

IN-VITRO FERTILIZATION (IVF) OUTCOMES USING EMBRYO MORPHOLOGY AND PATIENT FERTILITY DATA FROM INDIAN COHORTS: A FEASIBILITY STUDY AND PROPOSED MULTIMODAL FRAMEWORK

Vijya Tulsani^{1*}, Dr. Kaushal Jani²

¹Research Scholar, Indus University & Assistant Professor, Parul Institute of Computer Applications, Parul University

²Associate Professor & HOD, Department of Computer Science and Engineering, Indus University, Ahmedabad

*Corresponding Author: Vijya Tulsani, vjyulsani@gmail.com

ABSTRACT

In-vitro fertilization (IVF) outcome assessment depends heavily on the manual morphological grading of embryos, a process that is subjective, prone to inter-observer variability, and difficult to reproduce across clinics. Artificial intelligence (AI) can support more objective and consistent embryo assessment, yet many models remain image-only, opaque, and validated on populations that under-represent Indian cohorts. This paper is organised in two clearly separated parts. The *implemented* part is a feasibility study: a transfer-learning baseline based on a ResNet34 convolutional neural network pretrained on ImageNet was trained to classify embryo morphology images as good or not-good using the publicly available Hung Vuong Hospital embryo dataset (840 labelled images, a single stratified 80/20 split). The baseline reached approximately 98% training accuracy and 86% validation accuracy (95% confidence interval 80–90%), with an area under the receiver operating characteristic curve of 0.77. Because the validation set is markedly imbalanced (143 not-good versus 25 good), we report metrics that remain informative under imbalance: the majority not-good class is classified well (F1-score 0.92) while the minority good class is weak (F1-score 0.45), giving a balanced accuracy of 0.67, a macro-averaged F1-score of 0.69, a Matthews correlation coefficient of 0.38, and a Cohen's κ of 0.37; the diagnostic summaries derived from the same confusion matrix (specificity 0.94, Youden's J 0.34, geometric mean 0.61, positive and negative likelihood ratios 6.36 and 0.64) are reported so that the baseline is characterised beyond headline accuracy. The trained model was deployed in an interactive web interface that returns the predicted class, a confidence score, and the class-probability distribution. The *proposed* part, which is not yet implemented, outlines a multimodal and explainable framework that would fuse embryo images with patient fertility data from Indian cohorts, add gradient- and Shapley-based explanations, and deliver predictions through a clinician-facing decision-support interface; Grad-CAM and Shapley attribution are described only as components of that proposed methodology, and the trained explanations, additional architectures, and cross-validated multimodal results all belong to the same stage. By keeping the implemented feasibility baseline distinct from the proposed framework, the study confirms the practical feasibility of AI-assisted embryo assessment and sets out a concrete, transparent path toward clinically meaningful IVF outcome prediction.

KEYWORDS: In-vitro fertilization; Embryo morphology; Deep learning; Transfer learning; ResNet34; Clinical decision support

1 INTRODUCTION

In-vitro fertilization (IVF) is one of the most widely used assisted reproductive technologies (ART) for the treatment of infertility, and its global utilisation continues to grow, with millions of treatment cycles performed annually [1, 3]. A central determinant of IVF success is the selection of the embryo with the highest implantation potential. In routine practice, this selection relies on the visual morphological grading of embryos by embryologists, combined with clinical judgement and patient history [6]. Although effective, this manual process is inherently subjective: the same embryo may be graded differently by different experts, and prediction quality depends strongly on individual experience and on conditions that vary across centres [7, 8].

Artificial intelligence (AI), and deep learning in particular, has emerged as a promising means of making embryo assessment more objective, consistent, and scalable [4, 11]. Convolutional neural networks (CNNs) learn discriminative morphological features directly from embryo images, and more recent work fuses imaging with structured clinical variables to predict clinical pregnancy and live birth [17, 15]. Nevertheless, three issues limit clinical adoption: many systems use a single data modality; most operate as opaque “black boxes” that provide little reasoning for their predictions; and few are validated on Indian IVF cohorts, which restricts their applicability to the Indian population [5, 10].

1.1 Problem statement

Current IVF outcome assessment and embryo selection depend primarily on manual morphological evaluation and the limited use of available clinical and embryological information. Such approaches are associated with subjective decision-making, inter-observer variability, and inconsistent outcome prediction. There is therefore a need for an intelligent,

data-driven framework that can assist embryologists in embryo assessment and outcome prediction through objective analysis of embryo images and clinical information, with explainable outputs and particular attention to Indian IVF cohorts.

1.2 Motivation

Two considerations motivate this work. First, embryo selection is a high-stakes decision in a costly and emotionally demanding treatment; objective and reproducible support can reduce variability and assist less-experienced centres. Second, a prediction without justification is difficult for clinicians to trust and act upon, so explainability and a usable interface are as important as predictive accuracy. As a necessary first step, the present study establishes the technical feasibility of image-based embryo assessment before extending it toward a multimodal, explainable, and clinically validated system.

1.3 Contributions

This paper makes the following contributions:

1. A feasibility study that implements a ResNet34 transfer-learning baseline for the binary classification of embryo morphology on a public embryo dataset, with a full evaluation using accuracy, precision, recall, F1-score, ROC–AUC, and the confusion matrix.
2. An honest performance analysis that quantifies the effect of class imbalance through imbalance-aware and chance-corrected measures, namely balanced accuracy, macro-F1, MCC, Cohen's κ , Youden's J , the G-mean, and the likelihood ratios, with confidence intervals where applicable, showing strong majority-class performance but weak minority-class recall.
3. A positioning of the baseline against published embryo-imaging studies that benchmark alternative architectures, clarifying how much of the remaining gap is attributable to architecture choice rather than to data and modality.
4. An interactive, browser-based decision-support prototype that returns the predicted embryo class, confidence, and class probabilities.
5. A proposed multimodal and explainable framework that integrates embryo images with patient fertility data from Indian cohorts and embeds gradient- and Shapley-based explanations within a clinician-facing interface.

1.4 Background: IVF workflow and embryo grading

In a typical IVF cycle, controlled ovarian stimulation is followed by oocyte retrieval, fertilisation in the laboratory, and several days of embryo culture, after which one or more embryos are selected for transfer or cryopreservation. Embryo selection is the step most amenable to AI support because it currently rests on the visual assessment of morphological features such as cell number, symmetry, fragmentation, and, for blastocysts, the quality of the inner cell mass and trophoctoderm [6, 11]. This assessment is performed either on static microscope images or on time-lapse recordings. Despite standardised grading schemes, agreement between observers is imperfect, and embryos that receive the same grade can lead to different implantation and pregnancy outcomes because biological, clinical, and laboratory factors all contribute to success [7, 8]. Large volumes of embryo images, treatment records, and outcomes accumulate in IVF centres but remain under-used by manual workflows. These observations frame embryo-quality classification as a well-posed first target for AI and explain why objective, reproducible, and interpretable models are valuable in this domain.

2 Literature Review

Recent literature on AI in IVF spans reviews, multimodal models, and clinical prediction studies. Comprehensive reviews establish that AI is increasingly used for embryo grading, ovarian-stimulation optimisation, and outcome prediction, and they consistently identify explainability, multimodal integration, and clinical validation as open challenges [1, 3, 4, 6]. Surveys of fertility specialists report rapidly growing adoption of AI for embryo selection, alongside cost and trust as the main barriers [2].

Early landmark work by Khosravi et al. showed that deep learning can assess human blastocyst quality at near-expert level from static images [20], and subsequent studies have extended this feasibility to time-lapse and multimodal settings. On the modelling side, time-lapse imaging systems such as iDAScore use 3D-CNNs to score embryos, with moderate correlation to euploidy and live birth [7, 9, 8]. Clinical-only models based on gradient boosting and related machine-learning methods predict pregnancy and live birth from tabular features such as age, body mass index, and hormone levels [12, 13]. Image-only ensembles have reported very high accuracy on public embryo datasets [14], while explainable-AI reviews map Shapley- and gradient-based explanation families onto the ART workflow [5]. Multimodal approaches that fuse imaging with clinical data report the strongest outcome prediction: a cross-attention video–tabular model achieved competitive performance for several outcomes [15], and a systematic review concluded that multimodal models attain the highest live-birth discrimination but lack explainability and deployed clinical tools [16, 11]. The selected base study fuses raw blastocyst images with sixteen clinical variables through late fusion and reports an area under the curve of 0.91 for pregnancy and 0.88 for live birth, with modality-specific saliency explanations, using data from three Asian centres including India [17]. Table 1 summarises representative studies and the gaps that motivate the present work.

Beyond these representative studies, several recent works sharpen the case for the proposed direction. Time-lapse approaches based on three-dimensional convolutional scoring, such as iDAScore, have been validated pre-clinically and related to embryonic euploidy and reproductive outcomes, but they report only moderate discrimination and depend on

specialised incubators [8, 7, 9]. Clinical-only machine-learning models predict pregnancy and live birth from tabular variables such as age, body mass index, and hormone levels with strong reported performance, yet they omit the morphological information on which embryologists rely [12, 13]. Image-only ensembles achieve high accuracy on curated public datasets but neither incorporate clinical context nor provide explanations [14]. Methodological and ethical reviews converge on the same conclusion: multimodal fusion, interpretability, external validation, and deployable interfaces are the decisive missing ingredients for clinical adoption [4, 5, 10, 11, 6]. This convergence motivates a framework that is multimodal, explainable, and deployable rather than a marginally more accurate image-only classifier.

Table 1: Representative studies on AI for IVF and the gaps motivating this work.

Study	Year	Data	Approach	Gap relevant to this work
Khosravi et al. [20]	2019	Blastocyst images (~10k)	CNN (Inception) quality scoring	Image-only; no clinical fusion or interface
Cimadomo et al. [9]	2023	Time-lapse (PGT-A)	DL grading (pre-clinical)	Incubator-dependent; not deployed
Salih et al. [17] (base)	2025	Images + 16 clinical variables	ResNet34 + MLP late fusion; Score-CAM/LRP	Single architecture; siloed explanations; no interface
Bai et al. [12]	2025	2,625 cycles, clinical only	Gradient boosting (XGBoost, LightGBM)	No embryo images; single centre
Wan et al. [13]	2025	Clinical (endometriosis IVF)	ML pregnancy prediction	Tabular only; no morphology
Mansour et al. [14]	2025	Public embryo images	Ensemble deep learning	Image-only; no clinical data or XAI
Selntigia et al. [7]	2025	Time-lapse	DL vs. euploidy/outcomes	Moderate discrimination; special hardware
Moysis et al. [11]	2025	Methodological review	Embryo-selection ML survey	No deployed, externally validated tool
Aufieri & Mastrocola [10]	2025	Review (ethics)	AI embryo selection	Ethics/governance gaps
Gao et al. [15]	2026	Time-lapse video + tabular	Conv3D + transformer cross-attention	No XAI; costly hardware; no interface
Ouyang & Wei [4]	2026	Review (multimodal)	Grading + pregnancy prediction	Calls for fusion, XAI, validation
Kritsotaki et al. [5]	2026	Review (XAI in ART)	SHAP/LIME, Grad-CAM taxonomy	Conceptual; no end-to-end predictor
Mrugacz et al. [6]	2026	Review (routine IVF)	AI in clinical practice	Adoption and validation barriers
Yadav et al. [16]	2026	Systematic review	Multimodal vs. unimodal	Confirms missing XAI, stacking, and GUI

2.1 Datasets and evaluation in IVF artificial intelligence

The datasets used across the reviewed studies fall into three groups: static embryo or blastocyst image collections, including public datasets and institutional archives; time-lapse video datasets captured by incubator systems such as EmbryoScope; and tabular clinical datasets containing demographic, hormonal, and treatment variables, sometimes drawn from national registries. Multimodal studies combine static or time-lapse images with clinical tables [17, 15]. Reported evaluation typically uses accuracy, the area under the receiver operating characteristic curve, sensitivity, and specificity, and the most informative studies report performance for clinically defined endpoints such as clinical pregnancy and live birth rather than embryo grade alone [16]. A recurring methodological caution is that high accuracy on small, single-centre, or class-imbalanced datasets can overstate clinical readiness, which is why external validation and calibration-aware metrics are emphasised in recent reviews [11, 16]. The public embryo dataset used in the present feasibility study belongs to the first group; it is therefore suitable for validating an image-based baseline but not for outcome-level prediction, which requires the multimodal data described in the proposed framework.

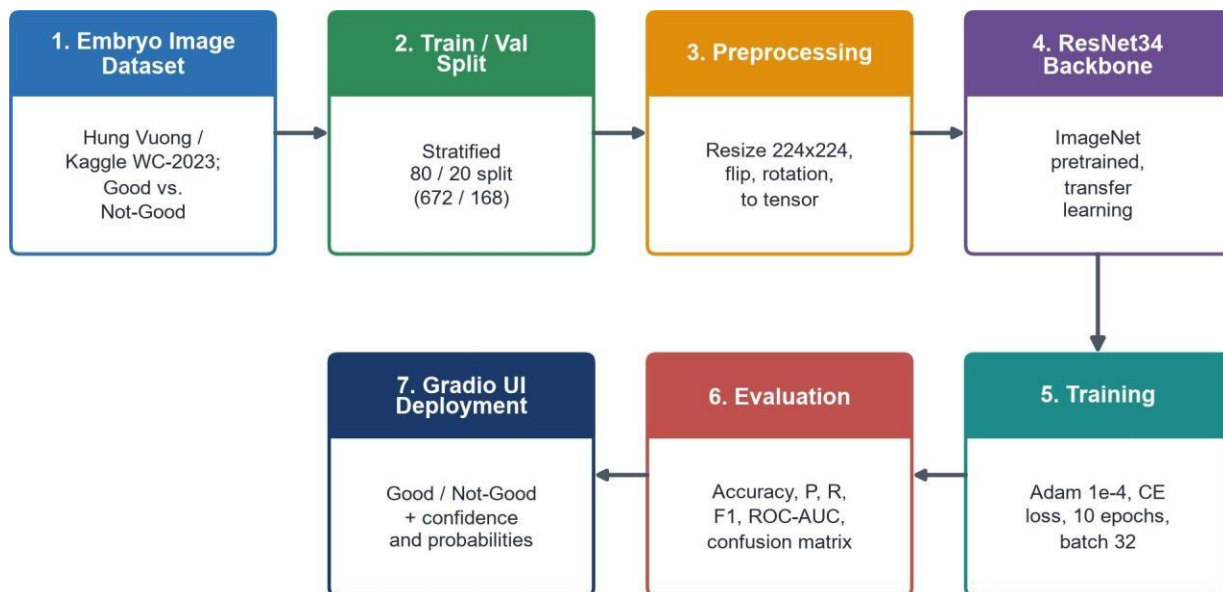


Figure 1: Implemented feasibility pipeline for ResNet34-based embryo-morphology classification, from dataset preparation to interface deployment.

2.2 Research gaps

Synthesising the reviewed literature yields five gaps that this research targets. (i) *Limited multimodal integration*: most studies use either embryo images or clinical parameters in isolation, leaving the complementary value of fusing them under-exploited [4, 15]. (ii) *Limited explainability*: many high-performing models are opaque, which hinders clinical interpretation and physician trust [5]. (iii) *Limited generalisability and Indian-cohort validation*: external validation is scarce and very few studies involve Indian IVF cohorts, reducing applicability to the Indian population [17, 16]. (iv) *Limited decision-support deployment*: most work optimises accuracy but offers little support for workflow integration and clinician decision-making. (v) *Limited feature representation*: predominantly CNN-based models emphasise local patterns and may not capture global visual relationships, motivating attention-based encoders. These gaps directly shape both the feasibility study and the proposed framework.

2.3 Research trends and positioning

Read together, the literature shows a clear trajectory. Early work established that deep networks can grade static embryo images at near-expert level [20]; the field then moved to time-lapse video and 3D-CNN scoring for richer temporal information [7, 9], and most recently to multimodal fusion of images with clinical variables, which currently reports the strongest outcome discrimination [17, 15]. Running in parallel is a second trend: explainability and clinical deployment have moved from afterthoughts to stated requirements, with recent reviews explicitly calling for interpretable, externally validated, and deployable systems [4, 5, 11, 16]. The present work is positioned at the intersection of these trends—validating a reproducible image-based baseline with a working interface and proposing a multimodal, explainable extension—rather than pursuing marginal accuracy gains on a single modality.

3 MATERIALS AND METHODS

The methodology comprises an implemented image-based feasibility study and a proposed multimodal, explainable framework. Figure 1 shows the implemented pipeline, from dataset preparation to evaluation and deployment.

Table 2: Comparison of datasets commonly used in AI-based IVF and embryo-assessment research; the dataset used in this study is shown in bold.

Dataset / source	Modality (scale)		Access	Representative use
Kaggle Hung Vuong 2023 (this work)	Static	Day-3/Day-5 embryo images (840 labelled)	Public	Reproducible image-based baseline [18]
STORK / Cornell	Static	blastocyst images (~10k)	Restricted	CNN quality scoring [20]
Public blastocyst image sets	Static	blastocyst images	Public	Ensemble image classifiers [14]
EmbryoScope time-lapse	Time-lapse	video (10 ³ –10 ⁴)	Mostly vatepri-	3D-CNN scoring (iDAScore) [7, 9]

Multi-centre image + clinical	Images + 16 clinical variables	Private	Late-fusion pregnancy/live-birth [17]
Time-lapse + tabular	Video + clinical tables	Private	Cross-attention multimodal [15]
National clinical registries	reg-Tabular clinical (10^3 – 10^5 cycles)	Restricted	ML outcome prediction [12, 13]

3.1 Dataset

Datasets for IVF artificial intelligence vary widely in modality, scale, and accessibility. Table 2 compares the main dataset families used across the literature. Most resources that support outcome-level prediction are private, time-lapse, or multimodal and are therefore difficult to reproduce, whereas public static-image collections enable reproducible image-based baselines. The present feasibility study deliberately adopts a public static-image dataset for this reason, and positions it against the richer multimodal datasets that the proposed framework will ultimately require.

The feasibility study uses the Hung Vuong Hospital embryo classification dataset, released through the Kaggle “World Championship 2023 Embryo Classification” competition [18]. The competition distributes 1,020 embryo morphology images (JPEG) of Day-3 and Day-5 embryos, of which 840 form the labelled training collection (with class labels in train.csv) and 180 form the unlabelled competition test set. Only the 840 labelled images were used here, because the unlabelled test images carry no ground truth for supervised evaluation. Each image is annotated with a binary label, *good* (class 1) or *not-good* (class 0). The labelled set is strongly imbalanced: 715 not-good (85.1%) versus 125 good (14.9%) images, a ratio that is central to the results reported below. The 840 images were divided into training and validation subsets by a single stratified 80/20 split (scikit-learn train_test_split, random_state=42), preserving the class ratio and producing 672 training images (572 not-good, 100 good) and 168 validation images (143 not-good, 25 good). The task is binary classification of embryo quality. Figure 2 shows representative embryo images.

3.2 Preprocessing

All images were resized to 224×224 pixels and converted to RGB tensors. To improve generalisation, the training set was augmented with random horizontal flips and random rotations of up to $\pm 10^\circ$. Validation images were resized without augmentation. Pixel intensities were scaled to the range expected by the pretrained backbone.

3.3 Implemented model architecture

The baseline (Fig. 3) is a ResNet34 CNN pretrained on ImageNet, used through transfer learning. The convolutional backbone provides residual feature extraction, a global average-pooling layer aggregates spatial features, and the final fully connected layer was replaced with a two-unit classifier for binary embryo-quality prediction. This choice follows the image branch of the selected base study [17] and provides a strong,



Figure 2: Representative Day-3 embryo morphology images from the dataset used in the feasibility study.

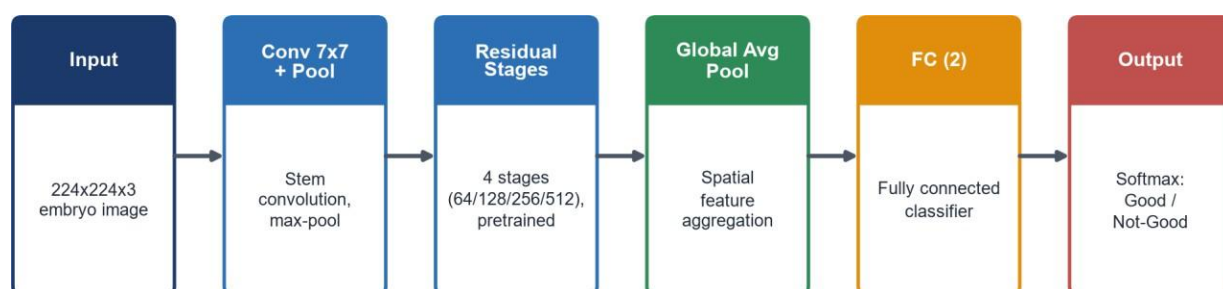


Figure 3: ResNet34 transfer-learning architecture used for embryo-morphology classification.

reproducible starting point.

3.4 Choice of backbone and alternative architectures

ResNet34 was selected for the feasibility baseline for three reasons. First, it is the image encoder used by the multimodal base study of Salih et al. [17], so adopting it keeps the present image-only result directly comparable with the modality that the proposed framework will extend. Second, its residual connections make transfer learning stable and sample-efficient, which matters on a dataset of only 840 images where larger models overfit quickly. Third, at roughly 21.8 million parameters it is light enough to fine-tune in a few minutes on free-tier cloud hardware, keeping the baseline easy to reproduce.

To place this choice in context, Table 3 contrasts ResNet34 with three architectures that the planned benchmarking will target—EfficientNet, DenseNet, and the Vision Transformer (ViT)—in terms of capacity and inductive bias rather than assumed accuracy. EfficientNet [23] and DenseNet [24] are parameter-efficient convolutional networks whose compound scaling and dense feature reuse, respectively, are attractive for small-data medical imaging, whereas ViT [22] relies on global self-attention that typically needs large-scale pretraining to compete on datasets of this size. The present study reports a measured result only for ResNet34; a controlled head-to-head comparison of these backbones, under an identical stratified protocol and imbalance handling, is a defined component of the planned work rather than a claim made here, so that any architecture ranking is established on like-for-like evidence. Published benchmarks that have already run such comparisons on embryo images are reviewed in Section 5.4, and they indicate that modern backbones differ by only a few accuracy points on this class of task.

Table 3: Backbone used in this study (ResNet34, measured) and alternative architectures targeted by the planned benchmarking. Only the ResNet34 row reports an in-house result; the remaining rows summarise architectural properties, not assumed performance.

Architecture	Params (approx.)	Inductive bias / data efficiency	Role in this study
ResNet34 [21]	21.8 M	Residual CNN; stable, sample-efficient transfer on small sets	Implemented baseline (86% val. accuracy, AUC 0.77)
EfficientNet-B0 [23]	5.3 M	Compound-scaled CNN; strong accuracy-per-parameter	Planned comparator
DenseNet121 [24]	8.0 M	Dense connectivity; feature reuse aids small-data training	Planned comparator; published embryo evidence in Table 7
ViT-B/16 [22]	86 M	Global self-attention; data-hungry without large-scale pretraining	Planned comparator

3.5 Proposed multimodal and explainable framework

The proposed framework (Fig. 4) extends the image-only baseline into a multimodal, explainable system. Embryo images are encoded by a hybrid CNN–Vision-Transformer branch, while patient fertility data from Indian cohorts (for example age, hormonal indicators, and treatment history) are encoded by a tabular branch. A cross-attention fusion module aligns the two modalities, and a prediction head estimates embryo quality and downstream outcomes such as clinical pregnancy and live birth. An explainability layer combines gradient-based image explanations (Grad-CAM/Score-CAM) [25] with Shapley-value feature attribution for clinical variables [26], and a clinician-facing interface presents the prediction, its confidence, and the supporting evidence. This framework is presented as the planned research direction; the present paper validates only the image-based baseline and the interface.

Methodologically, the framework is designed to remain lightweight enough for resource-limited clinics: a compact image encoder and a small tabular branch are combined by a cross-attention module, which retains the benefit of attention-based fusion while avoiding the computational cost of three-dimensional time-lapse encoders. Training is planned with stratified, patient-level data splits to prevent leakage between training and test partitions, with class-imbalance handling through resampling or cost-sensitive losses and repeated cross-validation reported with confidence intervals. Explanations are produced per case so that an embryologist can confirm whether the model attends to clinically meaningful evidence—for example the inner cell mass and trophoctoderm in the image and maternal age in the clinical data—and disagreements between the image and clinical branches are surfaced rather than hidden. These design choices keep the eventual system auditable and aligned with how embryologists already reason about embryo quality.

The framework can be described as five cooperating stages. (1) *Data acquisition* collects embryo images together with patient fertility variables (for example age, hormonal indicators, and treatment history) from Indian IVF cohorts. (2) *Modality-specific encoding* passes images through a hybrid CNN–Vision-Transformer branch [21, 22] and clinical variables through a tabular encoder. (3) *Cross-attention fusion* aligns the two representations so that morphological and clinical evidence reinforce one another. (4) *Prediction* estimates embryo quality and downstream outcomes such as clinical pregnancy and live birth. (5) *Explanation and decision support* attaches Grad-CAM/Score-CAM image saliency [25] and Shapley clinical attributions [26] to every prediction and presents them, with a calibrated confidence, through a

clinician-facing interface. Designing the stages as modular components allows each to be developed, validated, and audited independently before integration.

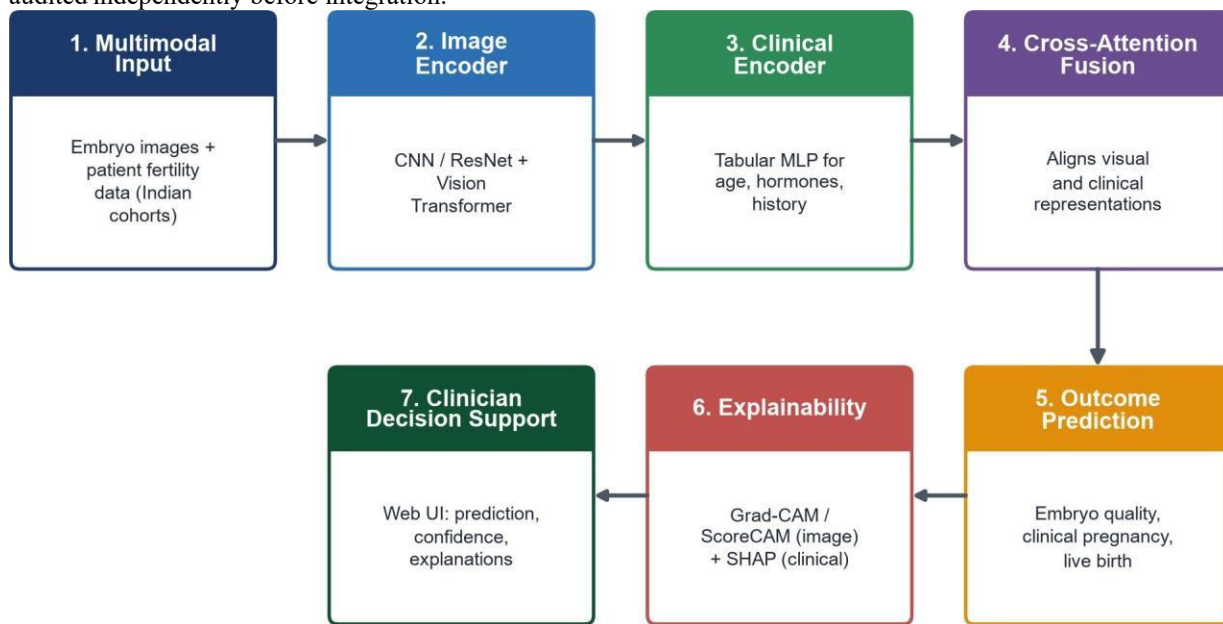


Figure 4: Proposed multimodal and explainable IVF outcome-prediction framework that fuses embryo images with patient fertility data and embeds explainability within a clinician-facing interface.

3.6 Interpretability within the proposed framework

Interpretability is a component of the proposed methodology only; *no explanation method is implemented or evaluated in the feasibility study reported in this paper*, and the following description should be read as the design that the framework stage will realise. On the image branch, gradient-weighted class-activation mapping (Grad-CAM) [25] is the natural explanation: it back-propagates the gradient of the predicted class to the last convolutional block of the encoder, weights the resulting feature maps by their pooled gradients, and produces a coarse heat map indicating which regions of the embryo image most influenced the decision. Figure 5 shows where this computation is intended to sit in the pipeline and what a clinician would ultimately see—a saliency overlay concentrated, in a clinically useful model, on the inner cell mass and trophectoderm rather than on background or dish artefacts. Shapley-value attribution [26] plays the analogous role for the clinical-variable branch, which likewise does not exist yet because no tabular data are used in the feasibility study.

Two reasons make deferring explanation the honest choice rather than an omission. First, the feasibility run did not persist its fine-tuned weights, so any heat map shown here would not correspond to the model whose numbers are reported in Section 5. Second, and more importantly, a saliency map is clinically meaningful only when it can be checked against expert annotation: an explanation that no embryologist has validated is decoration, not evidence. The framework stage therefore treats Grad-CAM/Score-CAM saliency and SHAP attributions as first-class, per-case outputs whose faithfulness is evaluated on Indian-cohort data with clinician review, and that evaluation—not an illustrative overlay—is what would justify describing the resulting system as explainable.

4 Experimental Setup

4.1 Hardware and software

Experiments were conducted in Python 3 using the PyTorch deep-learning framework with Torchvision on a GPU-enabled cloud environment (Google Colab, NVIDIA Tesla T4-class GPU). Torchvision supplied the ImageNet-pretrained ResNet34 weights, Pillow handled image loading, NumPy and pandas supported data preparation, scikit-learn computed the evaluation metrics and confusion matrix, and Matplotlib produced the

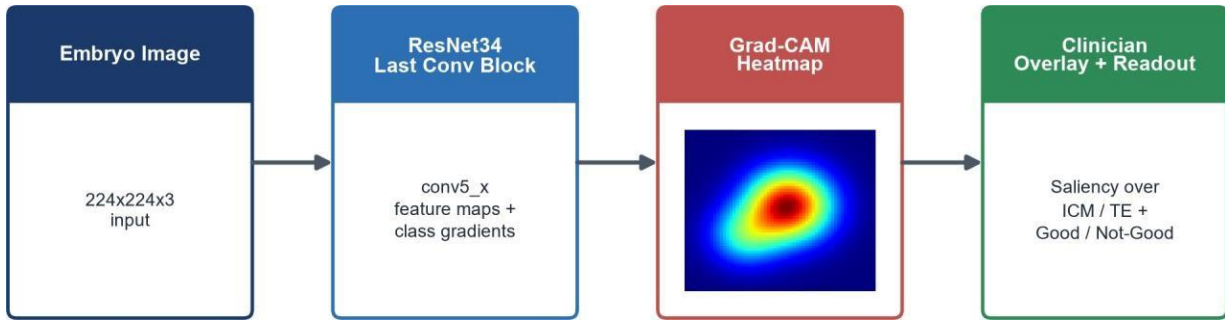


Figure 5: Grad-CAM interpretability workflow planned for the proposed framework (schematic, not implemented in this study): class gradients at the last convolutional block yield a class-activation heatmap that is overlaid on the embryo image and presented to the clinician alongside the predicted class. The heatmap panel illustrates the method; validated per-case saliency over the inner cell mass and trophectoderm is produced at the framework stage.

Table 4: Training configuration of the implemented ResNet34 baseline.

Parameter	Value	
Backbone	ResNet34 (ImageNet pretrained)	Input resolution
$\times 224$		224
Optimiser	Adam	
Learning rate	1×10^{-4}	
Loss function	Cross-entropy	Batch size
Epochs	10	32
Framework	PyTorch	

training curves and ROC plot. Because the dataset is small (840 images) and the backbone is pretrained, a full ten-epoch training run completed in a few minutes of GPU time, so the entire pipeline is straightforward to reproduce on free-tier cloud hardware.

4.2 Hyperparameters

The model was trained with the Adam optimiser at a learning rate of 1×10^{-4} , cross-entropy loss, a batch size of 32, and 10 epochs. Adam used its default moment coefficients ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with no weight decay and no learning-rate schedule; all backbone layers were fine-tuned rather than frozen; and class predictions were obtained by taking the arg-max of the two-way soft-max (equivalently, a 0.5 threshold on the good-class probability). A single fixed random seed (42) governs the split so that the exact numbers reported here are reproducible. Table 4 lists the configuration.

4.3 Training procedure and data splitting

The 840 labelled images were partitioned into training and validation subsets using a stratified 80/20 split that preserves the good/not-good class ratio in both partitions, yielding 672 training and 168 validation images; a single fixed split was used throughout so that the experiment is reproducible. Input tensors were normalised with the standard ImageNet channel statistics (mean [0.485, 0.456, 0.406], standard deviation [0.229, 0.224, 0.225]) expected by the pretrained backbone, and the training augmentations described earlier (random horizontal flips and $\pm 10^\circ$ rotations) were applied online. Transfer learning was used by initialising from ImageNet weights and fine-tuning all layers—rather than freezing the convolutional trunk—so that the network could adapt its mid- and high-level features to embryo morphology; the final fully connected layer was replaced by a two-unit classifier and optimised with Adam and cross-entropy as listed in Table 4. To expose the effect of the underlying class imbalance rather than mask it, the baseline was deliberately trained without class re-weighting or resampling, so that the resulting minority-class behaviour reported below is a faithful measurement that directly motivates the imbalance-aware training planned for the multimodal framework. Model selection used the epoch with the best validation accuracy, and a single training run is reported; repeated stratified cross-validation with confidence intervals is part of the planned extension.

4.4 Evaluation metrics

Performance was measured with accuracy, precision, recall, F1-score, the confusion matrix, and the area under the receiver operating characteristic curve (ROC-AUC). Because the validation set is imbalanced, per-class precision and recall and the confusion matrix are emphasised alongside overall accuracy. Class predictions were taken at the default decision threshold of 0.5 on the soft-max probability, while the ROC-AUC was computed from the continuous good-

class probability so that discrimination is assessed independently of any single operating point; the macro-averaged F1-score is reported alongside the weighted average so that the majority class does not dominate the summary. These quantities are defined, for the good (positive) class, as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall} + \text{F1}}$$

$$\text{macro-F1} = \frac{1}{2}(\text{F1}_{\text{good}} + \text{F1}_{\text{not-good}}) \quad (2)$$

To summarise performance in a way that does not let the majority class dominate, two further imbalance-aware measures are reported: the balanced accuracy, the mean of the per-class recalls, and the Matthews correlation coefficient (MCC), a $[-1, 1]$ -bounded correlation between predictions and labels that stays meaningful under strong imbalance,

$$\text{balanced acc} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3)$$

where TP , TN , FP , and FN are the true/false positives and negatives, and the ROC-AUC is the area under the true-positive-rate versus false-positive-rate curve traced as the decision threshold varies. On an imbalanced set the macro-F1, balanced accuracy, and MCC are more honest single summaries than raw accuracy, because they weight the rare good class comparably with the majority class.

Because the confusion matrix determines the complete set of counted outcomes, a further group of established measures can be computed exactly from it without any additional training run, and all of them are reported in Section 5. Chance-corrected agreement is quantified by Cohen's κ ; the two components of balanced accuracy are summarised jointly by Youden's index J and by the geometric mean (G-mean) of the per-class recalls, which—unlike an arithmetic mean—collapses towards zero if either class is ignored; and the clinical diagnostic convention is captured by the positive and negative likelihood ratios and the diagnostic odds ratio, which express how much a prediction shifts the odds that an embryo is truly of good quality,

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad J = \text{TPR} + \text{TNR} - 1, \quad \text{G-mean} = \sqrt{\text{TPR} \cdot \text{TNR}} \quad (4)$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{LR}^+ = \frac{\text{TPR}}{1 - \text{TNR}}$$

$$\text{LR}^- = \frac{1 - \text{TPR}}{\text{TNR}}$$

$$\text{LR} = \frac{\text{LR}^+}{\text{LR}^-}$$

$$\text{DOR} = \frac{\text{LR}^+}{\text{LR}^-}$$

$$\text{LR}^-$$

, (5) where p_o is the observed agreement (the accuracy), p_e the agreement expected by chance from the marginal totals, $TPR = TP/(TP + FN)$ the good-class recall (sensitivity), and $TNR = TN/(TN + FP)$ the not-good recall (specificity). Reporting this group costs no extra experiment yet materially widens the characterisation of the baseline; measures that genuinely require re-training or stored per-image probabilities—repeated k -fold cross-validation, the area under the precision–recall curve, and calibration—are consequently identified as planned work rather than quietly omitted. As a reference point, the no-skill average precision for the good class equals its prevalence ($25/168 = 0.149$), against which the good-class operating point is compared below. For the multimodal extension, calibration measures (for example the Brier score and reliability diagrams), the area under the precision–recall curve, and decision-curve analysis are additionally recommended so that clinically relevant operating points can be selected and compared fairly.

4.5 Reproducibility

To support reproducibility, the feasibility pipeline relies only on publicly available data and a standard ImageNet-pretrained backbone, uses a single fixed stratified data split, and records the complete training configuration in Table 4. The preprocessing, training, and evaluation steps described above are sufficient to reproduce the reported results on free-tier cloud hardware, and the deployed browser interface can be regenerated directly from the trained model weights. Reporting the pipeline at this level of detail is intended to make the baseline a dependable reference point for the multimodal extension and for independent replication.

5 RESULTS AND DISCUSSION

5.1 Performance analysis

The model learned the training data well, reaching approximately 98% training accuracy, while validation accuracy stabilised at 85–87% (Fig. 6); the gap indicates mild overfitting that is expected given the small dataset. On the validation set the overall accuracy was 86% and the ROC–AUC was 0.77, indicating satisfactory but not excellent discrimination (Fig. 7, right). The confusion matrix (Fig. 7, left) and the per-class metrics in Table 5 reveal a clear asymmetry: the majority not-good class is classified well (precision 0.90, recall 0.94, F1-score 0.92), whereas the minority good class is classified poorly (precision 0.53, recall 0.40, F1-score 0.45). The imbalance-aware summaries confirm the asymmetry: the macro-averaged F1-score is 0.69, the balanced accuracy 0.67, and the Matthews correlation coefficient only 0.38, all substantially below the 86% headline accuracy and therefore a more honest picture of a model that has largely learned the majority class. The chance-corrected and diagnostic measures agree: Cohen's κ of 0.37 places agreement in the fair range once the majority prior is discounted, Youden's J of 0.34 and a G-mean of 0.61 confirm that the two class recalls are far apart, and the likelihood ratios (LR^+ 6.36, LR^- 0.64, diagnostic odds ratio 9.93) show the clinically relevant asymmetry directly—a positive prediction raises the odds of a genuinely good embryo appreciably, whereas a negative prediction barely lowers them, so the baseline is far more useful for confirming quality than for ruling it out. This behaviour is a direct consequence of class imbalance (143 not-good versus 25 good validation images) and is an important caution for clinical use, where reliably identifying good embryos is the objective.

5.2 Statistical analysis

Because a single fixed split was used, the reported point estimates are accompanied by Wilson 95% confidence intervals so that their uncertainty is explicit rather than implied. Overall accuracy is 0.86 (95% CI 0.80–0.90; 144 of 168 images correct), and not-good recall is tightly estimated at 0.94 (95% CI 0.89–0.97, $n = 143$). The good class tells the opposite story: its recall of 0.40 carries a very wide interval (95% CI 0.23–0.59, $n = 25$) and its precision of 0.53 likewise (95% CI 0.32–0.73). These intervals are important for interpretation—the minority-class performance is not only low but also imprecisely estimated on 25 validation images, so a single split can neither be trusted to rank models nor to fix an operating point. This is precisely why the single fixed split, chosen here to make the exact numbers reproducible, is retained only as a feasibility snapshot: repeated stratified k -fold cross-validation with mean and confidence-interval reporting is part of the planned protocol (Section 5.9), and it is the appropriate tool for narrowing the wide minority-class intervals seen above before any comparative or clinical claim is made.

The learning curves in Fig. 6 are consistent with this picture. Training accuracy rises quickly and approaches saturation within the first few epochs, whereas validation accuracy plateaus at roughly 85–87% and then fluctuates—the characteristic signature of a high-capacity backbone fitting a small training set. The

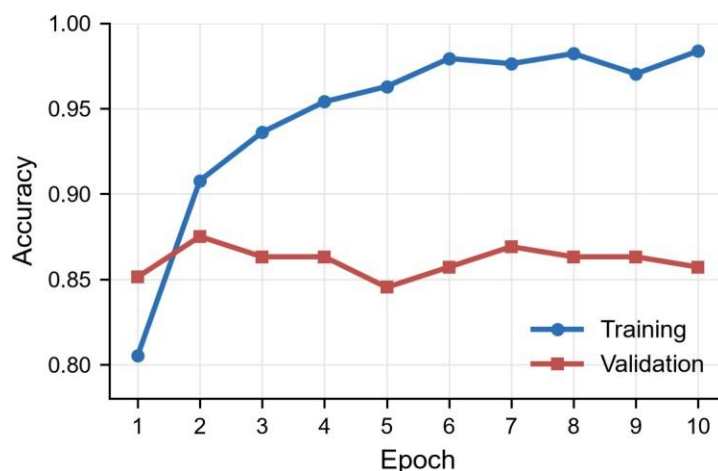


Figure 6: Training and validation accuracy across epochs for the ResNet34 baseline.

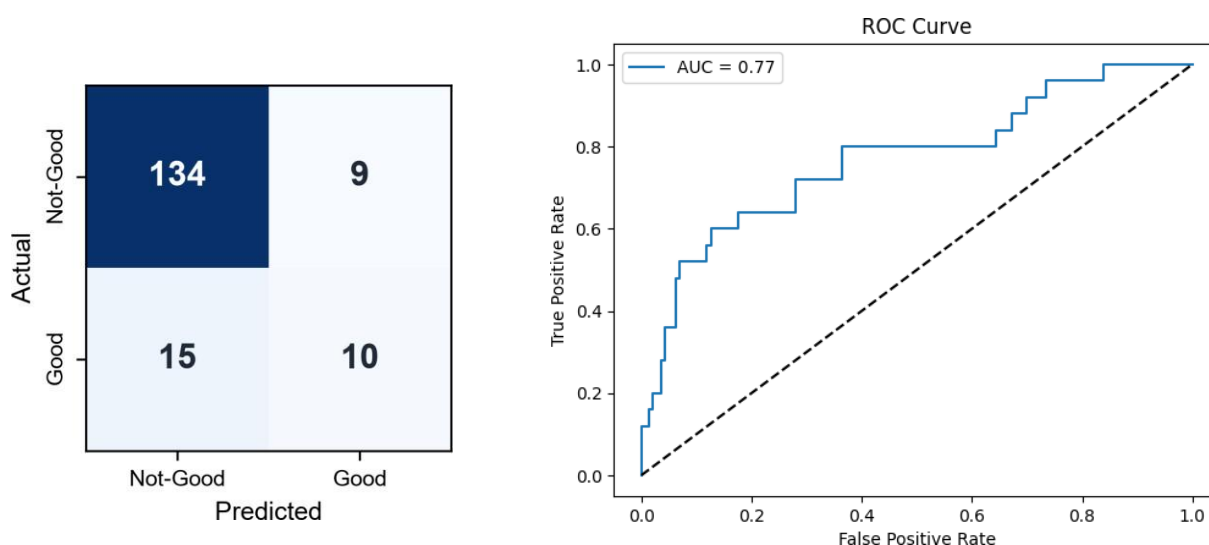


Figure 7: Left: confusion matrix on the validation set. Right: ROC curve (ROC-AUC = 0.77).

divergence between the two curves marks the onset of mild overfitting and indicates that, at this dataset size, longer training or a larger backbone would bring diminishing returns; the more productive levers are additional labelled data, explicit class balancing, and the complementary clinical modality of the proposed framework rather than further epochs on the same 672 images. This reading turns the modest baseline into concrete design guidance for the next stage rather than a result to be optimised in isolation.

5.3 Comparison with the base study

Table 6 compares the implemented baseline with the multimodal base study of Salih et al. [17]. The base study fuses embryo images with clinical variables and reports higher discrimination (AUC 0.91 for pregnancy), whereas the present feasibility study uses images only and reaches a comparable validation accuracy ($\approx 86\%$) with a lower AUC (0.77). The comparison is deliberately not a claim of parity: it shows that an image-only baseline is a sound starting point but that multimodal data and explainability, as in the proposed framework, are required to approach clinically useful performance. The present work additionally contributes a deployable interface, which the base study did not report. More importantly, the contrast between the image-only ROC-AUC of 0.77 and the base study's multimodal AUC of 0.91 illustrates the value of the missing modality. Although the two figures are measured on different endpoints—embryo grade here versus

Table 5: Classification performance of the ResNet34 baseline on the validation set, including imbalance-aware summaries (balanced accuracy, macro-F1, MCC), chance-corrected and diagnostic measures (Cohen’s κ , Youden’s J , G-mean, likelihood ratios, diagnostic odds ratio) and, where applicable, Wilson 95% confidence intervals.

Class	Precision	Recall	F1-score	Support
Not-good (0)	0.90	0.94	0.92	143
Good (1)	0.53	0.40	0.45	25
Accuracy	0.86	(95% CI 0.80–0.90; 144/168)		
Balanced accuracy		0.67		
Macro avg	0.71	0.67	0.69	168
Weighted avg	0.84	0.86	0.85	168
MCC		0.38		
Cohen’s κ		0.37		
Youden’s J		0.34		
G-mean of recalls		0.61		
LR ⁺ / LR [−]		6.36 / 0.64		
Diagnostic odds ratio		9.93		
ROC–AUC		0.77		

Sensitivity (good-class recall) = 0.40 and specificity (not-good recall) = 0.94 are the per-class recalls listed above; positive and negative predictive values are the corresponding per-class precisions (0.53 and 0.90). Wilson 95% CIs: good-class recall 0.23–0.59, good-class precision 0.32–0.73, not-good recall 0.89–0.97, not-good precision 0.84–0.94. Every quantity in this table is computed from the single validation confusion matrix of Fig. 7 and the good-class probabilities of that same run; none requires an additional training run.

clinical pregnancy in the base study—and are therefore not directly comparable, the consistent direction and size of the gap motivate the multimodal extension as the addition of a specific, potentially measurable source of signal rather than generic tuning of the image-only model.

A detail of the base study deserves emphasis because it calibrates the present result. Salih et al. report their image-only convolutional branch separately from the fused model, and that branch reaches an accuracy of 66.9% with an AUC of 0.73 [17]—below the fused model’s 82.4% and 0.91. The image-only baseline implemented here therefore sits in the same performance region as the published image-only branch rather than below it, which supports two conclusions at once: the feasibility implementation is sound, and the deficit that matters is the missing modality rather than the quality of the image model. That is precisely the deficit the proposed framework is built to close.

5.4 Comparison with published studies using other architectures

A controlled, in-house benchmark of alternative backbones is deferred to the framework stage, but the relevant evidence already exists in the literature and is summarised in Table 7 so that the architectural question is not left unaddressed. Thirumalaraju et al. [19] conducted exactly the comparison of interest on embryo images, evaluating Inception-v3, ResNet-50, Inception-ResNet-v2, NASNetLarge, ResNeXt-50, ResNeXt-101, and Xception on an independent test set of 742 embryos: the mainstream architectures cluster tightly between 89.1% and 90.8% accuracy, only NASNetLarge (78.4%) and a CNN trained from scratch separate significantly from the rest. On the DenseNet side, Mansour et al. [14] build on an improved DenseNet feature extractor within an ensemble and report 99.4% accuracy on a microscopic embryo image dataset.

Two cautions govern how Table 7 should be read, and they are the reason these numbers are presented as context rather than as a ranking. First, the studies differ in dataset, class definition, and endpoint—blastocyst versus non-blastocyst at 113 hours post insemination, ensemble-graded morphology, and the good versus not-good split used here are not the same task, so the percentages are not commensurable and a higher number does not establish a better architecture for the present problem. Second, and more informative than any

Table 6: Implemented feasibility baseline versus the multimodal base study.

Aspect	Base study [17]	This work (feasibility)
Data modality	Embryo images + 16 clinical variables	Embryo images only
Model	ResNet34 + MLP late fusion Score-CAM + LRP	ResNet34 (transfer learning) Explainability Planned in proposed framework
Output	Pregnancy / live birth	Good / not-good embryo
Reported performance	Fusion: accuracy 82.4%, AUC 0.91; image-only CNN: accuracy 66.9%, AUC 0.73	Accuracy 86%, AUC 0.77
Interface	Not reported	Interactive web interface

Table 7: Reported results from published embryo-image studies that benchmark other architectures, alongside the present baseline. Datasets, class definitions, and endpoints differ across rows, so the values give literature context and are not a like-for-like ranking.

Study	Architecture(s)	Task / data	Reported accuracy
Thirumalaraju et al. [19]	Xception; ResNeXt-101; Inception-ResNet-v2; Inception-v3; ResNeXt-50; ResNet-50; NASNetLarge	Blastocyst vs. non-blastocyst at 113 hpi; 1,188 train / 510 val. / 742 independent test embryos	90.5%; 90.8%; 90.2%; 89.7%; 89.9%; 89.1%; 78.4%
Mansour et al. [14]	Improved DenseNet + TCN/ENN/CVAE ensemble	Embryo grading on a microscopic image dataset	99.4%
Salih et al. [17] (branch)	ResNet34 (image)	Clinical pregnancy from blastocyst images; 1,503 cycles	66.9% (AUC 0.73)
This work (learning)	ResNet34 (transfer)	Good vs. not-good morphology; 840 images, single 80/20 split	85.7% (AUC 0.77)

single value, is the pattern: across the seven backbones benchmarked under one protocol the spread among modern architectures is on the order of two accuracy points, which indicates that architecture choice is not the dominant lever at this data scale. Dataset size, class balance, and the availability of a second modality are—consistent with the imbalance-driven behaviour observed in the present baseline, and a further reason to direct effort at the multimodal extension rather than at backbone substitution.

5.5 Deployment and decision support

The trained model was deployed in an interactive, browser-based interface (Fig. 8). The user uploads an embryo image and receives the predicted class (good or not-good), a confidence score, and the full class-probability distribution in real time. This transforms the offline experiment into a usable prototype and provides the foundation for the clinician-facing decision-support interface described in the proposed framework.

5.6 Clinical implications and robustness

From a clinical standpoint, the asymmetry between majority- and minority-class performance is the most important result. In embryo selection the costly error is failing to recognise a genuinely good embryo, which

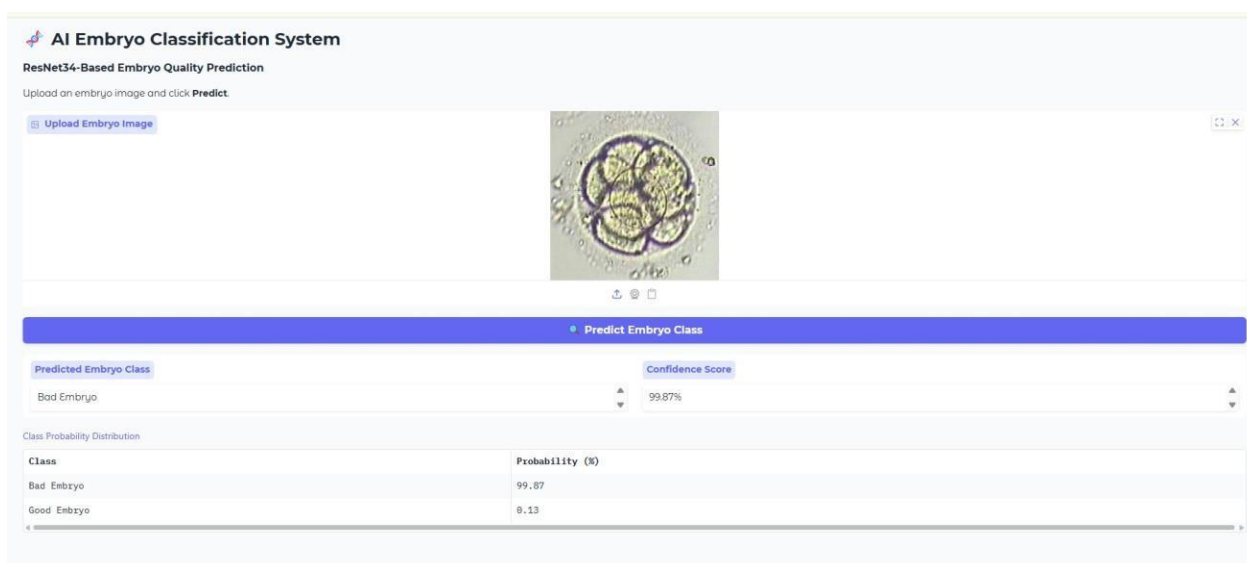


Figure 8: Browser-based prototype for AI-assisted embryo assessment, showing the predicted class, confidence score, and class-probability distribution.

corresponds to the low recall (0.40) observed for the good class. Headline accuracy is therefore a misleading summary for an imbalanced, safety-relevant task, and future development must report and optimise minority-class recall, calibration, and decision thresholds rather than accuracy alone. Robustness also requires evaluation beyond a single random split: stratified cross-validation, sensitivity to image acquisition conditions, and external validation on independent centres are needed before any clinical claim. Because embryo grade is only a proxy for the outcomes that matter to patients, the proposed framework targets clinical pregnancy and live birth and couples every prediction with an explanation, so that an embryologist can audit whether the model attends to clinically meaningful regions and variables. These requirements motivate the multimodal, explainable design rather than further tuning of the image-only baseline.

5.7 Threshold and operating-point analysis

The reported metrics use the default decision threshold of 0.5, but for an imbalanced, safety-relevant task this operating point is unlikely to be optimal. Because the ROC–AUC of 0.77 reflects discrimination across all thresholds, the good-class recall of 0.40 can be raised by lowering the decision threshold on the good-class probability, at the cost of lower precision and more false positives—a trade-off that, in embryo selection, may be acceptable, since flagging a borderline embryo for expert review is safer than discarding a potentially viable one. In practice the threshold should be selected with embryologists from the precision–recall curve rather than fixed at 0.5, and reported together with the resulting sensitivity and specificity. This threshold dependence reinforces the paper’s central methodological message: on imbalanced clinical data, a single accuracy figure at a default threshold is an incomplete description of model behaviour, and the operating point must be treated as a clinical decision in its own right.

5.8 Clinical translation, ethics, and governance

Translating an embryo-assessment model into practice raises requirements beyond accuracy. Predictions must be calibrated so that reported confidences are trustworthy, and decision thresholds should be chosen together with embryologists rather than fixed at the default of 0.5. Because embryo selection influences treatment decisions, models should be externally validated across clinics and imaging devices, and their performance audited across patient subgroups to avoid inequitable behaviour. Data governance is equally important: embryo images and patient fertility records are sensitive, so anonymisation, access control, informed consent, and adherence to local ethical approval are prerequisites for the multimodal extension that uses Indian-cohort clinical data. Finally, an AI output should be presented as decision support that an embryo-

Table 8: Limitations of the feasibility study and the corresponding component of the proposed framework that addresses each one.

Limitation (feasibility study)	How the proposed framework addresses it
Image-only, single modality	Cross-attention fusion of embryo images with patient fertility (clinical) data
Indian-cohort or external data validation	No Indian-cohort collection through clinical collaboration; external multi-centre
Class imbalance → weak, imprecise good-class recall	Cost-sensitive loss and resampling; operating point chosen from the precision–recall curve with clinicians
Single fixed split; wide confidence intervals	Repeated stratified <i>k</i> -fold cross-validation with mean and confidence-interval reporting and across-fold statistical testing
Single architecture (ResNet34) with a measured result; other backbones covered only through published studies	Head-to-head benchmark of EfficientNet, DenseNet, and ViT under an identical protocol
Explainability not implemented; described as proposed methodology only	Grad-CAM/Score-CAM and SHAP as per-case outputs, evaluated for faithfulness against expert annotation
Fine-tuned weights and probabilities not persisted	Versioned checkpoints and stored probability outputs enabling reproducible saliency and PR-AUC
Proxy target (embryo grade)	Prediction of clinical pregnancy and live birth as the clinical endpoints

ologist can interpret and override, not as an autonomous verdict; the explainability and interface components of the proposed framework are intended to make this human-in-the-loop use practical.

5.9 Limitations

The feasibility study has several limitations, and it is deliberately reported as a narrow, honest baseline rather than a finished system. It uses a small, public, single-source image dataset and an image-only modality; it does not yet include clinical or patient fertility data, nor Indian-cohort data, and therefore the title’s full multimodal and Indian-cohort scope is addressed by the proposed framework rather than by the present experiments. Performance on the minority good class is weak because of class imbalance, and—as the wide confidence intervals in the statistical analysis show—also imprecisely estimated on a single split. Explainability is not implemented at all in this study: Grad-CAM and SHAP appear only as components of the proposed methodology (Section 3.6, Fig. 5), which is why explainability is absent from the paper’s title and claims. Only one architecture is benchmarked with an in-house measured result—alternative backbones are discussed through published benchmarks (Section 5.4) rather than run here—and the predicted target is embryo quality rather than clinical pregnancy or live birth. Rather than leaving these as open caveats, Table 8 maps each limitation to the specific component of the proposed framework that is designed to resolve it, which is also the concrete response to the review points on cross-validation, additional architectures, and explainability.

6 CONCLUSION AND FUTURE SCOPE

This paper presented a feasibility stage toward a multimodal IVF outcome-prediction system, kept deliberately separate from the framework it proposes. A ResNet34 transfer-learning baseline classified embryo morphology images with 86% validation accuracy (95% CI 80–90%) and an ROC–AUC of 0.77, but imbalance-aware metrics—a balanced accuracy of 0.67, a macro-F1 of 0.69, and a Matthews correlation coefficient of only 0.38, together with a wide confidence interval on minority-class recall—showed that class imbalance severely limited performance on the good class, underlining the need to look beyond headline accuracy. A working browser-based prototype demonstrated end-to-end, real-time embryo assessment. On this basis, a

multimodal and explainable framework was proposed that fuses embryo images with patient fertility data from Indian cohorts and provides interpretable, clinician-facing predictions.

Building directly on these findings, future work will implement the proposed framework in full: integrating embryo images with patient fertility data collected from Indian IVF cohorts through planned clinical collaboration; addressing class imbalance with resampling and cost-sensitive learning; adding hybrid CNN–Vision-Transformer image encoding with cross-attention fusion; embedding Grad-CAM/Score-CAM and SHAP explanations; predicting clinically meaningful outcomes such as clinical pregnancy and live birth; and validating the system prospectively, with calibration and external multi-centre evaluation, before any clinical use and in accordance with applicable ethical approvals. Taken together, the feasibility study and the proposed roadmap confirm the practical viability of AI-assisted embryo assessment and set a transparent, clinically grounded agenda for trustworthy IVF outcome prediction.

List of abbreviations

AI: Artificial intelligence; ART: Assisted reproductive technology; AUC: Area under the curve; CNN: Convolutional neural network; Grad-CAM: Gradient-weighted class activation mapping; IVF: In-vitro fertilization; ICSI: Intracytoplasmic sperm injection; ROC: Receiver operating characteristic; SHAP: SHapley Additive exPlanations; ViT: Vision Transformer; XAI: Explainable artificial intelligence.

Declarations

Ethics approval and consent to participate: Not applicable. The feasibility study used only publicly available, de-identified embryo images and involved no human participants recruited by the authors. The proposed Indian-cohort extension will be conducted under appropriate institutional ethical approval and informed consent.

Consent for publication: Not applicable.

Availability of data and materials: The embryo dataset analysed in this study is publicly available through the Kaggle World Championship 2023 Embryo Classification competition [18]. No proprietary data were used in the reported feasibility experiments.

Competing interests: The authors declare that they have no competing interests.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions: All authors contributed to the study conception, design, implementation, analysis, and writing, and approved the final manuscript.

Acknowledgements: Not applicable.

REFERENCES

- [1] Shoham Z (2025) Artificial intelligence in reproductive medicine: transforming assisted reproductive technologies. *J IVF-Worldwide* 3(2). <https://doi.org/10.46989/001c.137620>
- [2] Shoham G, Alexandroni H, Weissman A, Mizrahi Y (2025) Global trends in the use of artificial intelligence in reproductive medicine. *J IVF-Worldwide* 3(3). <https://doi.org/10.46989/001c.140673>
- [3] Cohen J et al (2025) Artificial intelligence in assisted reproductive technology: separating the dream from reality. *Reprod Biomed Online* 50(4):104855. <https://doi.org/10.1016/j.rbmo.2025.104855>
- [4] Ouyang X, Wei J (2026) Multi-modal artificial intelligence of embryo grading and pregnancy prediction in assisted reproductive technology: a review. *Ann Biomed Eng* 54(1):74–101. <https://doi.org/10.1007/s10439-025-03865-1>
- [5] Kritsotaki N, Diamantidis D, Koutlaki N, Machairiotis N, Tsikouras P (2026) Explainable artificial intelligence in assisted reproductive technology: bridging prediction and clinical judgment. *Biomedicines* 14(5):1024. <https://doi.org/10.3390/biomedicines14051024>
- [6] Mrugacz G et al (2026) Artificial intelligence in routine IVF practice. *Biology* 15(1):42. <https://doi.org/10.3390/biology15010042>
- [7] Selntigia A, Maresca L, Montanino Oliva D, Coianiz C, Galliano D (2025) Deep learning models: their relationship with embryonic euploidies and reproductive outcomes. *Genes* 16(8):981. <https://doi.org/10.3390/genes16080981>
- [8] Cimadomo D et al (2023) Towards automation in IVF: pre-clinical validation of a deep learning-based embryo grading system during PGT-A cycles. *J Clin Med* 12(5):1806. <https://doi.org/10.3390/jcm12051806>
- [9] Mrugacz G et al (2025) Time-lapse imaging in IVF: bridging the gap between promises and clinical realities. *Int J Mol Sci* 26(19):9609. <https://doi.org/10.3390/ijms26199609>
- [10] Aufieri R, Mastrocola F (2025) Balancing technology, ethics, and society: a review of artificial intelligence in embryo selection. *Information* 16(1):18. <https://doi.org/10.3390/info16010018>
- [11] Moysis L et al (2025) Artificial intelligence-empowered embryo selection for IVF applications: a methodological review. *Mach Learn Knowl Extr* 7(2):56. <https://doi.org/10.3390/make7020056>
- [12] Bai R, Li J, Hong X, Xuan X, Li X, Tuo Y (2025) Predictive modeling of pregnancy outcomes utilizing multiple machine learning techniques for in vitro fertilization-embryo transfer. *BMC Pregnancy Childbirth* 25(1):316. <https://doi.org/10.1186/s12884-025-07433-2>
- [13] Wan X, Yu M, Wu X, Huang Z, Tan J (2025) Machine learning prediction of clinical pregnancy in endometriosis patients following fresh IVF/ICSI-ET. *Eur J Med Res* 30(1):838. <https://doi.org/10.1186/s40001-025-03113-1>
- [14] Mansour S, Almofarreh M, Khan J et al (2025) Ensemble deep learning with advanced feature engineering for embryo evaluation on in-vitro fertilisation procedures using biomedical images. *Sci Rep*. <https://doi.org/>

10. 1038/s41598—025—31772—x
- [15] Gao Q, Yao S, Du D et al (2026) Multimodal intelligent prediction model for in vitro fertilization. *npj Digit Med* 9:147. <https://doi.org/10.1038/s41746-025-02331-5>
- [16] Yadav R, Banerjee D et al (2026) AI-based live birth prediction in IVF cycles: a systematic review without meta-analysis of model performance and validation. *Middle East Fertil Soc J* 31(1). <https://doi.org/10.1186/s43043-026-00334-0>
- [17] Salih M et al (2025) Deep learning classification integrating embryo images with associated clinical information from ART cycles. *Sci Rep* 15(1):17585. <https://doi.org/10.1038/s41598-025-02076-x>
- [18] Kaggle (2023) World Championship 2023 embryo classification (Hung Vuong Hospital embryo dataset). <https://www.kaggle.com/competitions/world-championship-2023-embryo-classification/data>. Accessed 15 June 2026
- [19] Thirumalaraju P, Kanakasabapathy MK, Bormann CL, Gupta R, Pooniwala R, Kandula H, Souter I, Dimitriadis I, Shafiee H (2021) Evaluation of deep convolutional neural networks in classifying human embryo images based on their morphological quality. *Heliyon* 7(2):e06298. <https://doi.org/10.1016/j.heliyon.2021.e06298>
- [20] Khosravi P et al (2019) 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>
- [21] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp 770–778
- [22] Dosovitskiy A et al (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: *International conference on learning representations (ICLR)*

- [23] Tan M, Le QV (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th international conference on machine learning (ICML), pp 6105–6114
- [24] Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4700–4708
- [25] Selvaraju RR et al (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 618–626
- [26] Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in neural information processing systems (NeurIPS), pp 4765–4774