Contents lists available at ScienceDirect

Heliyon

Research article

Evaluation of deep convolutional neural networks in classifying human embryo images based on their morphological quality

Prudhvi Thirumalaraju [a,1], Manoj Kumar Kanakasabapathy [a,1], Charles L. Bormann [b,c,1], Raghav Gupta [a], Rohan Pooniwala [a], Hemanth Kandula [a], Irene Souter [b], Irene Dimitriadis [b], Hadi Shafiee [a,c,*]

[a] Division of Engineering in Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[b] Division of Reproductive Endocrinology and Infertility, Department of Obstetrics & Gynecology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[c] Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA

A R T I C L E   I N F O

A B S T R A C T

A critical factor that influences the success of an in-vitro fertilization (IVF) treatment cycle is the quality of the transferred embryo. Embryo morphology assessments, conventionally performed through manual microscopic analysis suffer from disparities in practice, selection criteria, and subjectivity due to the experience of the embryologist. Convolutional neural networks (CNNs) are powerful, promising algorithms with significant potential for accurate classifications across many object categories. Network architectures and hyper-parameters affect the efficiency of CNNs for any given task. Here, we evaluate multi-layered CNNs developed from scratch and popular deep-learning architectures such as Inception v3, ResNET-50, Inception-ResNET-v2, NASNetLarge, ResNeXt-101, ResNeXt-50, and Xception in differentiating between embryos based on their morphological quality at 113 h post insemination (hpi). Xception performed the best in differentiating between the embryos based on their morphological quality.

1. Introduction

Infertility is an underestimated healthcare problem that affects over 48 million couples globally and is a cause of distress, depression, and discrimination (Mascarenhas et al., 2012; Turchi, 2015). Although assisted reproductive technologies (ART) such as in-vitro fertilization (IVF) have alleviated the disease burden to an extent, it has been inefficient with an average success rate of approximately 30% reported in 2015 in the US (CDC, 2015). IVF remains an expensive therapy costing $7000 - $20,000 per ART cycle in the US, most of which is not covered by insurance (Birenbaum-Carmeli, 2004; CDC, 2015; Toner, 2002) with many patients requiring numerous cycles to achieve a successful pregnancy. Multiple factors such as maternal age, medical diagnosis, gamete and embryo quality, and endometrium receptivity determine the success of ART cycles (Barash et al., 2017; Demko et al., 2016; Einarsson et al., 2017; Erenus et al., 1991; Hill et al., 1989; Osman et al., 2015; Paulson et al., 1990). However, non-invasive selection of the highest available

quality from a patient's cohort of embryos (top-quality embryo) for transfer, remains one of the most important factors in achieving successful ART outcomes, yet this critical step remains a significant challenge (Barash et al., 2017; Conaghan et al., 2013; Filho et al., 2010; Machtinger and Racowsky, 2013; Racowsky et al., 2015; Vaegter et al., 2017; Wong et al., 2013).

Embryo transfers are performed at the cleavage or blastocyst stage of development. Embryos are at the cleavage stage 2–3 days after fertilization and may reach the blastocyst stage 5–7 days after fertilization. Traditional methods of embryo selection rely on visual embryo morphological assessment and are highly practice-dependent and subjective. Emulating the skill of highly trained embryologists in efficient embryo assessment in a fully automated system is a major unmet challenge in all of the previous work done in computer-aided embryo assessments due to focus on measuring specific expert-defined parameters such as zona pellucida thickness variation, number of blastomeres, degree of cell symmetry and cytoplasmic fragmentation, etc (Rocha et al.,

---

2017a, 2017b). Computer vision methods for embryo assessment are semi-automated, limited to measuring specific parameters providing metrics that require further analysis by embryologists and strictly controlled imaging systems (Filho et al., 2010). Previous attempts in developing systems using traditional machine-learning approaches require intensive image preprocessing followed by human-directed segmentation of embryo features for classification (Matos et al., 2014; Rocha et al., 2017a, 2017b). Owing to the dependency of these approaches on image processing and segmentation, such methods suffer from the same limitations as computer vision techniques.

Convolutional neural networks (CNN), which work on the principles of representation learning have already received significant attention from the medical community in evaluating, through embryo morphology, the embryo quality and implantation potential, and for other applications for clinical IVF practices such as quality control of systems and embryologists (Dimitriadis et al., 2019a, 2019b; Hariton et al., 2019; Kanakasabapathy et al., 2019a, 2019b, 2019c; Khosravi et al., 2019; Thirumalaraju et al., 2019a, 2019b, 2019c; Tran et al.,

2019). However, all of these studies provide limited information on the neural networks themselves and the effect of hyper-parameters on the task of embryo morphological assessments. Different architectures and hyper parameters achieve varying performances on the same task and are non-universal. To the best of our knowledge, no comparative studies of architectures or methods have been performed using a standardized set of clinical embryo images. Furthermore, domain-shifted embryo data has never been evaluated using any of the previous studies. Therefore, the primary goal of this study was to evaluate popular CNN approaches using a dataset of day 5 embryo (113 h post insemination) images in classifying embryos based on their morphological quality. Day 5 embryos were used for the network evaluation studies given their importance to the field of embryology and since most published studies are focused on this day of embryo development (Dimitriadis et al., 2019b; Hariton et al., 2019; Kanakasabapathy et al., 2019a, 2019b, 2019c; Khosravi et al., 2019; Thirumalaraju et al., 2019b, 2019c).

Embryos with normal fertilization were evaluated based on their morphology at 113 h post insemination (hpi) (Figure 1). At the blastocyst
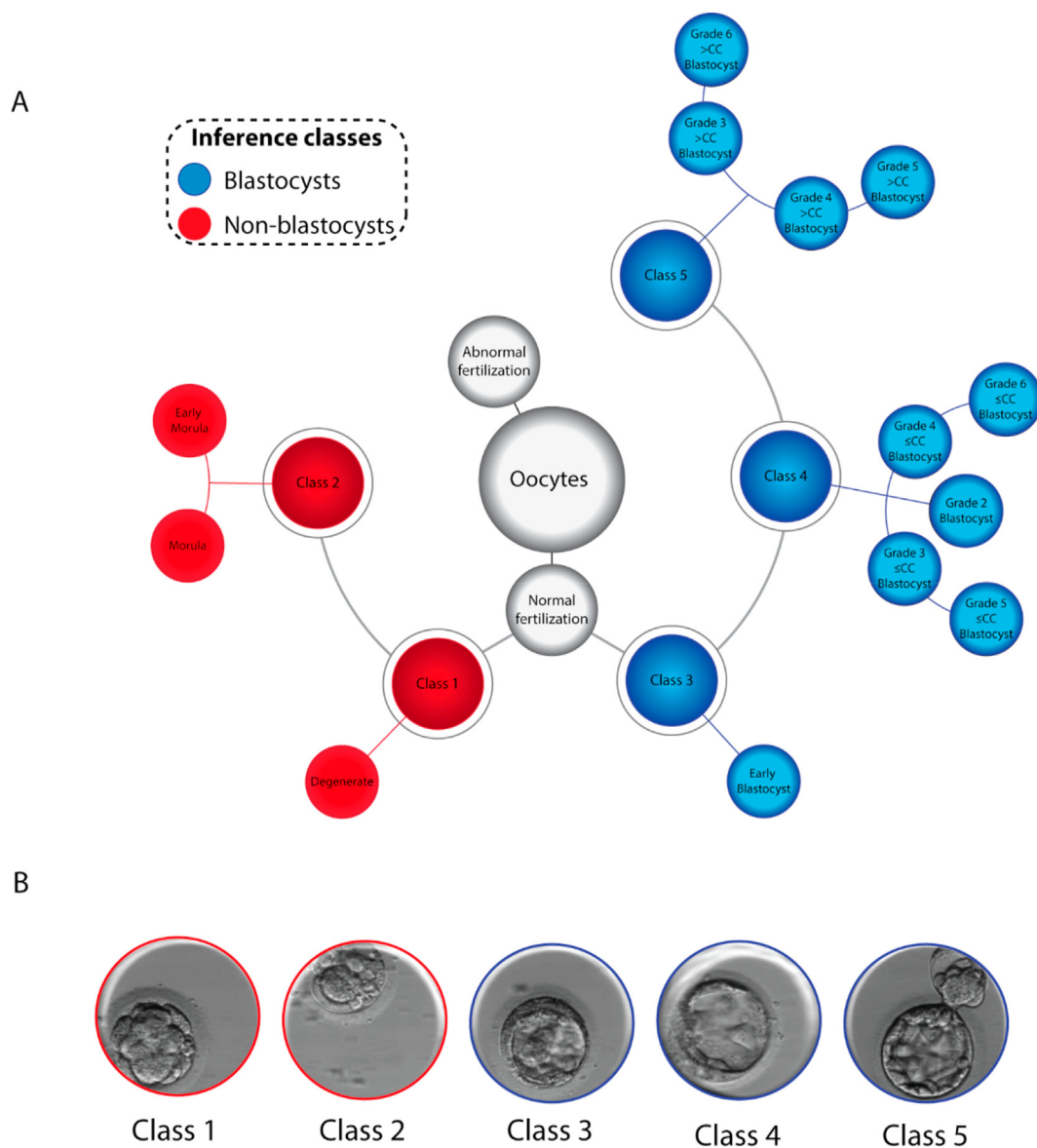
Figure 1. Embryo hierarchy used by the neural network. (A) Following insemination, pronuclear stage embryos are categorized into two classes and based on their 113-hours morphologies are sorted into 2 major classes (blastocysts and non-blastocysts) subdivided into 5 classes. Embryos with abnormal fertilization (non-2PN) were not tracked further and thus were not considered in 70 hpi and 113 hpi assessments. Classes 1 and 2 were composed of non-blastocysts and classes 3, 4, and 5 were composed of blastocysts. Class 5 composed of blastocysts that met the clinical criteria for cryopreservation. (B) Representative images for each class of embryo.

stage (at 113 hpi), embryos are conventionally graded through 83 classes of blastocysts based on the combinations of (i) the degree of blastocoel expansion (grades 1–6), (ii) inner cell mass quality (grades 1–4), and (iii) trophectoderm quality (grades 1–4) along with 3 classes of non-blastocysts. For the CNN classification algorithm, the grading system was simplified to encompass all 86 classes within a 2-level hierarchy of training and inference classes (Figure 1). Thus, the embryos were evaluated at 113 hpi stage using multilayered CNNs (5–43 layers), Inception v3 (Szegedy et al., 2015), ResNET-50 (He et al., 2015), Inception-ResNET-v2 (Szegedy et al., 2016), NASNetLarge (Zoph et al., 2017), ResNeXt-101 (Xie et al., 2016), ResNeXt-50 (Xie et al., 2016), and Xception (Chollet, 2016). The two major categories of non-blastocysts and blastocysts included the training classes 1, 2, and 3, 4, 5, respectively. Using a retrospective dataset comprising of 2,440 embryos, the deep CNN models were trained and tested to primarily classify between two classes (non-blastocysts and blastocysts) using images of embryos captured at 113 hpi. The best performing model was then used to evaluate an independent test set of 742 embryos in differentiating blastocysts based on their morphological quality. In addition, we tested the networks (Inception-v3, ResNET-50, Inception-ResNET-v2, multilayer CNN, NASNetLarge, ResNeXt101, ResNeXt50, and Xception) in evaluating embryo images of shifted domains. Xception performed the best in differentiating between the embryos based on their morphological quality.

2. Materials and methods

2.1. Data collection and preparation

Data was collected at the Massachusetts General Hospital (MGH) fertility center in Boston, Massachusetts. We used 3,469 recorded videos of embryos collected from 543 patients under institutional review board approvals (IRB#2017P001339; IRB#2019P002392). The retrospective image data used for this study were collected as part of routine clinical practice using an Embryoscope time-lapse system (Vitrolife). These instruments use Hoffman modulated contrast optics with $20\times$ objective to image each embryo. Images were acquired at a resolution of 1280×1024 pixels every 10 min at 7 focal planes, to generate videos. Videos were fragmented to extract the frames at a single focal plane and linked to a specific time point (113 hpi) using a custom python script, which made use of the OpenCV and Tesseract libraries. Machine-generated timestamps available on each frame of the video was used to identify the images associated with 113 hpi. All embryos used in the study were annotated using images from the fixed time-points by senior-level embryologists with a minimum of 5 years of human IVF training. Out-of-focus images were included in the datasets and used for both testing and training. Only images of embryos that were completely non-discernable were removed as part of the data cleaning procedure.

2.2. Data organization and hierarchical structuring

Embryo images collected at 113 hpi were separated prior to evaluation. Only embryos with normal fertilization were used for evaluations. The embryo images at 113 hpi time points were categorized between training classes 1 through 5 (Figure 1). The embryo class categorizations were based on the embryos' developmental state achieved by 113 hpi. Class 1 comprised of degenerated and arrested embryos, which did not begin compaction while class 2 comprised of embryos that were at the morula stage at 113 hpi. Classes 1 and 2 together formed the inference class of 'non-blastocysts'. Class 3 comprised of embryos exhibiting features of an early blastocyst such as the presence of a blastocoel cavity and a thick zona pellucida with lack of overall embryo expansion. Class 4 was made up of embryos, which were blastocysts with blastocoel cavities occupying over half of the embryo volume and possessed either poor inner cell mass (ICM) or poor trophectoderm (TE). These embryos were overall considered to fall below 113 hpi cryopreservation quality criteria based on the MGH fertility center guidelines ($>$3CC), where 3 represents

the degree of expansion (range 1–6) and C represents the quality of ICM and TE (range A-D), respectively (Table 1). Class 5 on the other hand comprised of all embryos, which met cryopreservation criteria and included full blastocysts to hatched blastocysts (Table 1). Classes 3, 4, and 5 together formed the inference class of 'blastocysts' that was used in this study.

The 113 hpi evaluation dataset included images of 2,440 embryos categorized across five classes post-cleaning based on their clinical annotations. Our training set for this classification task used 1,188 images (Class 1: 19.36%; Class 2: 17.68%; Class 3: 20.12%; Class 4: 16.92%; Class 5: 25.92%) with a validation dataset of 510 images (Class 1: 19.41%; Class 2: 18.43%; Class 3: 20.59%; Class 4: 15.69%; Class 5: 25.88%) obtained at 113 hpi. The independent non-overlapping test set consisted of 742 images (Class 1: 19.41%; Class 2: 14.42%; Class 3: 17.38%; Class 4: 11.46%; Class 5: 37.33%). All training was performed within the Keras environment, a popular open-source neural network library designed for python. With the availability of unskewed validation sets prior to augmentation, we used a data generator within Keras for batch generation during training that performed random rotations and flips across all classes on the fly.

2.3. Non-embryoscope image dataset

258 embryo images (Non-blastocysts: 54.65%; Blastocyst: 45.35%) collected through the Society for Reproductive Biologists and Technologists (SRBT) for the Embryo ATLAS project, which were imaged using standard inverted bright-field microscopes annotated by 8 director level embryologists from 8 different fertility practices across the United States, were used for the network evaluation. The threshold for classification was optimized for each architecture but no additional training was performed using the SRBT dataset.

2.4. CNN architectures evaluated in this study

Multiple CNN architectures were trained and tested in embryo assessments to identify the best suited network for the task of evaluating embryos. Inception-v3, ResNET-50, Inception-ResNET-v2, NASNetLarge, ResNeXt-101, ResNeXt-50, and Xception architectures along with a 40-layer CNN were tested by training them on 113 hpi embryo images for classification. The Inception-v3 and Inception-ResNET-v2 were trained with the Stochastic Gradient Descent (SGD) optimizer and with learning rates set to 0.0004 and 0.0005, and decay factors of 0.75 and 0.5 for every 10 epochs, respectively. ResNET-50 was trained with Adam optimizers and with learning rates set to 0.001. We trained NASNetLarge, ResNeXt-101, and ResNeXt-50 with SGD optimizer and with learning rates set to 0.00001, 0.01, and 0.01, respectively and with a learning rate scheduler. The input size of the embryo image used was 311×311 pixels in case of NASNetLarge and 210×210 pixels for all other architectures. The 40-layer CNN, similarly, used a learning rate of 0.001 and momentum of 0.5 with an SGD optimizer that had a decay of 0.5 for every 40 epochs. The input size of the embryo image used was 210×210 pixels and each image was convoluted through 64, 128, 256, and 512 feature maps using 3×3 filters with padding and Rectified Linear Unit (ReLU) activation. A dropout layer was also used with dropout probability set at 0.5 along with a flattened second-last layer which was connected to the 5-neuron output layer. A few models of ResNET-50 and de-novo CNNs were highlighted here to elucidate the effect of hyperparameters on such networks for the embryo dataset used (Table 2).

6 models for ResNET-50 trained using 113 hpi embryo images are presented here (Table 2). ResNET-50 architecture was used for models 1, 2, 3, and 4, where only model 1 possessed no dropout while the rest (2, 3, and 4) had a layer set at 0.5 probability. Three extra fully connected layers with 1024, 1024, and 512 neurons between the ResNET-50 bottleneck layer and the final classification layer with an additional 20 trainable layers were used for model 5. Model 6 possessed 1024 neurons in a single fully connected layer between the bottleneck layer and the

Table 1. Blastocyst grading system used by the Massachusetts General Hospital fertility center. The table shows how the graded embryos were categorized into 5 classes. Classes 1 and 2 primarily consisted of non-blastocysts while classes 3, 4, and 5 consisted of blastocysts. Only embryos belonging to class 5 met the freezing criteria employed at the MGH fertility clinic.

Day 5/6 Stage (>113 hpi)	Score	Class	Description
Degenerate or Arrested	D	1	Embryo failed to develop to at least the morula stage
Morula	M-A	2	More than 50% of the embryo has undergone compaction; no ICM or TE cells evident
Morula	M-B	2	Incomplete compaction (less than 50% compaction)
Early Blastocyst	1*	3	Blastocoele less than half the volume of the embryo, little or no expansion in overall size; ZP thick
			(1A = good quality; 1B = moderate quality, 1C = poor quality)
Blastocyst	2	4	Blastocoele more than half the volume of the embryo, some expansion in overall size; ZP beginning to thin
Full Blastocyst	3	4 or 5**	Blastocoele completely filling embryo; ZP not completely thinned
Expanded Blastocyst	4	4 or 5**	Blastocoele completely filling embryo; fully expanded embryo and ZP very thin
Hatching Blastocyst	5	5	Hatching blastocyst, TE starting to herniate through the ZP
Hatched Blastocyst	6	5	Blastocyst completely hatched (i.e. completely out of the ZP)
ICM Grade			**Description**
A			ICM prominent & easily discernible with many cells, and cells compacted and tightly adhered together
B			ICM discernible but with fewer cells, and loosely adherent together
C			Very few cells visible, either compacted or loose, may be difficult to distinguish completely from TE
D			No ICM cells discernible in any focal plane or ICM cells appear degenerate or necrotic
Trophectoderm Grade			**Description**
A			A continuous layer of small uniform eye-shaped cells bordering the blastocoele
B			Fewer, larger cells that may not form a continuous layer
C			Sparse TE cells, may be large
D			All TE cells degenerate

Freezing/Biopsy Criteria: Stage 3 or above with a quality score greater than CC (i.e. do NOT freeze or biopsy embryos with a quality score of CC or any embryo with a D (for ICM or TE).

ZP: Zona Pellucida; **ICM**: Inner cell mass; **TE**: Tropechtoderm.

* No ICM or TE score is given for Stage 1 Early Blastocysts.

** Class 5 consists of embryos which meet the freezing criteria only.

Table 2. Models of ResNET-50, multi-layer CNN, and Xception along with their hyperparameters. The ResNET-50 and multi-layered CNN models were trained using 113 hpi embryo images. Accuracy and loss represent the validation accuracy and loss.

Model	Layers between bottleneck and classification layer	Learning rate	Optimizer	Decay	Loss	Accuracy
ResNET-50						
113 hpi						
1	Base architecture	0.001	Adam	ND	0.8825	0.6011
2	Base architecture + dropout (0.5)	0.01	SGD	ND	0.9665	0.6137
3	Base architecture + dropout (0.5)	0.005	Adam	ND	0.9112	0.6039
4	Base architecture + dropout (0.5)	0.01	SGD	DwS	0.9269	0.6000
5	Base architecture + dropout (0.5), additional layers 3 (1024, 1024, 512) + 20 trainable layers	0.001	Adam	ND	1.6235	0.2058
6	Base architecture + dropout (0.5) + additional layers (1024)	0.001	SGD	DwS	0.9323	0.5847
CNN						
113 hpi						
1	2 layer3-3 + flatten + 1000 + 5	0.0005	SGD	Decay	1.4641	0.3549
2	2 layer5-5 + flatten + 1000 + 5	0.0005	SGD	Decay	1.4183	0.3843
3	2 layer3-5 + flatten + 1000 + 5	0.0005	SGD	Decay	1.4198	0.3803
4	2 layer5-5 + BN-BN + flatten + 1000 + 5	0.0005	Adam	Decay	1.3510	0.4058
5	2 layer5-5 + BN-BN + global average pooling + 32 + 5	0.0005	Adam	Decay	1.5231	0.3333
6	2 layer5-5 + global average pooling + 32 + 5	0.0005	SGD	Decay	1.5956	0.2588
7	2 layer5-5 + BN-BN + flatten + 1000 + 5	0.0005	SGD	ND	1.4178	0.3647
8	2 layer5-5 + BN-BN + flatten + 1000 + 5	0.0005	Adam	Decay	1.8889	0.3200
9	5 layer5-5 + BN + flatten + 64 + 5	0.005	Adam	Decay	1.2799	0.4313
10	5 layer5-5 + BN + same padding + flatten + 512 + 5	0.0005	SGD	Decay	1.2255	0.4294
11	5 layer5-5 + BN + global average pooling + 16 + 5	0.0005	SGD	Decay	1.2730	0.4215
12	40 layer3-2 + global average pooling + flatten + dense	0.0005	SGD	Decay	1.1581	0.4830
13	40 layer3-3 + global average pooling + flatten + dense	0.0005	Adam	ND	1.1689	0.5304
Xception						
113 hpi						
1	Base architecture	0.0005	SGD	DwS	0.8601	0.6373
2	Base architecture + dropout (0.5)	0.001	SGD	DwS	0.8866	0.6333
3	Base architecture	0.0001	SGD	DwS	0.9087	0.6235
4	Base architecture + additional layers 3 (1024,1024,512) + dropout (0.5)	0.0005	SGD	DwS	0.8732	0.6549
5	Base architecture + dropout (0.5)	0.0007	SGD	DwS	0.8704	0.6078
6	Base architecture + additional layers (1024) + dropout (0.5)	0.001	Adam	ND	0.8668	0.6372
7	Base architecture	0.006	SGD	ND	0.8651	0.6588
8	Base architecture	0.0008	SGD	ND	0.8850	0.6196

ND: No decay; DwS: Decay with scheduler; BN: Batch Normalization; SGD: Stochastic descent gradient.

final classification layer, and with a dropout layer set at 0.5. Although only two models are presented here with extra layers, in our overall evaluations adding extra layers to the network did not help the network to learn better. The network was trained by optimizing categorical cross-entropy loss using an SGD optimizer with Nesterov momentum of 0.9 in models 4 and 6 and Adam optimizer in other models. We used learning rates of 0.01 for model 2 and 4, 0.005 for model 3, and 0.001 for all other models. Models with extra layers, dropouts and with different optimizers did not help the network to learn better. Model 1 without any extra layers, decay and dropouts performed better than the models evaluated in our study.

For de novo CNN training, 13 models were trained using a 2 layer architecture for models 1 through 8, a 5 layer architecture for models 9, 10, and 11, and a 40 layer architecture for models 12 and 13 (Table 2). The models were tested with different combination of architectural modifications such as batch normalization, dropouts, global average pooling, padding, and dense layers. A learning rate of 0.0005 was used for all presented models except for model 9 that was trained at 0.005. SGD and Adam optimizers were used with and without decay. Even though increasing the number of layers helped in reducing validation loss and improving validation accuracy, the confusion matrices showed that the results were always skewed towards class 5. Therefore, the multi-layer CNNs were not able to learn to classify well between different embryo classes.

The Xception architecture pre-trained with 1.4 million images of ImageNet was used, which performed with a top-1 accuracy of 79% and top-5 accuracy of 94.5% across 1,000 classes of ImageNet database and fine-tuned the pre-training weights across all layers through transfer learning to fit our dataset and differentiate across the categories of embryos by recognizing relevant features (Chollet, 2016). During the transfer learning process, we discarded the last fully connected layer of the original network and added a new fully connected layer, which classifies the features into the defined five categories. The whole network was trained by optimizing categorical cross entropy loss using an SGD Optimizer with Nesterov Momentum of 0.9 and a learning rate of 0.0005 for 113 hpi. The network was trained over 200 epochs and model weights were saved when the lowest validation loss was achieved (early stoppage). The Xception architecture trained over 200 epochs for tasks of evaluating embryo morphologies between the 5 training classes achieved validation loss of 0.8601 and validation accuracy of 63.73%. The

dimensions of all embryo images used during training were resized to 210×210 pixels using computer vision libraries (OpenCV).

2.5. Classification at the inference level

For classification at the inference level, the algorithm outputs five confidence values mapping the probabilities of the tested embryo associated with each of the five training classes. The embryo is categorized into the training class with the highest confidence. Embryos are assigned to the inference classes based on the highest confidence score, which was obtained through the summation of all confidence values associated with the sub-classes of each inference class.

2.6. Xception: the effect of hyperparameters

Several trained models of the Xception architecture were evaluated for embryo classifications at 113 hpi and to show the performance related to the choice of learning rate, architectural parameters, and other hyperparameters, we presented 8 models here for visualization (Table 2).

We used the vanilla Xception architecture for models 1, 3, 7, and 8, and with dropouts set to 0.5 for models 2 and 5. An extra fully connected layer with 1024 neurons between the Xception bottleneck layer and the final classification layers with dropout was used for model 6. Model 4 possessed three extra fully connected layers with 1024,1024, and 512 neurons between the Xception bottleneck layer and the final classification layer and also possessed dropouts. In models 1, 2, 3, 4, 5, and 8 the network was trained by optimizing categorical cross entropy loss using an SGD optimizer with Nesterov momentum of 0.9 with a learning rate of 0.0005, 0.001, 0.0001, 0.0005, 0.0007 and 0.0008, respectively. Models 6 was trained with Adam optimizer with a learning rate of 0.001, while model 7 was trained with SGD and learning rate set at 0.006 with no decay.

2.7. Data visualization techniques

Keras vis environment was used for data visualization. Saliency maps were used to visualize the pixels involved in the networks during the decision-making process. We mapped the activations of the activation layer prior to bottleneck. We used the test set images in the generation of the saliency maps. t-distributed stochastic neighbor embedding (t-SNE) was performed to observe the distribution of the test dataset and verify if the CNN was able to isolate embryos into clusters based on their features. We used the fully connected layer after global average pooling which has 2,048 dimensional vectors in visualizing the similarities between the embryo images, as understood by the trained network, using the respective datasets. Initially, a principal component analysis (PCA) was performed to reduce 2,048 dimensions to 50 and then t-SNE was performed to reduce the 50 dimensions into 2 dimensions for visualization. We have utilized PCA for the initial dimensionality reduction to 50 from 2,048 dimensions, since it helps in suppressing noise while improving computational speed (van der Maaten and Hinton, 2008).

3. Results

3.1. Selection of the optimal neural network

Depending on the complexity of the problem of interest, CNNs generally require large amounts of annotated image data to accurately learn features and differentiate between the categories of classification. However, high-quality medical datasets are scarce and thus, we have transfer learned our networks over ImageNet weights. Proven high-performance CNN architectures such as Inception-v3, ResNET-50, Inception-ResNET-v2, NASNetLarge, ResNeXt-101, ResNeXt-50, and Xception were retrained using a dataset of 1,188 embryos imaged at 113 hpi and validated using 510 embryo images recorded at 113 hpi. The same dataset was used to train a 40-layer CNN de novo. All networks
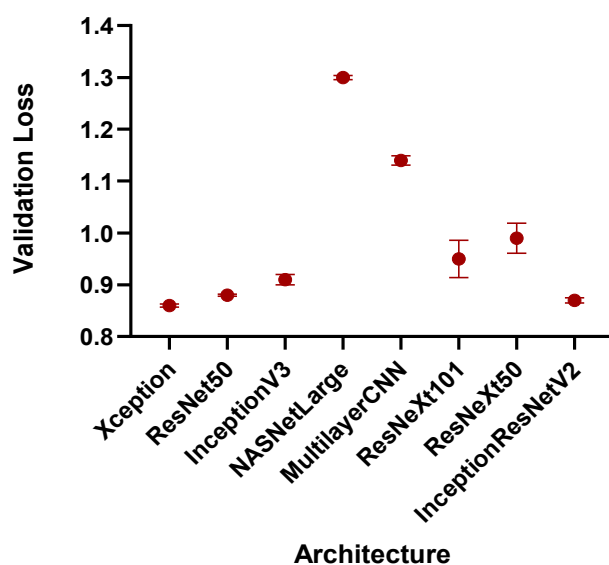
Figure 2. Comparison of different CNN architectures. Xception, ResNET50, Inception v3, NASNetLarge, 40-layer CNN, ResNeXt-101, ResNeXt-50, and Inception-ResNET v2 were used for embryo classification (5 classes) using 113 hpi embryo images. The error bars are standard errors of mean.

Table 3. Validation losses and accuracies of deep-convolutional neural networks. Each architecture was transfer learned with a dataset of blastocysts and non-blastocysts imaged at 113 hpi. The error values reported are standard errors of mean.

Architectures	Validation losses	Validation accuracies (%)
Xception	0.86 ± 0.003	63.53 ± 0.631
ResNET-50	0.88 ± 0.002	59.97 ± 1.08
Inception v3	0.91 ± 0.01	61.57 ± 0.689
NASNetLarge	1.3 ± 0.004	45.75 ± 1.052
Multilayer CNN	1.14 ± 0.009	49.17 ± 1.108
ResNeXt-101	0.95 ± 0.036	58.17 ± 1.2
ResNeXt-50	0.99 ± 0.029	60.07 ± 2.076
Inception ResNET-V2	0.87 ± 0.005	62.09 ± 1.342

Table 4. Performance of different architectures on embryoscope and SRBT datasets. All models tested were optimized through tuning their hyper-parameters. The error values reported are standard errors of mean.

Architecture	Test set accuracies (%)	
	Embryoscope (n = 742)	SRBT (n = 258)
Xception	90.48 ± 0.273	89.53 ± 2.134
ResNET-50	89.71 ± 0.72	84 ± 5.01
Inception v3	89.08 ± 0.812	85.9 ± 1.873
NASNetLarge	78.44 ± 0.233	72.48 ± 3.552
Multilayer CNN	82.2 ± 0.546	69.38 ± 8.067
ResNeXt-101	90.75 ± 0.273	77.52 ± 4.084
ResNeXt-50	89.94 ± 0.574	84.75 ± 0.904
Inception ResNET-V2	90.21 ± 0.518	85.01 ± 1.68

were trained with early stoppage rules prioritizing lowest validation loss to minimize overfitting. After training over 200 epochs, the lowest validation loss achieved by these networks were compared. After fine-tuning the hyperparameters for all evaluated networks with 3 different seeds each, the average 5-class validation loss along with the standard error, of the best models from Xception, ResNET-50, Inception-v3, NASNetLarge, multilayer CNN, ResNeXt-101, ResNeXt-50, and Inception-ResNET-v2 were 0.86 ± 0.003, 0.88 ± 0.002, 0.91 ± 0.01, 1.3 ± 0.004, 1.14 ± 0.009, 0.95 ± 0.036, 0.99 ± 0.029, and 0.87 ± 0.005, respectively and their 5-class validation accuracies, along with the standard errors were 63.53% ± 0.631%, 59.97% ± 1.08%, 61.57% ± 0.689%, 45.75 ± 1.052, 49.17% ± 1.108%, 58.17% ± 1.2%, 60.07% ± 2.076%, and 62.09% ± 1.342%, respectively (Figure 2, Table 3). Xception architecture achieved the lowest mean loss for embryo assessments.

Here, we also report 8 models of Xception to demonstrate the effect of different hyperparameters in learning blastocyst data (Figure 3A, Table 2). Models benefited from lower learning rates when trained with our dataset. In our tests, SGD performed better for models that evaluated these embryos.

We also evaluated simple CNNs designed from scratch, with an increasing number of layers starting from 5 layers to 40 layers (Figure 3B, Table 2). Evaluations with multi-layered CNNs indicated that as the complexity of networks increased, better classification performance can be achieved with our embryo dataset as was observed with most of the tested popular neural networks. Interestingly, however, ResNET-50 did not train well regardless of the hyper parameter optimizations employed (Figure 3C, Table 2).

3.2. Day 5 embryo developmental stage classification

In clinical practice, extending culture to Day 5 and transferring high quality embryos at the blastocyst stage has been effective in improving embryo selection and thus increasing implantation rates. We, therefore, evaluated the models with 3 different seeds each, with an independent test set of 742 embryos imaged at 113 hpi. The average accuracy along with the standard error of the networks in categorizing the embryos into two classes of blastocyst and non-blastocyst of the best models from Xception, ResNET-50, Inception-v3, NASNetLarge, multilayer CNN, ResNeXt-101, ResNeXt-50, and Inception-ResNET-v2 were 90.48% ± 0.273%, 89.08% ± 0.812%, 89.71% ± 0.72%, 78.44% ± 0.233%, 82.2% ± 0.546%, 90.75% ± 0.273%, 89.94% ± 0.574%, and 90.21% ± 0.518%, respectively (Table 4). For the best performing Xception model with an independent test set of 742 embryos imaged at 113 hpi, the accuracy in categorizing the embryos into two classes of blastocyst and non-
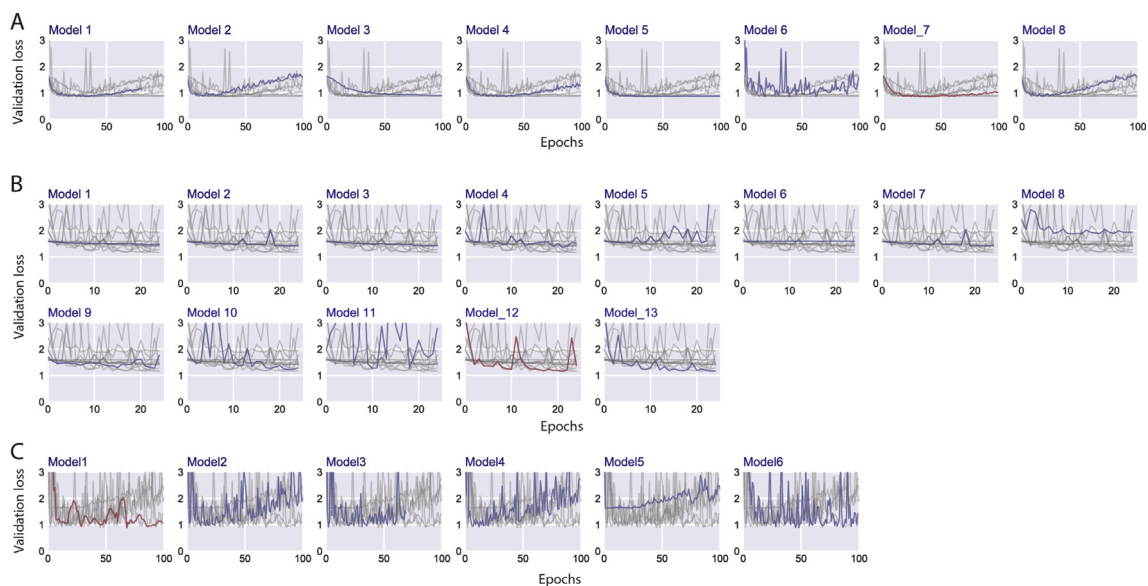
Figure 3. Comparison of models of ResNET-50, CNN, and Xception with different hyperparameters. (A) Validation losses for Xception trained with different hyperparameters using embryo images at 113 hpi. (B) Validation loss of different models of multi-layer CNN with varying hyperparameters trained with 113 hpi embryo images. (C) Validation loss of different models of ResNet to compare the effect of hyperparameters in training embryo images at 113 hpi. The plot with red curve represents the loss curve for the model that achieved the lowest loss among the evaluated models.
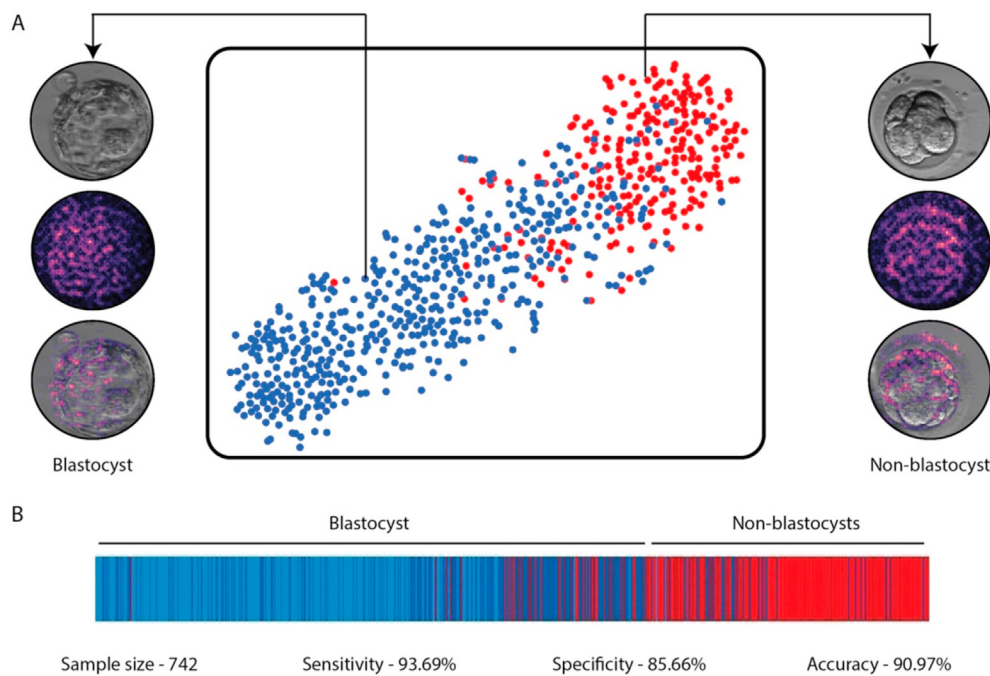
Figure 4. Evaluation at the blastocyst stage. (A) The t-SNE plot for the Xception model trained to classify between non-blastocysts (classes 1 and 2) and blastocysts (classes 3, 4, and 5). The saliency map of the two embryos provides an example of the features that the network use to classify embryos on day 5. (B) The composite of bars illustrates the system's performance in evaluating embryos (n = 742) from the test set of 97 patients. Each blue bar represents blastocysts and red bar represents non-blastocysts, while the color gradients differentiate the subclasses. The bars are sorted from blastocysts to non-blastocysts (blue-red; classes 5-1) based on their actual labels.

blastocyst was 90.97% (CI: 88.67%–92.93%) (Figure 4) where CI is the binomial 95% confidence interval. t-SNE visualization and saliency maps showed a clear separation between the two inference groups and reliance on the embryo morphological features by the network for classification (Figure 4, Figure 5), which further indicate that the model makes use of relevant embryo features that are distinct. The highlighted regions by the saliency maps, more specifically, included regions of cellular fragmentation, blastomeres (in cleavage stage embryos/underdeveloped embryos), cavitation, vacuoles, the inner cell mass, and trophectoderm. The confusion matrix of the network in classification between the five training classes (Figure 6) confirmed the model's ability to differentiate between the blastocysts and non-blastocysts and confusions were usually between classes of adjacent quality level.

The model's micro-average and macro-average area under the curve (AUC) values at the 5-class training level was calculated to be 0.91 and 0.89 (Figure 7), respectively. The AUC values for the classes 1, 2, 3, 4, and 5 were 0.94, 0.85, 0.88, 0.82, and 0.95, respectively.

The sensitivity and specificity of the Xception model in embryo classification between the two inference classes (blastocysts and non-blastocysts) at 113 hpi were 93.69% (CI: 91.16%–95.67%) and 85.66% (CI: 80.70%–89.75%), respectively (n = 742 embryos). The positive predictive value (PPV) and negative predictive value (NPV) of the algorithm were 92.74% (CI: 90.42%–94.54%) and 87.40% (CI: 83.09%–90.73%), respectively. The AUC metric of the model when evaluated for 2-class classification performance was of 0.963 (CI: 0.947 to 0.975) (Figure 7).

In a comparative analysis with the other architectures evaluated in this study, Xception performed better than most architectures with mean differences in classification accuracies of 0.7652%, 1.393%, 12.04%, 8.273%, 0.5390%, and 0.2695% when compared to ResNet-50, Inception-v3, NASNet, multilayer-CNNs, ResNeXt-50, and InceptionResNet-v2, respectively (Table 4). Xception performed marginally poorer than ResNeXt-101 with a mean difference of 0.2697% (Table 4). However, an ANOVA test with Tukey's multiple comparison correction, revealed that only the performance gains over NASNet and multilayer CNNs were statistically significant (P < 0.05).

3.3. Model performance on domain shifted data

We also evaluated the performance of models using embryo datasets recorded with imaging systems other than the Embryoscope used for collecting the training dataset used in this study. For this assessment, we used embryo images submitted from 8 fertility practices to the Society for Reproductive Biologists and Technologists (SRBT) for the Embryo ATLAS project (Figure 8), which were imaged using standard inverted brightfield microscopes and annotated by 8 director level embryologists. Using a test set of 258 embryo images and without any additional training or image pre-processing, the average accuracy along with the standard error of the networks in categorizing the embryos into two classes of blastocyst and non-blastocyst of the best models from Xception, Inception-v3, ResNET-50, NASNetLarge, multilayer CNN, ResNeXt-101, ResNeXt-50, and Inception-ResNET-v2 were 89.53% ± 2.134%, 85.9% ± 1.873%, 84% ± 5.01%, 72.48% ± 3.552%, 69.38% ± 8.067%, 77.52% ± 4.084%, 84.75% ± 0.904%, and 85.01% ± 1.68%, respectively (Table 4). The best Xception model performed with an accuracy of 91.47% with a CI of 87.37%–94.58% in classifying between blastocysts and non-blastocysts (Table 4). Xception performed with a sensitivity of 92.31% (CI: 85.90%–96.42%) and a specificity of 90.78% (CI: 84.75%–95.00%). The PPV and NPV of the CNN were 89.26% (CI: 83.15%–93.33%) and 93.43% (CI: 88.34%–96.39%), respectively with an AUC of 0.975 (CI: 0.948 to 0.990). Saliency maps were also mapped to confirm that the model was using morphological features of the embryos in these images for its classifications. Interestingly, in our evaluations the performance of the Xception model trained on Embryoscope data tested on the domain-shifted SRBT data was similar to its original performance. However, all other networks showed a drastic drop in performance in comparison to their original performance, when tested on the domain-shifted SRBT dataset (Table 4). In a comparative analysis with the other architectures evaluated in this study, Xception performed better than all architectures with mean differences in classification accuracies of 5.534%, 3.631%, 17.05%, 20.16%, 12.01%, 4.779%, and 4.522% when compared to ResNet-50, Inception-v3, NASNet, multilayer-CNNs, ResNeXt-101, ResNeXt-50, and InceptionResNet-v2, respectively

(Table 4). However, an ANOVA test with Tukey's multiple comparison correction, revealed that only the performance gain over multilayer CNNs were statistically significant (P < 0.05).

Images collected using the Embryoscope system was used in training the network and the SRBT images collected at different clinical laboratories using inverted-brightfield microscopes were used for our
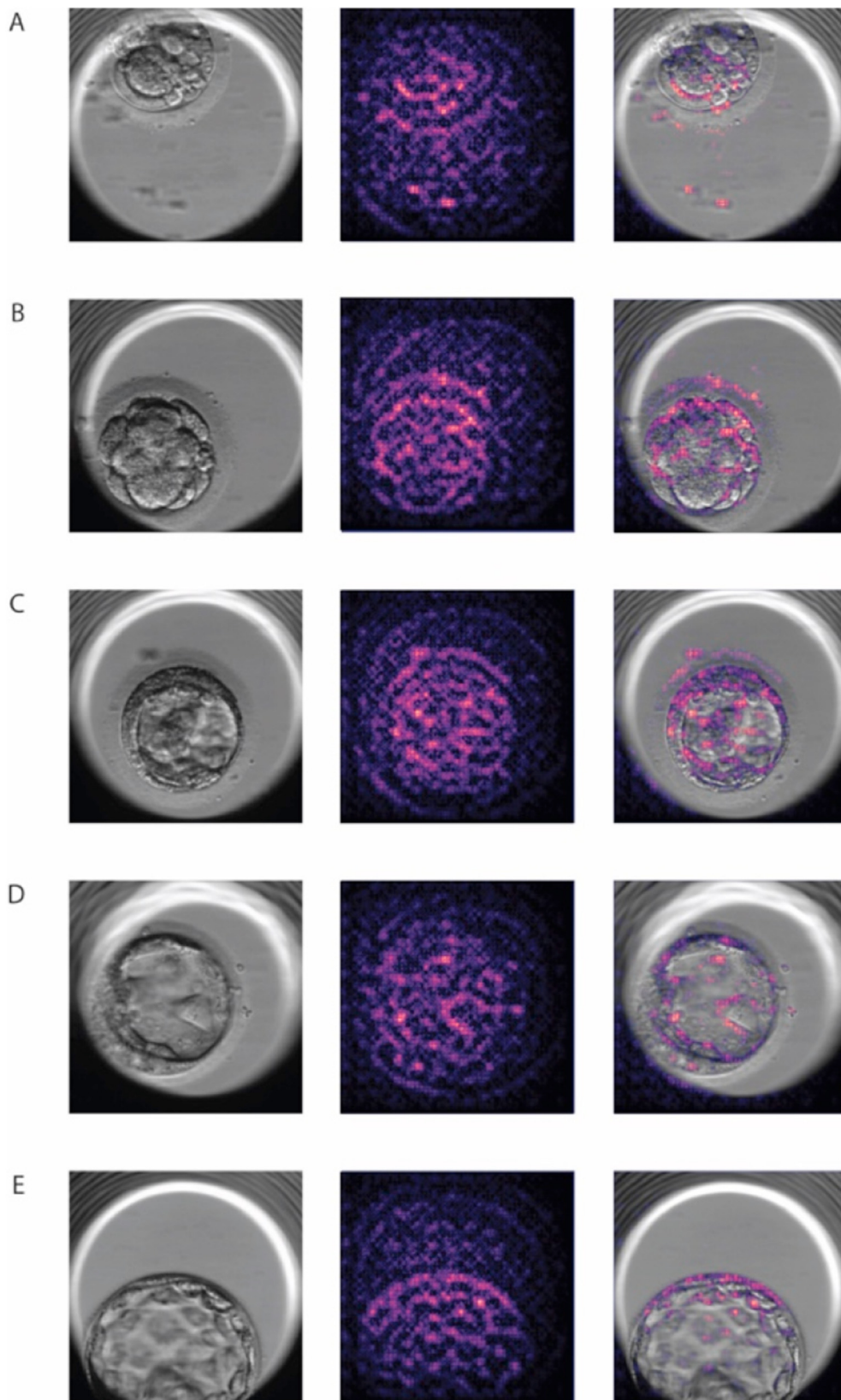
Figure 5. Saliency maps of embryos assessed at 113 h post-insemination. The saliency map was extracted from the network to highlight the highest weighted features for the embryo image. (A) A class 1 embryo category at 113 hpi along with its respective saliency map and saliency map overlaid on the bright-field embryoscope image (B) A class 2 embryo category at 113 hpi along with its respective saliency map and saliency map overlaid on the bright-field embryoscope image. (C) A class 3 embryo category at 113 hpi along with its respective saliency map and saliency map overlaid on the bright-field embryoscope image. (D) A class 4 embryo category at 113 hpi along with its respective saliency map and saliency map overlaid on the bright-field embryoscope image. (E) A class 5 embryo category at 113 hpi along with its respective saliency map and saliency map overlaid on the bright-field embryoscope image. The highlighted regions included regions of cellular fragmentation, blastomeres (in cleavage stage embryos), cavitation, vacuoles, the inner cell mass and trophectoderm.
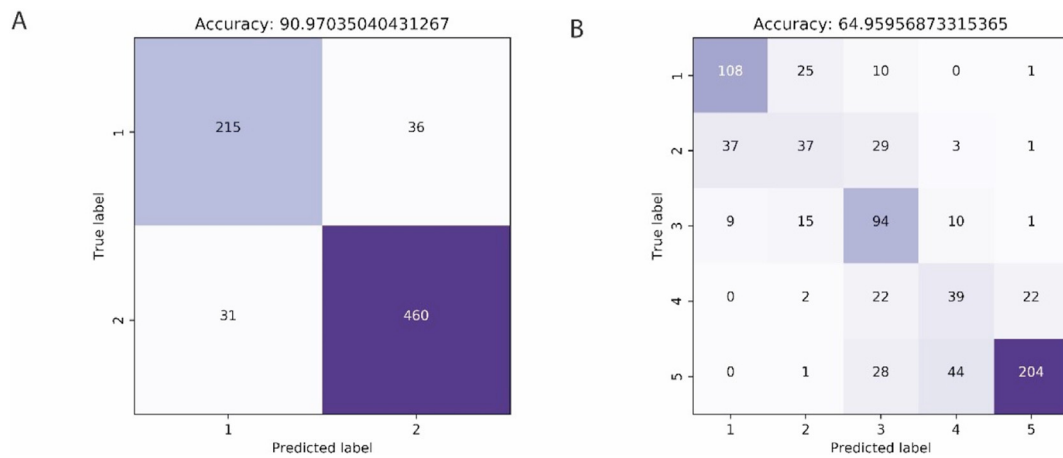
Figure 6. Confusion Matrices of the best Xception model in embryo classification tasks. (A) Confusion matrix for predicting embryos between 2 classes using 113 hpi embryo images. (B) Confusion matrix for classifying embryos between 5 classes using 113 hpi embryo images. Rows represent the historic clinical annotation while the columns indicate the network's predictions.
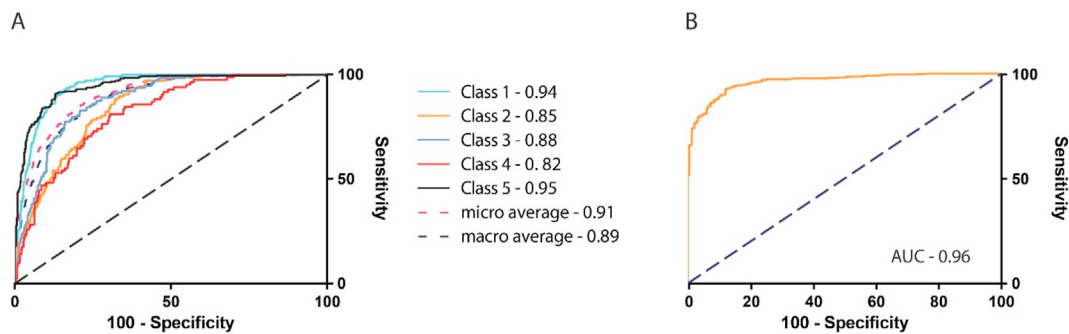

Figure 7. ROC analysis of the classification task performed by the best Xception model. (A) ROC analysis performed for all 5 classes of embryo images imaged at 113 hpi. (B) ROC curves for blastocyst and non-blastocyst classification and prediction tasks 113 hpi.

experiments with domain-shifted data. All images here are resized to 210 × 210 pixels to reflect the input of the neural networks.

4. Discussion and conclusions

Here, we report the development and evaluation of an AI-based approach for automated human embryo assessment of embryo development at 113 hpi. In the recent years, due to the upsurge in deep-learning research, various complex neural network architectures have been proposed and used for image recognition tasks and performance of these architectures is highly dependent on the task. In our study, CNN approaches were evaluated using a dataset of 113 hpi embryo images in classifying embryos based on their morphological quality. Firstly, our evaluations of whether a simple deep CNN, of up to 40 layers, was
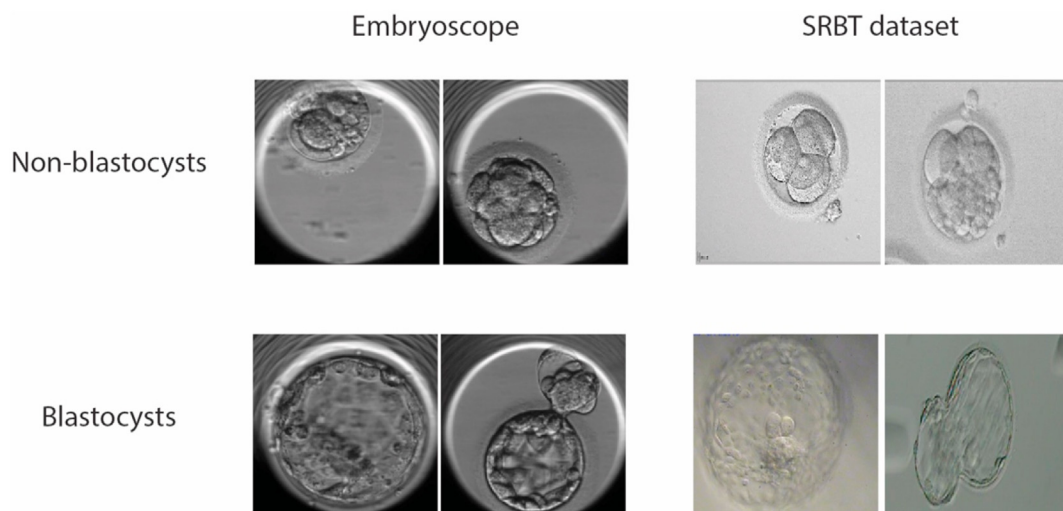
Figure 8. Representative images of Embryoscope and SRBT datasets.

sufficient for efficiently assessing embryos indicated that more sophisticated networks may be preferable than simply stacked deeper networks. Inception, (48 layers) which uses inception modules that are composed of multiple filters of different sizes, over simple convolution layers showed significantly better performance than our de novo CNNs. We evaluated the performance of other popular networks such as ResNET, which originally introduced residual blocks, and the hybrid Inception-ResNET. We also evaluated the performance of networks such as ResNeXt-101 and ResNeXt-50 that make use of the repeating layers, similar to ResNet, along with the split-transform-merge strategy exploited in Inception. Finally, NASNets, which utilize a reinforcement learning search method to optimize architecture configurations with recurrent neural network controller, were included in our study to understand the suitability of frameworks that made use of partial self learning strategy. Xception was developed with Inception as its base architecture, while adding residual blocks and replacing the convolutions in original inception modules with depthwise seperable convolutions.

Our observations have shown that Xception performed best in learning the categorical embryo data and was able to classify them based on their morphological quality. Xception, interestingly, performed well on domain shifted data (SRBT dataset) that was acquired through different imaging systems, which is uncommon with machine-learning- or computer vision-based approaches (Tommasi et al., 2015). Saliency maps of embryos imaged at the blastocyst stage (113 hpi), highlight the whole embryo as regions of interest which is further indicative of well-trained model that utilizes features that are relevant for classification (Figure 5). In this study, we made use of clinical data in evaluating shifted images and observed that the performance of different models of the same network tends to vary more in the shifted data domain, similar to prior reports (D'Amour et al., 2020). While the focus of this work was more specific to the utility of the various neural network architectures in embryology, further studies are needed to better understand the effect of domain shifts when using real-world clinical data. Even though our findings in this study are suggestive that the thus trained Xception model is more robust even with domain shifted data, additional evaluations on network consistency are required to be conclusive. It is, however, encouraging for future studies that make use of Xception for embryo morphological analyses.

The primary goal in an IVF procedure is to culture and transfer an embryo that will result in a healthy baby. Embryologists, therefore, try to identify the embryo of highest potential for transfer and to avoid embryos of the lower quality from a cohort for patients with good prognosis. A neural network's raw ability of separating embryos between five classes does not directly benefit such clinical processes. Furthermore, the 5-class classification accuracy should be taken with caution since it also affected by the annotating embryologist's ability to repeatably and consistently categorize embryos based on their morphology, which has been observed to be not ideal. However, the evaluations are useful in understanding the network learning. These networks, to be clinically viable, need to be modified to suit the applications. For example, Xception correctly classified >99.5% of the highest quality blastocysts as good embryos (blastocysts) which is of critical importance, clinically, when identifying embryos suited for transfer. In this work, primarily to minimize sparsity of data in limited dataset, we have classified embryos based on hierarchical classification system that consolidates the MGH blastocyst categorization into 5 classes though embryologists have highlighted that 5-class system for embryo morphology-based classification may be more beneficial over commonly used 3 class classification. For the study, we have consolidated our network's 5-class output to 2 inference classes and differentiated embryos between blastocysts and non-blastocysts to highlight its performance on a more universal classification system (blastocysts and non-blastocysts) since embryo categorization criteria tends to vary with each clinic.

The work presented here is an example of demonstrating how deep-learning techniques can be used in medicine particularly in an IVF procedure. The IVF community can greatly benefit from the modern advancements in machine learning. Our future work will be focused on many valuable applications and goals, such as predicting embryo developmental outcomes at earlier timepoints, studying the use of neural networks in aiding routine clinical tasks, and in predicting the eventual outcome of embryos. Our presented results show the promise of using neural networks in accurate embryo assessments with the potential to eventually improve IVF practices in both resource-rich and resource-poor settings regardless of the center's experience and infrastructure.

Declarations

Author contribution statement

P. Thirumalaraju, M. K. Kanakasabapathy, C. L. Bormann, H. Shafiee: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

R. Gupta, R. Pooniwala: Performed the experiments; Analyzed and interpreted the data.

H. Kandula: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

I. Souter, I. Dimitriadis: Contributed reagents, materials, analysis tools or data.

Data availability statement

Restrictions apply to the availability of the medical training/validation data, which were used with permission for the current study, and so are not publicly available. Some data may be available from the authors upon reasonable request and with permission of the Massachusetts General Hospital.

Declaration of interests statement

Prudhvi Thirumalaraju, Manoj Kumar Kanakasabapathy, Charles L Bormann Hadi Shafiee have a patent WO2019068073A1 pending.

Additional information

No additional information is available for this paper.

References

Barash, O., Ivani, K., Huen, N., Willman, S., Weckstein, L., 2017. Morphology of the blastocysts is the single most important factor affecting clinical pregnancy rates in IVF PGS cycles with single embryo transfers. Fertil. Steril. 108, e99.

Birenbaum-Carmeli, D., 2004. 'Cheaper than a newcomer': on the social production of IVF policy in Israel. Sociol. Health Illness 26, 897–924.

CDC, 2015. Fertility Clinic Success Rates Report.

Chollet, F., 2016. Xception: Deep Learning with Depthwise Separable Convolutions arXiv, 1610.02357.

Conaghan, J., Chen, A.A., Willman, S.P., Ivani, K., Chenette, P.E., Boostanfar, R., Baker, V.L., Adamson, G.D., Abusief, M.E., Gvakharia, M., et al., 2013. Improving embryo selection using a computer-automated time-lapse image analysis test plus day

3 morphology: results from a prospective multicenter trial. Fertil. Steril. 100, 412–419 e415.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D., et al., 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning arXiv:2011.03395.

Demko, Z.P., Simon, A.L., McCoy, R.C., Petrov, D.A., Rabinowitz, M., 2016. Effects of maternal age on euploidy rates in a large cohort of embryos analyzed with 24-chromosome single-nucleotide polymorphism–based preimplantation genetic screening. Fertil. Steril. 105, 1307–1313.

Dimitriadis, I., Bormann, C.L., Kanakasabapathy, M.K., Thirumalaraju, P., Gupta, R., Pooniwala, R., Souter, I., Rice, S.T., Bhowmick, P., Shafiee, H., 2019. Deep convolutional neural networks (CNN) for assessment and selection of normally fertilized human embryos. Fertil. Steril. 112, e272.

Dimitriadis, I., Bormann, C.L., Thirumalaraju, P., Kanakasabapathy, M., Gupta, R., Pooniwala, R., Souter, I., Hsu, J.Y., Rice, S.T., Bhowmick, P., et al., 2019. Artificial intelligence-enabled system for embryo classification and selection based on image analysis. Fertil. Steril. 111, e21.

Einarsson, S., Bergh, C., Friberg, B., Pinborg, A., Klajnbard, A., Karlström, P.-O., Kluge, L., Larsson, I., Loft, A., Mikkelsen-Englund, A.-L., et al., 2017. Weight reduction intervention for obese infertile women prior to IVF: a randomized controlled trial. Hum. Reprod. 32, 1621–1630.

Erenus, M., Zouves, C., Rajamahendran, P., Leung, S., Fluker, M., Gomel, V., 1991. The effect of embryo quality on subsequent pregnancy rates after in vitro fertilization. Fertil. Steril. 56, 707–710.

Filho, E.S., Noble, J.A., Wells, D., 2010. A review on automatic analysis of human embryo microscope images. Open Biomed. Eng. J. 4, 170–177.

Hariton, E., Dimitriadis, I., Kanakasabapathy, M.K., Thirumalaraju, P., Gupta, R., Pooniwala, R., Souter, I., Rice, S.T., Bhowmick, P., Ramirez, L.B., et al., 2019. A deep learning framework outperforms embryologists in selecting day 5 euploid blastocysts with the highest implantation potential. Fertil. Steril. 112, e77–e78.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. ArXiv e-prints.

Hill, G.A., Freeman, M., Bastias, M.C., Jane Rogers, B., Herbert III, C.M., Osteen, K.G., Wentz, A.C., 1989. The influence of oocyte maturity and embryo quality on pregnancy rate in a program for in vitro fertilization-embryo transfer. Fertil. Steril. 52, 801–806.

Kanakasabapathy, M., Dimitriadis, I., Thirumalaraju, P., Bormann, C.L., Souter, I., Hsu, J., Thatcher, M.L., Veiga, C., Shafiee, H., 2019. An inexpensive, automated artificial intelligence (AI) system for human embryo morphology evaluation and transfer selection. Fertil. Steril. 111, e11.

Kanakasabapathy, M.K., Thirumalaraju, P., Bormann, C.L., Kandula, H., Dimitriadis, I., Souter, I., Yogesh, V., Kota Sai Pavan, S., Yarravarapu, D., Gupta, R., et al., 2019. Development and evaluation of inexpensive automated deep learning-based imaging systems for embryology. Lab Chip 19, 4139–4145.

Kanakasabapathy, M.K., Thirumalaraju, P., Gupta, R., Pooniwala, R., Kandula, H., Souter, I., Dimitriadis, I., Bormann, C.L., Shafiee, H., 2019. Improved monitoring of human embryo culture conditions using a deep learning-derived key performance indicator (KPI). Fertil. Steril. 112, e70–e71.

Khosravi, P., Kazemi, E., Zhan, Q., Malmsten, J.E., Toschi, M., Zisimopoulos, P., Sigaras, A., Lavery, S., Cooper, L.A.D., Hickman, C., et al., 2019. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. NPJ Digit. Med. 2, 21.

Machtinger, R., Racowsky, C., 2013. Morphological systems of human embryo assessment and clinical evidence. Reprod. Biomed. Online 26, 210–221.

Mascarenhas, M.N., Flaxman, S.R., Boerma, T., Vanderpoel, S., Stevens, G.A., 2012. National, regional, and global trends in infertility prevalence since 1990: a systematic analysis of 277 Health surveys. PLoS Med. 9, e1001356.

Matos, F.D., Rocha, J.C., Nogueira, M.F.G., 2014. A method using artificial neural networks to morphologically assess mouse blastocyst quality. J. Anim. Sci. Technol. 56, 15.

Osman, A., Alsomait, H., Seshadri, S., El-Toukhy, T., Khalaf, Y., 2015. The effect of sperm DNA fragmentation on live birth rate after IVF or ICSI: a systematic review and meta-analysis. Reprod. Biomed. Online 30, 120–127.

Paulson, R.J., Sauer, M.V., Lobo, R.A., 1990. Embryo implantation after human in vitro fertilization: importance of endometrial receptivity. Fertil. Steril. 53, 870–874.

Racowsky, C., Kovacs, P., Martins, W.P., 2015. A critical appraisal of time-lapse imaging for embryo selection: where are we and where do we need to go? J. Assist. Reprod. Genet. 32, 1025–1030.

Rocha, J.C., Passalia, F.J., Matos, F.D., Takahashi, M.B., Ciniciato, D.d.S., Maserati, M.P., Alves, M.F., de Almeida, T.G., Cardoso, B.L., Basso, A.C., et al., 2017. A method based on artificial intelligence to fully automatize the evaluation of bovine blastocyst images. Sci. Rep. 7, 7659.

Rocha, J.C., Passalia, F.J., Matos, F.D., Takahashi, M.B., Maserati Jr., M.P., Alves, M.F., de Almeida, T.G., Cardoso, B.L., Basso, A.C., Nogueira, M.F.G., 2017. Automatized image processing of bovine blastocysts produced in vitro for quantitative variable determination. Sci. Data 4, 170192.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2016. Inception-v4, inception-ResNet and the impact of residual connections on learning. ArXiv e-prints.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the inception architecture for computer vision. ArXiv e-prints.

Thirumalaraju, P., Bormann, C.L., Kanakasabapathy, M.K., Kandula, H., Shafiee, H., 2019. Deep learning-enabled prediction of fertilization based on oocyte morphological quality. Fertil. Steril. 112, e275.

Thirumalaraju, P., Hsu, J.Y., Bormann, C.L., Kanakasabapathy, M., Souter, I., Dimitriadis, I., Dickinson, K.A., Pooniwala, R., Gupta, R., Yogesh, V., et al., 2019. Deep learning-enabled blastocyst prediction system for cleavage stage embryo selection. Fertil. Steril. 111, e29.

Thirumalaraju, P., Kanakasabapathy, M.K., Gupta, R., Pooniwala, R., Kandula, H., Souter, I., Dimitriadis, I., Bormann, C.L., Shafiee, H., 2019. Automated quality assessment of individual embryologists performing ICSI using deep learning-enabled fertilization and embryo grading technology. Fertil. Steril. 112, e71.

Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T., 2015. A deeper look at dataset bias. arXiv e-prints.

Toner, J.P., 2002. Progress we can be proud of: U.S. trends in assisted reproduction over the first 20 years. Fertil. Steril. 78, 943–950.

Tran, D., Cooke, S., Illingworth, P.J., Gardner, D.K., 2019. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Hum. Reprod. 34, 1011–1018.

Turchi, P., 2015. Prevalence, definition, and classification of infertility. In: Cavallini, G., Beretta, G. (Eds.), Clinical Management of Male Infertility. Springer International Publishing), Cham, pp. 5–11.

Vaegter, K.K., Lakic, T.G., Olovsson, M., Berglund, L., Brodin, T., Holte, J., 2017. Which factors are most predictive for live birth after in vitro fertilization and intracytoplasmic sperm injection (IVF/ICSI) treatments? Analysis of 100 prospectively recorded variables in 8,400 IVF/ICSI single-embryo transfers. Fertil. Steril. 107, 641–648 e642.

van der Maaten, L., Hinton, G., 2008. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Wong, C., Chen, A.A., Behr, B., Shen, S., 2013. Time-lapse microscopy and image analysis in basic and clinical embryo development research. Reprod. Biomed. Online 26, 120–129.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2016. Aggregated Residual Transformations for Deep Neural Networks arXiv:1611.05431.

Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2017. Learning Transferable Architectures for Scalable Image Recognition arXiv:1707.07012.