scientific reports

OPEN

Deep learning classification integrating embryo images with associated clinical information from ART cycles

Mohamed Salih[1], Christopher Austin[2], Krishna Mantravadi[3], Eva Seow[4], Sutthipat Jitanantawittaya[5], Sandeep Reddy[6], Beverley Vollenhoven[1,7], Hamid Rezatofighi[2] & Fabrizzio Horta[1,8,9,10]✉

An advanced Artificial Intelligence (AI) model that leverages cutting-edge computer vision techniques to analyse embryo images and clinical data, enabling accurate prediction of clinical pregnancy outcomes in single embryo transfer procedures. Three AI models were developed, trained, and tested using a database comprised of a total of 1503 international treatment cycles (Thailand, Malaysia, and India): 1) A Clinical Multi-Layer Perceptron (MLP) for patient clinical data. 2) An Image Convolutional Neural Network (CNN) AI model using blastocyst images. 3) A fused model using a combination of both models. All three models were evaluated against their ability to predict clinical pregnancy and live birth. Each of the models were further assessed through a visualisation process where the importance of each data point clarified which clinical and embryonic features contributed the most to the prediction. The MLP model achieved a strong performance of 81.76% accuracy, 90% average precision and 0.91 AUC (Area Under the Curve). The CNN model achieved a performance of 66.89% accuracy, 74% average precision and 0.73 AUC. The Fusion model achieved 82.42% accuracy, 91% average precision and 0.91 AUC. From the visualisation process we found that female and male age to be the most clinical factors, whilst Trophectoderm to be the most important blastocyst feature. There is a gap in performance between the Clinical and Images model, which is expected due to the difficulty in predicting clinical pregnancy from just the blastocyst images. However, the Fusion AI model made more informed predictions, achieving better performance than separate models alone. This study demonstrates that AI for IVF application can increase prediction performance by integrating blastocyst images with patient clinical information.

Keywords IVF/ICSI outcome, Infertility, Embryo selection, Embryology, Deep learning

It is estimated that 1 out of 6 couples present infertility in their lifetime[1]. Assisted reproductive technologies (ART) have been established to assist these couples. However, despite more than 40 years since the first IVF baby Louise Brown[2] was born, the treatment success rate is still below 30% to achieve a healthy live birth[3], therefore, the majority of patients may need to go through multiple cycles of treatment before achieving a successful outcome[4]. This leads to significant physical, emotional and financial burden on the patient, therefore improving the success rate of treatment is essential[5]. Infertility treatment plans follow a structured workout to find the correct treatment for each patient[6]. This starts with understanding the patient characteristics such as age, body mass index (BMI), ovarian reserve and sperm parameters, among several variables. Despite this, most of the treatments are not personalised to the individual circumstances, rather, following a standardised treatment plan.

[1]Department of Obstetrics and Gynaecology, Monash University, 246 Clayton Road, Clayton, VIC 3168, Australia. [2]Dept of Data Science and Artificial Intelligence, Faculty of Information Technology, Monash University, Clayton, VIC , Australia. [3]Oasis Fertility, Hyderabad, India. [4]IVF Bridge Fertility Center, Johor, Malaysia. [5]Ibaby Fertility, Bangkok, Thailand. [6]School of Medicine, Deakin University, Geelong, VIC, Australia. [7]Women's and Newborn Program, Monash Health, Melbourne, VIC, Australia. [8]Monash Data Future Institute, Monash University, Clayton, VIC, Australia. [9]Discipline of Women's Health, Fertility & Research Centre, Royal Hospital for Women & School of Clinical Medicine, University of New South Wales, Randwick, NSW, Australia. [10]City Fertility, Sydney, NSW, Australia. ✉email: fabrizzio.horta@unsw.edu.au

In vitro fertilisation (IVF) and Intracytoplasmic Sperm Injection (ICSI) are the two most important treatments of ART. Through these procedures, multiple embryos are produced with the selection of a single embryo to be transferred to the uterus[7]. Such processes rely on the quality assessment of the embryo which is visually determined by the embryologist. Considering that visual analyses by embryologists could be prone to bias[8], the use of artificial intelligence (AI) have been introduced[9] aiming to eliminate the subjectivity associated with these assessments. Indeed, AI has excelled in the medical field through different applications that require assessment, diagnosis and predictions[10]. The best example is medical images where Convolutional Neural Network (CNN) has helped diagnose medical conditions in brain, breast, lung and other organs with high accuracy[11]. AI has also helped in genetic studies making genetic equerries efficient[12].

AI has been also utilised recently using machine learning (ML) to address ART predictions. Mainly used to predict embryo morphology to improve the subjectivity associated with embryo selection and clinical pregnancy prediction[13]. Moreover, there has also been other studies involving AI in other aspects of ART, e.g.: Oocyte assessment, sperm selection, etc[14,15]. Part of these studies have used the patients' clinical data collected during the treatments to make predictions[16], and some have used embryo images paired with clinical data[17]. However, we hypothesise that the performance of separated models for clinical data and embryo images could be enhanced by integrating the models together. Indeed, we found that models integrating embryo images and clinical information presented higher accuracies for reproductive outcomes[13]. Salih et al.[13] found that out of twenty validated studies only four studies combined clinical information and embryo images in their AI models, however, these studies did not integrate clinical information with embryo images for their predictive AI model capabilities[17–20]. Thus, we hypothesise that integrating clinical data into models with embryo images at the embryo transfer stage could assist in connecting the features of importance from both clinical and embryo characteristics that are linked to successful reproductive outcomes. This could strengthen the decision making of the model which could become more confident in the prediction[13].

In this study, we aimed to develop a fully integrated clinical information and embryo morphology model, using both clinical records of patients who underwent to IVF/ICSI treatments and still images of single blastocyst embryos at embryo selection procedures in one system. We tested both model situations separately before combining them together in one single model (fusion model). This gave us three models to compare the performance against each other. Thus, we investigated how different features aid reproductive outcome prediction of clinical interest in ART. Additionally, we also investigated the performance of the system through Machine Learning visualisation to determine how the AI used clinical information and embryo images aid in the decision-making process.

Materials and methods

Data curation

A total of 1503 treatment cycles, which after analysis a total of 1394 IVF/ICSI treatment cycle information was made available along with their corresponding clinical data and blastocyst still images at embryo transfer procedures from 3 IVF groups: Oasis Fertility in India, Fertility Bridge in Malaysia, and iBaby Fertility in Thailand. From a total of 1980 embryo blastocyst images, 1190 of these samples were single embryo transfers. Subsequently, there were 599 samples that were discarded due to artifacts or incomplete data; thus, a total of 1585 images and treatments were used. The samples for clinical pregnancy prediction were 1503 samples (238 multiple embryo transfer and 1265 single embryo transfer), which was then divided into 1048 samples for training, 154 samples for validation and 301 samples for testing. For the live birth outcome predictions there were 1585 samples (230 multiple embryo transfers and 1355 single embryo transfers, where successful cases were connected to the higher scoring embryo morphology images) which were then divided into 1109 samples for training, 159 samples for validation and 137 for blind testing.

The clinical features from records were categorized as either "clinical features", "treatment features" or "ART and embryo transfer features" depending on what point of the treatment process they were collected (Table 1). The clinical features included patient features, with no personalised data, that would be captured early in the treatment process or prior to the first consultation. This included male and female features to understand the clinical background of the patients. The treatment features were the decisions of what type of treatment was chosen; for instance, IVF or ICSI. The treatment category of fresh embryo transfer was defined as single embryo transfer performed at the end of embryo culture on either day 5 or day 6, while frozen embryo transfer was defined as embryos frozen on either day 5 or 6 that were then thawed later for embryo transfer.

Information category	Attributes considered
Clinical features	Female and Male age, Infertility Diagnosis, Female and Male BMI
Treatment features	Cycle Number, Treatment Type (IVF/ICSI), Treatment category (Fresh/Frozen), E2 and P4 Prior to Trigger
ART and embryo transfer features	Total Number of Oocytes Inseminated, Total Oocytes Fertilised (2pn), Sperm Motility, Type of media (Single step/Sequential), Embryo Count on Day 5 and 6 (Blastocyst count), Embryo age at Day of Embryo Transfer (Day5 or day6), Number of Embryo Transferred
Clinical outcomes	Pregnancy status (Clinical pregnant/not pregnant), Pregnancy Outcome (Ongoing, delivered, miscarriage)

Table 1. Clinical data incorporated in the AI models. Note: BMI: Body Mass Index; E2: oestradiol; P4: progesterone.

The ART and embryo transfer features were collected during the ART laboratory processes (Table 1). The embryo quality grading followed the Istanbul consensus[21], however, as these features required an embryologist to manually determine them, they were not included in any of the final models.

Human ethics

This study was conducted under Monash Health ethics RES-21-0000236L for a low-risk study and was undertaken using data obtained from various IVF groups: Oasis Fertility in India, Fertility Bridge in Malaysia, iBaby Fertility in Thailand, and city fertility in Australia. The project was run in accordance with the National Statement on Ethical Conduct in Human Research (NHMRC, 2007). This prospective study was also approved by the city fertility group, part of the Global CHA IVF partners, as the representative Internal Review Board (IRB), including all the boards of each clinic involved. Consent was acquired by the respective clinic for the unidentifiable samples to be retrieved and utilised for the study.

Reproductive outcomes

The pregnancy status and outcome were used as the output classes of the model; the samples that resulted in clinical pregnancy were labelled as "Pregnant" while the samples that failed to result in pregnancy were labelled as "Not pregnant". Furthermore, those samples that resulted in a healthy live birth after clinical pregnancy were labelled as "Delivered", while the ones that resulted in a miscarriage after the pregnancy were labelled as "Miscarriage". The samples that resulted in a clinical pregnancy, but the pregnancy outcome was not complete at the time of data collection were labelled as "Ongoing".

AI data sampling and training

All samples were split into 3 sets; 70% were placed in a training set, 10% in a validation set and 20% in a blind test set. The blind dataset was used to simulate new samples that the model is expected to see in a clinical setting. The blind test set was completely hidden with new data, while the validation set assists the model's accuracy with each step to correct the direction of training the model will take to avoid any overfitting or negative learning curve[22,23]. These data sets were split so that the proportion of reproductive outcomes, such as clinical pregnancy and live birth, were distributed evenly across all three sets. During training, each batch was randomly selected with an additional probability (weight) to make each batch approximately evenly balanced. In order to ensure proper data presentation, a random selection of samples was chosen to train the models using an equal number of potential output prediction samples. Furthermore, this weighted batch sampling allowed the models to learn the differences between the classes[24]. During training, each step of training was temporarily saved for further analysis and supervision. After training was completed, the step that performed best on the validation dataset was selected for evaluation on the test set.

AI models

AI models were built using Python and the open source PyTorch framework[25]. Three types of models were created to compare the performance of clinical outcome prediction;1.- clinical data model using a multi-layer perceptron (MLP) model;2.- embryo images using a convolutional neuronal network model (Resnet 34) and a 3.- fusion model including both clinical data and embryo images models by a costume made model that integrated these two models (described below).

MLP models

The clinical data models used a MLP[26] architecture. Various experiments were performed to justify the performance of the models, these include different hyperparameters of the model: the number of layers, number of neurons in each hidden layer, activation functions, batch sizes, learning rate and optimisation function. The input to the final version of the MLP model had 16 clinical data features (Fig. 1) and fed them through 3 fully connected layers of 16 × 1024 neurons, 1024 × 1024 neurons and finally 1024 × 2 neurons representing a prediction of each class (clinical pregnancy, live birth and miscarriage). The models were fine-tuned through various experimentations to define the hyper parameters of the models. These hyper parameters aided in the accuracy and precision of the models by altering the learning rate and the method used in forward and backward propagation. In order to improve reasonability and predictive power, experimentation resulted in the activation function LeakyReLU[27]and the optimisation function AdamW[28] for training the MLP model. The experiments ran for 400 epochs, batch size of 50 samples, learning rate of 0.001 and the use of cross entropy loss.

Convolutional neuronal network models

Residual Network (ResNet 34)[29] was used as the backbone for the embryo image classification system. ResNet includes skip connections which allows the network to converge more quickly, while stabilising the training. Experiments were run with different variants of ResNet, 18, 34, 50. ResNet34 was determined to be the best performing variant. Experiments were also conducted comparing ResNet pretrained on ImageNet[30] with frozen and fine-tuning convolutional weights as well as training from scratch.

Fusion models

The final type of model was a Fusion model. It used a ResNet34 backbone[29] to extract the features from the blastocyst images and an MLP backbone[31]for the clinical information. The basic architecture for these backbones were the same as those used in the clinical and image models. However, for both ResNet34 and MLP backbones, the fully connected layer was removed and instead a concatenating layer was formed that connects to a classifier layer for the final output prediction. The image features were extracted by feeding the blastocyst image into its ResNet34 backbone giving 512 outputs. The clinical data features were extracted by feeding the clinical data to

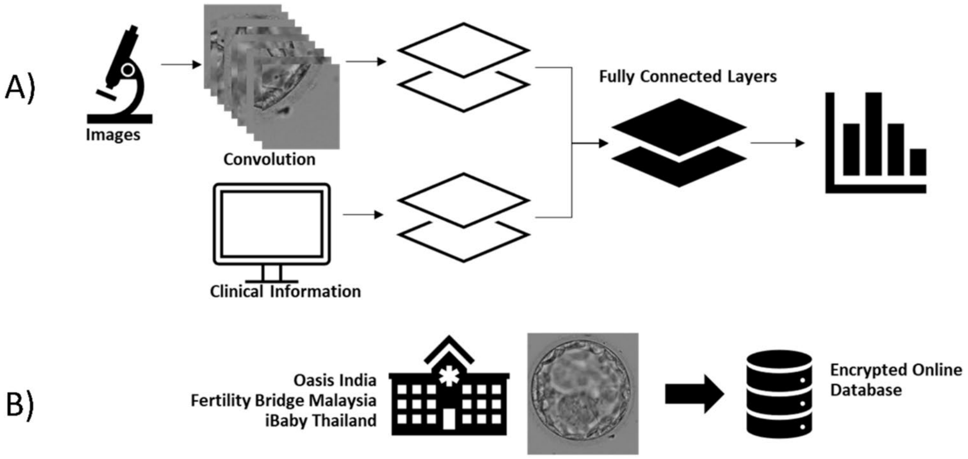

Fig. 1. **A**.- Fusion model high level architecture. The convoluted images' model output is concatenated with the clinical information model output to provide an overall probability of the output of the sample. **B**.- Data collection process through encrypted online database system.

the MLP backbone giving 1024 outputs. These data features were then concatenated and fed to a fully connected layer of shape 1536 × 2, giving a prediction of each class (Fig. 1).

Model performance metrics

All the results were performed on the blind test set that was assigned to be completely hidden from the model in the training phase. The F1-Score was calculated as the harmonic mean of the model's precision and recall. Precision was defined as the number of true positives divided by the number of true and false positives. Recall was the number of true positives divided by the number of true positives and false negatives. As the F1-Score was calculated with the number of false positives and false negatives, it is a better assessment of the performance of these models due to the imbalance of the dataset. The Average Precision was calculated as the weighted mean of precisions achieved by the model at each threshold, with the increase in recall from the previous threshold used as the weight. The Average Precision value ranges from 0 to 1, with higher values presenting better results. The AUC is the area under the ROC curve, which was calculated by plotting the true positive rate against the false positive rate. The AUC value ranges from 0 to 1, with higher values indicating greater performance. However, a completely uninformative model would achieve an AUC score of 0.5. Any model with a lower AUC can be "negatively" informative. Inverting any prediction made by the model would therefore result in better predictions.

Model performance visualisation

Visualisation process has been used to identify how each feature of the input data helps influence the output of the model. For the clinical information model, a back propagation was utilised to assess each clinical feature's importance in relevance to the output, which was further experimented in two specific situations to compliment the performance. Both tests required filtering of samples to comply, one test was for female samples above 37 years old while samples from females under the age of 30 years old were also compared. Furthermore, a mapping testing tool was used to identify the areas on the embryo image where the model has identified as important locations for the prediction.

ScoreCAM[32], a type of class activation mapping that identifies discriminative regions of an image, was used to create saliency maps on the embryo images. These generated saliency maps assign energy to the regions of the embryo images based on their contribution to the model's prediction for that sample. The areas highlighted in red shows the most important contributions to the model predictions relatively and those highlighted in blue being not employed by the models. Bayesian inference was used to assess whether these heatmap proportions could be used to predict the reliability of a prediction by the model.

In a similar manner, Layer-wise Relevance Propagation, LRP[33,34]was used to determine the importance of each of the clinical features in the clinical model's predictions. LRP works by propagating the prediction of the model backwards through the model's layers, where the neurons each layer split their relevance among those neurons that feed into them. As the total relevance was conserved in each layer, the final step determines the relevance of each of the input features to the model's prediction.

For further quantitative assessment of the visualisation heatmaps, masks were created for each embryo image to separate the embryo from the background during test procedures. These masks were used to determine the proportion of the heatmaps assigned to the embryos versus the background by the models. The masks were dilated to create three concentric ring-shaped segmentations of the embryos as well as the remaining central area of segmentation. The segmentations were created such that the outermost contained approximately 40% of the area of the embryo, the second 30% of the area, and the innermost ring and central region each contained 15% of the area. The first ring mostly contains the Zona Pelucida, while the second ring mostly contains the trophectoderm, while the third ring and innermost area mostly contain the cavity and inner cell mass.

Further reliability of prediction was experimented by manually creating masks of the embryos. These masks were dilated to create four regions, background, outer ring of embryo, inner ring of embryo, centre of embryo. The percentage of weight on each region from the saliency maps was used as input for Bayesian inference model[35]. While ScoreCam was used for assessing the image model, the algorithm of it was adapted for the reliability check. Furthermore, the final values were taken and propagate backwards to the influence of the weights, since Baysian inference is readily available and was the algorithm used to build ScoreCam. As for the fusion model the factor between image and clinical information models was split, and the percentage value for the image was back propagated to derive the accuracy of the maps.

Results

Model performance outcomes

The overall results of the models presented in this study included prediction testing for clinical pregnancy status (classes: Clinical Pregnancy and Not Pregnant) and Live birth (classes: Not pregnant, Miscarriage and Live Birth) (Table 2). Models were divided into three: Models using clinical information, embryo images, and end-to-end fusion model which helped oversee the performance of the generated models over the span of the full sample (Clinical information, embryo image and fused sample of both characteristics). Experiments for prediction of both clinical pregnancy and live birth showed an increase in the overall area under the curve compared to the embryo images model (73% Pregnancy status model, and 80% live birth model); it maintains the overall performance for the live birth model over the clinical information model (88%) while performing better for the pregnancy status model (91%) over the clinical information model (87%).

Training of the models was closely supervised where checkpoints at key iterations of the training were evaluated. Accuracy and training loss which are key points of training evaluation were output at certain iterations to closely monitor the models' training and confirm that it is going in the right directions. To confirm the correctness of training, the trained model was able to carry out prediction on the fused model with an average accuracy of 88% after the fivefold cross validation (88%, 88%, 88%, 88%, 87%) for the live birth model. Similarly, for the clinical pregnancy fused model an average of 89% was obtained after fivefold cross validation test (91%, 90%, 89%, 88%, 88%), where the best result model was used for further experiments. The average curves can be seen in Supplementary Figs. 1 and 2.

Clinical pregnancy

Figure 2. shows the performance of the model when presented with two class possibility output: Clinical Pregnancy and Not pregnant. The figure shows the three models involved in the system across ROC and precision graphs. The graphs show the Clinical information model scoring 91% accuracy and 90% precision, Images model preforming at 73% accuracy and 74% precision, and fusion model accuracy of 91% and precision of 91%.

Live birth

Figure 3. shows the performance of the model when presented with three class possibility output: Live Birth, Miscarriage, and Not pregnant. The figure shows the three models involved in the system across ROC and precision graphs. The graphs show the Clinical information model scoring 87% accuracy and 77% precision, Images model preforming at 80% accuracy and 66% precision, and fusion model accuracy of 88% and precision of 77%.

Visualisations

Embryo images model

The saliency heatmaps produced by ScoreCAM showed that most of the model's importance was assigned to the trophectoderm cells of blastocyst embryos (Fig. 4). Additionally, the importance assigned by the model included the overall shape of the embryo. Furthermore, the visualisations indicate specific cases where the AI used flawed decision making by determining the important features in the models' predictions. For instance, Fig. 4 shows two instances where the image model has assigned most of the importance of the embryo images on the empty background rather than the embryo itself.

Model	AUC (%)	Average precision (%)	F1-Score
Clinical pregnancy prediction			
Clinic data only	0.91	0.90	0.78
Embryo images only	0.73	0.74	0.57
End-to-End Fusion clinical and embryo data	**0.91**	0.91	0.76
Live birth prediction			
Clinic Only	0.87	0.77	0.64
Image Only	0.80	0.66	0.37
End-to-End Fusion	**0.88**	0.77	0.64

Table 2. Comparison of performance metrics for the different models. Note: F1-Score: Weighted average of precision and recall; AUC: Area Under the Curve.

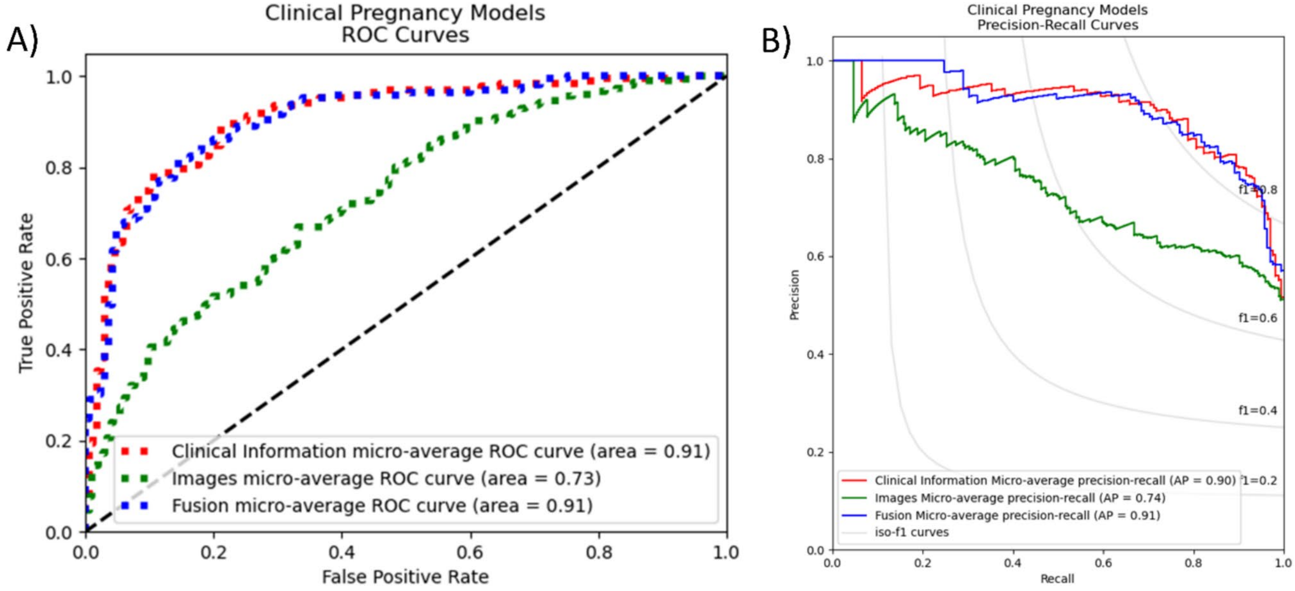

Fig. 2. ROC (receiver operating characteristic curve) and Precision-Recall graphs for the Pregnancy Status models.

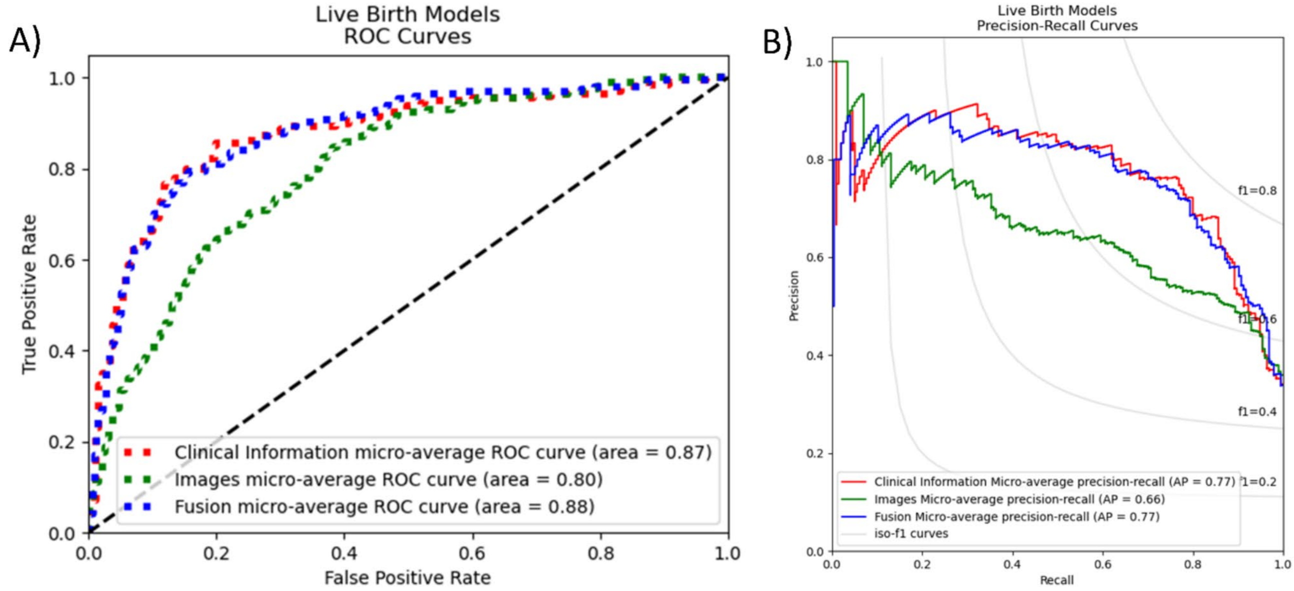

Fig. 3. ROC (receiver operating characteristic curve) and Precision-Recall graphs for the Pregnancy Status models.

Clinical data model

The average feature importance [Fig. 5] shows that the age of the female and male were the most important clinical characteristic in the model's predictions, followed by Sperm motility and female partner BMI., from 889 samples. Their high importance indicates the potential for strong performance of a model trained with only basic clinical information. The average importance showed that both female and male patient age were highly important factors to achieve successful live birth. Interestingly, the Embryology images weight played the highest weight compared to the rest of the of the variables. Two more experiments were conducted to check how the female age would affect the feature importance of the model. Although, the sample size used for this study is limited, there is a slight change in the percentage of the feature importance as can be seen in [Fig. 5—B and C].

Fusion model

For the fusion model, both ScoreCAM and LRP techniques were used [Fig. 5—A]. The overall feature importance showed that the most important was the embryo images. However, all the clinical features combined were more

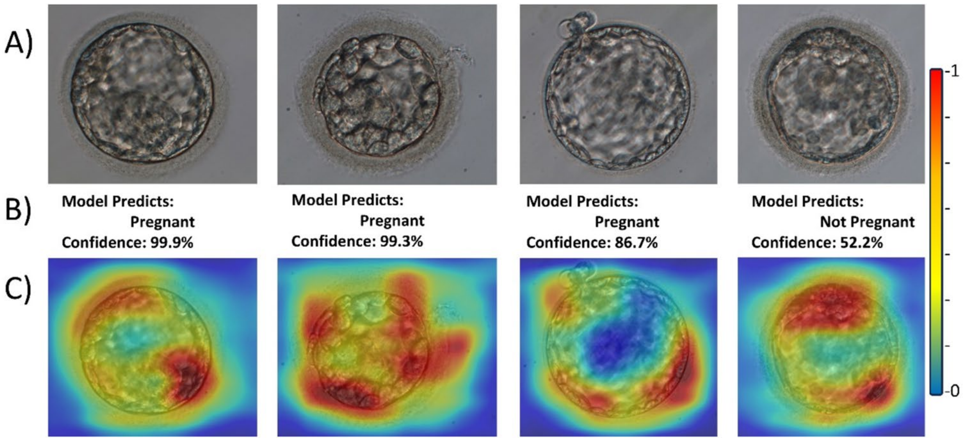

Fig. 4. Saliency Maps through ScoreCAM. **A**.- Human blastocyst embryo images before embryo transfer at day 5 of development. **B**.- Model prediction (clinical pregnancy/not clinical pregnancy) and prediction confidence based on visualisation performance. **C**.- Heatmaps to assess AI visualisation performance on each blastocyst embryo images.

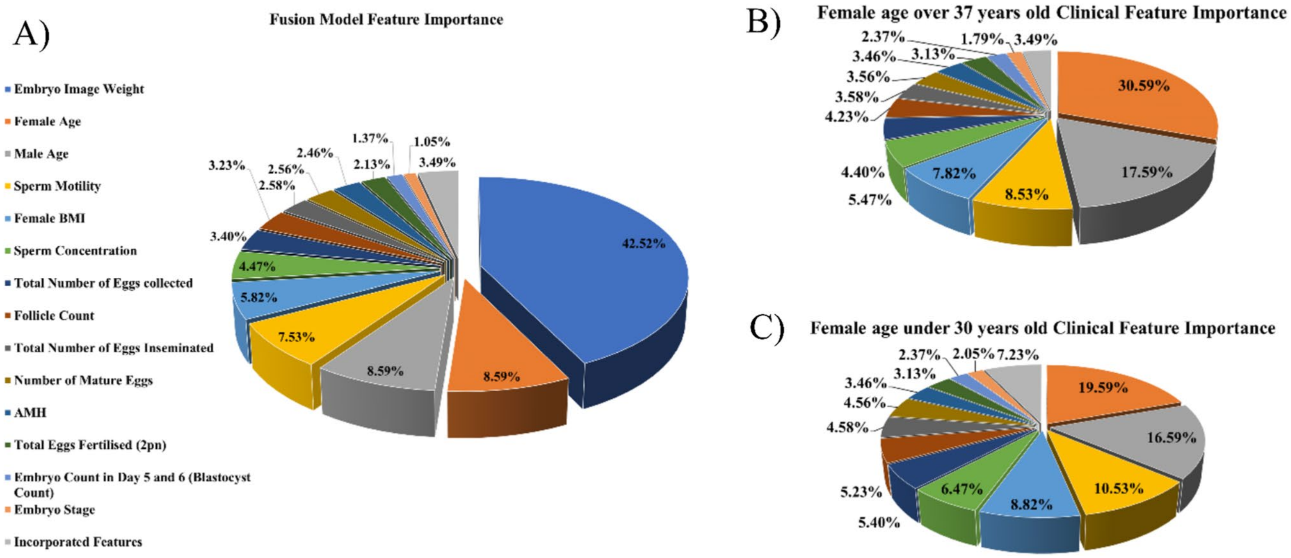

Fig. 5. **A**- Feature Importance of the fusion model including the images and the highly important features of the clinical model for live birth prediction. The chart shows that the images take a high importance in the prediction compared to each feature individually. The clinical information weight outranks the images model in the predictions. **B**- Feature importance of samples with female age over 37. **C**- Feature importance of samples with female age under 34.

important for the model's predictions than the images alone. This aligns with the relative performance of the models trained with only the image and only the clinical data. It also demonstrates that future AI systems should consider clinical factors in their predictions, as the clinical features of a patient could affect the aspects of an embryo which are most indicative of clinical and live birth outcome potential. The use of these visualisation techniques allows the performance to be assessed on a case-by-case basis in real time. This could allow a CDSS to indicate personalised treatment suggestions to improve the chance for successful implantation.

Reliability of predictions

Bayesian inference was used to assess the reliability of the fusion model's predictions for both Clinical pregnancy and live birth, whether the model was correct or incorrect, with an accuracy of 68.5% and AUC of 0.72 [Fig. 6]. This indicates that there is a recognisable difference in the produced saliency maps for samples the model makes correct predictions for, and for samples it makes incorrect predictions for. A more sophisticated segmentation of the embryos could further improve this reliability estimation. As these visualisation techniques can be applied to any sample, they allow a user to understand the behaviour of the model and critically assess its prediction.

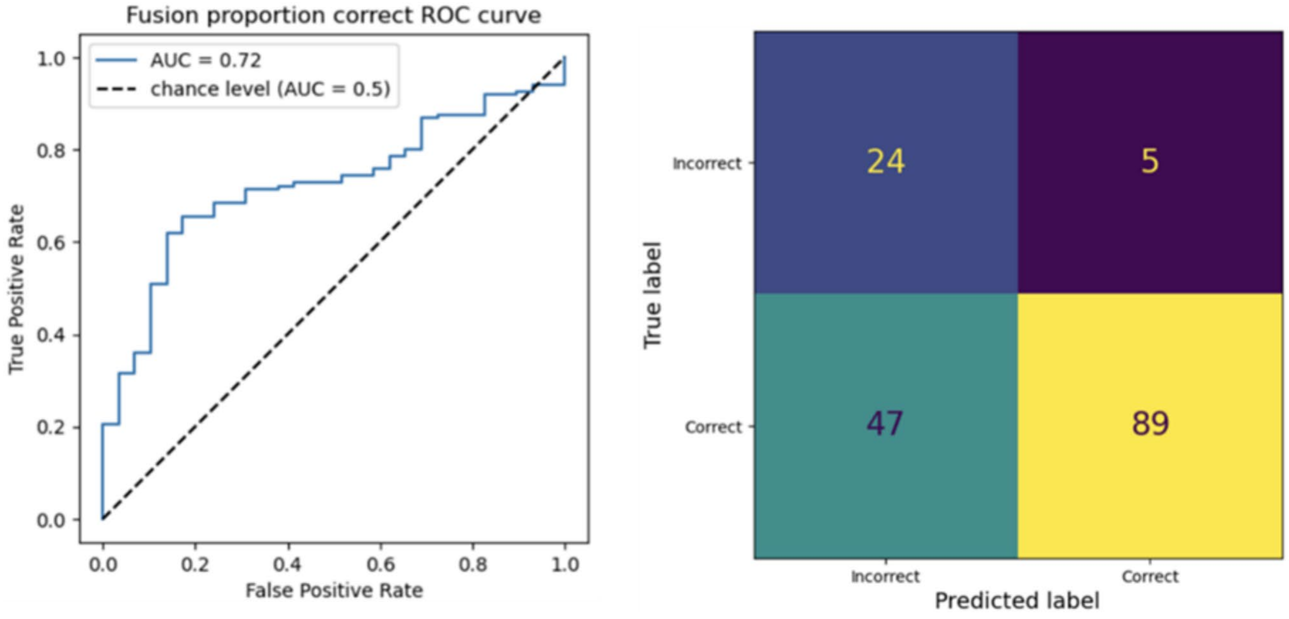

Fig. 6. Reliability measurements for area under the curve and confusion matrix.

Discussion

The proposed model in this study has addressed a critical gap in the field to standardise clinical pregnancy and live birth prediction and shorten time to pregnancy. The proposed model sets the building blocks for a clinical decision support system to aid the embryologists in treatment and choosing the right embryo for transfer. This gap was expected due to the inability of predicting clinical pregnancy and live birth considering clinical factors of relevance just by using embryo still images. The current recent studies include limited clinical information integrated into AI models, the incorporation of patients and IVF treatment information into AI models could lead to the creation of various additional suggestions to prompt embryologists and clinicians to support clinical decisions at the time of embryo transfer. The clinical model produced a strong performance in the F1-Score, average precision and AUC metrics, indicating that the clinical information features that AI can utilise increases the chances of correct clinical pregnancy and live birth outcome prediction. Although the image model performed slightly worse in terms of accuracy, the F1-Score maintained the same trend. This means that although the model is making fewer correct predictions for all samples, it was able to learn some distinguishing characteristics of the samples. However, the fusion model achieved even better performance than either the Clinical or Image model could achieve alone. With much higher performance on all metrics compared to the Image only model and higher accuracy, average precision and AUC and a similar F1-Score compared to the clinical information model.

Artificial Intelligence has revolutionised prediction and decision making and has proven the ability to learn from big datasets such as medical imaging in radiography and genomics[36–40]. This can also be a great asset in embryo selection and IVF treatment outcome prediction due to the ability to weigh variables and features leading to the proper outcome[41]. AI has also shown the flexibility to define and adapt the models created to the proper application. In this study, we have developed an AI model consisting of two essential areas in the field of IVF, a clinical model and an embryo images model, which are then combined to create the final model. As most currently proposed AI systems only use embryo images[13,42–47], the boosted performance of the fusion model presented in this study demonstrated a significant value of including the patients' clinical data in the model to predict clinical outcomes.

Combining clinical information and embryo images into one AI model has the potential to provide a tailored approach specific to the patients delivering a more personalised treatment. Incorporating this information into AI models provides insights about the patients' infertility history, which would have implications regarding the prognosis of patients. For instance, number of cycles a patient has attempted ART, plus increased reproductive age in the couple would mean a poor prognosis, which is known to negatively impact live birth outcomes in ART[48]. Sawada et al.[19] presented an approach to the gap by focusing on the time-lapse images of embryo morphology using ResNet and attention models. The highest accuracy reported was 67% when predicting live birth. However, the clinical information utilised was limited to age, method of fertilization and the type of culture media. While the embryo morphology feature was the mode of embryo transfer. In comparison for the same outcome of live birth our AI achieved an 88% prediction accuracy because the model considered more clinical features. Another model that was developed was reported by Loewke et al.[18] were the reported accuracy of 75% was achieved over image segmentation of embryo images. The outcome of the model was over ongoing pregnancy with combined input by choosing the embryos depending on the certain categories. Patil et al.[20] also approached the idea of combining clinical information with embryo images however focused on training their model on clinical information. The achieved accuracy was 86% using CNN to predict clinical pregnancy however embryo information was utilised as clinical records in the model. In comparison, our model achieved a 91% accuracy for the same outcome of clinical pregnancy likely due to the use of a larger database and the stronger ResNet architecture. Khosravi et al.[17] developed a model that took in embryo images as an input to rank the embryos on a three-class output. The model performed at 98% accuracy with the embryo grades grouped in their three-class system, however the model was only predicting the grade group unlike our model.

Although the previously discussed models show that there is an interest in combining clinical information with embryo images and/or time-lapse information, a gold standard has not been perfected yet. The AI model presented in this work, accompanied by the visualisation method to assess the AI performance, could lead to future developments for improving the embryo selection process by considering the individual clinical characteristics of patients. Indeed, our findings show how crucial the clinical information is for the embryo selection process, which it could facilitate new clinical strategies for patients during treatments[13]. Similarly, incorporating time-lapse videos in the proposed model as a next step in future studies will facilitate the study of early embryo development and biological milestones that could have influence on pregnancy outcomes.

Considering this approach, Liu et al.[49] have described a similar approach to the developed study. The study showed two models, patient background information model, and an embryo transfer model, that have been combined via an addition method to arrive at a prediction of live birth positive or negative outcome. A raw data input of 16 clinical patient information was used to train a multi-layer preceptor model, while padded embryo images are used to train the image model that utilises a readily available CNN backbone, EfficientNetV2-S[50]. Although, the study had a generous dataset of 17,580 samples size, this simple model could have had noise when combining unfocused focal planes with focused ones, leading to just a 77% accuracy for binary output (positive/negative) unlike the three-class output presented in this study; Live Birth/Miscarriage/No Pregnancy. Regarding visualisation, embryo heatmaps on both studies have used the backbone of XGradCam[51], however, feature importance used by the developed model has more real-time representation of the importance of clinical information compared to the logistic regression used by Liu et al[49]..

Liu et al.[49] model demonstrated a significant and important attempt to integrate clinical information with embryo images. Firstly, both inputs in the developed model accepts are raw data, hence the model has access to full availability of the data, compared to processed clinical data form the output of logistic regression and pre-analysis that Liu et al.[49] utilise on their samples. Secondly, the developed model accepts any raw embryo images where the model performs a minor image size pre-processing to setup the database before initialising the training, compared to adding padding which produces a lot of empty data that overwhelmed the model. Thirdly, we test the model's output over clinical pregnancy and live birth, including miscarriage which confirms the model's viability over more lab possible scenarios rather than a limited positive and negative outcome only. Concluding all these points have led the developed model to a better understanding of the samples which in turn led to a better accuracy outcome.

The proposed model combines clinical patient background with embryo images into one model. Although, the results presented showed that there was no drastic increase in the performance of the fusion model when embryo images are added. This could present a strong benefit in IVF laboratories since embryo selection is performed without the consideration of patient background and clinical information[8]. Furthermore, recent studies of AI in IVF applications lack real-time incorporation of clinical information, but in this study, we observed an improvement in the predictive capabilities of critical reproductive outcomes such as live birth. While the clinical model's strength lies in overall pregnancy and live birth prediction, the fused approach addresses the specific need to integrate clinical and imaging data to better identify viable embryos such that they will increase the likelihood of pregnancy and live birth. Thus, the Fusion model provides a more holistic tool for clinical practice by supporting embryo selection with equivalent accuracy across both types of inputs[13,52].

Personalised approach can play an important role in healthcare, and medical technologies have started taking this approach[53]. In IVF, recent studies have shown that combining clinical information with embryo images for each study can personalise the decision and prediction per samples[17–20]. However, these studies had a limited approach on the integration where the characteristics used relied on either the clinical information or the embryo images[13]; the models used lacked the full integration of both features. In this study, a fused model of both features can aid in various ways: whether to transfer fresh embryos or freeze, checking the number of cycles for suggestions on treatments among several clinical features. Importantly, this study approached training the model in an unbiased practice to include the broader population by including different ethnic groups[54]. However, all the samples used were collected retrospectively. This exposes the model and the data to data shifting, where the older samples in the dataset become more negligible compared to the newer collected data[55]. This can be done through testing the model in a clinical setting and survey it against embryologists to correctly assess the outcome.

Our choice of fusion approach based on a careful evaluation of existing literature[56], including the potential benefits and limitations of both early, mid and late fusion techniques. While early fusion can enhance feature extraction by integrating clinical data with imaging features[57], our preliminary experiments indicated that the marginal improvements were a significant concern. On another hand, late fusion outperformed all other fusion methods in our initial setup[58]. Hence, we opted for a hybrid strategy that balances the strengths of both early and late fusion, allowing for a more nuanced integration of clinical data while retaining the integrity of the imaging features. Our findings suggest that this approach helps in capturing complementary information from both data types, essentially important clinical information with high influence on prediction, and embryo image features the reflect a healthy embryo development. Recently, fusion of clinical information with produced images to produced specific outcome predictions could be seen heavily in the medical imaging field[59–63]. Patient specific treatment and diagnosis can be seen as a key element in the medical industry to minimise any risk taken by the medical staff and any risk taken towards the patient. Combining the clinical information with produced imaging through a decision support model provides the medical staff with all the information in one display. Importantly, although it could be argued that the integration of images to clinical data does not improve predictive performance as observed in different disciplines[49,59,60,62], this could of benefit in the IVF industry as current AI applications are currently only considering images or timelapse videos without including patient's clinical data[13]. Thus, considering our current results, future studies could explore if any benefits could be observed by integrating clinical data in the IVF clinical application.

As limitations, a challenge faced during this project was the distribution of samples between "pregnant" and "non-pregnant". In the provided data, we faced an imbalance of 69% towards positive pregnancy, which it is above what is normally expected in reproductive outcomes through IVF. In order to account for this, various strategies were used during training and evaluation of the produced models such as: dropping samples to have equal percentages or replicating a clinical situation by having less than 30% positive pregnancy samples or balancing the samples to have better representation of all possible cases. Moreover, the collected data included only information from embryos that were successfully transferred. Therefore, our models were unable to account to cycles that fail to achieve embryo transfer at the blastocyst stage, leading to a need for adjustment in the collected samples ratio to achieve the correct database usability as performed in the methodology. Importantly, diversity of the samples can play a critical role in a study of this proportion. Without testing the discarded embryos there was no indication of how much weight the samples could provide for training and prediction[24]. This is important as the inclusion of samples that were overseen for embryo transfer or were chosen for embryo freezing also contain both important clinical and embryo quality information that can impact reproductive outcomes. This can be addressed in future work by including time-lapse videos of all discarded and transferred embryos[64]. In addition, AI can be data hungry, the models developed and tested in this study suffered from a limited database. This can be addressed using a database with a larger diverse sample size which will potentially introduce more generalisability that will help the AI make more confident decisions. To help redeem the feasibility of the study is to test the concept in a clinical setting against embryologists[13,65–67]. The translation of a model at this magnitude would have a massive application in the field of embryology[68]. Moreover, this study will open the possibility to tailor decision making towards patients' requirement in each the treatment[69]. Although generalisability helps train the AI, the learnt features will suggest precise treatment and direction towards specific cases[70]. By using visualisations and other AI explainability techniques, a clinical decision support system can improve trust in its predictions as well as provide a warning when the system is making flawed predictions[71]. There is much more that can be done to improve the interpretability and explainability of this system such as prediction uncertainty estimations and active learning[72]. Due to the potential impact, for an AI system to be adopted in IVF it must have a very high predictive power as well as a high level of trust in its behaviour.

One key area arose with our model was data bias and shifting. Our approach on negating the bias was to shuffle together all the data types and possibilities into one database, as mentioned in the methods and materials. This was to assist the model to generalise the input to any possible data type input and camera/microscope, while setting the ground for future state-of-art model work and architecture such as transformers[73]. With the model setting the basic requirements of acquiring clinical features, it could be used as the standard of data collection and eliminating the chance of losing or missing any data points to avoid any possible multi-centre bias[74]. Hence the model was required to have the capability to accept any type of image that could be captured by any type and form of image capturing technique, by converting the images to the correct size and imaging clarity. Lastly, for future potential upgrades the model could be used to test continuous learning to improve data shifting[75].

Although the study results showed promising potential of integrating clinical information with embryo images, future studies requires that AI models are tested against embryologists[13]. Furthermore, prospective implementation studies that can be carried out with embryologists simultaneously, to authenticate the predictions as well as confirm the practical application of AI as a tool in decision making will also facilitate clinical validation of this present work[76]. Comparatively, using blind dataset is a standard practice to derive the performance of the model[77–79], for external validation and performance confirmation is an essential step to confirm the performance of AI models in a clinical setting[80]. Even though the model has a high accuracy against the blind test set, in a clinical setting there could be a performance drop due to the exposure to newer samples[81], thus, it is critical to assess the validity of AI models in a real-world scenario. Interestingly, the multi-modal approach presented in this study could be also explored in the future including biomarkers such as genetics, proteomics, metabolomics and even novel imaging methods such as metabolic imaging[82–84].

Ethical considerations and potential risks

The use of AI in IVF presents promising advancements but necessitates careful consideration of ethical concerns. These concerns include the potential dehumanization of the reproductive process, the introduction of algorithmic bias, a lack of transparency due to AI being a "black box"[85]. Considering this, we introduced a method to assess potential AI issues by having a multi-centre approach, producing visualisation elements of feature importance to detect AI bias. Furthermore, questions surrounding efficacy, patient autonomy and changes to informed consent[86] are of substantial relevance to be considered in future applications. Importantly, strategies to account for data privacy such as data collection, analysis of sensitive personal and genetic data, equitable access, the grading and moral status of embryos should be carefully considered before AI technologies are clinically introduced[87]. This was also an important point for the study, hence the use of encrypted databases for transferring data from clinics making access limited to principal investigators, as well as providing a common ground for embryo grounding and clinical information were key milestones for the initial study set up.

Conclusion

In conclusion, the integration of two distinct types of data; clinical information and embryo images, into AI models is a potentially worthwhile approach to embryo selection in ART. Embryo images contain a wealth of valuable features that AI algorithms can extract and analyse, such as morphological characteristics, developmental patterns, and cellular dynamics. These features, when interpreted in the context of relevant clinical data, could enable AI models to gain a comprehensive understanding of the factors influencing successful reproductive outcomes. Clinical information, including patient demographics and treatment protocols, could provide a framework for the AI to contextualize and interpret the visual features extracted from embryo images. As observed in this study; by leveraging the synergy between these two data sources, AI models could potentially achieve higher accuracy in predicting clinical pregnancy and live birth rates following single embryo transfer procedures. However, it is important to recognise that the efficacy of these AI models should be further investigated in real life applications by assessing the AI models performance against embryologist in a clinical setting, exposing both AI models and embryologist to new datasets from clinical cases.

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to patient related information but are available from the corresponding author on reasonable request.

Received: 15 April 2024; Accepted: 12 May 2025
Published online: 21 May 2025

References

1. Njagi, P. et al. Financial costs of assisted reproductive technology for patients in low-and middle-income countries: A systematic review. *Hum.Reprod. Open* **2023**, hoad007. https://doi.org/10.1093/hropen/hoad007 (2023).
2. Fishel, S. First in vitro fertilization baby-this is how it happened. *Fertil. Steril.* **110**, 5–11. https://doi.org/10.1016/j.fertnstert.2018.03.008 (2018).
3. Lazzari, E., Potancokova, M., Sobotka, T., Gray, E. & Chambers, G. M. Projecting the contribution of assisted reproductive technology to completed cohort fertility. *Popul. Res. Policy Rev.* **42**, 6. https://doi.org/10.1007/s11113-023-09765-3 (2023).
4. Martin, L. R., Williams, S. L., Haskard, K. B. & DiMatteo, M. R. The challenge of patient adherence. *Ther. Clin. Risk Manag.* **1**, 189–199. https://doi.org/10.2147/tcrm.s12160382 (2005).
5. Copp, T., Kvesic, D., Lieberman, D., Bateson, D. & McCaffery, K. J. Your hopes can run away with your realistic expectations: A qualitative study of women and men's decision-making when undergoing multiple cycles of IVF. *Hum. Reprod. Open* **2020**, hoaa059. https://doi.org/10.1093/hropen/hoaa059 (2020).
6. Kawwass, J. F., Penzias, A. S. & Adashi, E. Y. Fertility-a human right worthy of mandated insurance coverage: the evolution, limitations, and future of access to care. *Fertil. Steril.* **115**, 29–42. https://doi.org/10.1016/j.fertnstert.2020.09.155 (2021).
7. Jarvis, G. E. (2016) Early embryo mortality in natural human reproduction*Therapeutics* https://doi.org/10.12688/f1000research.8937.2
8. Kragh, M. F. & Karstoft, H. Embryo selection with artificial intelligence: How to evaluate and compare methods?. *J. Assist. Reprod. Genet.* **38**, 1675–1689. https://doi.org/10.1007/s10815-021-02254-6 (2021).
9. Sadeghi, M. R. Will artificial intelligence change the future of IVF?. *J. Reprod. & Infertil.* **23**, 139–140. https://doi.org/10.18502/jri.v23i3.10003 (2022).
10. Bates, D. W. et al. The potential of artificial intelligence to improve patient safety: A scoping review. *NPJ Digit. Med.* **4**, 54. https://doi.org/10.1038/s41746-021-00423-6 (2021).
11. Sarvamangala, D. R. & Kulkarni, R. V. Convolutional neural networks in medical image understanding: A survey. *Evol. Intell.* **15**, 1–22. https://doi.org/10.1007/s12065-020-00540-3 (2022).
12. Mishra, R. & Li, B. The application of artificial intelligence in the genetic study of Alzheimer's Disease. *Aging. Dis.* **11**(1567), 1584. https://doi.org/10.14336/AD.2020.0312 (2020).
13. Salih, M. et al. Embryo selection through artificial intelligence versus embryologists: A systematic review. *Hum. Reprod. Open* https://doi.org/10.1093/hropen/hoad031 (2023).
14. Nesvadbová, A., Hynečková, E., Halatová, M., Hoduláková, V. & Wiemer, K. The impact of a non-invasive artificial intelligence (Ai) oocyte scoring system on subsequent embryo development in group culture. *Fertil. Steril.* **120**, e43–e44. https://doi.org/10.1016/j.fertnstert.2023.05.086 (2023).
15. Cherouveim, P., Velmahos, C. & Bormann, C. L. Artificial intelligence for sperm selection-A systematic review. *Fertil. Steril.* **120**, 24–31. https://doi.org/10.1016/j.fertnstert.2023.05.157 (2023).
16. Goyal, P. et al. Clinical Characteristics of Covid-19 in New York City. *N. Engl. J. Med.* **382**, 2372–2374. https://doi.org/10.1056/NEJMc2010419 (2020).
17. Khosravi, P. et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ Digit. Med.* **2**, 21. https://doi.org/10.1038/s41746-019-0096-y (2019).
18. Loewke, K. et al. Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. *Fertil. Steril.* **117**, 528–535. https://doi.org/10.1016/j.fertnstert.2021.11.022 (2022).
19. Sawada, Y. et al. Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth. *Reprod. Biomed. Online* **43**, 843–852. https://doi.org/10.1016/j.rbmo.2021.05.002 (2021).
20. Patil, S. N., Wali, U. V. & Swamy, M. K. in *2019 International Conference on Communication and Signal Processing (ICCSP)* 0881–0886 (2019).
21. Medicine, A. S. I. R. & Embryology, E. S. I. G. Istanbul consensus workshop on embryo assessment: Proceedings of an expert meeting. *Reprod. Biomed. Online* **22**, 632–646. https://doi.org/10.1016/j.rbmo.2011.02.001 (2011).
22. Nguyen, Q. H. et al. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Math. Probl. Eng.* **2021**, 1–15. https://doi.org/10.1155/2021/4832864 (2021).
23. Xu, Y. & Goodacre, R. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J. Anal. Test* **2**, 249–262. https://doi.org/10.1007/s41664-018-0068-2 (2018).
24. Simopoulou, M. et al. Discarding IVF embryos: Reporting on global practices. *J. Assist. Reprod. Genet.* **36**, 2447–2457. https://doi.org/10.1007/s10815-019-01592-w (2019).
25. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).
26. Murtagh, F. Multilayer perceptrons for classification and regression. *Neurocomputing* **2**, 183–197. https://doi.org/10.1016/0925-2312(91)90023-5 (1991).
27. Xu, B., Wang, N., Chen, T. & Li, M. Empirical evaluation of rectified activations in convolutional network.*arXiv preprint* arXiv:1505.00853 (2015).
28. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization.*arXiv preprint* arXiv:1412.6980 (2014).
29. He, K., Zhang, X., Ren, S. & Sun, J. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

30. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2012).
31. Taud, H. & Mas, J. F. in Geomatic Approaches for Modeling Land Change Scenarios. Lecture Notes in Geoinformation and Cartography Ch. Chapter 27, 451–455 (2018).
32. Wang, H. *et al.* in *Proc. of the IEEE/CVF conference on computer vision and pattern recognition workshops.* 24–25.
33. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, e0130140. https://doi.org/10.1371/journal.pone.0130140 (2015).
34. Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. 193–209 (2019).
35. Tipping, M. E. In *advanced lectures on machine learning lecture notes in Computer Science* Ch. 3, 41–62 (2004).
36. Rahmani, A. M. et al. Artificial intelligence approaches and mechanisms for big data analytics: A systematic study. *PeerJ. Comput. Sci.* **7**, e488. https://doi.org/10.7717/peerj-cs.488 (2021).
37. Sutton, R. T. et al. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digit. Med.* **3**, 17. https://doi.org/10.1038/s41746-020-0221-y (2020).
38. Rosenberg, B. Challenges for radiologists dealing with clinical decision support systems (CDSS) from a legal point of view. *Eur. Radiol.* **33**, 7794–7795. https://doi.org/10.1007/s00330-023-10206-0 (2023).
39. Xu, B. et al. Distributed gene clinical decision support system based on cloud computing. *BMC Med. Genom.* **11**, 100. https://doi.org/10.1186/s12920-018-0415-1 (2018).
40. Gim, J. A. A genomic information management system for maintaining healthy genomic states and application of genomic big data in clinical research. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms23115963s (2022).
41. Fernandez, E. I. et al. Artificial intelligence in the IVF laboratory: Overview through the application of different types of algorithms for the classification of reproductive data. *J. Assist. Reprod. Genet.* **37**, 2359–2376. https://doi.org/10.1007/s10815-020-01881-9 (2020).
42. VerMilyea, M. et al. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. *Hum. Reprod.* **35**, 770–784. https://doi.org/10.1093/humrep/deaa013 (2020).
43. Kragh, M. F., Rimestad, J., Berntsen, J. & Karstoft, H. Automatic grading of human blastocysts from time-lapse imaging. *Comput. Biol. Med.* **115**, 103494. https://doi.org/10.1016/j.compbiomed.2019.103494 (2019).
44. Coticchio, G. et al. Cytoplasmic movements of the early human embryo: Imaging and artificial intelligence to predict blastocyst development. *Reprod. Biomed. Online* **42**, 521–528. https://doi.org/10.1016/j.rbmo.2020.12.008 (2021).
45. Bormann, C. L. et al. Deep learning early warning system for embryo culture conditions and embryologist performance in the ART laboratory. *J. Assist. Reprod. Genet.* **38**, 1641–1646. https://doi.org/10.1007/s10815-021-02198-x (2021).
46. Wu, C. et al. A classification system of day 3 human embryos using deep learning. *Biomed. Signal Process. & Control* https://doi.org/10.1016/j.bspc.2021.102943S (2021).
47. Liao, Q. et al. Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring. *Commun. Biol.* **4**, 415. https://doi.org/10.1038/s42003-021-01937-1 (2021).
48. Horta, F. et al. Male ageing is negatively associated with the chance of live birth in IVF/ICSI cycles for idiopathic infertility. *Hum. Reprod.* **34**, 2523–2532. https://doi.org/10.1093/humrep/dez223 (2019).
49. Liu, H. et al. Development and evaluation of a live birth prediction model for evaluating human blastocysts from a retrospective study. *Elife* https://doi.org/10.7554/eLife.83662 (2023).
50. Tan, M. & Le, Q. In *International conference on machine learning.* 10096–10106 (PMLR).
51. Fu, R.*et al.*Axiom-based grad-cam: Towards accurate visualization and explanation of cnns.*arXiv preprint* arXiv:2008.02312 (2020).
52. Bhide, P. et al. Clinical effectiveness and safety of time-lapse imaging systems for embryo incubation and selection in in-vitro fertilisation treatment (TILT): A multicentre, three-parallel-group, double-blind, randomised controlled trial. *Lancet* **404**, 256–265. https://doi.org/10.1016/S0140-6736(24)00816-X (2024).
53. Thimbleby, H. Technology and the future of healthcare. *J. Public Health Res.* **2**, e28. https://doi.org/10.4081/jphr.2013.e28 (2013).
54. Fremont, A., Weissman, J. S., Hoch, E. & Elliott, M. N. When race/ethnicity data are lacking: Using advanced indirect estimation methods to measure disparities. *Rand. Health Q.* **6**, 16–16 (2016).
55. Racowsky, C. et al. Standardization of grading embryo morphology. *J. Assist. Reprod. Genet.* **27**, 437–439. https://doi.org/10.1007/s10815-010-9443-2 (2010).
56. Chen, B., Huang, B. & Xu, B. Comparison of spatiotemporal fusion models: A review. *Remote Sens.* **7**, 1798–1835. https://doi.org/10.3390/rs70201798 (2015).
57. Sun, H.*et al.*Open domain question answering using early fusion of knowledge bases and text.*arXiv preprint* arXiv:1809.00782 (2018).
58. Gadzicki, K., Khamsehashari, R. & Zetzsche, C. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)* 1–6 (2020).
59. Yala, A., Lehman, C., Schuster, T., Portnoi, T. & Barzilay, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**, 60–66. https://doi.org/10.1148/radiol.2019182716 (2019).
60. Bhagwat, N. et al. Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Comput. Biol.* **14**, e1006376. https://doi.org/10.1371/journal.pcbi.1006376 (2018).
61. Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digit. Med.* **3**, 136. https://doi.org/10.1038/s41746-020-00341-z (2020).
62. Yap, J., Yolland, W. & Tschandl, P. Multimodal skin lesion classification using deep learning. *Exp. Dermatol.* **27**, 1261–1267. https://doi.org/10.1111/exd.13777 (2018).
63. Jabbour, S., Fouhey, D., Kazerooni, E., Wiens, J. & Sjoding, M. W. Combining chest X-rays and electronic health record (EHR) data using machine learning to diagnose acute respiratory failure. *J. Am. Med. Inform. Assoc.* **29**, 1060–1068. https://doi.org/10.1093/jamia/ocac030 (2022).
64. Wang, X., Wei, Q., Huang, W., Yin, L. & Ma, T. Can time-lapse culture combined with artificial intelligence improve ongoing pregnancy rates in fresh transfer cycles of single cleavage stage embryos?. *Front. Endocrinol. (Lausanne)* **15**, 1449035. https://doi.org/10.3389/fendo.2024.1449035 (2024).
65. Jacobs, C. K. et al. Embryologists versus artificial intelligence: predicting clinical pregnancy out of a transferred embryo who performs it better?. *Fertil. Steril.* **118**, e81–e82. https://doi.org/10.1016/j.fertnstert.2022.08.248 (2022).
66. Dimitriadis, I., Zaninovic, N., Badiola, A. C. & Bormann, C. L. Artificial intelligence in the embryology laboratory: A review. *Reprod. Biomed. Online* **44**, 435–448. https://doi.org/10.1016/j.rbmo.2021.11.003 (2022).
67. Zaninovic, N. & Rosenwaks, Z. Artificial intelligence in human in vitro fertilization and embryology. *Fertil. Steril.* **114**, 914–920. https://doi.org/10.1016/j.fertnstert.2020.09.157 (2020).
68. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc. J.* **6**, 94–98. https://doi.org/10.7861/futurehosp.6-2-94 (2019).
69. The Lancet Digital, H. Enhancing the success of IVF with artificial intelligence. *Lancet Digit Health* **5** e1 https://doi.org/10.1016/S2589-7500(22)00235-7 (2023)

70. Johnson, K. B. et al. Precision medicine, AI, and the future of personalized health care. *Clin. Transl. Sci.* **14**, 86–93. https://doi.org/10.1111/cts.12884 (2021).
71. Karran, A. J., Demazure, T., Hudon, A., Senecal, S. & Leger, P. M. Designing for confidence: The impact of visualizing artificial intelligence decisions. *Front. Neurosci.* **16**, 883385. https://doi.org/10.3389/fnins.2022.883385 (2022).
72. Vishwarupe, V. et al. Explainable AI and interpretable machine learning: A case study in perspective. *Procedia Comput. Sci.* **204**, 869–876. https://doi.org/10.1016/j.procs.2022.08.105 (2022).
73. Dosovitskiy, A.*et al.* An image is worth 16x16 words: Transformers for image recognition at scale.*arXiv preprint* arXiv:2010.11929 (2020).
74. Kuvas, K. R. et al. The risk of selection bias in a clinical multi-center cohort study. Results from the norwegian cognitive impairment after stroke (Nor-COAST) study. *Clin. Epidemiol.* **12**, 1327–1336. https://doi.org/10.2147/CLEP.S276631 (2020).
75. Localio, A. R., Berlin, J. A., Ten Have, T. R. & Kimmel, S. E. Adjustments for center in multicenter studies: An overview. *Ann. Intern. Med.* **135**, 112–123. https://doi.org/10.7326/0003-4819-135-2-200107170-00012 (2001).
76. Singh, M. et al. A new human embryonic cell type associated with activity of young transposable elements allows definition of the inner cell mass. *PLoS Biol.* **21**, e3002162. https://doi.org/10.1371/journal.pbio.3002162 (2023).
77. Evans, A. A. On the importance of blind testing in archaeological science: The example from lithic functional studies. *J. Archaeol. Sci.* **48**, 5–14. https://doi.org/10.1016/j.jas.2013.10.026 (2014).
78. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195. https://doi.org/10.1186/s12916-019-1426-2 (2019).
79. Sarker, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2**, 160. https://doi.org/10.1007/s42979-021-00592-x (2021).
80. Tran, D., Cooke, S., Illingworth, P. J. & Gardner, D. K. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Hum. Reprod.* **34**, 1011–1018. https://doi.org/10.1093/humrep/dez064 (2019).
81. Illingworth, P. J. et al. Deep learning versus manual morphology-based embryo selection in IVF: A randomized, double-blind noninferiority trial. *Nat. Med.* https://doi.org/10.1038/s41591-024-03166-5 (2024).
82. Chow, D. J. X. et al. Quantifying DNA damage following light sheet and confocal imaging of the mammalian embryo. *Sci. Rep.* **14**, 20760. https://doi.org/10.1038/s41598-024-71443-x (2024).
83. Morizet, J. et al. UVA hyperspectral light-sheet microscopy for volumetric metabolic imaging: Application to preimplantation embryo development. *ACS Photonics* **10**, 4177–4187. https://doi.org/10.1021/acsphotonics.3c00900 (2023).
84. Vargas-Ordaz, E. et al. Novel application of metabolic imaging of early embryos using a light-sheet on-a-chip device: A proof-of-concept study. *Hum. Reprod.* https://doi.org/10.1093/humrep/deae249 (2024).
85. Afnan, M. A. M. et al. Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Hum. Reprod. Open* **2021**, hoab040. https://doi.org/10.1093/hropen/hoab040 (2021).
86. Koplin, J. J., Johnston, M., Webb, A. N. S., Whittaker, A. & Mills, C. Ethics of artificial intelligence in embryo assessment: Mapping the terrain. *Hum. Reprod.* **40**, 179–185. https://doi.org/10.1093/humrep/deae264 (2025).
87. Warty, R. R., Smith, V., Patabendige, M., Fox, D. & Mol, B. Clarifying the unmet clinical need during medical device innovation in women's health - A narrative review. *Health Care Women Int.* **45**, 811–839. https://doi.org/10.1080/07399332.2023.2190983 (2024).

Acknowledgements

The authors would like to thank the international collaborators and clinics (iBaby Fertility and Genetics Centre, Thailand, Fertility Bridge, Malaysia, and Oasis Fertility, India) in providing enough samples for the AI development.

Author contributions

M.S. wrote the concept and overseen writing and edits of the manuscript. M.S. and C.A. carried out the experiments and prepared the figures and tables. K.M., E.S., and S.J. helped provide the database for training and testing of the AI. S.R, B.V., H.R., and F.H. assisted in the study direction and purpose. All Authors reviewed and accepted the manuscript.

Funding

Monash Data Future Institute, MDFI-2021-12, MDFI-2021-12, MDFI-2021-12.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-02076-x.

Correspondence and requests for materials should be addressed to F.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.