Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Explainable Embryo Viability Classification Using Multi-Scale Residual Attention Networks

Marium Basheer Mahmood[1], Ahmed Sabah Ahmed ALJumaili[2]

[1]Institute of Informatics for Postgraduate Studies, University of Information and Communications Technology, Baghdad, Iraq.
[1]Department of Financial Affairs, University of Baghdad, Baghdad, Iraq.
[2]Department of Business Information Technology (BIT), College of Business Informatics, University of Information Technology and Communications, Iraq.

E-mail: [1]ms202442002@uoitc.edu.iq, [2]asabahj@uoitc.edu.iq

Abstract

Embryo viability is essential for in vitro fertilization to achieve a successful transfer, implantation, and pregnancy outcome. However, automatic embryo classification remains a difficult task, as microscopy image collections are often limited in number, not balanced between classes, and challenging for clinical staff to read and trust completely. This paper presents a robust binary model for assessing the viability of embryos from images taken during IVF using a microscope. The proposed network is characterized by a hierarchical architecture with convolution, which consists of Conv Block modules, multi-scale residual blocks, depthwise separable convolutions, residual pathways, and spatial attention. The Hung Vuong Hospital Embryo Classification dataset was adopted, consisting of 801 cleaned and annotated images of embryos. Stratification and splitting were applied to the data to create training, validation, and test sets without any leakage. To balance the classes within the training dataset, only the minority class underwent data augmentation, while the other classes remained unchanged. The proposed algorithm was compared with EfficientNetB0, DenseNet121, and ResNet50 based on accuracy, precision, recall, F1 score, confusion matrix, ROC AUC, and PR AUC scores. The latter is particularly relevant when the data is not balanced. Additionally, the use of Grad-CAM heat maps and parameter randomization sanity tests has been applied to determine whether the generated explanations are grounded in consistent performance of the model or only seem reasonable visually. The proposed model achieved an accuracy of 0.96, precision of 0.91, recall of 0.94, F1 score of 0.92, ROC AUC of 0.92, and PR AUC of 0.90. Confusion matrix analysis revealed that the model obtained 11 true positives, 1 false negative, 2 false positives, and 67 true negatives. Consequently, it can be concluded that the proposed framework is capable of delivering accurate embryo viability classifications supported by images that can be interpreted in the context of medical imaging studies with insufficient amounts of data.

Keywords: Embryo Viability Classification, In Vitro Fertilization, Convolutional Neural Networks, Multi-Scale Feature Fusion, Spatial Attention.

DOI: https://doi.org/10.36548/jiip.2026.2.009
Received: 01.03.2026, received in revised form: 07.04.2026, accepted: 24.04.2026, published: 06.05.2026

1. Introduction

Infertility is among the most concerning issues in reproductive medicine, and in vitro fertilization (IVF) is believed to be one of the most widespread assisted reproductive technologies used in the management of this issue [1], [2]. Embryo selection is one of the most important steps that influence the success of an IVF procedure, as the correct embryo selection directly impacts the likelihood of successful implantation and pregnancy [3], [4]. Nevertheless, the ability to select embryos that have the greatest potential for development, using visual inspection alone, is a notable clinical challenge [5].

Morphological assessment is currently a major form of embryo evaluation used in clinical practice. Although the blastocyst stage is used to measure the level of expansion, inner cell mass (ICM) and trophectoderm (TE) quality, cell number, blastomere symmetry, and fragmentation are assessed in day 3 embryos [1], [2], [6], [7]. Despite the clinical utility of these criteria, manual assessment is subjective and hence introduces variability subject to the observer, which presents a challenge for standardization [6], [8]. This has sparked interest in artificial intelligence and, more specifically, deep learning methods for the evaluation of embryos using images [6], [9].

In recent years, convolutional neural networks (CNNs) and transfer learning-based models have shown promising results in tasks such as embryo grading, blastocyst classification, and predicting various clinical outcomes [10], [11]. Nevertheless, even with this development, the heterogeneity of datasets, the imbalance of classes, insufficient external validation, and the uninterpretable results of these methods make it difficult to confidently apply them in clinical practice [10], [11], [12]. Especially in the field of medical imaging, not only high performance but also transparency, reliability, and clinical interpretability are among the essential requirements [13].

In this context, interpretable artificial intelligence (XAI) methods offer an important tool for making the decision mechanisms of deep learning models more understandable. Grad-CAM is one of the methods widely used for this purpose because it can visualize class-specific attention regions in CNN-based systems [13], [14]. However, it is not enough to simply present visual explanations; it is also necessary to evaluate whether these explanations are actually based on the learned representations of the model. Therefore, it is of great importance to support interpretability analyses with reliability checks [15], [16].

A small deep learning architecture is presented in this paper, and it integrates hierarchical feature extraction, multiscale information integration, and spatial attention processes to categorize embryo suitability. The suggested model was tested in comparison with EfficientNet-B0, DenseNet-121, and ResNet-50. Accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC measures were used to perform the performance analysis, which are more informative in the presence of class imbalance [17], [18]. Moreover, Grad-CAM was used to analyze the visual foundation of the model choices, and the consistency of the received explanations was assessed through further consistency tests [19]. In this regard, the study seeks to fill the gap between predictive and clinical interpretability in embryo image classification.

The remainder of this paper is organized as follows. Section II (Related Work) reviews prior studies on embryo viability assessment and explainable deep learning. Section III (Methodology) presents the proposed framework and its main components. Section IV (Experimental Results) describes the dataset and pre-processing pipeline, the experimental

setup and evaluation metrics, and includes a results and discussion that analyzes the obtained findings and highlights their implications and limitations. Finally, Section V (Conclusion) summarizes the key contributions and outlines directions for future work.

2. Related Work

In Assisted Reproductive Technologies (ART), there is a direct correlation between embryo selection, success rates, and clinical costs. Thus, the automatic extraction of embryo viability and quality based on imaging has become a new target of intense research. Deep learning-based solutions are being developed to reduce subjectivity in classification and segmentation by humans, improve predictive power, and strengthen decision support systems. However, the literature presents a disjointed view, which can be explained by several factors, such as scarce and unbalanced datasets, distribution biases caused by different centers or devices, variability in training evaluation procedures, and a lack of validation of clinical trust and explainability. The analysis below summarizes existing research, clarifying its benefits and limitations, and thus contextualizes the intended methodological progress.

The architecture for classifying embryo fertility under a severe imbalance of classes was proposed by Panaite et al.[20], utilizing a pipeline that incorporated the use of lightweight CNN models in combination with hybrid CNN-ViT backbone models. The design included batch balancing, weighted loss functions, extensive augmentation, test-time augmentation, and voting ensembles. The objective was aimed at minimizing overfitting in the smaller positive minority class and formalizing the impact of the imbalance. However, minor increases in accuracy and F1 measures show that the limitations intrinsic to the dataset and imbalance constrained the potential gains further. In the work carried out by Borna et al.[6], pregnancy outcome prediction was investigated, where a deep learning-based segmentation approach based on U-Nets was created under the Deep Embryo framework using transfer learning on three timelines post-insemination. The temporal redundancy helped avoid any single-frame biases while embedding developmental cues in the latent space.

However, the small sample size and lack of mechanisms to explain decisions, which can provide visual justification for the decisions, limit the level of clinical trust that can be inspired. Zhang et al.[21] introduced a protocol for segmentation based on the SAM architecture, utilizing a dual-branch format to deal with cleavage-stage embryos. The detection-guided instance branch addresses the boundaries of blastomeres, and the semantic branch addresses the coverage of the fragments. The reported results perform excellently in terms of mean average precision and Dice coefficient scores, and the design choices seem to have taken into account some real-life problems such as cell overlapping and prompt sensitivity. Nonetheless, the requirement for an intermediate-sized dataset of novel data instances, the complicated process of multi-stage training, as well as the absence of information about clinically significant clues, all negatively affect scalability. Kodali Radha et al. [3] sought to mitigate the problems associated with the black box nature of deep neural networks by incorporating a hybrid CNN-LSTM network with LIME for blastocyst quality prediction tasks. It is clear from this integration that visual explanations can provide a solid foundation that would be considered by clinicians as well. Yet, the results achieved from a small-sized, highly augmented dataset do not instill any trustworthiness. The FSBS-Net by Ishaq et al.[22] provides better multi-classification of TE, ICM, ZP, and BC components with multi-scale feature supplementation and multi-channel blocks. The network model is lightweight and adheres to the requirements of clinical application. However, this is offset by the fact that the

accompanying dataset is small and generalizes widely, making it likely that generalization in the real world has not been fully realized yet. Miled and Chakroun et al.[13] have expanded the concept of interpretable segmentation by including spatial, channel, and scale attention modules atop an encoder-decoder architecture based on ResNet50. The attention modules provide visual explainability while simultaneously improving the IoU and F1 measures by emphasizing relevant features on the embryo and minimizing false positives. However, limitations in terms of the narrowness and single-source nature of information acquisition restrict transferability across institutions. Wang et al.[9] presented a potential structural remedy to the issues of prognostic assessment and reproducibility at large by offering a clinical embryo dataset, collected on a large scale via standard imaging and protocol adaptation and annotated by an expert. Although the dataset's size allows for consistent model performance and reduced metric variability, the morphological nature of annotations, the imbalance among classes, and the lack of contextual factors can somewhat diminish the model's generalizability and the importance of explanation in predictive cases.

The proposed model introduces a novel approach for classification in IVF microscopy images and addresses the limitations of the aforementioned works. Motivated by these limitations, the proposed model is designed to: (i) encourage more robust representation learning under small, imbalanced data regimes and (ii) provide clinically oriented explainability with a basic level of verification. Representation learning is carried out by means of a hierarchical stacking of ConvBlocks (3×/2×/2×). Additionally, the MS-Res block employs multi-scale morphological information through two streams—a common ConvBlock stream of size 3×3 and an economical stream of size 5×5 depthwise separable convolution—prior to feature concatenation. The spatial attention mechanism further adjusts the concatenated features in order to concentrate spatially on more credible evidence. Lastly, the utilization of Grad-CAM techniques aids in boosting the interpretability of the network by identifying areas of the images corresponding to the predicted category, which helps to avoid any possibility of overfitting.

3. Materials and Methods

3.1 Overview

The assessment of viability for the embryo is considered to be a crucial phase in IVF because here the criteria, which should be met by the embryo before implantation takes place, are analyzed based on morphological features that are highly sensitive and time-based. Although various breakthroughs have been achieved in developing machine learning models, their application is quite limited because of the low interpretability of internal representations as well as the absence of proof about the relationship between visual features and classification. The proposed model in this paper can address such issues through incorporating interpretable information into the representation using multi-scale analysis, residual learning, and attention modulation. As shown in Fig. 1(a), the proposed model involves a hierarchical stack of convolutional layers along with multi-scale residual modules and spatial attention modules. In terms of functionality, this model can facilitate the process of encoding both coarse context development and fine-grained details of cells; at the same time, features extracted from images will be adaptively re-weighted and projected via probabilistic approach. The depth-wise separable convolution makes it possible to avoid over-parametrization while working with small amounts of data. The three-layer structure was chosen so that multi-scale attention operates according to the image interpretation mechanism, where the lower layer refers to

visual attributes, the middle one to patterns of embryonic morphology, and the upper layer includes unique characteristics of each class. Thus, the use of additional layers and blocks was avoided so that there were fewer parameters in the model.

Normalization and adaptive pooling techniques are employed to standardize the intermediate activations and develop some degree of tolerance to scale differences between different embryo stages. Channel and spatial attention techniques control the output features and mitigate sensitivity to lighting differences and textures.

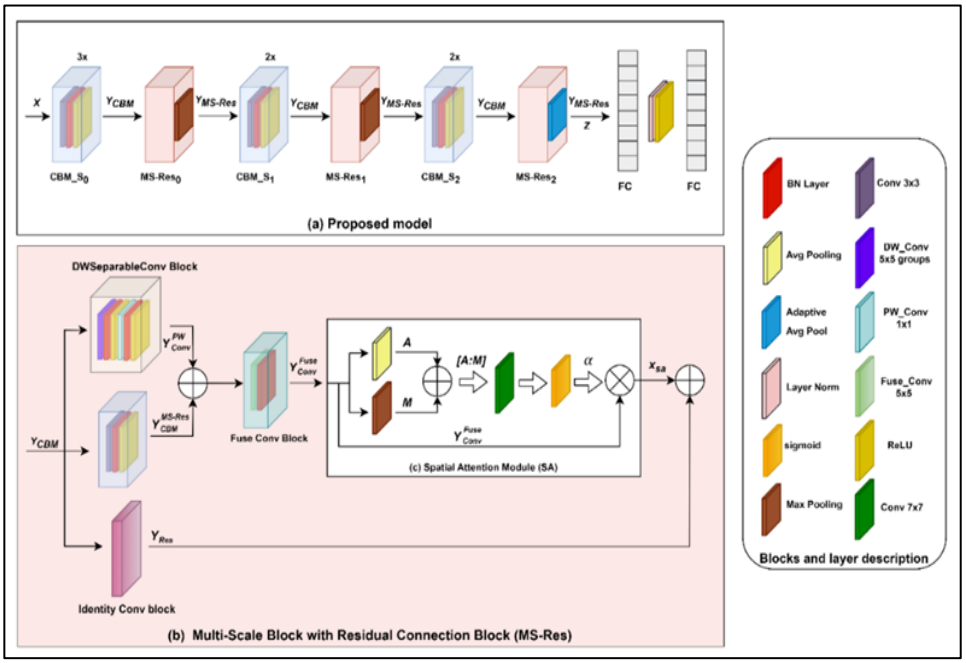

Figure 1. (a) Proposed Model Framework, (b) Multi-Scale Block with Residual Connection (MS-Res), (c)Spatial Attention Module (SA)

3.2 Conv Block Module (CBM)

Figure 1 (a) shows that the Conv Block Module serves as the basic feature-extraction block that will be repeated in the 3 hierarchical levels: 3x, 2x, and 2x. With an input tensor, $x \in R^{B\times C\times H\times W}$, a ConvBlock yields an output, Y_{CBM}, with a preservation of spatial resolution but changing the channel dimension, C, to C`. This is achieved using a bias-free 3 x 3 convolution followed by BatchNorm2d and ReLU activation. The formula for the Conv Block Module is given as;

$$Y_{CBM} = ReLU\left(BN\left(Conv_{3x3}(x)\right)\right) \quad (1)$$

In line with the Figure. 1(a), Stage 0 *includes* includes an initial ConvBlock with a channel progression of 3 → 32 and provides activations of size [B,32,224,224] followed by two ConvBlocks with a channel progression of 32 →64 → 64, whichprovides activations of size [B,64,224,224]. Stage S_1has two ConvBlocks having a progression of 64 to 128, and generates [B,128,111,111] after the initial pooling operation. Stage S_2 also contains two ConvBlocks, with a sequence of 128 →256 →256 and returns [B,256,55,55] after the subsequent pooling. This hierarchical design gradually transforms crude embryonic images into more abstract feature representations while maintaining intra-stage spatial localization; empirical evaluation on the target data is required to confirm any performance gains. The

3×/2×/2× CBM block layout was also tested again in the Section 3.5 ablation experiment. In that comparison, the same MS-Res modules, spatial attention layer, and residual connection settings were left in place, and this deeper CBM arrangement performed better than the shallower CBM alternatives.

3.3 Multi-Scale Block with Residual Connection (MS-Res)

Figure. 1(b) is the extension of the Multi-Scale Block with Residual Connection (MS-Res), which is added after each hierarchical CBM stage in Figure. 1(a) to combine multi-branch processing, feature fusion, spatial attention (Figure 1(c)), and residual addition. Given an input, Y_{CBM}, The MS-Res block produces an output, Y_{CBM}^{MS-Res},of the same spatial dimensions and channels C to C`. The former branch uses a ConvBlock Eq. (1) in order to generate a feature map $Y_{CBM}^{MS-Res} \in R^{B\times C`\times H\times W}$.

$$Y_{CBM}^{MS-Res} = ReLU(BN(Conv_{3x3}(Y_{CBM}))) \quad (2)$$

The second branch uses a depth-wise separable transform (DWSeparableConv Block), which begins with a depth-wise convolution, denoted asDW_{Conv}, with groups G = C (and groups of the same size) each, one group per input channel, followed by BatchNorm and ReLU:

$$Y_{Conv}^{DW} = ReLU(BN(Conv_{5x5}(Y_{CBM}))) \quad (3)$$

Channels and maps are then mixed to the size of the output, and then a point-wise 1 x 1 convolution, known as PW_Conv, is applied:

$$Y_{Conv}^{PW} = ReLU(BN(Conv_{1x1}(Y_{Conv}^{DW}))) \quad (4)$$

The results of the two branches (depth-wise separable transform and Y_{CBM}^{MS-Res})are concatenated, and the concatenated feature map is fused along the channel dimension, by a 1x1 convolution with BatchNorm (no activation is indicated in the fusion step), resulting in.

$$Y_{Conv}^{Fuse} = BN\left(Conv_{1x1}^{Fuse}([Y_{CBM}^{MS-Res}; Y_{Conv}^{PW}])\right) \quad (5)$$

where [.;.] denotes channel concatenation. A spatial attention is then performed to generate a gating signal to re-weight the fused features, resulting in X_{sa}(listed in section 2.4). The residual shortcut can be defined as the identity mapping, otherwise as a 1x1 convolution for dimension matching.

$$Y_{Res} = \begin{cases} Y_{CBM}, & c = c` \\ Conv_{1x1}(Y_{CBM}) & c \neq c` \end{cases} \quad (6)$$

The MS-Res block output is then calculated via combination the residual with the output of the spatial attention, and this combination followed by ReLU:

$$Y_{MS-Res} = (ReLU(X_{sa} + Y_{Res})) \quad (7)$$

Architecturally, the parallel $Conv_{1x1} \rightarrow Y_{CBM}^{MS-Res}$ and depthwise-separable $(Conv_{5x5} \rightarrow Conv_{1x1}) \rightarrow Y_{Conv}^{PW}$ paths are intended to expose the representation to multiple local spatial contexts, and the residual mapping Y_{Res} is intended to preserve information flow and may ease optimization. The three-stage MS-Res layout was used to build a step-by-step feature ladder, while still keeping the network small enough for the limited embryo microscopy

image set. In embryo microscopy frames, useful signs are not found at only one scale. Some signs are low-level, such as embryo borders, edges, and small texture changes. Other signs are mid-level, such as cellular appearance and visible morphology. Higher-level signs include wider embryo-region patterns. For this reason, the MS-Res blocks were placed after the three main convolutional stages, so that multi-scale feature fusion and the spatial attention unit could work on shallow maps, middle feature maps, and deeper feature maps, instead of being used only near the last classification layer. The first MS-Res block mainly serves to clean and strengthen early local texture and edge information; the second MS-Res block helps combine mid-level morphology after the first spatial downsampling step; and the third MS-Res block processes the smaller, higher-level feature maps before global pooling and the final classifier. This staged setup also controls extra model growth because the channel depth is increased gradually, and the wider-context branch uses depthwise-separable convolution rather than a heavier standard convolution. In this way, the three-stage structure was selected as a workable trade-off between richer feature representation, easier visual interpretation, and lower computational cost. Still, this study does not state that this exact setting is the best possible design in all cases; future work should run a complete ablation study with one-stage, two-stage, three-stage, and deeper MS-Res versions, so that the separate effect of the selected depth can be measured more directly.

3.4 Spatial Attention Module (SA)

Figure. 1(c) shows the Spatial Attention (SA) module applied in MS-Res (Figure 1(b)) to create a spatial gate with channel-collapsed descriptors. Given$Y_{\mathrm{Conv}}^{\mathrm{Fuse}}$, channel mean, and maximum projections are calculated as:

$$A = AvgPool_{ch}(Y_{Conv}^{Fuse}), M = MaxPool\_ch\ (Y_{Conv}^{Fuse}) \tag{8}$$

The two descriptors are combined and then run through a bias-free 7 x 7 convolution with a sigmoid gating function σ (.) to obtain a spatial attention mask:

$$\alpha = \sigma\big(Conv_{7x7}^{sa}([A; M])\big) \tag{9}$$

Finally, the mask is applied to re-weight the fused features:

$$x_{sa} = Y_{Conv}^{Fuse} \odot \alpha, \tag{10}$$

The Spatial Attention (SA) module reduces the influence of background and low-contribution areas by weighting more spatially informative regions in intermediate feature maps. This enhances the representational strength of the model by emphasizing unique local structures and is part of a more selective utilization of cues in multi-scale fusion. This reweighting aims to produce a qualitatively consistent focusing behaviour.

3.5 Classification Head

Figure 1(a) represents the classification head that converts the final backbone tensor to logits through global pooling and a two-layer MLP. With the input of the third MS-Res block feature map $Y_{\mathrm{MS-Res}}$, the head initially uses AdaptiveAvgPool2d followed by a flatten layer to produce a pooled embedding, then applies dropout with a probability of 0.5. It then produces feature map h which is followed by an affine mapping normalized by LayerNorm, and then ReLU to generate a hidden representation h, which is passed to the second FC layer, giving

$$logits = FC_2\,(ReLU(LN(FC_1\,(h))) \tag{11}$$

The classification head reduces the high-level feature maps from the backbone to a fixed-size feature vector per sample using global average pooling, thus summarizing spatial information in a compact form. It then regularizes with dropout and produces class logits with a two-layer MLP to generate final decision scores. The purpose of this structure is to directly project the learned representation of the backbone into class differentiation and more stable generalization behaviour.

3.6 Evaluation Metric

Model performance was evaluated on the held-out test set using confusion-matrix-based metrics and threshold-independent measures. For the binary embryo viability classification task, the confusion-matrix counts, true positives ((TP)), true negatives ((TN)), false positives ((FP)), and false negatives ((FN)), were used to calculate accuracy, precision, recall, and F1-score.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

$$F1\text{-}score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{15}$$

ROC-AUC and PR-AUC were also reported, so the model's class-separation ability could be checked across different decision thresholds. PR-AUC was included because it provides more useful evidence in an imbalanced binary classification setting. All test-set results were produced from the checkpoint selected by validation loss, and the test set was not used for any threshold adjustment.

4. Experimental Results

4.1 Dataset and Pre-processing Steps

4.1.1 Dataset Acquisition and Scope Definition

The dataset used in this study was the public Hung Vuong Hospital Embryo Classification dataset on Kaggle [23]. The entire dataset consists of 1,020 files of microscopy images of embryos, 840 of which are labelled and 180 of which are not. In the case of the supervised two-class classification experiment, the files of labelled images were chosen. Good-quality embryos were designated as Class 1, and poor-quality embryos were designated as Class 0. The prefixes of the files are D3 and D5, which represent the developmental stages of the Day-3 and Day-5 embryos, respectively, and were kept as metadata in contextual image interpretation but were never used as additional prediction labels.

4.1.2 Initial Quality Control and Sample Filtering

The 840 labelled embryo images were screened prior to the training, validation, and test subsets in terms of simple image usability and label appropriateness. An image file was eliminated when it had one or more problems, such as a damaged file format, an inaccessible image file, missing label information, ambiguous label information, excessive cropping that lost clinically useful areas of the embryo, severe blur, severe illumination artifacts, or ambiguous visual content. Following this quality-control screening, 801 labelled images remained to be analyzed. This also implies that there were 39 labeled images that were not counted during the checking process. Image removal was conducted prior to dataset partitioning to avoid introducing a selection bias associated with the training set, validation set, or test set.

4.1.3 Leakage-Safe Stratified Partitioning

The training set was highly class-imbalanced, with 546 Class-0 images, and 94 Class-1 images, thus augmentation was only done to Class 1 within the training split. This was done to minimize imbalance when optimizing the model, and the validation and test distributions were left in their original, cleaned state to perform a less biased analysis. The minority class was increased until the number of its training samples was equal to that of the majority-class. More precisely, 452 additional augmented Class-1 samples were created, resulting in a balanced training set of 546 Class-0 images and 546 Class-1 images or 1,092 training samples in total. They did not augment, copy, duplicate or synthetically balance the validation and test sets since these subsets were employed to model the cleaned data distribution to measure the final performance.

4.1.4 Training-Only Imbalance Mitigation via Selective Augmentation

To maintain the balance of evaluation without biasing, only in the training set and only in Class 1 (n = 94), augmentation was performed with the creation of 452 additional samples through gamma correction (0.8, 1.5), light Gaussian blur (5,5), and horizontal/vertical flipping. This produced a balanced training set (Class 0 = 546, Class 1 = 546) with 1,092 training images; the validation/test set was not augmented at all and did not have augmented derivatives. An example of data augmentation is provided below.

The augmentation pipeline employed moderate changes in photometry and geometric alterations. The gamma correction values were varied between 0.8 and 1.5 to introduce controlled brightness and contrast variations that may occur when the embryo images are taken under various microscope-lighting and acquisition settings. The original intensity pattern is mostly preserved by values close to 1.0, while moderate darkening or brightening is produced by values below or above 1.0, with no intent of altering embryo morphology. To this end, this range was considered a safe perturbation range to enhance training robustness, rather than suggesting that it is the most optimal photometric augmentation setting. Horizontal flips, vertical flips, and slight Gaussian blurring with a 5 x 5 kernel were also implemented, thus leaving the minority class with larger appearance variation.

4.1.5 Geometric Standardization and Intensity Normalization

The image files were downsized to 224 x 224 pixels to ensure that the size of the images matched the input size of the proposed model and the pretrained baseline networks. The pixel intensity values were then normalized using one common pre-processing rule throughout the training set, validation set, and test set. This is a typical normalization operation to stabilize the numerical training process and to diminish intensity differences due to variations in the image-acquisition process.

4.2 Experimental Setup

All experiments were run in a Python setup, using the PyTorch deep learning framework, CUDA acceleration, and an A100 GPU provided through Google Colab. The proposed model and the pretrained baseline networks were trained and tested with the same preprocessing steps, the same 224 × 224 image input size, the same data split, the same training augmentation strategy, and the same training procedure, so that the comparison would be carried out under a shared experimental setting. The hyperparameter values used during training are listed in Table 1.

Table 1. Training Hyperparameters Used in the Experimental Setup

Component	Setting / Description
Number of epochs	50
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Learning-rate scheduler	Factor = 0.5, patience = 2
Batch size	16
$p_{t,i}$	Predicted probability assigned to the ground-truth class for sample (i)
γ	$\gamma = 2$ Focal loss focusing parameter
W_{y_i}	Weight assigned to the ground-truth class (y_i)
Class weights	$(W_0 = 0.822884), (W_1 = 2.5485437)$

To handle class imbalance during model optimization, focal loss was used together with class-weighted cross-entropy. For a mini-batch with (B) samples and (K) classes, let (x_i) represent the input image at index (i), let ($y_i \in 0,1, \ldots, K-1$) represent the true class label for that image, and let ($z_i \in R^K$) represent the model output logits. The predicted probability for class (c) is obtained with the softmax operation:

$$p_{i,c} = \frac{exp(z_{i,c})}{\sum_{j=1}^{K-1} exp(z_{i,j})} \tag{16}$$

The probability assigned to the true class of sample (i) is written as:

$$p_{t,i} = p_{i,y_i} \tag{17}$$

For sample (i), the class-weighted cross-entropy part is defined as:

$$CE_i^w = -W_{y_i}\, log\, (p_{t,i}) \tag{18}$$

where W_{y_i} is the class weight given to the ground-truth class y_i. In this work, the class weights were fixed at (W_0=0.822884) for Class 0 and (W_1=2.5485437) for Class 1. These weight values were calculated only from the training split. Validation data and test data were

not used in this calculation, so that information from the evaluation subsets would not leak into the training loss. The focal loss term for sample (i) is then written as:

$$FL_i = \left(1 - p_{t,i}\right)^{\gamma} CE_i^w \tag{19}$$

where (γ) is the focusing parameter used in focal loss. This factor lowers the effect of samples that the model already classifies well, and it puts relatively more training pressure on difficult samples. The final loss for the mini-batch is calculated as:

$$\mathcal{L}\mathrm{FL} = \frac{1}{\mathrm{B}} \sum_{\mathrm{i}=1}^{\mathrm{B}} FL_i \tag{20}$$

This loss formulation was selected because the cleaned training split was imbalanced before augmentation, and the minority class was still important for the embryo viability classification task. The class weights and focal modulation were therefore used to reduce the influence of easy majority-class images and to help the model learn from harder samples or samples from the underrepresented class. To make the workflow clearer and easier to reproduce, the training process is described in Algorithms 1 and 2.

Algorithm 1: Proposed Hybrid MS-Res Attention Model for Embryo Viability Classification

Input: Resized and normalized embryo image ($I \in R^{3\times224\times224}$)
Output: Class logits ($\hat{y} \in R^{2}$)

1. Set the input tensor:
$$X \leftarrow I$$
2. Apply the stem convolutional block:
$$X \leftarrow Y_{CBM}(X; C_{in} = 3, C_{out} = 32)$$
3. For each hierarchical feature-extraction stage ($s \in 1,2,3$), apply two sequential convolutional blocks:
$$X \leftarrow Y_{CBM}(X; C_{in} = C_s, C_{out} = C_{s+1})$$
$$X \leftarrow Y_{CBM}(X; C_{in} = C_{s+1}, C_{out} = C_{s+1})$$
4. Apply the multi-scale residual attention block:
$$Y_{MS-Res} \leftarrow MS\text{-}Res(X; C_{in} = C_{s+1}, C_{out} = C_{s+1})$$
6. Apply max-pooling after the first and second hierarchical stages only:
$$X \leftarrow MaxPool_{3\times3,s=2}(X), \quad s \in 1,2$$
Apply the classification head:
$$h \leftarrow Flatten\big(Global\ Averag\ Pooling\ (Y_{MS-Res})\big)$$
$$h \leftarrow FC(h; d_{in} = 256, d_{out} = 128)$$
$$h \leftarrow ReLU\big(LayerNorm(h)\big)$$
$$\hat{y} \leftarrow FC(h; d_{in} = 128, d_{out} = 2)$$
7. Return (\hat{y}) for binary embryo viability classification.

The implemented channel progression is $(32 \rightarrow 64 \rightarrow 128 \rightarrow 256)$, where$(C_s)$denotes the input channel size of stage (s), and (C_{s+1}) denotes the corresponding output channel size after the stage convolutional blocks.

Algorithm 2: Multi-Scale Residual Attention Block

Input: Feature map ($Y_{CBM} \in R^{C_{in} \times H \times W}$)
Output: Refined feature map($Y_{MS\text{-}Res}$)

1. Extract features from the standard convolution branch:
$$Y_{CBM}^{MS\text{-}Res} = ReLU\left(BN\left(Conv_{3\times3}(Y_{CBM})\right)\right)$$
2. Extract features from the depthwise-separable branch:
$$Y_{Conv}^{DW} = ReLU\left(BN\left(DWConv_{5\times5}(Y_{CBM})\right)\right)$$
$$Y_{Conv}^{PW} = ReLU\left(BN\left(PWConv_{1\times1}(Y_{Conv}^{DW})\right)\right)$$
3. Concatenate and fuse the two branch outputs:
$$Y_{Conv}^{Fuse} = BN\left(Conv_{1\times1}([Y_{CBM}^{MS\text{-}Res}; Y_{Conv}^{PW}])\right)$$
4. Generate the spatial attention mask:
$$M_s = \sigma\left(Conv_{7\times7}([AvgPool_c(Y_{Conv}^{Fuse}); MaxPool_c(Y_{Conv}^{Fuse})])\right)$$
5. Apply spatial reweighting:
$$Y_{SA} = Y_{Conv}^{Fuse} \odot M_s$$
6. Define the residual shortcut:
$$R(Y_{Res}) = \begin{cases} Y_{CBM}, & c = c` \\ Conv_{1x1}\,(Y_{CBM}) & c \neq c` \end{cases}$$
7. Compute the final MS-Res output:
$$Y_{MS\text{-}Res} = ReLU\left(Y_{SA} + R(Y_{Res})\right)$$

Return ($Y_{MS\text{-}Res}$).

The computational cost and inference behaviour of the proposed model were measured in the same PyTorch environment, with CUDA acceleration on an A100 GPU in Google Colab. As shown in Table 2, the proposed model has 2,244,616 parameters in total, and all of these parameters are trainable, giving about 2.245 million trainable parameters. For one 224 × 224 input image, the estimated computation cost is 15.63 GFLOPs-equivalent. The model produced an average inference time of 10.185 ms/image, with a standard deviation of 0.082 ms. The minimum measured latency was 9.967 ms, and the maximum measured latency was 10.618 ms. The related throughput was 98.19 images/second. These measurements show that the proposed architecture stays relatively compact while still providing a usable inference speed in the tested GPU-based setting.

Table 2. Computational Complexity of the Proposed Model

Model	Total parameters	Trainable parameters	Parameters (M)	Computational cost	Average latency	Throughput
Proposed model	2,244,616	2,244,616	2.245 M	15.63 GFLOPs-equivalent	10.185 ms/image	98.19 images/s

The latency values were recorded with CUDA acceleration on an A100 GPU in the Google Colab runtime. During repeated inference runs, the timing results varied only slightly, giving a standard deviation of 0.082 ms, a minimum latency of 9.967 ms, and a maximum latency of 10.618 ms. Since inference latency depends on the GPU device, batch size, numerical precision mode, memory bandwidth, and runtime configuration, these reported timings should be read as measurements from this specific experimental setup, not as fixed model properties that remain the same across all hardware environments.

4.3 Ablation Study of the Hierarchical CBM Structure

To justify using the hierarchical 3×/2×/2× CBM arrangement, an ablation experiment was carried out with five architecture variants. In this experiment, the MS-Res blocks were held constant. The spatial attention module was also left unchanged, and the residual connection paths were kept fixed in every variant. With these components controlled, the comparison focused mainly on the effect of changing the CBM repetition count across the three hierarchical stages.

As reported in Table 3, the shallowest version, A1 1×/1×/1×, gave the weakest overall results. It reached 0.75 accuracy, 0.63 F1-score, 0.74 ROC-AUC, and 0.60 PR-AUC. When more CBM blocks were added, the performance increased step by step. A2 2×/1×/1× raised the F1-score to 0.66 and the accuracy to 0.78. Then A3 2×/2×/1× gave another increase, with the F1-score moving to 0.71 and the accuracy moving to 0.81. A4 3×/2×/1× added more improvement again, reaching 0.75 F1-score, 0.85 ROC-AUC, 0.75 PR-AUC, and 0.84 accuracy.

The proposed 3×/2×/2× arrangement gave the strongest results for every reported metric. It achieved 0.91 precision, 0.94 recall, 0.92 F1-score, 0.92 ROC-AUC, 0.90 PR-AUC, and 0.96 accuracy. It also had the smallest error numbers, with FN=1 and FP=2, when compared with the shallower architecture variants. These findings suggest that adding the last CBM repeat in the third stage helps the network refine stronger high-level features before the global pooling layer and the final classification step. For this reason, the selected hierarchical layout is supported by the ablation study results, and it offers a practical middle point between progressive representation learning and classification performance.

Table 3. Ablation Study of the Hierarchical CBM Structure

Variant	Hierarchical CBM structure	P	R	F1	AUC	PR-AUC	Acc	TP	FN	FP	TN
A1	(1 ×/1 ×/1 ×)	0.62	0.68	0.63	0.74	0.60	0.75	7	5	15	54
A2	(2 ×/1 ×/1 ×)	0.65	0.73	0.66	0.78	0.65	0.78	8	4	14	55
A3	(2 ×/2 ×/1 ×)	0.69	0.79	0.71	0.82	0.70	0.81	9	3	12	57
A4	(3 ×/2 ×/1 ×)	0.72	0.84	0.75	0.85	0.75	0.84	10	2	11	58
Proposed model	(3 ×/2 ×/2 ×)	0.91	0.94	0.92	0.92	0.90	0.96	11	1	2	67

4.4 Extraction Feature Process

Given an RGB embryo image batch($x \in R^{B\times C\times H\times W}$). The forward pipeline builds a hierarchical representation upon low-level appearance cues to class-discriminative features in an entirely stages-based fashion (Figure 1(a)).The input first passes through; Stage 0; this stage represents CBM_{s_0} composes of initial stem (3x3) Conv. produces low-level features ($Y^{S_0}_{CBM-stem} \in R^{B\times 32\times H\times W}$), intended to capture basic edges, contrast transitions, and fine texture while preserving spatial localization for later analysis, and followed by two Conv. sequential layers to obtain feature map ($Y^{S_0}_{CBM} \in R^{B\times 64\times H\times W}$) represented the final output of the CBM_{s_0}. The refines with a MS-Res (64→64) to ($Y^{S_0}_{MS-Res} \in R^{B\times 64\times H\times W}$); this step is supposed to stabilize initial patterns in imaging variability and to induce an inductive bias in the locally consistent morphology. A max pooling operation (kernel 3, stride 2) reduces resolution, yieldng ($Y^{S_0}_{MS-Res \to MaxPooling} \in R^{B\times 64\times 111x111}$improving robustness by compressing spatial redundancy. Stage 1; it includes $CBM_{(s_1)}$ with two (3×3) convolutional

sequential layers to produce ($Y_{CBM}^{S_1} \in R^{B\times128\times111x111}$) and then applies a MS-Res (128→128) to obtain $Y_{MS-Res}^{S_1} \in R^{B\times128\times111x111}$), targeting mid-level morphological patterns and spatially localized structures relevant to viability cues. A second pooling step (kernel 3, stride 2) yields 55x55, ($Y_{MS-Res \rightarrow MaxPooling}^{S_1} \in R^{B\times128\times55x55}$). Stage 2; In these stages two (3×3) convolutional blocks are applied to reach ($Y_{CBM}^{S_2} \in R^{B\times256\times55x55}$) followed by a MS-Res (256→256) producing ($Y_{MS-Res}^{S_2} \in R^{B\times256\times55x55}$), intended to encode higher-level semantics in a compact but spatially structured form. The aggregation of global average pooling at$\left(Y_{MS-Res}^{S_2}\right)$ is regularized by dropout (p=0.5) and the embedding is aggregated at (256→128) followed by a two-layer classifier that maps (128→num classes) to generate logits. To be interpretable, Grad-CAM heatmaps are produced based on feature maps of selected convolutional or fusion features to give explanations about the selected feature's spatial localization in relation to the internal evidence on which the model is based.

4.5 Main Results

Figure 2 shows the training and validation curves for the proposed model across 50 epochs. During training, both the training loss and the validation loss decreased and then settled into a low, flat region in the later epochs. The training accuracy and validation accuracy increased gradually, and after the middle part of training, the two curves remained close to each other. These final curves indicate that the proposed model converged in a stable manner under the selected data split, augmentation procedure, and training protocol.

Figure 3 shows the training and validation loss curves for the pretrained baseline networks. EfficientNetB0, DenseNet121, and ResNet50 all exhibited a rapid drop in training loss, but their validation losses remained higher and changed unevenly across epochs. This pattern created a clear train–validation loss gap for the pretrained baseline models.

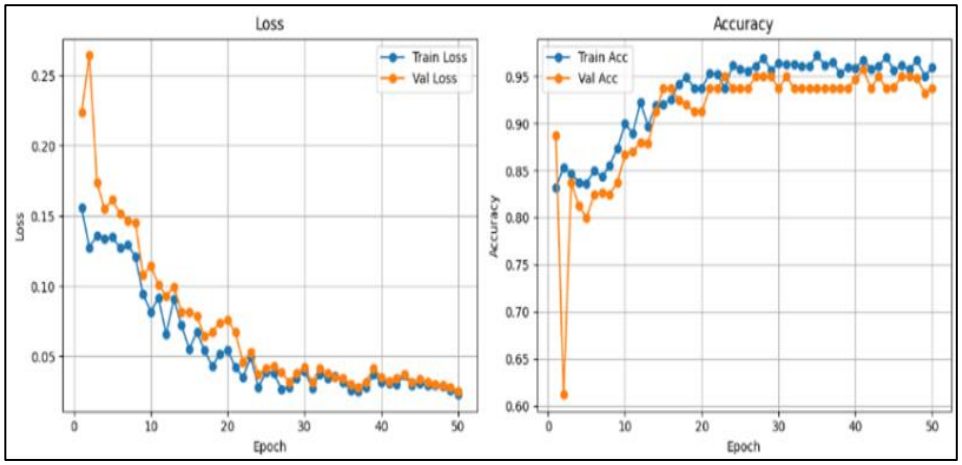

Figure 2. Evaluation Training Curves of the Proposed Model

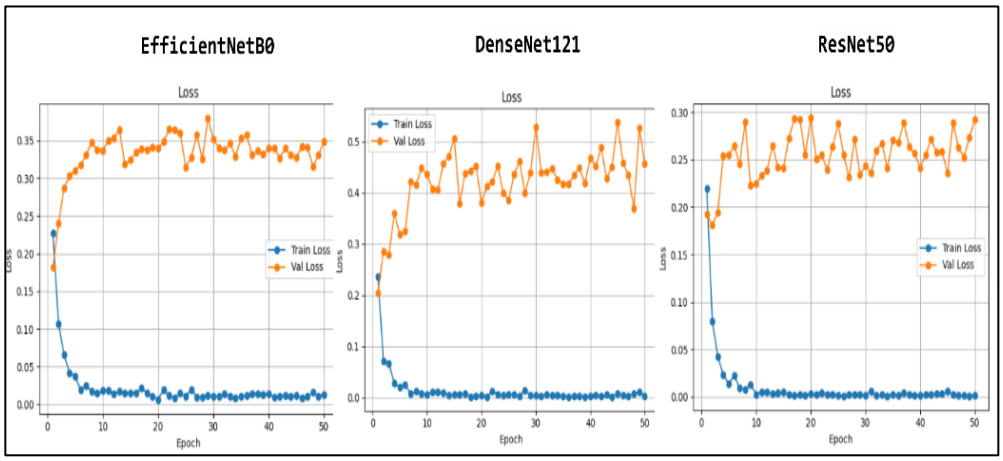

Figure 3. Training and Validation Loss Curves of Pretrained Baselines (Evidence of Overfitting)

Table 4 presents the test-set results for the proposed model and the pretrained baseline architectures. The proposed model recorded the best values for the listed measures, with 0.91 precision, 0.94 recall, 0.92 F1-score, 0.92 ROC-AUC, 0.90 PR-AUC, and 0.96 accuracy. EfficientNetB0 achieved 0.63 precision, 0.72 recall, 0.65 F1-score, 0.79 ROC-AUC, 0.63 PR-AUC, and 0.76 accuracy. DenseNet121 reported 0.68 precision, 0.78 recall, 0.71 F1-score, 0.84 ROC-AUC, 0.57 PR-AUC, and 0.81 accuracy. ResNet50 reached 0.68 precision, 0.72 recall, 0.69 F1-score, 0.78 ROC-AUC, 0.51 PR-AUC, and 0.82 accuracy. The confusion-matrix counts in Table 1 indicate that the proposed model produced TP=11, FN=1, FP=2, and TN=67. The pretrained networks exhibited larger false-positive counts, with FP=15 for EfficientNetB0, FP=12 for DenseNet121, and FP=9 for ResNet50. The proposed model also had the smallest false-negative count among all compared models.

Table 4. Comparative Performance Evaluation of the Proposed Model Against Pretrained Architectures

Method	P	R	F1	AUC	PR	Acc	TP	FN	FP	TN
EfficientNetB0	0.63	0.72	0.65	0.79	0.63	0.76	8	4	15	54
DenseNet 121	0.68	0.78	0.71	0.84	0.57	0.81	9	3	12	57
ResNet50	0.68	0.72	0.69	0.78	0.51	0.82	7	5	9	60
Proposed model	0.91	0.94	0.92	0.92	0.90	0.96	11	1	2	67

Figure 4 shows the ROC curves and precision–recall curves for the proposed model and the pretrained baseline models. The proposed model obtained the highest ROC-AUC score of 0.92, and also the highest PR-AUC score of 0.90. Among the pretrained baseline models, DenseNet121 had the strongest ROC-AUC value at 0.84, while EfficientNetB0 had the strongest baseline PR-AUC value at 0.63.

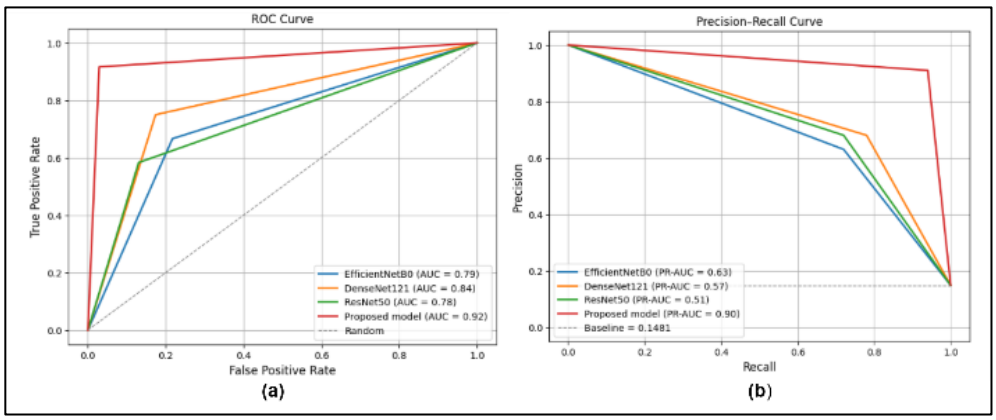

Figure 4. (a) Roc Curve, (b) Precision–Recall

Figure 5 displays Grad-CAM visualizations taken from different layers of the proposed model for selected Day-3 and Day-5 embryo images. The early stem-convolution output in Figure 5(b) produced broader and more spread-out activation areas, while the later feature maps in Figure 5(c)–(h) became more localized step by step. The outputs linked to the MS-Res blocks and the spatial attention component showed stronger focus on embryo regions than the earlier convolution outputs. The randomized-weight heatmaps in Figure 5(i) looked weaker and had less spatial coherence than the heatmaps from the trained model.

The Grad-CAM sanity check used parameter randomization on the final MS-Res block and the classification head. Across n=50 test images, the Pearson correlation between the trained-model heatmaps and the randomized-weight heatmaps was low, with a mean ± SD of -0.096±0.187 and a range of −0.364 to 0.497. The randomized-weight heatmaps are presented in Figure 5(i).

Figure 6 shows Grad-CAM heatmaps for the pretrained baseline models, namely EfficientNetB0, DenseNet121, and ResNet50. The baseline heatmaps had less stable localization behaviour than the deeper MS-Res-related visualizations produced by the proposed model. These visual results were used as qualitative evidence for interpretability, not as numerical clinical validation.

Table 5 provides a contextual comparison between the proposed model and selected studies that used the same public embryo dataset. The proposed model achieved 0.9630 accuracy, 0.9157 precision, 0.9438 recall, and 0.9291 F1-score. Saraniya et al. [7] reported 94.3 accuracy and 0.874 F1-score. Kalatehjari et al. [24] reported 93.09 ± 0.0015 accuracy, 95.26 ± 0.0022 precision, 97.25 ± 0.0022 recall, and 97.04 ± 0.0028 F1-score. Panaite et al. [20] reported F1-scores of 0.75 on the public test set and 0.61 on the private test set. Since the studies used different sample filtering steps, class-balancing methods, and data-splitting protocols, Table 5 is used as a contextual comparison, not as a direct one-to-one benchmark.

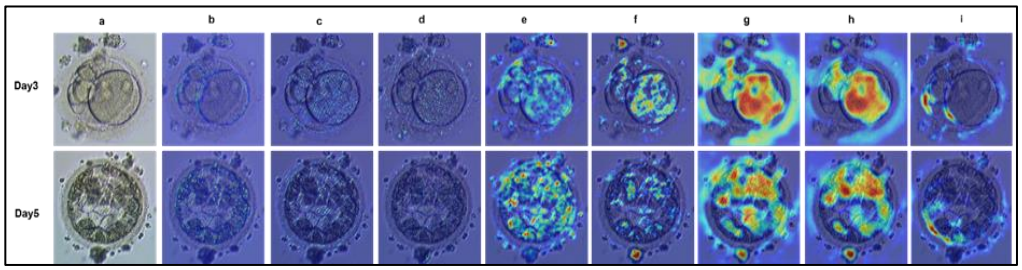

Figure 5. (a) Original Image, (b) $Y_{CBM-stem}^{S_0}$, (c) $Y_{CBM_{(CONV2)}}^{S_0}$, (d) $Y_{MS-Res}^{S_0}$, (e) $Y_{CBM}^{S_1}$, (f) $Y_{MS-Res}^{S_1}$, (g) $Y_{CBM}^{S_2}$, (h) $Y_{MS-Res}^{S_2}$, (i) Randomized Weights

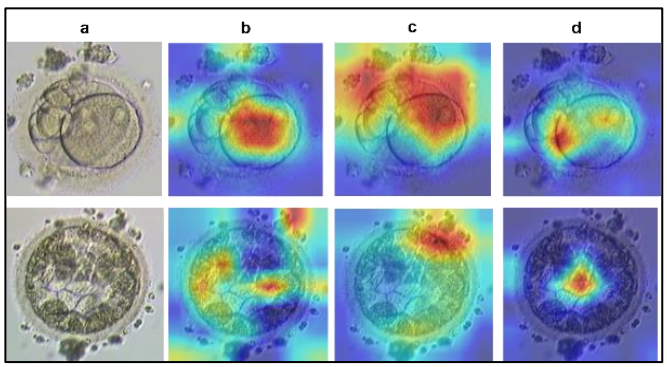

Figure 6. Grad-CAM Heatmaps of Baseline Models, (a) EfficientNet B0, (b) DenseNet121, (c) ResNet50

Table 5. Comparative Analysis of the Proposed Model and State-of-the-Art Approaches

Model	Accuracy	Precision	Recall	F1- Score
Proposed Model	0.9630	0.9157	0.9438	0.9291
Saraniya et. al. [7]	94.3	0.849	0.900	0.874
Kalatehjari et. al. [24].	93.09 ± 0.0015	95.26 ± 0.0022	97.25 ± 0.0022	97.04 ± 0.0028
Panaite et. al [20]	N. A	N. A	N. A	0.75 public, 0.61private test

5. Discussion

The results indicate that, under the same pre-processing workflow, minority-class augmentation plan, training configuration, and test-set procedure, the proposed model achieved the strongest performance among all evaluated architectures. The proposed model reached 0.96 accuracy, 0.91 precision, 0.94 recall, 0.92 F1-score, 0.92 ROC-AUC, and 0.90 PR-AUC. These metric values point to a steadier classification behaviour than the pretrained CNN baseline models. This is more noticeable for the positive class, because recall and PR-AUC are especially important when the embryo image dataset is imbalanced. The confusion-matrix results support the same interpretation. The proposed model produced fewer false-negative cases and fewer false-positive cases than EfficientNetB0, DenseNet121, and ResNet50.

The confusion matrix fits the size and class makeup of the test set used in this work. The independent test subset included 81 embryo images, so the TP, FN, FP, and TN values are small when reported as absolute counts. In this test subset, the proposed model identified 11 positive embryo images correctly as positive cases, which were counted as TP. It also identified 67 negative embryo images correctly as negative cases, which were counted as TN. The model made only 1 false-negative prediction and 2 false-positive predictions. This pattern of errors shows that the high accuracy was not produced only by sending most images into the majority class. Instead, the model maintained correct classification behaviour for both embryo-quality groups.

The weaker results from the pretrained baseline networks seem to be connected to the small dataset and the training setup used in this experiment. EfficientNetB0, DenseNet121, and ResNet50 all reduced their training loss quickly, but the validation loss stayed higher and changed unevenly from epoch to epoch. This behaviour shows a train–validation gap. It also suggests that the pretrained CNN models learned the training images more tightly than they generalized to the validation images and test images. A likely reason is the limited size of the cleaned labelled dataset, which included 801 embryo images. The number of images may not be enough for stable fine-tuning of large pretrained CNN backbones. Also, these networks were first trained on natural-image datasets, while embryo microscopy images have a different visual structure. The embryo images contain embryo morphology, faint texture areas, lighting variation, and small class-related image regions. Because of this gap between natural-image features and embryo-image features, the pretrained filters may be less useful when the target dataset is small and class-imbalanced.

The proposed model was built as a small task-specific network for embryo-image classification, rather than as a large general pretrained backbone. Its hierarchical CBM structure, MS-Res blocks, spatial attention module, and residual connection paths allow the network to collect image features at several representation levels. This lets the model work with low-, middle-, and high-level embryo-image patterns without depending on a heavy pretrained CNN backbone. The ablation analysis also supports the selected (3×/2×/2×) CBM hierarchy.

In that controlled experiment, the proposed layout performed better than the shallower variants, while the MS-Res blocks, spatial attention module, and residual connections were kept fixed. This result indicates that the added CBM depth in the last stage helped the model strengthen high-level feature maps before global pooling and the final classification layer.

The Grad-CAM maps provide qualitative support for how the proposed model behaves across image space. The heatmaps from the early layers were broad and not strongly focused, while the heatmaps from the deeper MS-Res blocks and spatial-attention-related parts became more concentrated over embryo regions. This pattern is consistent with the way the network was designed. The first convolutional layers capture basic visual cues, and the deeper layers transform these cues into more compact features that are more useful for class separation. The Grad-CAM maps generated from randomized weights were weaker and less spatially structured than the heatmaps generated from the trained model. The low Pearson correlation between trained-model and randomized-weight heatmaps (-0.096 + 0.187; range: 0.364 to 0.497) also suggests that the visual explanations were influenced by learned model parameters, rather than being produced only by the Grad-CAM method itself.

The Grad-CAM outputs of the pretrained baseline networks were less consistent than the later-layer heatmaps generated by the proposed model. This visual finding aligns with the numerical results, where the pretrained baselines produced lower F1-score and PR-AUC values. Still, these visual findings should be interpreted with caution. Grad-CAM is a gradient-based sensitivity method. It does not demonstrate that the highlighted image regions are causal biological markers. The heatmaps can support model inspection and make the prediction process more transparent, but any clinical interpretation would still require expert embryologist review and validation using external data.

The contextual comparison with earlier studies indicates that the proposed model achieved competitive scores across the reported evaluation measures, including accuracy, precision, recall, and F1-score. The model also produced a more balanced performance profile. It maintained high recall, high F1-score, and strong overall accuracy. This suggests that the proposed architecture did not improve one classification metric at the expense of the others. Instead, it performed consistently across the main evaluation measures. Based on this comparison, the results support the proposed model as an effective explainable approach for embryo viability classification. However, even with these encouraging results, several limitations related to the sample must be noted. First, after the quality-control filtering step, the dataset remained small, with only 801 labeled embryo images used in the analysis. The number of images may not encompass the full variation in embryo morphology, image quality, microscope configuration, and image-acquisition settings. Second, the held-out test set included only 81 images, and only 12 of them were positive images. Due to this small test-set size, the confusion matrix counts are low in absolute numbers and may shift significantly if only a few predictions change. Third, although the ablation results support the selected hierarchical CBM design, the analysis was still carried out on the same limited dataset. More ablation experiments should be conducted on larger and more varied embryo-image datasets so that the stability of the hierarchical CBM contribution can be tested under different data distributions. In future work, the proposed model can be further developed by validating it on larger multi-center embryo datasets and independent external cohorts.

6. Conclusion

The present research demonstrates that a small convolutional network trained to perform the specificembryo-viability task can provide an informative tradeoff between classification and interpretability in an IVF microscopy problem with a small amount of data. The model is a combination of hierarchical CBM feature extraction, multi-scale fusion based on MS-Res, residual learning paths, and spatial attention. Through these elements, the network is able to learn patterns of embryo-images at multiple levels of representation, without the huge parameter space associated with more complex pretrained backbone networks. The chosen 3×/2×/2× CBM hierarchy is also supported by the ablation results, as the refinement of features progressively assisted with the final classification performance. In the test set that was withheld, the proposed model had a balanced performance profile with high accuracy, recall, F1-score, ROC-AUC, and PR-AUC. The confusion-matrix distribution also demonstrated that the model did not improve its performance simply by being biased towards the majority class, since the number of false negatives and the number of false positives were lower than those of the pretrained baseline models. The poorer performance of EfficientNetB0, DenseNet121, and ResNet50 in the same experimental protocol indicates that large pretrained net capacities might require bigger and more diverse embryo data to stabilize fine-tuning. The interpretability analysis further provided insight into the decision behaviour of the model. Grad-CAM visualizations revealed that later MS-Res and spatial-attention layers had more localized activation patterns than early convolutional layers, and parameter-randomization tests revealed that the heatmaps were affected by learned model parameters. These results justify Grad-CAM as a model-inspection tool; however, the identified regions should not be viewed as causal embryological biomarkers until validated by experts. The primary weakness of this experiment is the small sample size following quality-control filtering, particularly the small sample size of positive samples in the held-out test set. Additionally, the results rely on one public dataset, which restricts the assumptions about generalization to imaging devices, clinical centers, and acquisition protocols. As a future direction, the proposed framework can be extended through validation on larger multi-center embryo datasets, independent external cohorts, and clinician-guided assessment of visual explanations.

References

[1] Barhoun, Abbas, Mohammad Ali Balafar, Amin Golzari Oskouei, and Leila Sadeghi. "MDMBG-Net: A Multi-Task Deep Learning Model Addressing Class Imbalance for Blastocyst Grading in IVF." Biomedical Signal Processing and Control 116 (2026): 109455.

[2] Chen, Kuo, Jing Zuo, Wei Han, and Jin-hong Guo. "Intelligent Assisted Reproduction: Innovative Applications of Artificial Intelligence in Embryo Health Assessment." LabMed Discovery 2, no. 2 (2025): 100075.

[3] Kodali, Radha, Venkata Rao Dhulipalla, Venkata Siva Kishor Tatavarty, Madhavi Nadakuditi, Bharadwaj Thiruveedhula, Suryanarayana Gunnam, Durga Prasad Bavirisetti, and Gogulamudi Pradeep Reddy. "Interpretation of Deep Learning Model in Embryo Selection for in Vitro Fertilization (IVF) Treatment." arXiv preprint arXiv:2506.06680 (2025).

[4] Saraniya, M., and J. Anitha Ruth. "EmbryoNet-VGG16 Framework for Deep Learning-Based Embryo Classification with Otsu Segmentation." Discover Artificial Intelligence 5, no. 1 (2025): 194.

[5] Kishida, Takuma, and Haruki Nishizawa. "Usefulness of Embryo Evaluation via Artificial Intelligence–Based Image Analysis." Fujita Medical Journal 12, no. 1 (2026): 29-32.

[6] Borna, Mahdi-Reza, Mohammad Mehdi Sepehri, and Behnam Maleki. "An Artificial Intelligence Algorithm to Select Most Viable Embryos Considering Current Process in IVF Labs." Frontiers in artificial intelligence 7 (2024): 1375474.

[7] Saraniya, M., and J. Anitha Ruth. "Deep Learning-Based Embryo Quality Assessment: A Dual-Branch CNN Model Integrating Morphological and Spatial Features." Intelligence-Based Medicine 12 (2025): 100273.

[8] Boucret, Lisa, Floris Chabrun, Magalie Boguenet, Pascal Reynier, Pierre-Emmanuel Bouet, and Pascale May-Panloup. "Deep-Learning Model for Embryo Selection Using Time-Lapse Imaging of Matched High-Quality Embryos." Scientific reports 15, no. 1 (2025): 28068.

[9] Wu, C-Y., W-C. Wang, and P-K. Yang. "Application of Hyperspectral Imaging Technique and Artificial Intelligence on Quality Prediction of Embryonic Cells." Experimental Mechanics 66, no. 1 (2026): 207-220.

[10] Mienye, Ibomoiye Domor, Theo G. Swart, George Obaido, Matt Jordan, and Philip Ilono. "Deep Convolutional Neural Networks in Medical Image Analysis: A Review." Information 16, no. 3 (2025): 195.

[11] Raghu, Maithra, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. "Transfusion: Understanding Transfer Learning for Medical Imaging." Advances in neural information processing systems 32 (2019).

[12] Kartik, Kunal, Tafeer Ahmed, Shantanu Ghosh, and Rajat Gupta. "Transfer Learning with CNNs in Small ML Datasets: Applying Pre-Trained CNN Models and Fine-Tuning them for Limited Data Scenarios." Transfer 7, no. 5 (2023).

[13] Miled, Wided Souid, and Nozha Chakroun. "Leveraging Attention Mechanisms for Interpretable Human Embryo Image Segmentation." In ICAART (2), pp. 872-879. 2025.

[14] Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." Information fusion 58 (2020): 82-115.

[15] Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. "Sanity Checks for Saliency Maps." Advances in neural information processing systems 31 (2018).

[16] Afnan, Michael Anis Mihdi, Yanhe Liu, Vincent Conitzer, Cynthia Rudin, Abhishek Mishra, Julian Savulescu, and Masoud Afnan. "Interpretable, not Black-Box, Artificial

Intelligence Should be Used for Embryo Selection." Human reproduction open 2021, no. 4 (2021): hoab040.

[17] Abdulaal, Alaa Hussein, Ali H. Abdulwahhab, Aqeel Majeed Breesam, Zahra Hasan Oleiwi, Riyam Ali Yassin, Morteza Valizadeh, and Saja Nafea Mohsin. "MEMF-Net: A Mega-Ensemble of Multi-Feature CNNs for Classification of Breast Histopathological Images." Iraqi Journal for Computer Science and Mathematics 6, no. 3 (2025): 36.

[18] Movahedi, Faezeh, Rema Padman, and James F. Antaki. "Limitations of Receiver Operating Characteristic Curve on Imbalanced Data: Assist Device Mortality Risk Scores." The Journal of thoracic and cardiovascular surgery 165, no. 4 (2023): 1433-1442.

[19] Citarella, Alessia Auriemma, Pietro Battistoni, Chiara Coscarelli, Fabiola De Marco, Luigi Di Biasi, and Mengyuan Wang. "EmbryoVision AI: An Explainable Deep Learning Framework for Enhanced Blastocyst Selection in Assisted Reproductive Technologies." Image and Vision Computing (2025): 105795.

[20] Panaite, Doru-Răzvan, Vlad Barbu, Radu-Andrei Rosu, George Stoica, Şerban-Gabriel Doncean, and Mihaela Elena Breabăn. "Advanced Methods for Dealing with High Data Imbalance for Embryo Fertility Classification." Procedia Computer Science 246 (2024): 82-90.

[21] Zhang, Chensheng, Xintong Shi, Xinyue Yin, Jiayi Sun, Jianhui Zhao, and Yi Zhang. "Cleavage-Stage Embryo Segmentation Using SAM-Based Dual Branch Pipeline: Development and Evaluation with the CleavageEmbryo Dataset." Bioinformatics 41, no. 4 (2025): btae617.

[22] Ishaq, Muhammad, Salman Raza, Hunza Rehar, Shan E. Zain ul Abadeen, Dildar Hussain, Rizwan Ali Naqvi, and Seung-Won Lee. "Assisting the Human Embryo Viability Assessment by Deep Learning for in Vitro Fertilization." Mathematics 11, no. 9 (2023): 2023.

[23] Sarunpm. "Embryo Viability Prediction Dataset." Kaggle, 2023. https://www.kaggle.com/datasets/sarunpm/embryo-prediction.

[24] Kalatehjari, Maryam, Younes Ghasemi, Shaghayegh Mahmoudiandehkordi, Fatemeh Afrazeh, Hossein Abbasi, and Fariba Ghasemi. "Human Embryo Quality Assessment with Deep Learning Models: M. Kalatehjari et al." The Journal of Obstetrics and Gynecology of India 75, no. 3 (2025): 227-232

Appendix:

Greek Symbols	
$\mathcal{L}$	Loss Function
α	Spatial attention mask/gate
γ	Focal loss focusing parameter
Abbreviation	
F1	F1-score
P	Precision
R	Recall
PR	Precision–Recall
PR-AUC	Area Under the Precision–Recall Curve
Acc	Accuracy