviewed studies demonstrate recurring use of transfer learning, one-stage object detection, U-Net- and Mask R-CNN-based segmentation, video-level convolutional models, transformer-based architectures, model ensembles, and generative data augmentation. Although many studies report high performance on curated internal datasets, comparisons across studies are limited by differences in species, imaging modality, annotation procedures, evaluation metrics, and data-partitioning strategies. External validation, patient- or donor-level leakage control, calibration, uncertainty estimation, and prospective assessment of clinical utility remain uncommon. Hybrid systems appear most justified when their components address complementary failure modes and when their incremental value is evaluated against an appropriately matched single-model baseline; however, such comparisons are not consistently reported. Overall, the available evidence supports the technical feasibility of artificial intelligence for automated semen analysis but does not yet establish readiness for routine clinical deployment. Translation will require harmonized multicenter datasets, consensus-based annotation, leakage-aware external validation, task-specific reporting standards, defined human oversight, and regulatory planning appropriate to the intended use and jurisdiction.

PAPER INFORMATION

HISTORY

Received: 25 May 2026
Revised: 19 June 2026
Accepted: 13 July 2026
Online: 28 August 2026

MSC

68T07; 68R10; 94A60; 68M15

KEYWORDS

Artificial Intelligence;
Automated Semen Analysis;
Clinical Translation;
Hybrid Systems;
Deep Learning.

[1]Corresponding author at Management Technology and Health Informatics Department, Faculty of Applied Health Sciences Technology, Beni-Suef University, Beni-Suef City, 62511, Egypt; E-mail: Shimaa.Bahy@fahst.bsu.edu.eg

1. INTRODUCTION

Male-factor infertility represents a substantial share of infertility cases and is estimated to affect approximately 7% of men of reproductive age [1]. According to the World Health Organization (WHO) 2021 laboratory manual, semen analysis remains the principal test for assessing male reproductive potential through measurements of sperm concentration, total and progressive motility, and morphology [2]. Manual assessment by trained andrologists is subject to considerable interobserver variability, particularly in morphological evaluation, where interpretation is inherently subjective [3]. These limitations have encouraged the development of more objective, standardized, and automated approaches to semen assessment. Computer-assisted sperm analysis (CASA) emerged in the 1980s as an early approach to quantitative assessment. By tracking sperm trajectories, CASA systems made it possible to calculate kinematic parameters systematically [4]. Modern CASA platforms have further improved the reproducibility of structural and functional measurements across laboratories [5]. Nevertheless, conventional CASA systems often depend on handcrafted features, fixed morphometric thresholds, and rule-based decision logic. Their performance may therefore be affected by differences in staining procedures, microscope optics, and image-acquisition settings, limiting generalizability across laboratories. Accurate analysis is particularly challenging in samples with high sperm concentration, overlapping cells, imaging artifacts, or complex morphological abnormalities. Deep learning, a subfield of artificial intelligence based on multilayer neural networks, provides a data-driven alternative that can learn hierarchical visual representations directly from microscopy images and videos, reducing the need for manual feature engineering [6]. Deep learning methods have shown strong performance on internal benchmarks for sperm detection [7], morphology classification [8, 9], and concentration estimation [10]. However, performance on a study's internal test partition does not necessarily indicate that the method will generalize to data from other laboratories, populations, imaging systems, or acquisition protocols. The distinction between internal performance and external generalizability is therefore considered throughout this review. Hybrid methods that combine deep learning with classical image processing or conventional machine learning have also been investigated, particularly in settings where the available training data are insufficient for fully end-to-end optimization [11]. When such methods are evaluated only through a single internal comparison, the results indicate improved performance under the reported conditions but do not, by themselves, demonstrate robustness to distributional change. Accordingly, this review interprets claims of improved robustness as evidence of higher internal performance unless the source study explicitly evaluates the method under an external or shifted data distribution. Although the number of reviews addressing artificial intelligence in sperm and semen analysis has increased, important gaps remain. Existing reviews often focus on individual components of the analysis pipeline, such as morphology assessment [10] or the historical development of CASA [4], with limited coverage of the complete computational workflow. Recent developments in transformer-based architectures, foundation models, self-supervised learning, federated learning, explainable artificial intelligence (XAI), and regulatory compliance have further broadened the field, but these topics are frequently examined separately. A review that connects the full process, from image acquisition and preprocessing to detection, segmentation, morphology classification, motility analysis, dataset characterization, and clinical translation, is therefore warranted. This review provides an end-to-end synthesis of deep learning and hybrid methods for automated semen analysis. The terminology used throughout the review is defined as follows. "Automated semen analysis" refers to the computational evaluation of a semen sample as a whole. "Sperm analysis" is reserved for cell-level tasks involving individual spermatozoa, including detection, segmentation, and morphology classification. This distinction follows common usage in the source literature and avoids implying that a cell-level prediction is equivalent to a patient-level semen-analysis diagnosis. "Computer-assisted sperm analysis" (CASA) refers specifically to instrumented or software-based systems that automate one or more components of semen evaluation. "Detection" denotes the localization of individual spermatozoa, typically with a bounding box, whereas "segmentation" denotes the pixel-level delineation of a cell or subcellular structure. The term "identification" is not used as a generic task label in this review, except when it appears in the title of a cited study, because studies using this term may perform detection or localization rather than identification in the strict sense. Because "hybrid system" has no single standardized meaning in the literature, the term is defined operationally here. Four non-exclusive forms of hybridization are considered:

1. *Architectural fusion*, in which two or more independently trained models generate predictions that are combined through averaging, voting, or a learned meta-classifier;
2. *Pipeline-level hybridization*, in which a deep learning component is combined sequentially with a classical image-processing or rule-based component;
3. *Feature-level fusion*, in which representations from structurally different encoders are concatenated or otherwise combined before a shared prediction head; and
4. *Generative augmentation*, in which a generative adversarial network or another generative model produces additional training examples for a separately trained detection or classification model.

This taxonomy is applied throughout the discussion of hybrid systems and, where appropriate, in the preceding technical sections. The categories are non-exclusive because a single study may combine more than one form of hybridization. For

example, a framework may use architectural fusion together with generative augmentation, or may combine feature-level fusion with a rule-based post-processing stage. The review has six main objectives. First, it examines the complete computational pipeline, including image acquisition, preprocessing, detection, segmentation, morphology classification, motility assessment, tracking, and clinical interpretation. Second, it organizes conventional machine learning, deep learning, transformer-based, and hybrid frameworks according to the taxonomy defined above. Third, it examines image-processing procedures as potential determinants of model robustness rather than treating them as a minor preprocessing step. Fourth, it evaluates the principal public and private datasets used in the field, with attention to dataset size, annotation quality, species, and suitability for specific analytical tasks. Fifth, it compares reported outcomes while identifying the methodological differences that limit direct quantitative comparison and presents a task-specific reporting framework for interpreting published results. Sixth, it discusses the main barriers to clinical translation, including intended use, human oversight, explainability, reproducibility, and regulatory requirements, and proposes a dependency-based translational framework that does not rely on fixed timelines.

The remainder of the review is organized as follows. Section 2 describes the literature-search methodology and review design. Section 3 examines sperm detection and segmentation. Section 4 discusses morphology classification and transfer learning. Section 5 addresses motility assessment and tracking. Section 6 reviews image acquisition and preprocessing. Section 7 considers pattern recognition and generative augmentation. Section 8 presents hybrid systems. Section 9 evaluates benchmark datasets and performance metrics. Section 10 discusses challenges and future directions. Section 11 concludes the review.

2. LITERATURE SEARCH METHODOLOGY

This article is a structured narrative review informed by PRISMA-style reporting, not a formal systematic review or scoping review, and this distinction is stated explicitly because these designs carry different evidentiary expectations. A formal systematic review would require a prespecified, typically registered protocol; complete, database-specific Boolean search strings reported in full; independent screening and eligibility adjudication by at least two reviewers with a documented conflict-resolution procedure; and a study-level risk-of-bias or quality assessment applied with a validated instrument. This review follows the PRISMA 2020 flow structure to report how records moved from identification through screening to inclusion, because that structure improves transparency, but it does not claim the full methodological apparatus of a registered systematic review: no protocol was registered in advance, dual independent screening with a formal adjudication log is not reported, and no validated risk-of-bias instrument was applied to individual studies. Readers should treat the search process described here as a transparently reported, reproducible-in-outline literature search supporting a structured narrative synthesis, and should not treat the included-study set as a registered systematic-review sample. A small number of terminology conventions are fixed here and applied consistently thereafter. Percentage figures denote proportions of a stated denominator and are reported to the precision given in the source publication; this review does not add spurious decimal precision beyond what the original study reported, and, where a metric's definition (macro- versus micro-averaged, IoU threshold, matching rule) is not stated in the source, this review flags the ambiguity explicitly rather than resolving it by assumption. The search was performed across five databases, PubMed/MEDLINE, IEEE Xplore, Scopus, Web of Science Core Collection, and Google Scholar, using Boolean combinations of the terms "sperm analysis," "semen analysis," "artificial intelligence," "machine learning," "deep learning," "convolutional neural network" (CNN), "computer-assisted sperm analysis," "object detection," "YOLO," "Faster R-CNN," "Mask R-CNN," "DETR" (Detection Transformer), "U-Net," "transfer learning," and named architectures including ResNet, VGG, EfficientNet, and Vision Transformer. Database-specific syntax was adapted while preserving the same conceptual search strategy across all sources; the exact per-database syntax strings, date-by-date search records, and a full-text exclusion log are not reproduced in this article, which is a specific respect in which it falls short of a registered systematic-review protocol, as stated above, and is revisited as a limitation. Studies were eligible for inclusion if they presented original research applying artificial intelligence, machine learning, or deep learning methods to automated sperm or semen analysis; reported quantitative evaluation metrics; addressed at least one analytical task among detection, segmentation, morphology classification, motility assessment, or concentration estimation; and were published in English as peer-reviewed journal articles or conference proceedings. Studies were excluded if they were duplicate publications, were not peer-reviewed, lacked quantitative evaluation, or addressed non-automated sperm analysis, or if they used a non-comparative, single-model design that additionally lacked any quantitative baseline, ablation, or externally reported benchmark against which the proposed model's performance could be contextualized; this last criterion was not applied to single-model studies that reported comparisons against prior published results on the same dataset. The study-selection process was summarized using the PRISMA flow structure to support methodological transparency. Database searching, screening, full-text retrieval, and eligibility assessment proceeded in the stages summarized in **Figure 1**; the exact stage-by-stage counts, including the number of records excluded under the single-model criterion above, are reported in the figure rather than restated as a running total in the prose, so that the figure remains the single source of truth for these counts and the two do not risk drifting out of agreement across revisions. Every numerical value reported in the sections that follow, including sample counts, video counts, annotated object counts, class definitions, magnification, image resolution, and train/test partitioning, was checked against the original

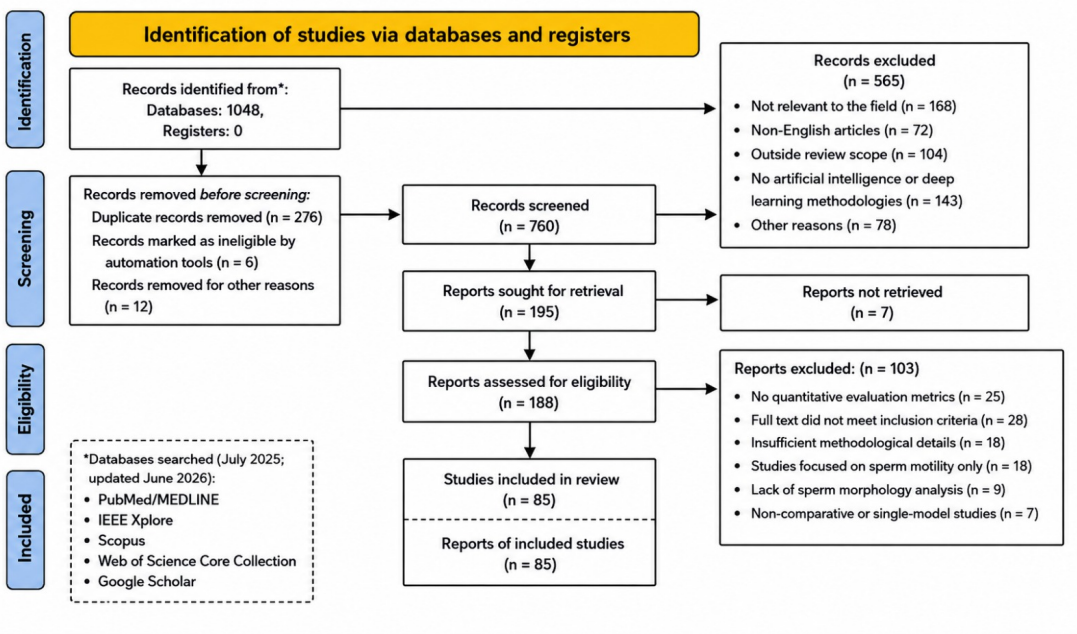

Figure 1: PRISMA flow diagram

source publication rather than an abstract, preprint, or secondary citation, to the extent verifiable from the publicly available version of each source at the time of writing. Where a dataset or method has been described in both a conference and a subsequent journal version, or where a preprint precedes a peer-reviewed version, the values reported here follow the peer-reviewed version, and this substitution is noted explicitly wherever it applies; values from different versions of the same dataset are not combined, and where the same underlying donor or patient cohort appears to underlie more than one cited publication, this is noted at the relevant point in the text rather than treated as independent evidence. Readers who identify a discrepancy between a citation in this review and the cited source are encouraged to consult the original publication directly, since a review article of this scope cannot substitute for primary-source verification of any individual claim; a claim-level extraction record, keyed to source table, figure, or page location for each high-impact numerical claim, is recommended as a supplementary companion to this review and is noted as a specific limitation where not yet available. This review did not conduct a formal, tool-based risk-of-bias assessment, such as QUADAS-2 for diagnostic-accuracy studies or PROBAST for prediction-model studies, applied individually to each included study. This decision was based on the absence of a validated instrument that is well suited to the heterogeneous study designs, tasks, and reported measures characteristic of AI-based sperm-analysis research. Instead, **Table 1** distinguishes three reporting states across the included studies: a check mark indicates that the source publication reports or addresses a given domain; a dash indicates that the domain is not reported; and "NA" indicates that the domain is not applicable to the study design. The table is descriptive rather than quantitative: it enables readers to distinguish studies that provide stronger evidence of external validity from those that demonstrate only internal technical feasibility, but it should not be interpreted as a pooled risk-of-bias score.

Table 1: Representative, illustrative evidence table for a subset of included studies, covering methodological domains relevant to the strength of the evidence

Study	Sampling	Reference standard	Annotator agreement	Leakage control	Split granularity stated	External validation	Class imbalance addressed	CI reported	Code/data available
Javadi & Mirroshandel [11]	✓ (235 patients)	WHO-aligned	–	✓ patient-stratified	✓	–	Partial	–	✓ (public)
Haugen et al. [12]	✓ (external QA program)	Multi-laboratory reference reads	–	✓ site-independent	✓	✓	NA	–	Partial
Wang et al. [13]	✓ (multi-clinic)	Internal expert annotation	–	✓ site-level	✓	✓	–	–	–
Valiuskaite et al. [14]	–	Reference vitality measure	–	Unclear	–	–	–	✓	–
Spencer et al. [15]	–	Expert-labeled (HuSHeM)	–	Unclear	–	–	–	✓ (significance test)	–

A methodological concern common to nearly all studies reviewed here is the level at which validation is performed and the resulting risk of data leakage. This issue is treated as a central methodological problem rather than a minor

limitation. **Table 2** records the validation designs of the representative studies summarized therein, as reported in the source publications. However, the label "internal split" alone does not indicate whether the training and test data were partitioned at the level of the individual image or frame, video, patient or donor, microscope slide, or acquisition site. This distinction is consequential rather than merely semantic. In video-based motility or tracking studies, adjacent frames from the same video sequence are highly correlated. Consequently, a frame-level or clip-level split that allows frames from the same video to appear in both partitions can substantially inflate the reported accuracy relative to a video-level or patient-level split. This occurs because the model may partially memorize sequence-specific visual characteristics rather than learn a generalizable representation of motility. In morphology-classification studies, an analogous risk arises when individual sperm-cell crops from the same slide or donor are distributed across different partitions. Cells from the same slide share staining, illumination, and preparation characteristics that a model can exploit without learning morphological features that generalize to a new patient. When a source publication explicitly states that the data were partitioned at the patient, donor, or site level, this information is recorded in the table. Examples include the predefined patient-stratified partition described for the MHSMA dataset [11] and the external, site-independent quality-assessment design reported by Haugen and colleagues [12]. When a source publication does not specify the partitioning unit, this review records the split as "internal, granularity not reported." When even the internal-versus-external distinction is ambiguous, it is recorded as "unclear," rather than being assigned a favorable interpretation. This is identified as a specific limitation of the underlying study, rather than of this review's synthesis. Readers comparing reported accuracies across studies should therefore treat any study that does not report its partitioning unit with additional caution. Based on the information provided, an unreported image-level or frame-level split cannot be distinguished from a leakage-prone split.

Three evidence levels recur throughout the sections that follow and are used consistently to qualify what a given result does and does not support: internal technical validation, meaning a result reported on an internal split of a single dataset with no independent test population; external analytical validation, meaning a result reported on data from an independent site, laboratory, or population not used in model development; and prospective clinical or workflow validation, meaning a result obtained from a study that evaluated the system's effect on an actual clinical or laboratory decision process rather than only its agreement with a retrospective reference label. The great majority of studies discussed in this review report internal technical validation only; this hierarchy is applied explicitly and is the basis for the review's overall conclusion that current evidence supports technical feasibility rather than demonstrated clinical utility.

3. SPERM DETECTION AND SEGMENTATION

Sperm detection, the localization of individual spermatozoa within a microscopic image or video frame, is a prerequisite for downstream morphology, motility, and concentration measurement; segmentation extends this task to pixel-level delineation of the head, acrosome, midpiece, and tail. **Figure 2** illustrates one representative shared-backbone architecture used by several of the convolutional detection and segmentation frameworks discussed below, in which a common feature-extraction backbone (for example, CSPDarknet, ResNet, EfficientNet, or MobileNet) feeds parallel detection and segmentation heads; this is a conceptual illustration of one common architectural family and does not represent every architecture discussed in this section, since one-stage detectors, two-stage detectors, transformer-based detectors, and classical pipeline approaches without a shared convolutional backbone are also reviewed below and do not share this exact structure. CNN architectures have been applied to both tasks across a range of microscopy modalities. Zou and colleagues [16] introduced TOD-CNN, an architecture optimized for tiny-object detection, trained on 111 videos containing more than 278,000 annotated objects, and reported an average precision at an intersection-over-union (IoU) threshold of 0.50, denoted AP_{50}, of 85.60 percent under offline conditions described as real-time by the study's authors; generalizability beyond the training distribution was not assessed. Wu and colleagues [17] developed a Faster R-CNN-based system for sperm identification in microdissection testicular sperm extraction biopsy samples, trained on 702 images from 30 patients, achieving a mean average precision of 0.741; the small patient sample limits the statistical reliability of this estimate for clinical validation. Object detection frameworks have also been evaluated directly against one another under matched conditions. Yuzkat and colleagues [24] compared single-stage and two-stage detectors for discriminating sperm from non-sperm structures and reported that YOLOv5 achieved the strongest single-stage performance on their dataset, while a fusion approach achieved the highest overall mean average precision; the study addressed detection alone and did not incorporate motility analysis. Hidayatullah and colleagues [23] proposed DeepSperm, a lightweight single-layer detector developed for bull semen video; because this study was conducted in bovine rather than human samples, its reported mean average precision of 94.11 percent, F1-score of 0.93, and processing speed of 51.9 frames per second, with a model roughly one-twentieth the size of YOLOv4, is read here as evidence of technical feasibility in an animal model rather than as evidence applicable to human semen, since sperm-head geometry, staining response, and imaging requirements differ between species and clinical intended use differs correspondingly. Sato and colleagues [7] applied a YOLOv3-based framework to clinical intracytoplasmic sperm injection video from the Jikei Sperm Dataset, a human clinical dataset, reporting sensitivity and positive predictive value of 0.881 and 0.853 for abnormal sperm and 0.794 and 0.689 for normal sperm; because the model detects cells without independently assessing Kruger or WHO morphological criteria, it is not by itself sufficient as a standalone selection tool. Zhu and colleagues [25] introduced YOLOv5s-SA, incorporating shuffle attention and depthwise separable convolutions to

Table 2: Representative study characteristics and evidence summary for detection, segmentation, morphology classification, and motility assessment tasks

Study	Year	Dataset	Modality	Task	Architecture	Partition unit / validation	External validation	Limitation
Zou et al. [16]	2022	111 videos, 278K+ objects	Brightfield video	Tiny-object detection	TOD-CNN	Internal, unit not reported	No	Single homogeneous source; generalizability untested
Wu et al. [17]	2021	702 images, 30 patients	Micro-TESE biopsy	Detection	Faster R-CNN	Internal, unit not reported	No	Very small patient sample
Shaker et al. [8]	2017	HuSHeM (216), SCIAN	Stained brightfield	Head morphology	APDL (dictionary learning)	Internal, unit not reported	No	Manual patch extraction step
Riordon et al. [9]	2019	HuSHeM, SCIAN	Stained brightfield	Head morphology	VGG16 (transfer learning)	Internal, unit not reported	No	Static images only, no motility
Javadi & Mirroshandel [11]	2019	MHSMA (1,540; 235 patients)	Unstained, low-res	Region abnormality detection	Custom CNN	Predefined train/val/test, patient-stratified	No	No motility assessment
Keller et al. [18]	2025	10,000 boar spermatozoa	Imaging flow cytometry	Morphology, acrosome integrity	CNN	Internal, unit not reported	No	Non-human species (boar); human transfer unestablished
Marin & Chang [19]	2021	SCIAN-SpermSegGS	Stained brightfield	Multi-part segmentation	U-Net, Mask R-CNN (transfer learning)	Internal, unit not reported	No	Dataset-specific; midpiece/tail lower accuracy
Lei et al. [20]	2025	Live unstained human sperm	Brightfield	Multi-part segmentation	Mask R-CNN, YOLOv8/11, U-Net (benchmark)	Internal, unit not reported	No	High compute cost limits real-time use
Haugen et al. [21]	2019	VISEM (85 donors)	Video, 400x	Motility proportion	ResNet variants	Internal, donor-linked metadata	No	Single-center population
Haugen et al. [12]	2023	65 videos, external QA	Video	WHO motility category	ResNet-50	10-fold CV; site-independent external QA benchmark	Yes	Coarse categories; no kinematic parameters
Sato et al. [7]	2022	Jikei Sperm Dataset (4,625)	Clinical ICSI video	Detection, tracking	YOLOv3	Internal, unit not reported	No	Detection only, no Kruger criteria
Zhang et al. [22]	2024	VISEM-Tracking	Video	Detection and tracking	YOLOv8 variant + DeepOCSORT	Internal, unit not reported	No	Single public tracking benchmark
Wang et al. [13]	2024	Multi-clinic data	Brightfield	Cross-site generalizability	DL generalizability study	Internal and external, site-level	Yes	Documents degradation under domain shift
Hidayatullah et al. [23]	2021	Bull semen video	Video	Dense/occluded detection	DeepSperm (single-layer detector)	Internal, unit not reported	No	Non-human species (bull); human transfer unestablished
Yuzkat et al. [24]	2023	Custom labeled dataset	Brightfield	Sperm vs. non-sperm discrimination	YOLOv5, single/two-stage fusion	Internal, unit not reported	No	Detection only, no motility
Zhu et al. [25]	2023	Semen microscopy images	Brightfield	Occluded, low-texture detection	YOLOv5s-SA (shuffle attention)	Internal, unit not reported	No	Localization only
Dobrovolny et al. [26]	2023	Custom labeled dataset	Brightfield	Sperm cell detection	YOLOv5	Internal, unit not reported	No	Small dataset, few annotators
Liu et al. [27]	2021	HuSHeM (216)	Stained brightfield	Head morphology (4-class)	AlexNet (transfer learning)	Internal, unit not reported	No	Limited dataset variability
Spencer et al. [15]	2022	HuSHeM (216)	Stained brightfield	Head morphology (4-class)	Stacked CNN ensemble	Internal, unit not reported	No	Ensemble complexity; single dataset
Mahali et al. [28]	2023	SVIA, HuSHeM, SMIDS	Mixed	Morphology classification	Swin Transformer + MobileNetV3 + autoencoder	Internal, unit not reported; 3 datasets	No	Complex fusion, possible information loss
Cansiz et al. [29]	2026	Sperm morphology datasets	Stained brightfield	GAN-based augmentation	LB-EGAN (ensemble dual-GAN)	Internal, unit not reported	No	Augmentation-only; downstream generalizability not externally tested
Valiuskaite et al. [14]	2020	VISEM	Video	Head segmentation, motility	R-CNN + centroid tracking	Internal, unit not reported	No	Small, non-diverse dataset
Somasundaram & Nirmala [30]	2021	Custom (3,200 images)	Video, 30 fps	Classification, motility	Faster R-CNN + ESA + THMA	Internal, unit not reported	No	Short controlled clips

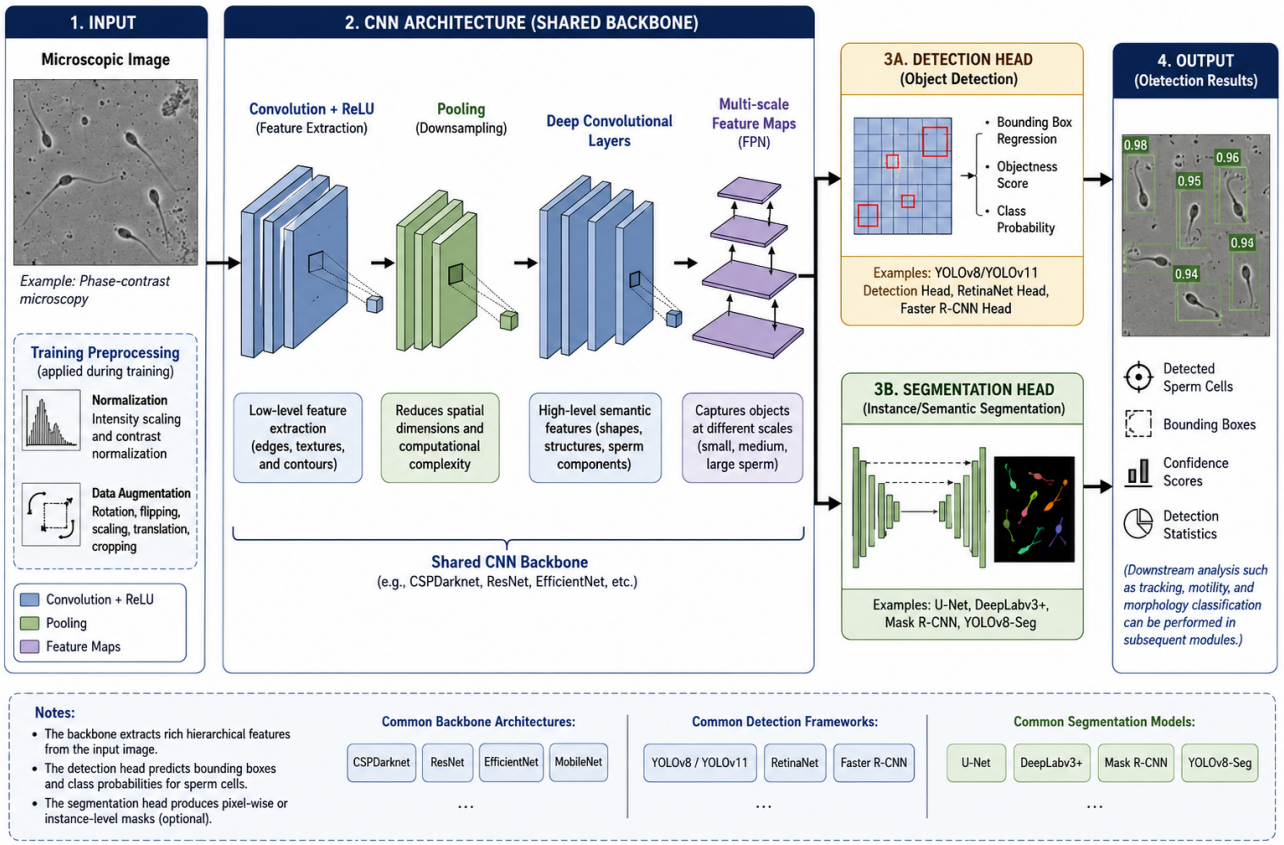

Figure 2: Conceptual example of a shared-backbone convolutional pipeline for sperm detection and segmentation

address occlusion and low-texture artifacts, and Dobrovolny and colleagues [26] compared YOLOv5 against YOLOv3 and YOLOv4 on a custom-labeled human sperm dataset, reporting accuracy differences of approximately 2 and 0.25 percentage points on their internal test set, respectively, together with faster inference than region-based detectors on the same hardware. Collectively, these studies indicate that single-stage detectors, particularly members of the YOLO family [31], and, to a lesser extent, single-shot detectors such as SSD [32], offer a favorable internal-benchmark balance of accuracy and inference speed for near-real-time laboratory workflows; reported margins between architectures should be read as internal-comparison evidence specific to each study's dataset and protocol, not as a ranked cross-study comparison.

Transformer-based detectors have more recently been adapted to sperm microscopy. Carion and colleagues [33] introduced the Detection Transformer (DETR), which removes hand-crafted components such as non-maximum suppression and anchor generation by combining a convolutional backbone with a transformer encoder-decoder that predicts bounding boxes through bipartite matching; on the COCO 2017 benchmark, a general-purpose natural-image dataset rather than a sperm-specific one, DETR matched the accuracy and runtime of an optimized Faster R-CNN baseline. Zhu and colleagues [34] introduced Deformable DETR to address slow convergence and limited multiscale resolution through deformable attention, again evaluated on general-purpose benchmarks rather than sperm microscopy. Prangemeier and colleagues [35] adapted the DETR architecture to biomedical microscopy in Cell-DETR, an end-to-end instance segmentation model for cells in microstructured environments. Zhao and colleagues [36] introduced RT-DETR, reporting 53.1 and 54.3 percent average precision on COCO val2017 at 108 and 74 frames per second for ResNet-50 and ResNet-101 backbones, respectively, again on a general-purpose benchmark. Zhang and colleagues [22] combined an attention-augmented YOLOv8 variant with an improved DeepOCSORT tracking algorithm for human sperm detection and tracking on the VISEM-Tracking dataset, reporting a higher-order tracking accuracy (HOTA) of 74.303 percent and a multi-object tracking accuracy (MOTA) of 71.167 percent on that dataset. **Figure 3** presents a conceptual example of one DETR-based pipeline of this kind, combining sample-level viability estimation with detection, tracking, and motility classification; this is a single conceptual illustration among several architectural families reviewed in this section, not a description of the architecture used by every cited study, since many detection and tracking systems reviewed here instead use one-stage or two-stage convolutional detectors, tracking-by-detection with a Kalman filter or SORT-family association step, or recurrent temporal models without a transformer encoder-decoder.

Pixel-level segmentation provides morphometric measurements, including head area, perimeter, and elongation, that bounding-box detection cannot supply, and is a prerequisite for detailed midpiece and flagellum analysis. U-Net [37], with its encoder-decoder structure and skip connections, is the most widely adopted backbone for sperm segmentation.

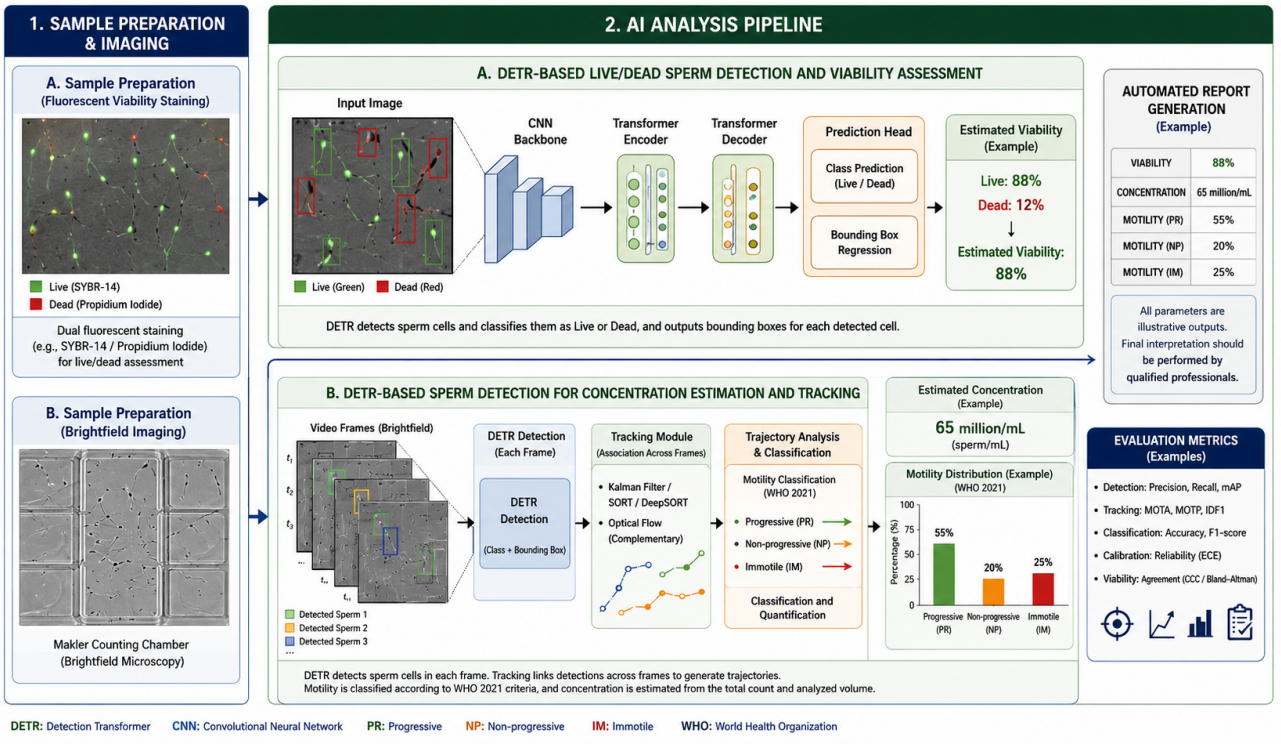

Figure 3: Conceptual example of a DETR-based workflow

Marin and Chang [19] reported that a transfer-learned U-Net achieved Dice coefficients (a measure of pixel-wise mask overlap, defined as twice the intersection divided by the sum of the two mask areas) of 0.96 for the head, 0.94 for the acrosome, and 0.95 for the nucleus on the SCIAN-SpermSegGS dataset; performance for the midpiece and tail was lower, with Dice coefficients of 0.64 and 0.75 reported by Movahed and colleagues [38] for these elongated structures on their own dataset, a gap generally attributed to the mismatch between the extreme aspect ratio of these components and the network receptive field. Mask R-CNN [39] extends detection to instance-level masks and has been used for concurrent localization and per-cell segmentation. Lv and colleagues [40] proposed a dilated U-Net with a modified skip-connection block, trained on 1,207 images from more than 20 patients, and reported a Dice coefficient of 95.14 percent for head segmentation on unstained, low-resolution images, with additional evaluation on an external prostate imaging dataset. Lei and colleagues [20] benchmarked Mask R-CNN, YOLOv8, YOLO11, and U-Net across five anatomical regions on a live, unstained human sperm dataset and found that Mask R-CNN performed best for the head, nucleus, and acrosome on this dataset, while U-Net achieved the highest IoU for the tail; because instance-level segmentation with Mask R-CNN carries substantially higher computational cost, a practical tension exists between the most accurate segmentation approach on this benchmark and real-time clinical deployment. Transformer-based segmentation architectures, including TransUNet [41], have also been applied to sperm head segmentation, although adoption remains comparatively limited. Across the segmentation literature, inconsistent use of the Dice coefficient, IoU, and pixel accuracy, three related but numerically distinct quantities, complicates direct cross-study comparison; Dice and IoU are mathematically related (Dice equals $2 \times \text{IoU}/(1 + \text{IoU})$) but not numerically interchangeable, and a Dice score will always exceed the corresponding IoU value for the same segmentation. This relationship is stated once here and referenced, rather than re-derived, where relevant later in this review. Classical segmentation techniques retain practical value in controlled laboratory settings because of their low computational cost. Global thresholding by Otsu's method [42] performs well on stained brightfield images with a bimodal intensity histogram but is less effective on phase-contrast images, where the halo artifact reverses local contrast relative to the background. Local adaptive thresholding [43] accommodates spatially varying illumination that defeats global methods. Morphological operations and the marker-controlled watershed transform [44] separate touching sperm heads, and active contour and level-set methods [45, 46] refine coarse boundaries with sub-pixel accuracy and accommodate topological changes from touching or merging cells. Deep learning-based segmentation has nonetheless become the preferred approach for heterogeneous clinical data, consistent with its reported robustness advantages over classical methods on the specific comparisons available in the reviewed literature.

4. MORPHOLOGY CLASSIFICATION AND TRANSFER LEARNING

Sperm morphology assessment in the studies reviewed here is typically anchored to the WHO Kruger strict criteria [2], which classify spermatozoa as normal or abnormal according to the shape and dimensions of the head, midpiece, and flagellum. Not every study applies this reference standard directly; where a study instead relies on laboratory-specific criteria or a locally defined consensus among annotators, this is noted explicitly, because "normal" and "abnormal" labels are only comparable across studies when the underlying labeling protocol is comparable. A cell-level "normal morphology" label in these studies should also not be conflated with a patient-level "normal semen analysis" result, which additionally requires concentration and motility to fall within WHO reference ranges [2]; the studies reviewed in this section address the cell-level task only. Because manual application of any morphology criteria is subject to interobserver disagreement, automated classification has attracted sustained research interest. Shaker and colleagues [8] proposed Adaptive Patch-based Dictionary Learning (APDL), constructing class-representative dictionaries from image patches, and reported 92.2 and 62.0 percent classification accuracy on the HuSHeM and SCIAN-Morpho datasets, respectively, using their introduced 216-image HuSHeM dataset; the method required a manual patch-extraction step that limits full automation. Riordon and colleagues [9] fine-tuned a VGG16 network pretrained on ImageNet on the HuSHeM and SCIAN benchmarks, reporting a true positive rate of 94.1 percent that exceeded a cascade ensemble support vector machine (SVM) baseline evaluated on the same data and split. Javadi and Mirroshandel [11] introduced a multiclass classifier evaluated on the 1,540-image MHSMA dataset with a predefined, patient-stratified split, reporting $F_{0.5}$ scores (a variant of the F-score weighting precision more heavily than recall) of 84.74, 83.86, and 94.65 percent for acrosome, head, and vacuole abnormality detection, respectively; because assessment was restricted to still images, motility, which the WHO manual treats as a jointly required parameter, was not evaluated. Keller and colleagues [18] trained a convolutional network on 10,000 annotated boar spermatozoa images acquired by imaging flow cytometry, reporting F1-scores between 96.73 and 99.31 percent across magnification levels and 99.8 percent for acrosome integrity; because this study used a non-human species with morphology criteria and defect categories that differ from human andrology practice, its results describe feasibility in a veterinary and agricultural imaging context and should not be read as evidence for human diagnostic performance. Iqbal and colleagues [47] proposed a lightweight custom architecture and reported 88 percent recall on SCIAN and 95 percent recall on HuSHeM, operating on pre-cropped human sperm images that require prior manual segmentation. Hernandez-Herrera and colleagues [48] applied ResNet50 for flagellum defect classification and U-Net for segmentation on human sperm, reporting a mean F1-score of 0.99 and a mean Dice coefficient of 0.81 on flow-cytometry images; performance on standard brightfield or phase-contrast microscopy was not evaluated in this study. Transfer learning from ImageNet-pretrained networks is the dominant strategy for addressing the limited scale of annotated sperm datasets, which seldom exceed a few thousand images [49]. In addition to the VGG16 transfer-learning result of Riordon and colleagues [9] above, Abbasi and colleagues [50] compared single-task and multitask transfer learning on the MHSMA dataset and found that a multitask VGG19 formulation was associated with higher internal performance for the head and acrosome labels, consistent with the studied inter-label correlations, while the benefit was smaller for the vacuole classification task; this result describes an internal-comparison association on this dataset rather than a demonstrated robustness advantage under distributional change. Liu and colleagues [27] adapted AlexNet with additional batch-normalization layers for four-class WHO morphology classification on HuSHeM, achieving 96.0 percent accuracy with a comparatively lightweight architecture. Spencer and colleagues [15] constructed a stacked ensemble of VGG16, VGG19, a modified ResNet-34, and DenseNet-161, all initialized with ImageNet weights, combined through a meta-classifier; the ensemble achieved an F1-score of 98.2 percent, a recall of 98.5 percent, and a precision of 98.0 percent on HuSHeM, a statistically significant improvement over individual base models on this dataset ($P = 0.005$, Mann-Whitney U test). Marin and Chang [19] additionally reported that, on their dataset, transfer-learned U-Net and Mask R-CNN segmentation models produced lower variance and fewer complete segmentation failures than models trained from randomly initialized weights; this is reported here as an internal-comparison finding on the source dataset, not as a general robustness claim. Because ImageNet pretraining is derived from natural images rather than microscopic cellular structures, the resulting representations may encode textures irrelevant to sperm microscopy, a domain gap that motivates continued interest in domain-adaptive pretraining and self-supervised learning on microscopy-specific corpora. Direct quantitative comparison of reported classification accuracies across studies is constrained by the absence of a universally accepted benchmark dataset and standardized evaluation protocol. Accordingly, this review avoids the phrase "superior performance" except when comparisons are based on the same dataset, data split, endpoint definition, and evaluation protocol. Ilhan and Serbes [51] reported classification accuracies of 90.87%, 92.1%, and 73.2% on the SMIDS, HuSHeM, and SCIAN-Morpho datasets, respectively, using a two-stage fusion of fine-tuned GoogleNet and VGG16. Dobrovolny and colleagues [52] reported a mean average precision of 72.15% for YOLOv5 on the VISEM dataset; this result is not directly comparable to those of detection studies that used different datasets and evaluation metrics. Shahzad and colleagues [53] reported accuracies of 89%, 90%, and 92% for acrosome, head, and vacuole classification, respectively, using a sequential deep neural network on MHSMA. Prabaharan and Saravanan [54] reported an overall accuracy of 98.99% using a three-stage CNN framework with entropic segmentation on an independently collected dataset. Shahali and colleagues [55] reported an accuracy and precision of 94% using an ensemble of DenseNet, ResNet, and Inception models for morphology classification in live, unstained sperm. Aristoteles and colleagues [56] reported an overall mean average precision of 79.58% using YOLOv4 for sperm identification, which is a detection task despite the study's title. Mahali and colleagues [28] proposed a

hybrid model combining a Swin Transformer, MobileNetV3, and an autoencoder, reporting accuracies of 95.4%, 97.6%, and 91.7% on the SVIA, HuSHeM, and SMIDS datasets, respectively. These studies address distinct prediction tasks, datasets, and evaluation protocols. Their results should therefore be interpreted as evidence that multiple architectural families can achieve high performance on internal benchmarks involving curated datasets, rather than as constituting a ranked comparison of algorithms. Where multicenter testing has been performed, differences in staining protocols, microscope configurations, and site-specific acquisition procedures have been shown to substantially degrade performance [13]. Within individual studies that report such comparisons, one-stage detectors offer a favorable balance between accuracy and computational efficiency for near-real-time laboratory workflows, whereas two-stage detectors, such as Faster R-CNN and Mask R-CNN, tend to provide more accurate localization of complex structures at a higher computational cost. The appropriate choice therefore depends on the intended application and available hardware, rather than on accuracy alone.

5. MOTILITY ASSESSMENT AND TRACKING

Sperm motility, defined by the WHO as progressive (PR, meaning progressive motility), non-progressive (NP, meaning non-progressive motility), or immotile (IM, meaning immotile), is among the most clinically consequential parameters in semen analysis, and its manual assessment is similarly subject to interobserver variability [2]. Automated motility assessment addresses both video-based trajectory analysis and direct classification of motility category from video clips. Haugen and colleagues [21] introduced the publicly available VISEM dataset, comprising videos from 85 donors at 400x magnification linked to biological metadata, and used a frame-stacking approach to predict the proportions of progressive, non-progressive, and immotile sperm. Valiuskaite and colleagues [14] presented a region-based convolutional network for sperm head segmentation combined with centroid tracking to estimate movement velocity, reporting 91.77 percent (95 percent confidence interval, 91.11 to 92.43 percent) detection accuracy on VISEM and a Pearson correlation coefficient of 0.969 between predicted and reference vitality measurements. A Pearson correlation quantifies the strength of a linear association between two variables and, on its own, does not establish that two measurements agree in absolute terms; this point is stated once here and applies to every correlation-based result discussed elsewhere in this review. This particular result should therefore be read as evidence of a strong linear relationship rather than as evidence of clinical-grade agreement, which would require a complementary agreement analysis, such as Bland-Altman limits of agreement or a reported mean absolute error, neither of which was provided in this study. Somasundaram and Nirmala [30] combined a Faster R-CNN detector with an Elliptic Scanning Algorithm and a Tail-to-Head Movement Algorithm for simultaneous classification and motility analysis, reporting 97.37 percent detection accuracy with a minimum per-frame processing time of 1.12 seconds; because 1.12 seconds per frame corresponds to fewer than one frame per second, well below the acquisition frame rate used in this and comparable motility studies, this result is read here as evidence of offline algorithmic feasibility rather than of real-time operation. Consistent with the terminology convention adopted throughout this review, “real-time” is reserved for systems demonstrated to meet or exceed the acquisition frame rate under stated hardware and preprocessing conditions, including any tracking or post-processing overhead; other results are described as offline, near-real-time, or low-latency, as appropriate, and are reported together with the hardware and batch conditions stated by the source study where available. Haugen and colleagues [12] subsequently evaluated a ResNet-50 deep convolutional neural network on 65 videos from an external quality-assessment program using tenfold cross-validation, training separate three-class and four-class WHO motility category models and benchmarking predictions directly against multiple reference laboratory assessments; this study is notable among those reviewed for its explicit external quality-assessment design, although it reports coarse categorical motility rather than continuous kinematic parameters such as curvilinear or straight-line velocity that are required for detailed andrological interpretation.

Across the studies reviewed, a transition from handcrafted kinematic feature engineering toward learned, video-level representations is evident, with convolutional architectures forming the dominant framework. Reported performance gains reflect improvements in dataset quality, transfer-learning strategy, and preprocessing in addition to architectural refinement, but direct comparison across studies remains constrained by variation in imaging modality, annotation protocol, dataset scale, and evaluation metric. Where available, calibration analysis, mean absolute error, limits-of-agreement analysis, and clinically relevant decision thresholds, for example the threshold used to flag a sample as oligospermic or asthenozoospermic, would provide a more clinically interpretable characterization of motility-prediction accuracy than correlation alone, and their more consistent reporting is identified as a priority for future work.

Domain shift affects motility estimation somewhat differently than it affects detection, morphology classification, or segmentation. Because motility labels derive from continuous trajectory dynamics rather than a single static image, motility models are particularly sensitive to frame rate and to differences in optical modality between training and deployment sites; a model trained on 50 frame-per-second phase-contrast video, such as VISEM, may not transfer reliably to a lower frame-rate or brightfield acquisition setting, independent of any change in the underlying patient population. This is a form of covariate shift specific to the temporal sampling of the input, distinct from the static-image covariate shift, such as staining or illumination variation, that primarily affects detection, segmentation, and morphology classification.

6. IMAGE PREPARATION AND FEATURE EXTRACTION

The imaging modality used for acquisition materially affects downstream model performance. Brightfield microscopy with Papanicolaou or Diff-Quik staining is the reference standard for WHO morphology assessment because it provides high-contrast visualization of chromatin and acrosomal structure [2], but staining is cytotoxic and therefore unsuitable for live-cell analysis. Phase-contrast microscopy converts the phase shifts produced by unstained specimens into amplitude variations, enabling live-cell motility assessment, and is widely used as an acquisition modality for CASA systems and video-based deep learning approaches [4]; the halo artifact characteristic of phase-contrast optics complicates boundary segmentation and is partially mitigated by digital correction techniques [57]. Fluorescence microscopy with targeted assays, including the terminal deoxynucleotidyl transferase dUTP nick end labeling (TUNEL) assay, supports assessment of DNA fragmentation. Because DNA-fragmentation prediction is not treated as a formal computational task elsewhere in this review, the following is noted briefly rather than as a dedicated subsection: recent work has explored predicting fragmentation from label-free quantitative phase imaging to reduce reliance on fluorescence staining [58, 59], in a body of work in which reactive oxygen species are implicated as a mechanistic contributor to DNA damage [60]. Standardization of acquisition parameters, including magnification, frame rate, and illumination, across laboratories remains an unresolved barrier to multisite model training [4].

Figure 4 summarizes a representative preprocessing and feature-extraction pipeline of the kind used across the studies discussed in this section, spanning contrast enhancement, denoising, illumination correction, morphological and texture feature extraction, and both classical and deep learning-based segmentation.

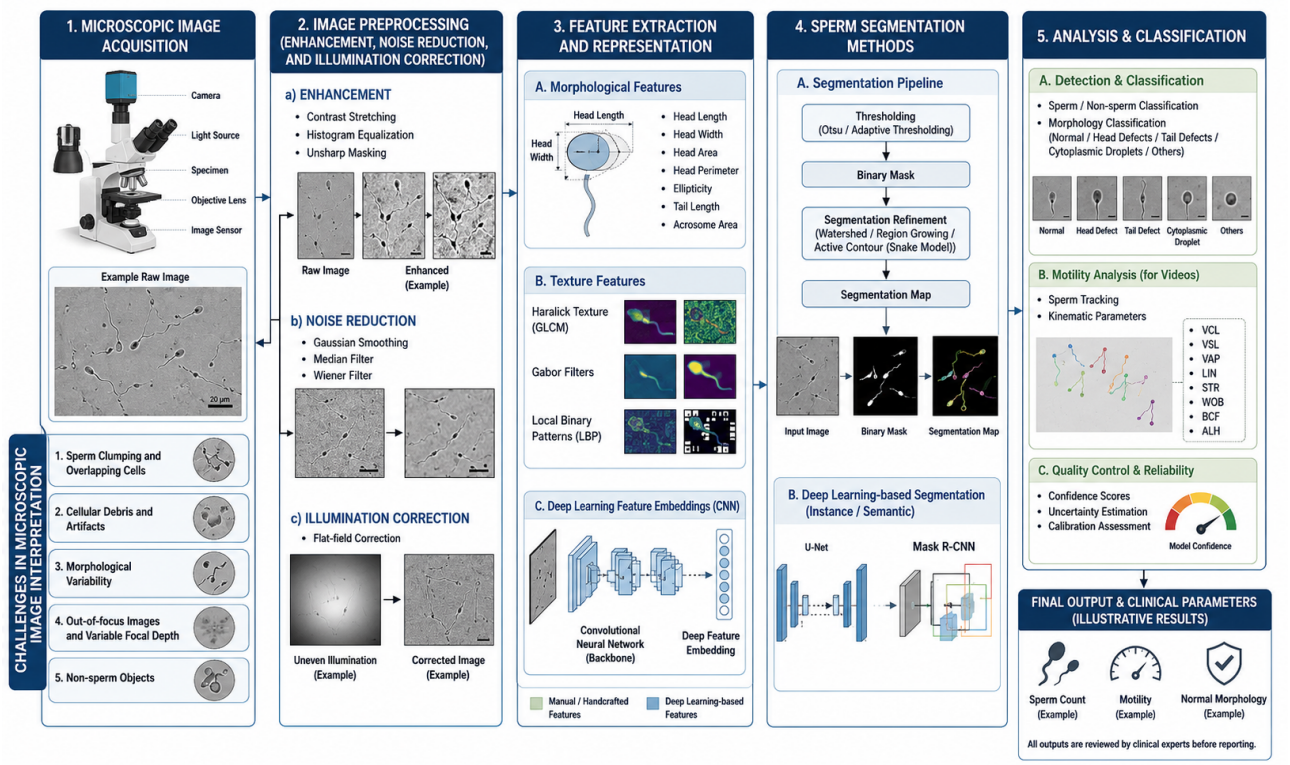

Figure 4: Representative preprocessing, feature-extraction, and segmentation pipeline for microscopic sperm image analysis

Preprocessing pipelines commonly include contrast enhancement, denoising, and illumination correction prior to detection or segmentation. Contrast-limited adaptive histogram equalization [61] improves local contrast while limiting noise amplification, and Gaussian or median filtering is applied depending on whether the dominant noise source is Gaussian or impulsive. Background subtraction removes slowly varying illumination from phase-contrast images [62]. Deep denoising networks such as DnCNN [63], originally developed for general-purpose image restoration, have been proposed for adaptation to biomedical microscopy because they preserve fine structural boundaries more effectively than conventional filters in comparisons reported in the general image-restoration literature. Direct ablation evidence isolating the contribution of individual preprocessing steps to downstream detection or segmentation accuracy is limited in the reviewed sperm-analysis literature; the studies cited above did not report formal ablation experiments for contrast normalization or background subtraction, so the expectation that omitting these steps degrades performance in low-contrast or phase-contrast imaging should be regarded as a plausible inference transferred from the broader image-processing literature, rather than as a result directly demonstrated within this review's sperm-analysis evidence base. Wang and colleagues [13] instead provided direct

empirical evidence, through internal and external multicenter testing, that the diversity of imaging conditions represented in the training dataset is a stronger determinant of cross-site generalizability than any single preprocessing step.

Feature representation converts image data into compact descriptors of morphology or kinematics. Geometric features, including head area, perimeter, width-to-length ratio, and acrosome area fraction, are computed from segmentation masks and correspond directly to WHO morphology criteria [2]. Texture descriptors, including local binary patterns [64] and Gabor filter responses [65], characterize chromatin condensation and acrosomal texture. Kinematic features derived from centroid trajectories, including curvilinear velocity, straight-line velocity, average path velocity, linearity, straightness, wobble, beat-cross frequency, and amplitude of lateral head displacement, have traditionally supported WHO motility classification and continue to complement deep learning-based prediction. Deep feature embeddings extracted from convolutional layers encode a hierarchy from low-level edges and textures to morphology-specific semantic patterns; Abbasi and colleagues [50] reported that such learned representations outperformed handcrafted feature baselines for morphology classification on MHSMA in their internal comparison, consistent with the broader shift toward learned representations documented throughout this review.

7. PATTERN RECOGNITION AND GENERATIVE AUGMENTATION

Classical pattern-recognition methods retain a role in sperm analysis, particularly for interpretable, feature-based classification. Support vector machines with radial-basis-function kernels have been widely applied to handcrafted morphology and motility features; Ilhan and colleagues [66] reported accuracies of 80.5 and 83.8 percent for wavelet-based and descriptor-based features, respectively, using this approach on their dataset. Goodson and colleagues [67] proposed CASAnova, a multiclass support vector machine that sequentially applies decision hyperplanes to classify sperm motility patterns into one of five classes, reporting an overall accuracy of 89.9 percent. Random forest classifiers offer internal feature-importance estimates that support clinical interpretation and have been proposed for motility classification from kinematic feature vectors. Recurrent architectures, and long short-term memory networks in particular, have been increasingly adopted for motility analysis because they capture long-range temporal dependencies in sperm trajectories while mitigating the vanishing-gradient limitation of standard recurrent networks.

Generative adversarial networks (GANs) [68] have been applied primarily as a data-augmentation strategy to address the pronounced class imbalance characteristic of annotated sperm datasets, in which morphologically abnormal subclasses are typically underrepresented relative to normal forms. A recognized limitation of conventional generative adversarial augmentation is mode collapse, in which the generator converges to a narrow range of near-identical synthetic images and fails to capture intraclass morphological variability. Cansiz and colleagues [29] proposed a loss-based ensemble generative adversarial network (LB-EGAN) that fuses two structurally distinct generator architectures based on discriminator loss, reporting that ensemble-based generative augmentation improved downstream classifier performance on minority morphological defect classes in their data-scarce internal setting; downstream generalizability to external data was not evaluated in this study.

Domain-specific challenges continue to complicate microscopic image interpretation, and these challenges affect the principal analytical tasks in distinct ways. Low contrast and halo artifacts around the sperm head in phase-contrast imaging, partially mitigated by digital halo-removal filtering [69], primarily degrade segmentation boundary accuracy rather than coarse detection or classification confidence. Overlapping and touching cells in high-concentration samples produce ambiguous pixel regions that instance-aware segmentation methods, including Mask R-CNN, address only partially; segmentation accuracy has been shown to degrade substantially in high-density fields, including intracytoplasmic sperm injection microscopy [70], whereas detection recall is comparatively more robust to overlap because a bounding box tolerates partial occlusion better than a pixel-accurate mask. Debris and non-sperm elements, such as white blood cells and epithelial cells, require multiclass detectors capable of distinguishing sperm from impurities, and the relative rarity of white blood cells in typical samples introduces label-shift-like class imbalance during training that is distinct from the covariate shift caused by staining variation [71]. Limited depth of field causes cells at different focal planes to appear with variable sharpness, a limitation that three-dimensional z-stacking or extended depth-of-field synthesis can address but that is not standard in routine clinical workflows [72]; this factor primarily affects morphology classification, since a defocused head can be misclassified as morphologically abnormal. Finally, variation in staining protocol, dye concentration, and slide preparation across laboratories constitutes a covariate shift that necessitates staining normalization for models intended for multisite deployment [13], and, because staining protocol also shapes how “normal” and “abnormal” are visually defined at the point of annotation, this source of variation can additionally introduce a degree of concept shift, when models trained under one staining and labeling convention are applied under another.

8. HYBRID DEEP-LEARNING AND MACHINE-LEARNING SYSTEMS

Because hybrid systems are a central focus of this review, the approaches identified in the preceding sections are consolidated here and classified according to the four-category taxonomy introduced above. **Table 3** maps each study to its hybridization strategy, component models, fusion stage, analytical task, comparator, dataset, validation design, and reported incremental benefit. This comparison is intended to clarify what each hybrid configuration contributes beyond its constituent or single-architecture baseline, while distinguishing genuine performance gains from improvements reported only under internal or incompletely specified validation conditions.

Table 3: Hybrid studies mapped onto the four-category taxonomy

Study	Operational category	Component models	Fusion location	Task	Comparator (same split)	Dataset	Validation design	Reported incremental benefit
Ilhan & Serbes [51]	Architectural fusion	Fine-tuned GoogleNet + VGG16	Late (prediction-level)	Morphology classification	Individual GoogleNet and VGG16 baselines	SMIDS, HuSHeM, SCIAN-Morpho	Internal, unit not reported	Higher accuracy than either constituent model on the same split
Yuzkat et al. [24]	Architectural fusion	Single-stage + two-stage detectors	Late (prediction-level)	Sperm vs. non-sperm detection	Individual single- and two-stage detectors	Custom labeled dataset	Internal, unit not reported	Highest overall mAP among compared configurations
Mahali et al. [28]	Feature-level fusion + pipeline hybridization	Swin Transformer + MobileNetV3 + autoencoder	Feature-level, pre-classification head	Morphology classification	Constituent single backbones (partial)	SVIA, HuSHeM, SMIDS	Internal, unit not reported; 3 datasets	Consistent accuracy across three datasets relative to reported single-backbone results
Hernandez-Herrera et al. [48]	Pipeline-level hybridization	U-Net (segmentation) → ResNet50 (classification)	Sequential	Flagellum defect classification	Not a single-model comparator; sequential design only	Flow-cytometry human sperm	Internal, unit not reported	Not evaluated against a single-model baseline on the same split
Ilhan et al. [66]	Pipeline-level hybridization	MobileNet features + conventional classifier	Sequential (feature extractor → classifier)	Detection and classification	Conventional handcrafted-feature classifiers	Internal dataset	Internal, unit not reported	Higher accuracy than conventional-feature baselines in the source comparison
Cansiz et al. [29]	Generative augmentation	Two structurally distinct GAN generators, loss-guided fusion	Discriminator-loss-guided parameter fusion	GAN-based augmentation for morphology classification	Single-generator GAN augmentation baseline	Sperm morphology datasets	Internal, unit not reported	Improved downstream minority-class performance versus single-generator augmentation on the source dataset

Across the reviewed studies, the evidence suggests that hybridization is most defensible when its constituent components address complementary failure modes, for example combining a detector optimized for small or occluded objects with a classifier optimized for fine-grained morphological discrimination, and when the source study includes an appropriate single-model baseline evaluated on the same split, rather than when two similar architectures are combined primarily to obtain an incremental accuracy gain without such a baseline. An ensemble that improves overall accuracy by pooling architecturally similar convolutional networks, as in the four-model stacked ensemble reported by Spencer and colleagues [15], demonstrates the value of variance reduction through model averaging within that study's own comparison; the CNN-transformer hybrids of Prangemeier and colleagues [35] and Mahali and colleagues [28] instead illustrate the value of combining architecturally distinct inductive biases (local convolutional feature extraction and global self-attention). A hybrid label is therefore not, by itself, evidence of superior performance. In particular, the pipeline-level design of Hernandez-Herrera and colleagues [48] does not report a single-model baseline on the same split; consequently, no incremental benefit attributable specifically to hybridization can be established. The reviewed approaches are best understood as three distinct design patterns: architecturally novel fusions that combine complementary inductive biases, ensembles that primarily reduce variance, and sequential pipeline compositions whose added value remains uncertain without a matched baseline. Neither fusion-based nor pipeline-based hybrids have yet been evaluated in a sufficiently broad multicenter setting to determine whether their additional architectural complexity produces a proportionate gain in real-world robustness.

9. DATASETS, BENCHMARKING, AND PERFORMANCE METRICS

The scale, annotation quality, and acquisition characteristics of available datasets directly constrain model generalizability and the validity of cross-study comparison. The most frequently used public benchmarks in the reviewed literature are HuSHeM, a 216-image, four-class human head-morphology dataset stained by the Diff-Quik method [73]; SCIAN-MorphoSpermGS, a gold-standard human dataset of 1,854 expert-validated head images spanning five morphological categories [74]; MHSMA, a grayscale human dataset of 1,540 images from 235 infertile patients with four annotated morphological regions and predefined training, validation, and test partitions [11]; VISEM, a multimodal human video dataset from 85 donors

linking motility and biological metadata [21]; SVIA, a large-scale human dataset exceeding 278,000 annotated objects spanning detection, tracking, and denoising subsets [75]; and SMIDS, a 3,000-image, three-class human dataset acquired through smartphone-based imaging [76]. Additional datasets used in individual studies include the Jikei Sperm Dataset of 4,625 unstained human images from Japanese donors [7] and the Somasundaram and Nirmala dataset of 3,200 human images across four morphological classes [30]. The non-human datasets discussed elsewhere in this review, namely the boar imaging-flow-cytometry dataset of Keller and colleagues [18] and the bull semen video dataset of Hidayatullah and colleagues [23], because biological structure, imaging conditions, and intended clinical use differ between species; animal-model results are treated in this review as evidence of algorithmic feasibility only and do not contribute to any statement about human diagnostic, clinical, or regulatory evidence. Several smaller private collections acquired by quantitative phase imaging, digital holographic microscopy, or image cytometry are also represented in the literature [77, 78, 79, 80, 81].

Earlier, smaller-scale human datasets remain relevant as historical benchmarks against which later, larger public releases can be compared: a 283-cell, 100-slide Papanicolaou-stained dataset acquired at 1000x magnification supported early evaluation of foundational morphological algorithms [82]; a 160-image, two-class dataset of isolated sperm heads shared by Tseng and colleagues became an early reference for computer-aided morphology classification [83]; and a 19-image, 264-cell gold-standard dataset with manually generated ground-truth masks for the head, nucleus, and acrosome has been used to benchmark detection, segmentation, and tail-point localization algorithms [84]. These early, small-scale resources illustrate the incremental growth in dataset scale that has accompanied the field's transition from classical to deep learning-based methods, but their limited size also means that models validated exclusively on them should not be assumed to generalize to larger or more heterogeneous clinical populations. Licensing and reuse terms are also inconsistently reported across the datasets summarized above: several public releases, including HuSHeM, VISEM, and SVIA, are distributed through open repositories with stated reuse conditions, whereas earlier datasets such as the Papanicolaou-stained and gold-standard segmentation collections were typically shared informally between research groups without a documented licensing framework, a further, largely overlooked barrier to standardized reuse and benchmarking.

Table 4 consolidates the species, imaging modality, staining condition, annotation unit, and public or private status of the principal datasets discussed in this section, so that the species- and modality-specific caveats raised throughout this review can be checked against a single reference rather than reconstructed from prose.

Table 5 summarizes descriptive quantitative patterns across the studies included in this review. These descriptive patterns were compiled by the reviewing authors during full-text coding according to the task and architecture categories defined in this review, applied to this review's own inclusion set rather than to an independently audited or externally validated bibliometric database; exact per-study numerators and the precise total denominator are not separately tabulated here, so the values should be read as an approximate descriptive summary of the literature captured by this particular search, not as a precise census of the field, and are not repeated elsewhere in this review as a marker of completeness.

Several considerations limit the interpretability of metric comparisons across studies and are made explicit here as a task-specific reporting framework, because accuracy, precision, recall, F1-score, mAP, AP_{50}, the Dice coefficient, Pearson correlation, and processing time are not directly interchangeable and several are task-dependent. Metric notation is standardized throughout this review as follows: mAP and AP_{50} are reported as percentages with the IoU threshold stated; IoU and Dice are reported as proportions between 0 and 1 unless the source publication reports a percentage, in which case that convention is preserved and noted; F1-score, precision, recall, and accuracy are reported as percentages; and correlation coefficients are reported as unitless values between -1 and 1.

For morphology classification, overall accuracy alone is an unreliable summary when the evaluation set is imbalanced, which is the norm rather than the exception in this literature given that normal forms are typically a minority class in samples from infertile patients (this prevalence pattern is a transferred clinical observation rather than a statistic independently re-derived by this review). Class-wise sensitivity and specificity, balanced accuracy, and macro-averaged F1-score, which weight each class equally regardless of its frequency, are preferable to a single pooled accuracy figure for characterizing performance on minority defect categories, and confidence intervals around these estimates are informative given the small test sets, often a few hundred images or fewer, that characterize much of this literature.

For detection, a mean average precision value is only interpretable alongside the IoU threshold at which it was computed, the matching rule used to associate a predicted box with a ground-truth box, the class definitions used, and, where relevant, the confidence threshold applied before precision and recall were computed. A study reporting mAP at a single IoU threshold of 0.50 (AP_{50}) is not directly comparable to one reporting mAP averaged across IoU thresholds from 0.50 to 0.95 in the manner used by the COCO benchmark, and studies applying different thresholds, matching rules, or confidence operating points should not be ranked against one another even when both report a single mAP figure.

For segmentation, the Dice coefficient and IoU are not interchangeable, and a study reporting only one of the two cannot be compared on that basis with a study reporting the other. Where fine sub-cellular morphology is clinically relevant, such as acrosome integrity or flagellum continuity, a pixel-overlap metric such as Dice or IoU may remain high even when the predicted boundary is locally inaccurate in the specific region of clinical interest.

Table 4: Dataset characteristics for the principal public and named private datasets discussed in this review

Dataset	Species	Modality	Staining	Size	Annotation unit	Status
HuSHeM [73]	Human	Brightfield	Diff-Quik	216 images	Whole-head, 4 morphology classes	Public (Mendeley Data)
SCIAN-MorphoSpermGS [74]	Human	Brightfield	Stained	1,854 images	Whole-head, 5 morphology classes, 3-expert consensus	Public
MHSMA [11]	Human	Brightfield	Unstained, low-resolution	1,540 images; 235 patients	Region-level (acrosome, head, vacuole, tail)	Public, predefined split
VISEM [21]	Human	Phase-contrast video, 400x	Unstained	85 donor videos	Video-level motility proportions; donor metadata	Public
SVIA [75]	Human	Video and image, mixed	Mixed	>278,000 objects	Object-level (detection, tracking, denoising subsets)	Public
SMIDS [76]	Human	Smartphone brightfield	Stained	3,000 images	Whole-image, 3 classes (normal, abnormal, non-sperm)	Public
Jikei Sperm Dataset [7]	Human	Clinical ICSI video	Unstained	4,625 images	Frame-level, focus-quality groups	Private (clinical)
Somasundaram–Nirmala [30]	Human	Video, 30 fps	Unstained	3,200 images	Whole-image, 4 morphology classes	Private
Boar IFC dataset [18]	Boar (non-human)	Imaging flow cytometry	Unstained	10,000 images	Whole-cell, species-specific defect categories	Private
Bull semen video [23]	Bull (non-human)	Brightfield video	Unstained	Not specified per-cell	Frame-level detection boxes	Private
Papanicolaou dataset [82]	Human	Brightfield, 1000x	Papanicolaou	283 cells, 100 slides	Whole-cell	Private, historical
Gold-standard segmentation [84]	Human	Brightfield	Stained	264 cells, 19 images	Pixel-level (head, nucleus, acrosome masks)	Private, historical

Table 5: Descriptive landscape of the studies included in this review

Category	Approximate distribution across the included studies
Model family	Convolutional (detection, classification, segmentation) predominates; transformer or DETR-family, hybrid convolutional-transformer or convolutional-classical, GAN-based augmentation, and classical machine learning-only approaches together account for a smaller but non-trivial and growing share
Primary task	Morphology classification and detection or localization are the most frequently addressed tasks, followed by motility or tracking, segmentation, and multitask or combined approaches
Most-used public datasets	HuSHeM, SCIAN-MorphoSpermGS, VISEM/VISEM-Tracking, and MHSMA are the most frequently reused benchmarks, followed by SVIA and SMIDS
Most-reported metrics	Accuracy and precision/recall are most frequently reported, followed by F1-score, mean average precision, Dice coefficient, and area under the receiver operating characteristic curve (AUC-ROC)
Publication trend	Publication volume has increased across the review's search window, with the largest share of included studies published from approximately 2023 onward, reflecting accelerating adoption of transformer-based and hybrid architectures

For motility and concentration estimation, a correlation coefficient demonstrates association but not agreement [14]. Bland-Altman limits-of-agreement analysis, mean absolute error, calibration analysis, and evaluation against a clinically relevant decision threshold, for example the WHO reference threshold used to flag a sample as oligozoospermic or asthenozoospermic, would each provide stronger evidence of clinical utility than correlation alone, and their reporting is recommended wherever feasible.

Finally, across all four tasks, image-level, cell-level, frame-level, and patient-level evaluation are not equivalent, and this distinction compounds the validation-granularity concern. A model achieving high image-level or cell-level accuracy may perform less consistently when evaluated at the level of the individual video frame or, most importantly for clinical use, the individual patient, and patient-level external validation is reported infrequently in the reviewed literature. For these reasons, this review avoids characterizing any single architecture as superior in an unqualified sense and instead reports comparisons only where the same dataset, split, and evaluation protocol were used within a given study.

10. CHALLENGES AND FUTURE DIRECTIONS

10.1 Dataset standardization and scale

The limited availability of large, standardized, and publicly released annotated datasets remains a principal constraint on the field. Most published studies rely on private clinical data that is not released publicly, which renders inter-study comparison largely irreproducible. Existing public morphology benchmarks typically contain a few thousand annotated images at most, several orders of magnitude smaller than datasets used to train comparable models in general computer vision (a transferred comparison rather than a claim specific to any single cited study), which constrains generalizability across patient populations and imaging conditions. Model generalizability is further affected by demographic diversity, imaging and microscope variation, staining protocol, and annotation quality; many public datasets originate from a single population or species, and differences in magnification, illumination, and staining can substantially alter the visual appearance of sperm across datasets. Federated learning, an approach that permits collaborative model training across distributed clinical centers without centralizing raw patient data, is a plausible path toward greater statistical diversity while respecting data-governance constraints; this is offered here as an author recommendation, since federated learning has seen limited application within the sperm-analysis literature reviewed here.

10.2 Label quality and interobserver agreement

Although WHO morphology criteria are widely accepted clinically, empirical studies in the broader andrology literature have reported only moderate interobserver kappa agreement among expert andrologists, and models trained on single-annotator labels risk encoding systematic interpretive biases that constrain cross-site generalizability; this is a transferred observation from the broader quality-control literature rather than a finding of the AI studies reviewed here specifically. Consensus-based annotation protocols, structured disagreement resolution, probabilistic soft-label approaches that represent annotation uncertainty, and noise-robust training techniques are author-recommended methodological directions that could improve label reliability. Because high-quality pixel-level annotation requires expert embryologists and andrologists, dataset construction remains time-consuming and costly, which in turn constrains the scale of data available for training and limits cross-study reproducibility.

10.3 Statistical reporting standards

A further, closely related barrier to synthesis is inconsistent statistical reporting. Among the studies reviewed, confidence intervals accompanied a reported performance metric in only a minority of cases, exemplified by the 95 percent confidence interval reported by Valiuskaite and colleagues [14] and the cross-validated, externally benchmarked design of Haugen and colleagues [12]; formal significance testing against a comparator model, as performed by Spencer and colleagues [15] using the Mann-Whitney U test, was reported even less frequently. Single-point accuracy or mean average precision estimates without an accompanying measure of uncertainty make it difficult to judge whether an apparent difference between two architectures reflects a genuine effect or sampling variability, particularly for the many studies evaluated on test sets of a few hundred images or fewer. Consistent reporting of confidence intervals, variance across cross-validation folds, and, where a comparator model is evaluated on identical data, an explicit statistical test, is an author recommendation that would substantially strengthen the evidentiary basis for architectural claims in future work.

10.4 Class imbalance

Morphologically normal spermatozoa are reported to constitute a minority of cells in samples from infertile patients in the broader andrology literature, and specific rare defect categories, including coiled-flagellum and pin-head morphology, occur at frequencies that make reliable data-driven classification difficult without targeted sampling or generative augmentation.

10.5 Real-time and point-of-care deployment

Deep learning models with large parameter counts may be difficult to deploy efficiently on resource-constrained laboratory hardware without model optimization or compression. Knowledge distillation, structured pruning, and low-precision quantization are reported in the general computer-vision literature to achieve substantial reductions in model size with limited accuracy loss; this is a transferred claim from that broader literature rather than a result demonstrated specifically for sperm-analysis models in the studies reviewed here, and lightweight architectures such as MobileNetV3 and EfficientNet-Lite have demonstrated real-time throughput, by the frame-rate-matched definition adopted in this review, for sperm detection on low-power devices in the cited studies. Preliminary evidence from smartphone-based analyzers combining low-cost optics with on-device inference has been reported to suggest that point-of-care concentration and motility estimation may fall within WHO-acceptable variability limits in specific reported comparisons [10]; this evidence base remains limited and has not been established under the multicenter conditions that would be required for a general clinical-equivalence claim. Processing speed alone does not establish clinical usefulness; sample preparation, focus control, operator interaction, quality assurance, and integration with laboratory information systems remain largely unresolved for point-of-care deployment.

10.6 Clinical translation: intended use and system categories

Discussion of clinical translation is more precise when it distinguishes among the categories of system represented in the reviewed literature, because the evidentiary and regulatory bar differs substantially across them. The great majority of the studies reviewed here describe research prototypes: models trained and evaluated by an academic or engineering research group, typically on a single internal dataset, without integration into a laboratory workflow or a stated intended use, corresponding to internal technical validation. A smaller number of studies, including the multicenter generalizability study of Wang and colleagues [13] and the external quality-assessment benchmark of Haugen and colleagues [12], provide external analytical validation, evaluated within or across specific clinical laboratory settings with a defined analytical task. None of the studies reviewed here reports prospective clinical or workflow validation, meaning none describes a system operating as a decision-support tool embedded in a clinical workflow with defined human-in-the-loop review, and none describes an autonomous diagnostic device intended to issue a clinical result without expert review. This distinction matters because a research prototype demonstrating strong internal-benchmark accuracy provides substantially weaker evidence for clinical deployment than a decision-support tool validated prospectively within a laboratory workflow, which in turn provides weaker evidence than an autonomous device validated for unsupervised diagnostic use; conflating these categories, or describing prototype-stage results using language that implies deployment readiness, risks overstating what the current evidence supports. The intended clinical decision that a proposed system aims to support, and the outcome that would constitute meaningful utility for that decision, for example a change in treatment selection, a reduction in time to result, or improved concordance with reference laboratory readings, is rarely specified in the studies reviewed here and is recommended as a standard element of future reporting.

10.7 Clinical translation: analytical, clinical, and regulatory validity

Clinical deployment requires validation beyond benchmark accuracy across three complementary stages: analytical validity, which demonstrates accurate measurement of semen parameters within an acceptable margin of uncertainty; clinical validity, which demonstrates an association between measured parameters and reproductive outcomes such as natural conception or in vitro fertilization success; and clinical utility, which demonstrates that the system improves clinical decision-making relative to standard practice. Most studies reviewed here address analytical validity by comparing model output to expert annotation; evidence for clinical validity and utility is comparatively limited, and high algorithmic accuracy on a benchmark dataset should not be equated with clinical benefit, particularly where predictions correlate only weakly with patient outcomes or carry systematic biases. A robust clinical validation framework, offered here as an author recommendation, would include multicenter evaluation across diverse laboratories, prospective study designs with prespecified endpoints, comparison against the WHO laboratory reference standard [2], assessment of intra- and inter-laboratory reproducibility, and adequately powered sample sizes.

Regulatory requirements for AI-based diagnostic systems vary substantially by jurisdiction and by the intended use of the specific system; the following is general orientation rather than exhaustive or jurisdiction-specific legal guidance, and readers seeking to deploy a specific system should consult the applicable national or regional regulatory authority

directly. The regulatory pathway cannot be determined from algorithm type alone: it depends jointly on the stated intended use and clinical claims, the jurisdiction or jurisdictions of deployment, the precise software function performed, the deployment model, and whether the system is positioned as research-use-only, as a laboratory-developed test used solely within the laboratory that created it, or as a commercially distributed product; a system that changes any one of these factors, for example moving from a single-laboratory tool to a multi-site commercial product with the same underlying algorithm, can change the applicable pathway without any change to the model itself. Several elements of a responsible deployment program are stable across jurisdictions regardless of specific regulatory classification and are largely absent from the studies reviewed here: intended use and risk classification, stated explicitly and matched to the evidentiary bar required; a documented quality management system and change-control procedure for model updates; analytical and clinical performance data appropriate to that intended use; defined human oversight, meaning a mechanism by which a qualified laboratory professional reviews and can override a model's output before it informs a clinical decision; failure detection and calibrated uncertainty estimation, for example via temperature scaling, isotonic regression, Monte Carlo dropout, or deep ensembles, so that low-confidence or out-of-distribution predictions are flagged for expert review rather than returned with unwarranted confidence; cybersecurity and data governance appropriate to the sensitivity of reproductive health data; human-factors and workflow-integration assessment, including operator burden and compatibility with existing laboratory information systems; and post-market surveillance and drift monitoring that continues for the operational lifetime of the system. None of the studies reviewed here reports having pursued a full regulatory pathway of this kind, consistent with the observation above that the current literature consists predominantly of internal technical validation. Explainability techniques, including Grad-CAM and its variants, integrated gradients, and model-agnostic methods such as LIME and SHAP, can support verification that predictions rely on biologically relevant structures rather than artifacts or debris, and can contribute usefully to failure detection and auditability, but none of these techniques by itself establishes analytical validity, clinical validity, clinical utility, or regulatory compliance; explainability is one input to a broader validation and governance program, not a substitute for it.

10.8 External validation and domain shift

Models that achieve high performance on internal test data frequently show reduced accuracy when evaluated on external data from different clinical sites, a pattern attributable to differences in patient population, imaging protocol, microscope configuration, and staining method, as documented directly by Wang and colleagues [13]. This performance degradation can arise from covariate shift, in which the distribution of input images differs between sites while the underlying task remains the same; from label shift, in which the prevalence of morphological or motility categories differs across populations; or from concept shift, in which the relationship between image features and the underlying diagnostic label itself varies because different laboratories apply different local conventions for borderline cases. These three sources of shift affect the four principal analytical tasks reviewed here unevenly. Detection is comparatively robust to label shift, because localizing a sperm cell does not depend on its morphological category, but is sensitive to the covariate shift introduced by staining and optical differences. Morphology classification is sensitive to all three forms of shift, consistent with the comparatively larger multicenter performance drop reported for classification tasks relative to detection in the generalizability literature. Motility estimation is additionally sensitive to the temporal covariate shift. Segmentation performance is most directly affected by covariate shift in staining and focus quality. Distinguishing among these sources of domain shift is useful because they call for different mitigation strategies, ranging from image-level normalization and domain adaptation for covariate shift, to recalibration or reweighting for label shift, to harmonized annotation protocols and consensus labeling for concept shift. External and multicenter validation, together with well-calibrated confidence estimates that permit low-confidence predictions to be flagged for expert review, is necessary to establish real-world robustness, and future studies are encouraged to report external validation, calibration assessment, and uncertainty alongside standard performance metrics.

10.9 Emerging directions

Foundation models for image segmentation, including the Segment Anything Model and its biomedical adaptations, offer a potential means of reducing the annotation burden associated with sperm morphology datasets, and self-supervised learning methods that exploit large volumes of unlabeled microscopy data may improve robustness in data-scarce settings; these are author recommendations for future work rather than results demonstrated in the sperm-analysis studies reviewed here. Vision Transformer architectures represent a further alternative to convolutional backbones, although their application to sperm analysis specifically remains comparatively limited relative to general biomedical imaging within the studies reviewed here. Multimodal approaches that integrate microscopy images with motility measurements, laboratory biomarkers, and clinical metadata may support more comprehensive semen evaluation, though this direction remains largely unexplored within the studies reviewed here. Current approaches predominantly treat detection, morphology classification, and motility assessment as separate tasks, whereas clinical reporting requires concentration, motility distribution, and normal-form percentage jointly from a single video sequence; multitask architectures sharing a common encoder with task-specific prediction heads offer a principled route toward this integration, and prompt-specified foundation models pretrained on

large-scale microscopy data may eventually support generalization to novel imaging conditions with reduced retraining burden, though this remains a prospective rather than demonstrated capability within the current literature.

11. CONCLUSION

Automated semen analysis has evolved from rule-based CASA methods toward deep-learning, transformer-based, and hybrid systems capable of addressing sperm detection, segmentation, morphology classification, motility assessment, and tracking. Although the literature demonstrates substantial technical activity and frequently reports high performance on internal benchmarks, it does not yet establish consistent clinical readiness. Most of the available evidence represents internal technical validation rather than external analytical validation or prospective clinical validation. The principal barriers are not architectural alone. They include limited and heterogeneous datasets, uncertain partition granularity, inconsistent reference standards, incomplete reporting of uncertainty, scarce external validation, and limited evidence of clinical utility. Hybrid systems are most promising when their components provide complementary capabilities and their incremental value is demonstrated against appropriately matched single-model baselines. The term "hybrid" should not, by itself, be interpreted as evidence of superior performance. Future studies should prioritize patient- or donor-level leakage-aware evaluation, multicenter external testing, calibration and agreement analyses, transparent reporting of annotation quality, and prospective assessment within defined laboratory workflows. Until analytical, clinical, workflow, and regulatory requirements have been demonstrated for a specific intended use, current systems should generally be described as research prototypes or decision-support candidates. This review identifies requirements for responsible translation but does not itself establish the explainability, generalizability, or deployment readiness of any specific system.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. Material preparation, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Ethics declaration: not applicable. This study did not involve human participants or animals. Therefore, ethical approval and consent to participate are not applicable.

CONSENT FOR PUBLICATION

Consent to Publish declaration: not applicable.

DATA AVAILABILITY

No new primary clinical datasets were generated during this study. This work is based on a structured review and comparative analysis of previously published literature.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the intellectual contribution of the broader interdisciplinary literature reviewed in this article. No external funding was received for this study. The authors also acknowledge the use of ChatGPT-5 for assistance in refining the manuscript.

FUNDING

No funding.

DISCLOSURE STATEMENT

The authors declare no conflict of interest. The article is based on a structured review of published literature and did not involve human participants, personal data collection, or experimental intervention.

REFERENCES

[1] A. Agarwal, A. Mulgund, A. Hamada, and M. R. Chyatte, "A unique view on male infertility around the globe," *Reproductive biology and endocrinology*, vol. 13, no. 1, p. 37, 2015.

[2] W. H. Organization *et al.*, *WHO laboratory manual for the examination and processing of human semen*. World Health Organization, 2021.

[3] K. Siddharth, T. Kumar, and M. Zabihullah, "Interobserver variability in semen analysis: findings from a quality control initiative," *Cureus*, vol. 15, no. 10, 2023.

[4] 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, p. 545, 2015.

[5] J.-w. Choi, L. Alkhoury, L. F. Urbano, P. Masson, M. VerMilyea, and M. Kam, "An assessment tool for computer-assisted semen analysis (casa) algorithms," *Scientific reports*, vol. 12, no. 1, p. 16830, 2022.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[7] T. Sato, H. Kishi, S. Murakata, Y. Hayashi, T. Hattori, S. Nakazawa, Y. Mori, M. Hidaka, Y. Kasahara, A. Kusuhara, *et al.*, "A new deep-learning model using yolov3 to support sperm selection during intracytoplasmic sperm injection procedure," *Reproductive Medicine and Biology*, vol. 21, no. 1, p. e12454, 2022.

[8] F. Shaker, S. A. Monadjemi, J. Alirezaie, and A. R. Naghsh-Nilchi, "A dictionary learning approach for human sperm heads classification," *Computers in biology and medicine*, vol. 91, pp. 181–190, 2017.

[9] J. Riordon, C. McCallum, and D. Sinton, "Deep learning for the classification of human sperm," *Computers in biology and medicine*, vol. 111, p. 103342, 2019.

[10] H. O. Ilhan and N. Aydin, "Smartphone based sperm counting-an alternative way to the visual assessment technique in sperm concentration analysis," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 6409–6435, 2020.

[11] S. Javadi and S. A. Mirroshandel, "A novel deep learning method for automatic assessment of human sperm images," *Computers in biology and medicine*, vol. 109, pp. 182–194, 2019.

[12] 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. 14777, 2023.

[13] J. Wang, Y. Jin, A. Jiang, W. Chen, G. Shan, Y. Gu, Y. Ming, J. Li, C. Yue, Z. Huang, *et al.*, "Testing the generalizability and effectiveness of deep learning models among clinics: sperm detection as a pilot study," *Reproductive Biology and Endocrinology*, vol. 22, no. 1, p. 59, 2024.

[14] V. Valiuškaitė, V. Raudonis, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Deep learning based evaluation of spermatozoid motility for artificial insemination," *Sensors*, vol. 21, no. 1, 2021.

[15] L. Spencer, J. Fernando, F. Akbaridoust, K. Ackermann, and R. Nosrati, "Ensembled deep learning for the classification of human sperm head morphology," *Advanced Intelligent Systems*, vol. 4, no. 10, p. 2200111, 2022.

[16] S. Zou, C. Li, H. Sun, P. Xu, J. Zhang, P. Ma, Y. Yao, X. Huang, and M. Grzegorzek, "Tod-cnn: An effective convolutional neural network for tiny object detection in sperm videos," *Computers in Biology and Medicine*, vol. 146, p. 105543, 2022.

[17] D. Wu, O. Badamjav, V. Reddy, M. Eisenberg, and B. Behr, "A preliminary study of sperm identification in microdissection testicular sperm extraction samples with deep convolutional neural networks," *Asian Journal of Andrology*, vol. 23, pp. 135–139, 10 2020.

[18] A. Keller, M. Maus, E. Keller, and K. Kerns, "Deep learning classification method for boar sperm morphology analysis," *Andrology*, vol. 13, no. 6, pp. 1615–1625, 2025.

[19] R. Marín and V. Chang, "Impact of transfer learning for human sperm segmentation using deep learning," *Computers in Biology and Medicine*, vol. 136, p. 104687, 2021.

[20] P. Lei, M. Saadat, M. G. Hassani, and C. Shu, "Deep learning models for multi-part morphological segmentation and evaluation of live unstained human sperm," *Sensors*, vol. 25, no. 10, p. 3093, 2025.

[21] T. B. Haugen, S. A. Hicks, J. M. Andersen, O. Witczak, H. L. Hammer, R. Borgli, P. Halvorsen, and M. Riegler, "Visem: A multimodal video dataset of human spermatozoa," in *Proceedings of the 10th ACM Multimedia Systems Conference*, pp. 261–266, 2019.

[22] 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, 2024.

[23] P. Hidayatullah, X. Wang, T. Yamasaki, T. L. Mengko, R. Munir, A. Barlian, E. Sukmawati, and S. Supraptono, "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, 2021.

[24] M. Yuzkat, H. O. Ilhan, and N. Aydin, "Detection of sperm cells by single-stage and two-stage deep object detectors," *Biomedical Signal Processing and Control*, vol. 83, p. 104630, 2023.

[25] R. Zhu, Y. Cui, J. Huang, E. Hou, J. Zhao, Z. Zhou, and H. Li, "Yolov5s-sa: light-weighted and improved yolov5s for sperm detection," *Diagnostics*, vol. 13, no. 6, p. 1100, 2023.

[26] 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, 2023.

[27] R. Liu, M. Wang, M. Wang, J. Yin, Y. Yuan, and J. Liu, "Automatic microscopy analysis with transfer learning for classification of human sperm," *Applied Sciences*, vol. 11, no. 12, p. 5369, 2021.

[28] M. I. Mahali, J.-S. Leu, J. T. Darmawan, C. Avian, N. Bachroin, S. W. Prakosa, M. Faisal, and N. A. S. Putro, "A dual architecture fusion and autoencoder for automatic morphological classification of human sperm," *Sensors*, vol. 23, no. 14, p. 6613, 2023.

[29] B. Cansiz, H. O. Ilhan, and G. Serbes, "Loss-based ensemble generative adversarial network model for enhancing the sperm morphology classification," *Advanced Intelligent Systems*, vol. 8, no. 2, p. e202500441, 2026.

[30] D. Somasundaram and M. Nirmala, "Faster region convolutional neural network and semen tracking algorithm for sperm analysis," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105918, 2021.

[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[32] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[34] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[35] T. Prangemeier, C. Reich, and H. Koeppl, "Attention-based transformers for instance segmentation of cells in microstructures," in *2020 IEEE international conference on Bioinformatics and Biomedicine (BIBM)*, pp. 700–707, IEEE, 2020.

[36] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974, IEEE, 2024.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[38] R. A. Movahed, E. Mohammadi, and M. Orooji, "Automatic segmentation of sperm's parts in microscopic images of human semen smears using concatenated learning approaches," *Computers in Biology and Medicine*, vol. 109, pp. 242–253, 2019.

[39] H. Kaiming, G. Georgia, D. Piotr, and G.-s. Ross, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, vol. 2017, pp. 2961–2969, 2017.

[40] Q. Lv, X. Yuan, J. Qian, X. Li, H. Zhang, and S. Zhan, "An improved u-net for human sperm head segmentation," *Neural Processing Letters*, vol. 54, no. 1, pp. 537–557, 2022.

[41] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[42] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[43] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[44] S. Beucher, "The watershed transformation applied to image segmentation," *Scanning microscopy*, vol. 1992, no. 6, p. 28, 1992.

[45] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[46] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations," *Journal of computational physics*, vol. 79, no. 1, pp. 12–49, 1988.

[47] I. Iqbal, G. Mustafa, and J. Ma, "Deep learning-based morphological classification of human sperm heads," *Diagnostics*, vol. 10, no. 5, p. 325, 2020.

[48] P. Hernández-Herrera, V. Abonza, J. Sanchez-Contreras, A. Darszon, and A. Guerrero, "Deep learning-based classification and segmentation of sperm head and flagellum for image-based flow cytometry," *Computación y Sistemas*, vol. 27, no. 4, pp. 1133–1145, 2023.

[49] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[50] A. Abbasi, E. Miahi, and S. A. Mirroshandel, "Effect of deep transfer and multi-task learning on sperm abnormality detection," *Computers in Biology and Medicine*, vol. 128, p. 104121, 2021.

[51] H. O. Ilhan and G. Serbes, "Sperm morphology analysis by using the fusion of two-stage fine-tuned deep networks," *Biomedical Signal Processing and Control*, vol. 71, p. 103246, 2022.

[52] M. Dobrovolny, J. Benes, O. Krejcar, and A. Selamat, "Sperm-cell detection using yolov5 architecture," in *International Work-Conference on Bioinformatics and Biomedical Engineering*, pp. 319–330, Springer, 2022.

[53] S. Shahzad, M. Ilyas, M. I. U. Lali, H. T. Rauf, S. Kadry, and E. A. Nasr, "Sperm abnormality detection using sequential deep neural network," *Mathematics*, vol. 11, no. 3, p. 515, 2023.

[54] L. Prabaharan and N. Saravanan, "A three stage framework for abnormality detection in sperm cell images using cnn," *Biomedical Signal Processing and Control*, vol. 99, p. 106827, 2025.

[55] S. Shahali, M. Murshed, L. Spencer, O. Tunc, L. Pisarevski, J. Conceicao, R. McLachlan, M. K. O'Bryan, K. Ackermann, D. Zander-Fox, *et al.*, "Morphology classification of live unstained human sperm using ensemble deep learning," *Advanced Intelligent Systems*, vol. 6, no. 11, p. 2400141, 2024.

[56] A. Aristoteles, A. Syarif, S. Sutyarso, and F. R. Lumbanraja, "Identification of human sperm based on morphology using the you only look once version 4 algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, pp. 424–431, 2022.

[57] Z. Yin, T. Kanade, and M. Chen, "Understanding the phase contrast optics to restore artifact-free microscopy images for segmentation," *Medical image analysis*, vol. 16, no. 5, pp. 1047–1062, 2012.

[58] K. Chatzimeletiou, A. Fleva, T.-T. Nikolopoulos, M. Markopoulou, G. Zervakakou, K. Papanikolaou, G. Anifandis, A. Gianakou, and G. Grimbizis, "Evaluation of sperm dna fragmentation using two different methods: Tunel via fluorescence microscopy, and flow cytometry," *Medicina*, vol. 59, no. 7, p. 1313, 2023.

[59] L. Noy, I. Barnea, S. K. Mirsky, D. Kamber, M. Levi, and N. T. Shaked, "Sperm-cell dna fragmentation prediction using label-free quantitative phase imaging and deep learning," *Cytometry Part A*, vol. 103, no. 6, pp. 470–478, 2023.

[60] A. Agarwal, C.-L. Cho, S. C. Esteves, and A. Majzoub, "Reactive oxygen species and sperm dna fragmentation," *Translational andrology and urology*, vol. 6, no. Suppl 4, p. S695, 2017.

[61] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV* (P. S. Heckbert, ed.), pp. 474–485, San Diego, CA, USA: Academic Press, 1994.

[62] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Upper Saddle River, NJ, USA: Prentice Hall, 2 ed., 2002.

[63] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[64] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[65] J. G. Daugman, "Complete discrete 2-d gabor transforms by neural networks for image analysis and compression," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 7, pp. 1169–1179, 1988.

[66] H. O. Ilhan, I. O. Sigirci, G. Serbes, and N. Aydin, "A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods," *Medical & biological engineering & computing*, vol. 58, no. 5, pp. 1047–1068, 2020.

[67] S. G. Goodson, S. White, A. M. Stevans, S. Bhat, C.-Y. Kao, S. Jaworski, T. R. Marlowe, M. Kohlmeier, L. McMillan, S. H. Zeisel, *et al.*, "Casanova: a multiclass support vector machine model for the classification of human sperm motility patterns," *Biology of reproduction*, vol. 97, no. 5, pp. 698–708, 2017.

[68] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[69] M. E. Kandel, M. Fanous, C. Best-Popescu, and G. Popescu, "Real-time halo correction in phase contrast imaging," *Biomedical optics express*, vol. 9, no. 2, pp. 623–635, 2018.

[70] E. Lewandowska, D. Węsierski, M. Mazur-Milecka, J. Liss, and A. Jezierska, "Ensembling noisy segmentation masks of blurred sperm images," *Computers in Biology and Medicine*, vol. 166, p. 107520, 2023.

[71] A. Chen, J. Zhang, M. M. Rahaman, H. Sun, T. Zeng, M. Grzegorzek, F.-L. Fan, C. Li, *et al.*, "Active: A deep model for sperm and impurity detection in microscopic videos," *arXiv preprint arXiv:2301.06002*, 2023.

[72] M. E. Kandel, M. Rubessa, Y. R. He, S. Schreiber, S. Meyers, L. Matter Naves, M. K. Sermersheim, G. S. Sell, M. J. Szewczyk, N. Sobh, *et al.*, "Reproductive outcomes predicted by phase imaging with computational specificity of spermatozoon ultrastructure," *Proceedings of the National Academy of Sciences*, vol. 117, no. 31, pp. 18302–18309, 2020.

[73] F. Shaker, "Human sperm head morphology dataset (hushem)," *Mendeley Data*, vol. 3, p. 2018, 2018.

[74] V. Chang, A. Garcia, N. Hitschfeld, and S. Härtel, "Gold-standard for computer-assisted morphological sperm analysis," *Computers in biology and medicine*, vol. 83, pp. 143–150, 2017.

[75] A. Chen, C. Li, S. Zou, M. M. Rahaman, Y. Yao, H. Chen, H. Yang, P. Zhao, W. Hu, W. Liu, *et al.*, "Svia dataset: a new dataset of microscopic videos and images for computer-aided sperm analysis," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 1, pp. 204–214, 2022.

[76] H. O. Ilhan and N. Aydin, "A novel data acquisition and analyzing approach to spermiogram tests," *Biomedical Signal Processing and Control*, vol. 41, pp. 129–139, 2018.

[77] V. Dubey, D. Popova, A. Ahmad, G. Acharya, P. Basnet, D. S. Mehta, and B. S. Ahluwalia, "Partially spatially coherent digital holographic microscopy and machine learning for quantitative analysis of human spermatozoa under oxidative stress condition," *Scientific reports*, vol. 9, no. 1, p. 3564, 2019.

[78] M. L. D. Garcia, D. A. P. Soto, and L. S. Mihaylova, "A bag of features based approach for classification of motile sperm cells," in *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 104–109, IEEE, 2017.

[79] 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," in *Scandinavian conference on image analysis*, pp. 397–406, Springer, 2017.

[80] S. K. Mirsky, I. Barnea, M. Levi, H. Greenspan, and N. T. Shaked, "Automated analysis of individual sperm cells using stain-free interferometric phase microscopy and machine learning," *Cytometry Part A*, vol. 91, no. 9, pp. 893–900, 2017.

[81] A. Butola, D. Popova, D. K. Prasad, A. Ahmad, A. Habib, J. C. Tinguely, P. Basnet, G. Acharya, P. Senthilkumaran, D. S. Mehta, *et al.*, "High spatially sensitive quantitative phase imaging assisted with deep neural network for classification of human spermatozoa under stressed condition," *Scientific reports*, vol. 10, no. 1, p. 13118, 2020.

[82] A. Bijar, M. Mikaeili, R. Khayati, *et al.*, "Fully automatic identification and discrimination of sperm's parts in microscopic images of stained human semen smear," *Journal of Biomedical Science and Engineering*, vol. 5, no. 2012, 2012.

[83] K.-K. Tseng, Y. Li, C.-Y. Hsu, H.-N. Huang, M. Zhao, and M. Ding, "Computer-assisted system with multiple feature fused support vector machine for sperm morphology diagnosis," *BioMed research international*, vol. 2013, no. 1, p. 687607, 2013.

[84] F. Shaker, S. A. Monadjemi, and A. R. Naghsh-Nilchi, "Automatic detection and segmentation of sperm head, acrosome and nucleus in microscopic images of human semen smears," *Computer methods and programs in biomedicine*, vol. 132, pp. 11–20, 2016.