Journal Pre-proof

Evaluation of an AI-facilitated sperm detection tool in azoospermic samples for use in ICSI

DM. Goss , SA. Vasilescu , PA. Vasilescu , S. Cooke , SHK. Kim , GP. Sacks , DK. Gardner , ME. Warkiani

PII: S1472-6483(24)00099-3
DOI: https://doi.org/10.1016/j.rbmo.2024.103910
Reference: RBMO 103910

To appear in: *Reproductive BioMedicine Online*

Received date: 7 September 2023
Revised date: 31 January 2024
Accepted date: 9 February 2024

Please cite this article as: DM. Goss , SA. Vasilescu , PA. Vasilescu , S. Cooke , SHK. Kim , GP. Sacks , DK. Gardner , ME. Warkiani , Evaluation of an AI-facilitated sperm detection tool in azoospermic samples for use in ICSI, *Reproductive BioMedicine Online* (2024), doi: https://doi.org/10.1016/j.rbmo.2024.103910

Evaluation of an AI-facilitated sperm detection tool in azoospermic samples for use in ICSI

DM. Goss[1,2,3,†], SA. Vasilescu[1,2,†], PA. Vasilescu[2], S. Cooke[3], SHK. Kim[3,4], GP. Sacks[1,4,3] DK. Gardner[2,5], ME. Warkiani[1,2,6*]

[1] School of Biomedical Engineering, University of Technology Sydney, Sydney, 2126, Australia

[2] NeoGenix Biosciences Pty Ltd, Sydney, 2126, Australia

[3] IVFAustralia, Sydney, 2015, Australia.

[4] University of New South Wales, Sydney, 2052, Australia

[5] Melbourne IVF, Melbourne, 3002, Australia

[6] Institute for Biomedical Materials & Devices (IBMD), University of Technology Sydney, Sydney, 2126, Australia

[†] The authors consider that the first two authors should be regarded as joint First Authors

*****Corresponding author:

Majid Ebrahimi Warkiani (majid.warkiani@uts.edu.au) (ORCID 0000-0002-4184-1944)

School of Biomedical Engineering, University Technology Sydney, Sydney, New South Wales 2007, Australia

Abstract

Research question: Can artificial intelligence (AI) improve efficiency and efficacy of sperm searches in azoospermic samples?

Design: This two-phase proof-of-concept study beginning with a training phase using 8 azoospermic patients (>10000 sperm images) to provide a variety of surgically collected samples for sperm morphology and debris variation to train a convolutional neural network to identify sperm. Secondly, side-by-side testing on 2 cohorts of non-obstructive azoospermia (NOA) patient samples, an embryologist versus the AI identifying all sperm in still images (cohort 1, N=4) and then a side-by-side test with a simulated clinical deployment of the AI model on an ICSI microscope and the embryologist performing a search with and without the aid of the AI (cohort 2, N=4).

Results: In cohort 1, the AI model showed improvement in time-taken to identify all sperm per field of view ($0.019\pm0.30 \times 10^{-5}$s versus 36.10±1.18s, P<0.0001) and improved recall (91.95±0.81% vs 86.52±1.34%, P<0.001) compared to an embryologist. From a total of 688 sperm to find in all samples combined, 560 were found by an embryologist and 611 were found by the AI in <1000th of the time. In cohort 2, the AI-aided embryologist took significantly less time per droplet (98.90±3.19s vs 168.7±7.84s, P<0.0001) and found 1396 sperm, while 1274 were found without AI, although no significant difference was observed.

Conclusions: AI-powered image analysis has the potential for seamless integration into laboratory workflows, and to reduce time to identify and isolate sperm from surgical sperm samples from hours to minutes, thus increasing success rates from these treatments.

Key words: Azoospermia, male infertility, mTESE, sperm, SSC

Introduction

Male infertility is increasing worldwide at an alarming rate, sperm counts declining by 50% over the past 50 years (Levine et al., 2023). 30% human infertility cases are caused solely by male infertility and 50% of cases being attributed to male infertility as a contributing factor (Agarwal et al., 2015). While assisted reproductive technologies (ART) have proven to be effective in treating infertile couples, some forms of male infertility remain difficult to treat. Azoospermia, defined as the absence of spermatozoa in centrifuged semen on at least two occasions, is the most severe form of male infertility, affecting 10-20% of infertile men and 1% of the general male population (Verheyen et al., 2017, Wosnitzer et al., 2014).

Azoospermia can be classified as either obstructive and/or non-obstructive. Obstructive azoospermia (OA) occurs due to obstruction of the reproductive tract and constitutes 40% of azoospermic cases, while non-obstructive azoospermia (NOA) results from either primary, secondary or incomplete/ambiguous testicular failure which compromises sperm production and constitutes 60% of azoospermic cases (Jarow et al., 1989, Wosnitzer, et al., 2014). Patients with OA can attempt for reconstruction (vasovasostomy, vasoepididymostomy or transurethral resection ejaculatory duct) when possible, or surgical sperm collection can be performed from the testis via testicular sperm aspiration (TESA), testicular sperm extraction (TESE) or microdissection testicular sperm extraction (mTESE) or the epididymis Microsurgical epididymal sperm aspiration (MESA) or percutaneous epididymal sperm aspiration (PESA) (Schrepferman et al., 2001) (Flannigan et al., 2017). NOA patients will require sperm extraction from the testis (TESA, TESE or Micro TESE) and surgically collected sperm is then used for ICSI.

The gold-standard for treating NOA patients, is mTESE, with a high sperm retrieval rate of up to 64% in suitable patients operated on (Deruyver et al., 2014, Ramasamy et al., 2005, Schiff et al., 2005). Although these rates seem promising, the current manual examination process to find sperm within tissue recovered from mTESE surgeries is time-consuming and inefficient, typically taking anywhere between 1-6 h of laboratory time, and in some cases even up to 14 h (Mangum et al., 2020, Ramasamy et al., 2011). This extended time is due to the requirement for

manual searching through prepared suspensions of testicular tissue with a microscope, before using isolated sperm for ICSI.

The outcome of such searching is heavily dependent upon the complexity and contamination of the suspension provided to them by the surgeon. Viable sperm are easily overlooked due to variables such as collateral cell density, resulting in a process that is prone to human error, combined with inexperience and fatigue of lab staff (Ramasamy, et al., 2011). For patients with NOA, if sperm are overlooked due to human error, this could wrongly indicate absolute infertility (Samuel et al., 2016). Similarly, for extended sperm searches in semen as a diagnostic or as a last check in the ejaculate before surgery, failure to identify sperm present could striate patients into surgery unnecessarily. Furthermore, prolonged sample examination procedures can have adverse effects on the viability of sperm, consequently affecting their potential for fertilization and thus undermining the efforts of sperm searches and the considerable cost and physical strain caused on patients during these procedures (Ouitrakul et al., 2018). For patients with NOA, a more efficient and higher throughput method capable of locating and isolating sperm from the suspension would therefore greatly benefit the clinical workflow of assisting severe forms of male infertility.

Panning through surgically collected sperm samples, under a microscope is a form of manual image analysis in which machine learning (ML) and artificial intelligence (AI) has the potential to automate and improve. Therefore, with preliminary works showing promising results (Goss et al., 2023), this study aims to comprehensively assess the use of an assistive convolutional neural network (CNN) AI to identify sperm in complex tissue suspensions in real time was developed and trained (Figure 1). Using a YOLO V8 model, an open-source, high-speed, high-accuracy object detection and image segmentation model, this software works in tandem with an embryologist to instantly identify and alert embryologists to sperm of interest for their assessment from the camera feed mounted into their microscope. The objective of this study was to compare the AI to embryologists without the AI's aid in terms of time, recall, and number of sperm found first using still images in cohort 1, and then in a simulated sperm search with the AI integrated into an ICSI microscope kit for cohort 2, to demonstrate its potential for clinical implementation.

Materials and methods

Ethical approval

Ethical approval for healthy sperm samples was received from the University of Technology Sydney ethics review board (ETH19-3677), and for the use of discarded testicular tissue samples from the IVFAustralia Human Research Ethics Committee (DG01192) and UTS ethics review board (ETH22-7189).

Specifically prepared sample preparation

Specifically prepared samples for initial training of the AI model were used prior to access to clinical testicular tissue samples to generate images for the training dataset (Figure 1A). These samples consisted of donor sperm, fingerprick blood and cells from cancer cell culture lines. Human semen samples were obtained through ejaculation after 2-5 days of sexual abstinence (WHO, 2021). Raw semen samples were left at room temperature for 20 minutes to allow for liquefaction. Samples were centrifuged for 8 minutes at 500xg to separate the sperm pellet from the seminal plasma. Red blood cells (RBCs) were obtained from whole blood specimens within three days of collection. Collected blood samples were also resuspended in GMOPS Plus (Vitrolife, Sweden) media. Mixed cell suspensions were created to simulate testicular tissues samples containing sperm, RBCs, white blood cells (WBCs), epithelial cells, C2C12 and THP1 cells. All cells were mixed in warmed GMOPS Plus (37^{o}C). Raw semen samples were diluted down to between $1x10^{7}$ and $1x10^{8}$ sperm/mL, RBC concentration ranged between 2-$15x10^{6}$ cells/mL (approximated ranges for a mTESE sample), WBCs (purchased from IQ Biosciences, $10x10^{6}$ cells/mL) were diluted to a concentration between $5x10^{5}$ and $1x10^{6}$ cells/mL, and epithelial cells were diluted to a concentration of between $7x10^{5}$ and $1x10^{6}$ cells/mL. To add extra complexity, background cells from sperm donors were isolated from donors with high concentrations of background cell populations and cryopreserved until needed. These cells helped simulate the conditions of poor-quality samples with high levels of collateral cell contamination from surgery and for infertile semen samples with high levels of contamination in the ejaculate.

Testicular biopsy retrieval and processing

Surgical sperm collection was performed in accordance with the routine workflow for each method of sperm collection (mTESE and TESA). Azoospermic patients scheduled for surgical sperm collection for both OA and NOA. Surgical sperm collections were performed under general anesthesia, and the samples were immediately placed in a sterile conical tube containing 1 mL of G-MOPS Plus (37^{o}C) and transported to the IVF laboratory. During mTESE, embryologists search through seminiferous tubules handed to them by the surgeon, with simultaneous further searching by the surgeon for dilated seminiferous tubules. Further samples are then sent to the IVF laboratory for further search before being placed in 1-2 mL of G-MOPS Plus in a sterile petri dish under a stereo microscope, to wash off excess blood from the tissue then moved to a new petri dish with 300 µL of G-MOPS Plus. Tissue was gently teased apart using sterile syringes to release potential sperm from tubules into the surrounding G-MOPS Plus media. The macerated tissue and large pieces were then removed and placed into a separate tube, and the remaining suspension used for the sperm search and treatment. In cases whereby imaging and/or testing was not possible on the same day or following day, samples were fixed with 4% formalin to preserve its morphological integrity and prevent any microbial growth until use in the study.

To prepare samples for comparison between AI-enabled sperm search and sperm search by an embryologist in cohort 1, samples that were recorded having no sperm found from clinical searches were spiked with low concentrations of sperm from semen donors (prepared as described in specifically prepared sample preparation). To help create a master count of total sperm in plated samples, spiked sperm were stained with propidium iodide (PI) and washed to remove excess stain before spiking. This was done to help identify the total number of sperm to be found in each sample for comparison with the AI and embryologist performance groups. Samples that had sperm present in the clinics were not spiked with donor semen and preserved in their clinical state for processing.

Image acquisition and processing

To train the model, specifically prepared samples containing mixtures of sperm, RBCs, WBCs and epithelial cells from cell culture media were prepared and plated in a similar manner to a

clinical sperm search, 10 long drops of GMOPS Plus of 2-3 mm in length under OVOIL (Vitrolife, Sweden) in an ICSI dish (Vitrolife, Sweden), and imaged at 200X magnification using cellSens Imaging Software (Olympus Life Science, Japan; Figure 1A). This approach was chosen to initiate training and once the model's ability to identify sperm was confirmed, clinically obtained testicular tissue samples from 8 azoospermic patients (6 NOA patients and 2 OA patients) were then used to train the model with more representative backgrounds. The training dataset comprises of 540 images (152 from specifically prepared samples and 388 from testicular tissue samples), containing 5624 unique sperm instances, duplicated, and augmented generating at least one augmented copy per image, resulting in over 10 000 sperm to train the identification function (Figure 1A). The use of synthetic data (duplication) during model training was performed to create more unique images for the model to learn from and is commonly performed to improve dataset fidelity (Chavez-Badiola et al., 2020, Cubuk et al., 2018, Trembley et al., 2018). By creating these flipped and augmented duplicate images, these images can be used in the training process as they may be considered functionally unique to their original copy (see supplementary Figure 1). Images were annotated using Computer Vision Annotation Tool (CVAT; Intel, USA) which is an open-source software with a web-based interface designed for image and video annotation for computer vision tasks. This software was chosen to create the annotated dataset of images whereby sperm in these images were annotated with simple bounding boxes (see supplementary Figure 2) that enclose the entire visible sperm including head and tail. If the sperm is partially occluded it is still bound by a single bounding box encompassing all visible areas. CVAT was chosen for this purpose due to collaborative annotation from multiple users (including the AI model) as well as the user-friendly interface.

Dataset preparation

Training images were 2456x1842px, JPEG with 95% compression. Images were saved in JPG format to better reflect real-world environments where images may be sent over a network and require rapid real-time feedback. These were resized to 1664x1664px with black fill. 85% of the images were used for training and 15% reserved for validation of the model's performance after training. Augmentations were applied to all images including duplicates from both the specifically prepared samples as well as the excess testicular tissue samples to inflate the dataset

and make the trained model more robust to variations in microscope camera images, such as compression artifacts, changing focal length, or lighting and colour variations. A vertical flip was applied to each duplicate image, ensuring it was unique from its source, then with various probability a series of augmentation techniques were employed using the Python-based Albumentations library (Buslaev et al., 2020). Initially, a blur effect with a kernel size of 2x2 pixels was applied to each image to simulate the effect of slight defocusing. Thereafter, JPEG compression was implemented, adjusting the compression quality to a range between 60% and 80%, to mimic the common lossy compression artifacts (features identifiable with the human eye) found in digital imaging. An example of these augmentations is shown in supplementary Figure 1.

Training of the AI model

Once the dataset of images was compiled for training, an open-source machine-learning model architecture, YOLO V8 (Ultralytics, USA), was chosen which provided the framework and tools to train our own model. YOLO V8 was selected due to it being a highly performant architecture for real-time object detection tasks and this suits the application of identifying sperm in highly complex tissue samples. YOLO V8 was used with the 'small' size architecture configuration with 225 layers and 1,1166,560 parameters to prioritize minimal inference time (i.e. speed of identifying potential sperm during searching) over potentially greater recall from more parameters (Jocher et al., 2023). Further image augmentations were applied by YOLO V8 during the training process, including horizontal flipping, scaling, translation and augments to hue, saturation, and value. The training setup was restricted to a modest <8GB VRAM as the which limits the size of the model and training image resolution. Thus, to maintain a high image resolution required to differentiate fine detail and the desired model size on this setup, we used a small batch size of 4 images being trained in parallel. We trained the model for 300 training iterations or epochs with a learning rate of 0.01. The stochastic gradient descent optimizer was used with 0.937 momentum and 0.005 weight decay. The trained model was then used to make inferences on unseen, unlabeled images from the 15% of images allocated for validation (Figure 1A). The performance of the model was validated on images with a ground truth sperm number showing 85% precision and 78% accuracy after 300 epochs, the model was then considered ready for side-by-side testing against an embryologist as the training dataset is purposely

compiled to validate performance on edge-cases and relatively difficult to identify sperm. This approach has been proven to produce robust and unbiased image detection models (Vabalas et al., 2018).

Comparison of AI model vs embryologist performance

Side-by-side testing was split into two cohorts both using immotile sperm for the ability to standardize sperm spatial detection. The first cohort was performed on fixed samples at UTS research laboratories and consisted of comparing the time, recall and precision of sperm detection on still images between the AI model and an embryologist (Figure 1B). The AI model was loaded onto a desktop computer (Intel Core i5-10600K CPU @ 4.10 GHz (6 cores), RTX 3070 graphics card) and annotated sperm in still images of plated discarded testicular tissue samples (N=4 NOA patients, 512 images acquired with a total of 2660 sperm to be found) in droplets at 200X magnification. The embryologist used CVAT to annotate the location of sperm independently in the same images while being timed. Sperm annotations from both the AI model and the embryologist (using the annotation software, CVAT) were then compared to a ground truth of verified sperm labels for each image to attain comparable metrics i.e. precision, recall, time per field of view (FOV) and total sperm found. Consensus for the ground truth annotation for each image was performed by two scientists independent of the embryologist used to test against the AI model. Precision is a measure of how many sperm detections are correct, i.e. the ratio of correctly predicted positive observations to the total number of predictions made, and recall (sensitivity or true positive rate) is a measure of how many of the sperm in a FOV the model finds, i.e. the ratio of correctly predicted positive observations to total of actual sperm in the FOV. Precision and recall are defined by:

$$Precision = \frac{\sum TP}{\sum (TP + FP)} \qquad Recall = \frac{\sum TP}{\sum (TP + FN)}$$

TP = True positives; FP = False positives; FN = False negatives

Potential sperm detections (bounding boxes) with significant overlap (>40% Intersection of Union) with confirmed sperm were counted as positive detections and those without as negatives. Sperm bordering the edge of an image are often cutoff and lack enough information to

distinguish them as either positive or negative, thus any potential sperm within 2px of the edge of the image were omitted.

The second cohort, to better simulate real-time clinical deployment, a side-by-side test of the AI comparing the performance of an embryologist with and without the AI was performed. Dishes were plated and prepared testicular tissue samples were added to dishes similar manner to a clinical sperm search, 10 long drops of GMOPS Plus of 2-3 mm in length under OVOIL (Vitrolife, Sweden) in an ICSI dish (Vitrolife, Sweden) per patient sample. The embryologist recorded and compared the number of sperm found per droplet for each tissue sample (N=4 NOA patients) they processed with (see supplementary video 1 and 2) and without AI, as well as the time taken to complete their assessment (Figure 1B). No ground truth total sperm number was acquired for each drop or dish therefore a direct comparison of sperm found per unit time in each media drop was compared between using the AI and not using the AI. Dishes were blinded to the embryologist and re-ordered to prevent any memory of sperm location by the embryologist when performing each search.

Statistical Analysis

All statistical analyses were performed using GraphPad Prism 9.0 (GraphPad Software). Normal distribution was assessed using the Shapiro-Wilk Test. The statistical significance of the differences between groups were tested using the two-tailed unpaired Student's t-test or Mann–Whitney U test if the data were not normally distributed. Two-way analysis of variance to assess the effects of the counting method and sample were performed. $P < 0.05$ was considered statistically significant and means are expressed with Standard Error of the Mean (SEM) as a measure of sample mean estimates.

Results

In the first cohort of this study (N=4 NOA patients), when assessing performance of sperm identification from still images, the AI model showed dramatic improvement in time taken to identify sperm per FOV, improved recall in identifying sperm as well as a high level of precision (Table 1). The AI was able to identify all sperm within each field of view (FOV) in significantly less time compared to the trained embryologist, with a duration of $0.019 \pm 0.3 \times 10^{-5}$s versus

36.10 ± 1.18s, respectively (P < 0.0001; Table 1). This represents an approximate 99.95% reduction in time per FOV. The AI model demonstrated a significant difference in recall compared to the trained embryologist (91.95 ± 0.81% vs 86.52 ± 1.34%, P < 0.001; Table 1). The model exhibited a precision of 89.58 ± 0.87%, considering the correct identification of sperm and false positives relative to the control count (Table 1). In contrast, the embryologist had a precision of 98.18 ± 0.38%. Out of a total of 2660 sperm, the embryologist identified 1937, while the AI model detected 1997 (Table 1).

In the second cohort of this study (N=4 NOA patients), a simulated deployment of the AI was performed in a research laboratory whereby the AI was used as an assistive tool to guide embryologists to identify sperm on a ICSI microscope kit (see supplementary video 1 and 2). Like cohort 1, the AI-assisted embryologist outperformed the individual assessment of an embryologist across all 4 samples. The embryologist using the AI took significantly less time to find all sperm per droplet (98.9 ± 3.19 vs 168.7 ± 7.84, P < 0.0001) and found a total of 1396 sperm while they found 1274 without the use of the AI (Table 1). There was no significant difference in the number of sperm found per droplet for the embryologist using AI versus without the use of AI although a slight trend of more sperm found, consistently, was observed (34.9 ± 3.23 vs 31.85 ± 3.09 sperm respectively).

Discussion

AI image analysis can identify sperm faster and with better recall than an embryologist in still images and significantly faster in a simulated sperm search scenario when integrated into an ICSI microscope. This is the first known application of ML AI for surgical sperm searches for the clinical treatment of azoospermia and results in a streamlining of a historically laborious process. ML is an algorithmic method of data analysis whereby a predictive model is trained to recognize patterns and associations from input data (Bannach-Brown et al., 2019). Supervised ML models can be trained on labelled images and/or video to understand how to predict the labels of unseen data. CNN algorithms are a type of deep-learning model that attempts, through iterative training, to transform input data into the desired output labels. There have been considerable studies on the utility of machine learning and AI-based image analysis on the selection of embryos for prediction of euploidy status, implantation potential and incidence of

miscarriage (Barnes et al., 2023, Diakiw et al., 2022, Duval et al., 2023, Hariharan et al., 2019, Tran et al., 2018, VerMilyea et al., 2020). Studies have also proven the application of ML in the selection and assessment of sperm for use in ICSI by tacking sperm correlated with better quality blastocysts (Joshi et al., 2023, Mendizabal-Ruiz et al., 2022). Furthermore, studies using images of sperm having been labelled as normal or abnormally shaped by a professional or stained for DNA integrity; given a sufficient volume and variety of these labelled images, ML models have been trained to label the morphology of predict DNA fragmentation of new, unseen, images of sperm (McCallum et al., 2019, Wang et al., 2019). Where the CNNs, commonly referred to as AI have largely looked at sperm in a clear environment, we have applied a CNN on complex, processed tissues from testicular sperm retrieval procedures and implemented it in a live video feed to real-time identification of sperm for use in ICSI.

The application of a computer vision-based ML model to identify sperm in real-time during sperm searches outperforms embryologists' manual searching in simulated searches using still images in time taken, recall and sperm count. The biggest noticeable difference is in the time reduction, where image analysis is almost instant (0.02 s per field of view) but does not consider clinical tasks such as dish setup, panning and magnification change, and collection of identified sperm using a micromanipulator needle. Recall and precision were measured as metrics of both the AI and the embryologists performance against a ground truth number of sperm per image. Significantly lower time taken to identify sperm per FOV, higher recall, and an increase in the total number of sperm found show clear superiority of AI image analysis compared to the eyes and focus of trained embryologists (Table 1). Although the AI had a lower precision value than embryologists in the first cohort, it is worth noting that this is a result of the annotation approach taken when training the AI and precision values are particularly relevant in applications when the cost of false positives is high. For the application of this AI model in sperm searching, the cost of false negatives is much higher whereby a potential sperm suitable for ICSI could be missed, as opposed to an extra two seconds of an embryologist's attention might be wasted in the case of a false positive. Recall however, is essential when the cost of false negatives is high, as is in sperm searches of NOA patient samples.

In the second cohort, testicular tissue samples with supplemented sperm (for better quantification of efficacy) were searched by an embryologist in plated ICSI dishes to better simulate a clinical sperm search on an ICSI kit with and without the aid of the AI (see supplementary video 1 and 2), it was determined that the AI reduced the time taken to identify all sperm in the droplet by around 50% (Table 1), with no drop in number of sperm identified per drop and a higher total number of sperm identified in total (Table 1).

Using an exhaustively trained image analysis model to identify sperm based on tens of thousands of sperm images has clinical utility in directing an embryologist's attention to what the AI deems may be of interest and can thus drastically reduce the time taken or manual extended sperm searches when integrated to a micromanipulator microscope. The model trained in this study is designed to cater for multiple clinics which may have different microscopes, light environments, filters, and cameras. These environmental and equipment factors may affect the performance of the AI and have thus been catered for. The image augmentations such as blur, colour variations, focus changes, image saturation, and colour balance changes and flipping of images used to train the AI model on, follows a common strategy in computer vision image analysis whereby these augmentations artificially replicate variant circumstances that may appear in images that were not necessarily widely represented in the training data that comes from a relative few, largely homogenous samples (Chavez-Badiola et al., 2020, Cubuk et al., 2018, Trembley et al., 2018). It is common for microscope images to be slightly blurry or have different lighting conditions and this is replicated in the training data through our choice of augmentations, such that the model is resilient to these conditions. This is another area that with further tuning could improve model performance in the future. The model was also trained using both epididymal and testicular sperm to broaden the sample dataset empowering the AI to broaden target sperm prompting. Importantly, the model can also identify sperm with the broad range of motility, from immotile to hyperactivated, and adjusts and adapts to magnification change and panning in real-time (see supplementary video 3).

The role of this model is not to replace an embryologist, but to be a guide towards sperm of interest, leaving the embryologist to make the final determination on the suitability of a sperm for ICSI. AI can negate the biological limits of human error and observation as well as the effects

of fatigue which have long been a limiting factor to extended sperm searches of heterogenous samples obtained via surgical sperm collection. It is important to remember however, the AI is limited to detection within the manually directed field of view, thus if the embryologist overlooked an area in the sample, the AI will not be able to detect sperm without having it within view.

This study was performed solely on immotile sperm for the most accurate quantification for spatially identifying and locating sperm, although the AI identifies motile sperm very well (see Supplementary video 3), a true clinical deployment will better prove the clinical utility of the model. This proof-of-concept study demonstrates the potential for AI-assisted sperm searches, both in semen for extended sperm searches and testicular tissue. While the results of this study are promising, continuing to improve the core data set and image variety will make the model more robust and adoptable for clinics with significantly different microscope arrangements as well as achieving a higher level of recall. The limitation of a simulated sperm search using an ICSI workstation with and without the use of the AI, using samples with spiked in sperm, is that it does not consider the time spent confirming the locations of sperm in the field of view during panning (as to not recount or miss sperm). This is a disadvantage of the testing method and can contribute to the lower difference in time taken per method in cohort 2. Therefore, a robust clinical deployment study has been planned for consenting in-treatment patients whereby embryologists can perform sperm searches with the aid of the AI model. Furthermore, there is potential for the expansion of this AI to include motility and morphological assessments of identified sperm to aid in the choice of sperm for insemination when sperm outnumbers the number of oocytes suitable for injection. Another useful addition to the AI, would be a sensitive measure of sperm 'twitching' in these cases. 'Twitching' sperm movement in severe NOA cases confirms the vitality of sperm without the need for other interventions to prove sperm vitality such as the hyperosmotic swelling (HOS) test, also reducing time taken when selecting sperm found.

In conclusion, azoospermia affects 10% of infertile men, with NOA, the most severe form, constituting 60% of these cases (Verheyen, et al., 2017). The current approaches to recover sperm from men who undergo surgery from this condition are antiquated and potentially

detrimental to the quality of the sperm found. In this study, we have successfully demonstrated a proof-of-concept application of an AI image analysis model to drastically reduce sperm search time in testicular tissue samples in simulated clinical sperm searches. When applying the AI to a simulated real-time search workflow, a 50% reduction in time taken to identify sperm has been demonstrated. This presents a potential to avoid or at least reduce the negative effect of extended exposure of sperm to biopsied testicular tissue containing a host of molecules capable of reducing sperm viability. Applying this approach with further development and ergonomic optimization, we believe it can result in a standardized and more efficient workflow, greatly improving the current processing procedure of all surgically retrieved samples and azoospermic ejaculates by increasing access to treatment for azoospermia and reducing staff time required, as well as increasing sample coverage to ultimately increase chances of finding sperm.

Authors roles

DMG, SAV, PAV, SC, SHK, DKG and MEW designed and conceptualized the study. DMG, SAV and PAV were responsible for all data acquisition. PV and SAV designed and trained the AI model. DMG, SHK and SAV facilitated clinical sample acquisition. DMG, SAV and PAV drafted the manuscript and all authors critically revised, finally approved and all agree to be accountable for academic integrity and accuracy of the research.

Acknowledgments

We would like to acknowledge the support and facilitation provided by IVFAustralia Eastern Suburbs doctors Dr Jeffrey Persson and Dr Shadi Khashaba, providing access to patients undergoing treatment for severe male factor infertility and technical support and input from the embryology team. M.E.W. would like to acknowledge the support of the Cancer Institute New South Wales through the Career Development Fellowship (2021/CDF1148).

References

Agarwal A, Mulgund A, Hamada A, Chyatte MR, 2015. A unique view on male infertility around the globe. Reproductive biology and endocrinology. **13**, 1-9.

Bannach-Brown A, Przybyła P, Thomas J, Rice AS, Ananiadou S, Liao J, Macleod MR, 2019. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Systematic reviews. **8**, 1-12.

Barnes J, Brendel M, Gao VR, Rajendran S, Kim J, Li Q, Malmsten JE, Sierra JT, Zisimopoulos P, Sigaras A, 2023. A non-invasive artificial intelligence approach for the prediction of human blastocyst ploidy: A retrospective model development and validation study. The Lancet Digital Health. **5**, e28-e40.

Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA, 2020. Albumentations: fast and flexible image augmentations. Information. **11**, 125.

Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J, 2020. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reproductive BioMedicine Online. **41**, 585-593.

Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV, 2018. Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501.

Deruyver Y, Vanderschueren D, Van der Aa F, 2014. Outcome of microdissection TESE compared with conventional TESE in non-obstructive azoospermia: a systematic review. Androl. **2**, 20-24.

Diakiw S, Hall J, VerMilyea M, Amin J, Aizpurua J, Giardini L, Briones Y, Lim A, Dakka M, Nguyen T, 2022. Development of an artificial intelligence model for predicting the likelihood of human embryo euploidy based on blastocyst images from multiple imaging systems during IVF. Human Reproduction. **37**, 1746-1759.

Duval A, Nogueira D, Dissler N, Maskani Filali M, Delestro Matos F, Chansel-Debordeaux L, Ferrer-Buitrago M, Ferrer E, Antequera V, Ruiz-Jorro M, 2023. A hybrid artificial intelligence model leverages multi-centric clinical data to improve fetal heart rate pregnancy prediction across time-lapse systems. Human Reproduction. **38**, 596-608.

Flannigan R, Bach PV, Schlegel PN, 2017. Microdissection testicular sperm extraction. Translational Androl. and Urology. **6**, 745.

Goss D, Vasilescu S, Vasilescu P, Sacks G, Gardner D, Warkiani M, 2023. O-136 Artificial intelligence to assist in surgical sperm detection and isolation. Hum. Reproduction. **38**, dead093. 163.

Hariharan R, He P, Meseguer M, Toschi M, Rocha JC, Zaninovic N, Malmsten J, Zhan Q, Hickman C, 2019. Artificial intelligence assessment of time-lapse images can predict with 77% accuracy whether a human embryo capable of achieving a pregnancy will miscarry. Fertility and Steril. **112**, e38-e39.

Jarow JP, Espeland MA, Lipshultz LI, 1989. Evaluation of the azoospermic patient. The J. of urology. **142**, 62-65.

Jocher G, Chaurasia A, Qiu J. YOLO by Ultralytics. 2023. Ultralytics, GitHub.

Joshi K, Simbulan RK, Rajah AM, Burd G, Gupta S, Behr B, Guarnaccia M, Singh G, 2023. A proof-of-concept prospective study of applying artificial intelligence for sperm selection in the IVF laboratory. Reproductive BioMedicine Online. 103329.

Levine H, Jørgensen N, Martino-Andrade A, Mendiola J, Weksler-Derri D, Jolles M, Pinotti R, Swan SH, 2023. Temporal trends in sperm count: a systematic review and meta-regression analysis of samples collected globally in the 20th and 21st centuries. Hum. reproduction update. **29**, 157-176.

Mangum CL, Patel DP, Jafek AR, Samuel R, Jenkins TG, Aston KI, Gale BK, Hotaling JM, 2020. Towards a better testicular sperm extraction: novel sperm sorting technologies for non-motile sperm extracted by microdissection TESE. Translational Androl. and Urology. **9**, S206.

McCallum C, Riordon J, Wang Y, Kong T, You JB, Sanner S, Lagunov A, Hannam TG, Jarvi K, Sinton D, 2019. Deep learning-based selection of human sperm with high DNA integrity. Communications biology. **2**, 250.

Mendizabal-Ruiz G, Chavez-Badiola A, Figueroa IA, Nuño VM, Farias AF-S, Valencia-Murilloa R, Drakeley A, Garcia-Sandoval JP, Cohen J, 2022. Computer software (SiD) assisted real-time single sperm selection associated with fertilization and blastocyst formation. Reproductive BioMedicine Online. **45**, 703-711.

Ouitrakul S, Sukprasert M, Treetampinich C, Choktanasiri W, Vallibhakara SA-O, Satirapod C, 2018. The Effect of Different Timing after Ejaculation on Sperm Motility and Viability in Semen Analysis at Room Temperature. J. of the Méd Association of Thail. **101**.

Ramasamy R, Reifsnyder JE, Bryson C, Zaninovic N, Liotta D, Cook C-A, Hariprashad J, Weiss D, Neri Q, Palermo GD, 2011. Role of tissue digestion and extensive sperm search after microdissection testicular sperm extraction. Fertil. and Steril. **96**, 299-302.

Ramasamy R, Yagan N, Schlegel PN, 2005. Structural and functional changes to the testis after conventional versus microdissection testicular sperm extraction. Urology. **65**, 1190-1194.

Samuel R, Badamjav O, Murphy KE, Patel DP, Son J, Gale BK, Carrell DT, Hotaling JM, 2016. Microfluidics: The future of microdissection TESE? Systems Biology in Reproductive Medicine. **62**, 161-170.

Schiff JD, Palermo GD, Veeck LL, Goldstein M, Rosenwaks Z, Schlegel PN, 2005. Success of testicular sperm injection and intracytoplasmic sperm injection in men with Klinefelter syndrome. The J. of Clinical Endocrinology & Metabolism. **90**, 6263-6267.

Schrepferman CG, Carson MR, Sparks AE, Sandlow JI, 2001. Need for sperm retrieval and cryopreservation at vasectomy reversal. The J. of urology. **166**, 1787-1789.

Tran A, Cooke S, Illingworth P, Gardner D, 2018. Artificial intelligence as a novel approach for embryo selection. Fertil. and Steril. **110**, e430.

Tremblay J, Prakash A, Acuna D, Brophy M, Jampani V, Anil C, To T, Cameracci E, Boochoon S, Birchfield S. Training deep networks with synthetic data: Bridging the reality gap by domain randomization Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018, pp. 969-977.

Vabalas A, Gowen E, Poliakoff E, Casson AJ, 2019. Machine learning algorithm validation with a limited sample size. PloS one. **14**, e0224365.

Verheyen G, Popovic-Todorovic B, Tournaye H, 2017. Processing and selection of surgically-retrieved sperm for ICSI: a review. Basic and Clinical Androl. **27**, 1-10.

VerMilyea M, Hall J, Diakiw S, Johnston A, Nguyen T, Perugini D, Miller A, Picou A, Murphy A, Perugini M, 2020. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. Hum. Reproduction. **35**, 770-784.

Wang Y, Riordon J, Kong T, Xu Y, Nguyen B, Zhong J, You JB, Lagunov A, Hannam TG, Jarvi K, 2019. Prediction of DNA integrity from morphological parameters using a single-sperm DNA fragmentation index assay. Advanced Science. **6**, 1900712.

WHO. *Laboratory manual for the examination and processing of human semen*. Sixth edn, 2021. World Health Organisation.

Wosnitzer M, Goldstein M, Hardy MP, 2014. Review of azoospermia. Spermatogenesis. **4**, e28218.

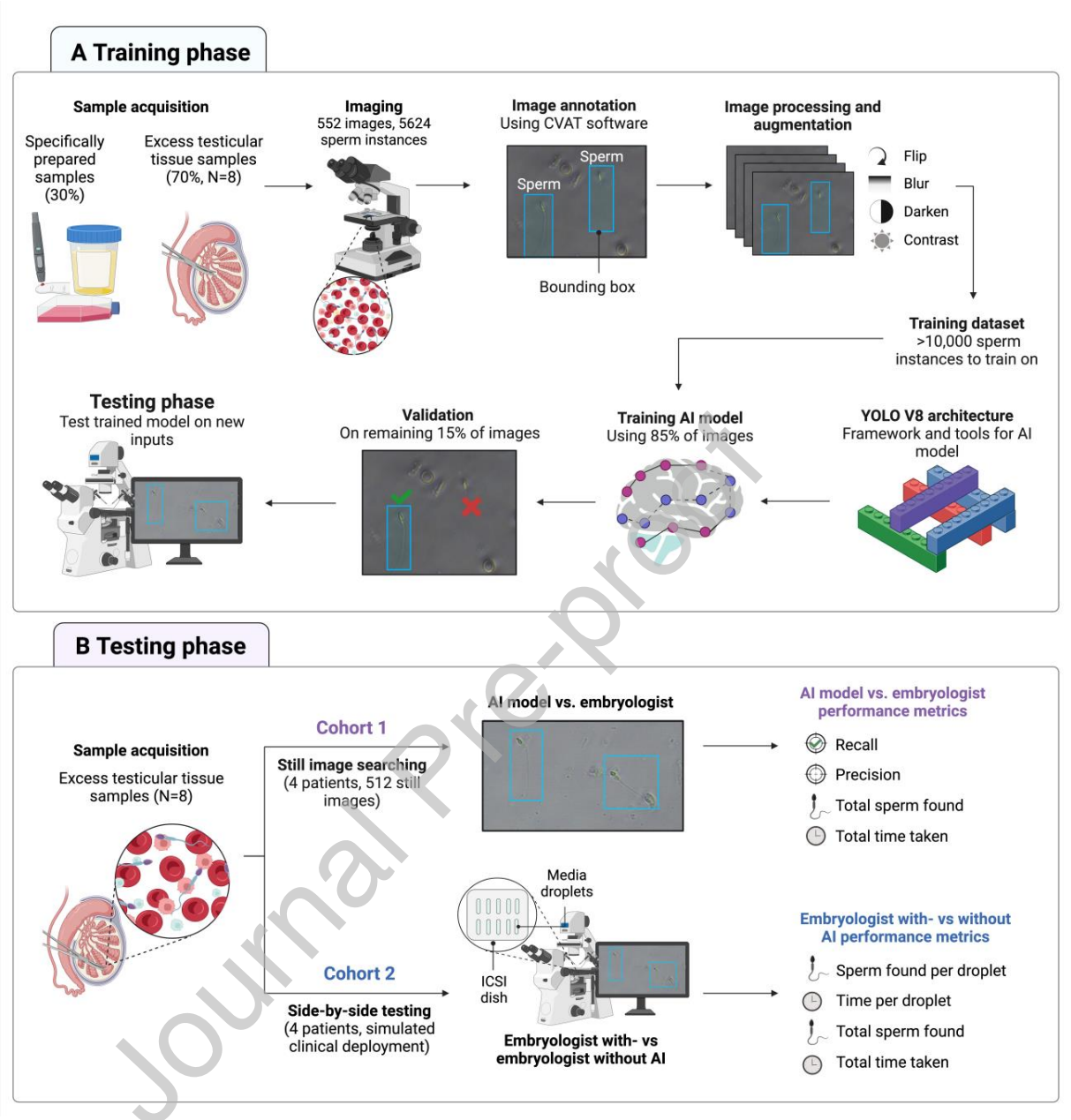

Figure 1 | Overview of the study phases. (A) The training phase begins with sample acquisition from 30% specifically prepared samples and 70% testicular tissue samples, which are plated in dishes and imaged at 200X. Images are then annotated using CVAT creating bounding boxes around all sperm in each image. Each image is processed and augmented to create a training dataset which is then used to train a model created using YOLO V8 architecture and tools. 85% of these images are used to train the model and 15% are used to validate the model performance before testing. **(B)** The testing phase begins with sample acquisition of excess testicular tissue

from NOA patients and testing performance of the model on still images (cohort 1, N=4) versus an embryologist and side-by-side testing of an embryologist with and without the aid of the AI in a simulated real-world sperm search using an ICSI microscope (N=4).

Table 1 | AI and embryologist sperm search performance metrics for comparison.

	Embryologist	**AI**	**P-value**
Cohort 1 (still images)			
Time per FOV (s)	36.10 ± 1.18	$0.02 \pm 0.3 \times 10^{-5}$	<0.0001
Recall (%)	86.52 ± 1.34	91.95 ± 0.81	0.0006
Precision (%)	98.18 ± 0.38	89.58 ± 0.87	<0.0001
No. of sperm found (of 2660)	1937	1997	N/A
Cohort 2 (side-by-side deployment)			
Time taken per drop (s)	168.7 ± 7.84	98.9 ± 3.19	< 0.0001
Total time taken (s)	6749.71	3955.89	N/A
Sperm found per drop	31.85 ± 3.09	34.9 ± 3.43	0.3843
Total no. of sperm found	1274	1396	N/A

Data are presented as the mean ± SEM or total. Between-group differences were tested using the two-tailed unpaired Student's t-test or Mann–Whitney U test if the data were not normally distributed. Two-way analysis of variance to assess the effects of the counting method and sample were performed. AI, artificial intelligence; FOV, field of view; N/A, not applicable; s, seconds

Author Biography:

Dale Goss is a PhD student at University of Technology Sydney, Australia and a graduate of Stellenbosch University, South Africa and Monash University Melbourne. He works as a clinical embryologist at IVF Australia and scientific advisor for NeoGenix Biosciences. He has experience in animal and human embryology research and his research interests are in human embryology, male infertility and technological advances used for improving assisted reproduction.

Key Message:

This proof-of-concept study shows an artificial intelligence image analysis tool can drastically improve sperm search times on testicular tissue samples, thus reducing physical strain and fatigue on embryologists and possibly improving chances of finding sperm. This is a highly translatable clinical tool for treatment of severe male-factor infertility.

Declarations of interest:

None