DEEP LEARNING DEVELOPMENTAL MODEL FOR EMBRYO GRADING IN THAI POPULATION

Budsaba Wiriyasirivaj[1], Sawit Limkiatsataporn[2], Apisit Pukinghin[2], Phatrapron Kuekulkomain[3], Pornprom Promrungrueng[1], Busaba Supawattanabodee[4] and Kietikul Jearanaitanakij[2,*]

Received: June 29, 2024; Revised: August 05, 2024; Accepted: August 08, 2024

Abstract

In light of the growing challenges associated with infertility, an increasing number of researchers are resorting to assisted reproductive technologies such as In Vitro Fertilization (IVF). Embryo grading is a crucial step in the IVF process that requires embryologists' expertise. However, their limited availability has led to the exploration of technological alternatives. This study aims to use deep learning for human embryo grading, with models specifically designed for the dataset in Thailand at Vajira Hospital. The process of IVF at Vajira Hospital presents its own set of challenges since its embryo classification extends beyond the Istanbul consensus. Furthermore, classes that occur infrequently are removed and classes with similarities are merged. We apply transfer learning to pretrained deep learning models like VGG19, VGG16, ResNet152, Resnet101, Xception, InceptionV3, and EfficientNet, to create a system capable of accurately classifying embryo quality scores. Experimental results from the embryo dataset collected from Vajira Hospital in Thailand demonstrate the proposed classifier's accuracy, precision, recall, f1-score, and AUC superiority. This research contributes to the field of IVF in Thailand by potentially reducing human errors and addressing the demand for skilled embryologists.

Keywords: Infertility; In Vitro Fertilization; Embryo Grading; Deep Learning; Classification

[1] *Department of Obstetrics and Gynecology, Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Dusit, Bangkok, 10300, Thailand. E-mail: budsaba@nmu.ac.th; pornprom.p@nmu.ac.th*
[2] *Department of Computer Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Ladkrabang, Bangkok, 10520, Thailand. E-mail: kietikul.je@kmitl.ac.th; 63011052@kmitl.ac.th; 63010960@kmitl.ac.th*
[3] *Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Dusit, Bangkok, 10300, Thailand. E-mail: 6301101070@nmu.ac.th*
[4] *Department of Research and Medical Innovation, Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Dusit, Bangkok, 10300, Thailand. E-mail: sbusaba@nmu.ac.th*
[*] *Corresponding author*

DOI: https://doi.org/10.55766/sujst-2024-04-e05589

Introduction

Creating a potent deep-learning model to classify human embryo images is difficult, especially when pretrained models do not have specific training on embryo datasets. The success of IVF significantly relies on accurately selecting embryos. This process has been subjective, which may introduce inconsistencies in classifying embryos. It's essential to explore objective and reliable methods for embryo grading to enhance IVF outcomes. Deep learning presents a promising instrument to transform this process by examining intricate morphological characteristics from images and time-lapse videos. The subsequent literature reviews of recent studies that employ artificial intelligence and deep learning to enhance different phases of IVF are presented in chronological order.

Basari and Gunawan (2019) suggested an automated approach for determining the quality of blastocyst grading, utilizing a Convolutional Neural Network (CNN) model in conjunction with image processing techniques, aiming to enhance accuracy. It was found that the CNN model combined with the Canny edge detector, greatly surpasses the performance of another model that does not incorporate image preprocessing. Harun *et al*. (2019) applied deep learning to segment irregularly shaped blastocysts (a cluster of dividing cells originating from a fertilized egg). The blastocysts assessed in this study have been subjected to laser ablation of the zona pellucida, a necessary procedure before conducting a trophectoderm biopsy. The high accuracy, precision, and recall in the experimental result indicate the success of their method. Rad *et al*. (2019) presented a semantic segmentation system to help embryologists consider human blastocyst component images. Their method achieves the best Jaccard Index at 82.85%, which reflects a good segmentation.

Kan-Tor *et al*. (2020) established a fully automated and standardized classification system by training deep neural networks on raw video files of more than 6200 embryos labeled for blastulation and over 5500 embryos labeled for implantation. The experimental findings indicate that their embryo implantation prediction surpasses the accuracy of the existing state-of-the-art Morphokinetic IVF classifier. Chavez-Badiola *et al*. (2020) trained the Embryo Ranking Intelligent Classification Algorithm (ERICA) to the blastocyst image database to predict ploidy and implantation. ERICA is designed to help embryologists decide on the embryo selection process leading to embryo transfer. It classifies the embryo image as euploid (embryo is more likely to implant) and aneuploid (there are an abnormal number of chromosomes in a cell). ERICA's performance is quite acceptable as it ranked a euploid blastocyst better than two senior embryologists. However, the experiment they conducted is limited to a small dataset.

Liao *et al*. (2021) presented a deep learning model capable of accurately forecasting the formation of blastocysts and viable blastocysts, utilizing time-lapse microscopy videos from the embryo's initial three days. They combined the DenseNet201 network, focal loss, a long short-term memory (LSTM) network, and a gradient-boosting classifier as the primary elements in their model. Hu *et al*. (2021) introduced a scale-attention deep learning network (SA-Net) that leverages a residual module to extract features at various scales and employs an attention module to enhance its scale-attention ability. SA-Net is adept at learning multi-scale features, leading to more precise segmentation across different medical images. Experimental results demonstrate that SA-Net excels in the application of blastocyst segmentation. Milyea *et al*. (2021) combined computer vision image processing methods and deep learning to develop the non-invasive Life Whisperer AI model. This model robustly predicts embryo viability, gauged by the clinical pregnancy outcome, using individual static images of Day 5 blastocysts captured from conventional optical light microscope systems.

Loewke *et al*. (2022) employed a combination of three Resnet-18 CNN models, each trained at varying image resolutions. Their approach was influenced by how an embryologist observes embryos at multiple magnifications. They underscored the possible advantages of the CNN model, such as enhanced prediction of clinical pregnancy, while also noting potential limitations related to image quality, bias, and the detail level of scores. Berntsen *et al*. (2022) examined the performance of an embryo selection model, based on deep learning and solely utilizing time-lapse image sequences, across patient ages and clinical scenarios. The experiment on the test set indicates the model generalization as its AUC is in the range of 0.60-0.75 for known implantation data. Chéles *et al*. (2022) proposed an image processing protocol to segment and extract the potential variables for embryo fitness prediction. Their approach demonstrates promise as a tool for embryo selection, effectively segmenting images of human blastocysts from two separate sources and deriving 33 variables. Arsalan *et al*. (2022) unveiled a sprint semantic segmentation network (SSS-Net) to identify blastocyst components for embryological studies. Their approach relies on a fully convolutional semantic segmentation framework that offers pixel-wise categorization of crucial blastocyst components, aiding in the automated examination of these elements' morphologies.

Liu *et al.* (2023) applied deep learning to grade the human blastocyst images. They utilized CNN along with the additional multilayer perceptron to process the dataset. The patient couple's clinical data, e.g., the mother's age, hormone level, endometrial thickness, and sperm quality, are also included in the features of their multimodal dataset. They reported an AUC (Area under the receiver operating characteristic curve) of 0.77 indicating a fairly high performance. Barnes *et al.* (2023) introduced an automatic and safe method to predict embryo ploidy by applying machine learning and deep learning techniques. Their dataset consists of images captured at 110-hour intracytoplasmic sperm injection (ICSI) and some external clinical data. Their method showed potential as an assistant tool to predict embryo ploidy of IVF in a non-invasive manner. Berman *et al.* (2023) conducted a study on the precision of CNN models in evaluating embryos through time-lapse monitoring. Their approach holds promise for delivering more efficient, precise, and unbiased evaluations of embryos. The models that focused on the classification of the blastocyst stage demonstrated the most accurate predictions.

In this study, we introduce a deep-learning model for grading human embryos, utilizing an embryo image dataset in Thailand gathered at Vajira Hospital. We evaluate various pretrained CNN models to identify the one that delivers the most effective performance. A sequence of dense, dropout, and batch normalization layers is appended to the pretrained model to finetune it with the training set. The suggested model demonstrates promising results in classifying the embryo test set and holds potential for further development as a supportive tool for embryologists in assessing embryo quality during the IVF process.

The challenge of our research is to improve the unique IVF process at the Vajira Hospital. Specifically, it extends the embryo classification beyond the Istanbul consensus. Three additional classes-Arrest, Early, and Morula - present difficulties in distinguishing them due to their subtle attributes and common characteristics. Additionally, infrequent classes are eliminated, and similar classes are merged.

This paper is structured as follows: The next section describes the materials and methods by introducing the essential concepts relevant to this study, the embryo dataset, and the proposed method. Then our experimental results and comparisons with other CNN models are given in the results and discussion section. Finally, the conclusion is provided in the last section.

Materials and Methods

In Vitro Fertilization (IVF): IVF is a groundbreaking procedure that has transformed the approach to infertility treatment. It is a process where an egg and sperm are combined outside the human body, usually in a lab environment, to create an embryo (Coward and Wells, 2013). This method offers a lifeline for couples who have difficulties conceiving naturally and has emerged as a commonly employed assisted reproductive technique. The IVF process involves several stages, including the stimulation of the ovaries, extraction of eggs, fertilization in a lab setting, cultivation of the embryo, and finally, the implantation of a chosen embryo into the woman's uterus. The effectiveness of IVF is dependent on a variety of elements, such as the quality of the reproductive cells, ideal lab conditions, and the successful selection of embryos. The constant evolution of IVF techniques, along with technological breakthroughs like artificial intelligence, strive to boost success rates and further the ongoing advancements in the field of reproductive medicine.

Embryo Selection: Embryo selection, a key component of IVF, entails the assessment and selection of embryos that have the greatest potential for successful implantation and robust growth (Coward and Wells, 2013). This procedure employs a variety of methods, with morphological grading used to evaluate an embryo's physical characteristics as a measure of its viability. In recent times, Preimplantation Genetic Diagnosis (PGD) has surfaced as a more sophisticated instrument, facilitating the detection of particular genetic disorders before implantation. Nevertheless, ethical considerations and constraints persist. For example, the precision of PGD for intricate genetic traits is contested, and there are apprehensions about the potential societal consequences of selection based on specific traits beyond the prevention of serious genetic illnesses. A balanced approach is essential since it harmonizes technological progress with ethical norms and continuous research for the conscientious application of embryo selection in assisted reproductive technologies.

Deep Learning: Deep Learning, as described by Goodfellow *et al.* (2016), is a distinct discipline within the realm of artificial intelligence that concentrates on training intricate neural networks to carry out complex tasks, often outperforming conventional machine learning methods. These neural networks, modeled after the structure of the human brain, are composed of multiple layers, enabling them to autonomously learn hierarchical features and representations from data. The training procedure entails tweaking the parameters of these networks via backpropagation, reducing errors, and boosting the model's predictive process. Convolutional Neural Networks (CNNs) are especially effective for image recognition tasks, while Recurrent Neural Networks (RNNs) are adept at handling sequential data. Recent progress includes attention mechanisms for enhanced concentration, transfer learning to utilize pre-trained models, and generative adversarial networks for generating realistic synthetic data. Despite significant achievements, challenges such as the requirement for large labeled datasets, computational resources, and the interpretability of complex models persist. Nevertheless, deep learning continues to catalyze innovations in various applications like computer vision, natural language comprehension, and autonomous systems.

Convolutional Neural Network (CNN): Convolutional Neural Networks (CNNs) signify a crucial progression in deep learning, specifically designed for the analysis of visual data and image processing tasks. Stemming from the foundational work by LeCun *et al*. (1998), CNNs have been consistently refined, forming the core of numerous applications related to images. The architecture utilizes convolutional layers to autonomously learn hierarchical features from input images, aided by the convolutional operation and pooling layers that capture spatial hierarchies and down-sample learned features. The pioneering AlexNet architecture (Krizhevsky *et al*., 2012) showcased the potential of CNNs in large-scale image recognition, sparking a flurry of innovations with subsequent architectures like VGG, GoogLeNet, and ResNet. CNNs have attained remarkable success in various fields, including medical imaging for tasks such as tumor detection and organ segmentation. Despite their significant achievements, challenges remain, such as the requirement for extensive labeled datasets and addressing ethical issues related to biases in training data. The literature persistently explores ways to enhance CNNs, with a focus on interpretability, transfer learning, and domain-specific adaptations to ensure improved performance and reliability (LeCun *et al*., 1998; Simonyan and Zisserman, 2014; He *et al*., 2016).

Transfer Learning: Transfer Learning is a concept in machine learning that utilizes the knowledge acquired from addressing one task to enhance performance on a separate but related task. In the realm of deep learning, it involves initially training a model on a substantial dataset for a specific task and subsequently fine-tuning it on a smaller dataset for a related task. This method takes advantage of the representational power learned during pre-training and tailors it to the specifics of the target task. The foundational work of Pan and Yang (2010) provides a classification of transfer learning into domain adaptation, where the source and target domains vary, and inductive transfer, where the source and target tasks are different. Transfer learning has shown to be particularly beneficial in situations with scarce labeled data for the target task. Prominent applications include natural language processing, image recognition, and healthcare, where pre-trained models, such as OpenAI's GPT (Radford *et al*., 2018) and various architectures like VGG and ResNet, have exhibited the adaptability and efficiency of transfer learning approaches.

Visual Geometry Group (VGG) Model: VGG, created by Simonyan and Zisserman (2014), is a notable architecture within the sphere of CNNs. Its distinguishing feature is its layered convolutional layers with compact filter sizes (3×3), yielding remarkable outcomes on image classification tasks such as ImageNet. Despite the advent of more complex architectures, VGG continues to be a fundamental model, spurring additional advancements and acting as a standard for comparison. Its triumph underscores the effectiveness of straightforward building blocks when properly layered and optimized, illustrating the significance of architectural design in the performance of CNNs.

Residual Neural Network (ResNet) Model: Residual Networks (ResNets), brought forth by He *et al*. (2016), brought about a revolution in deep learning by tackling the vanishing gradient problem, a significant hurdle in training extremely deep neural networks. Drawing inspiration from shortcut connections in Highway Networks (Srivastava *et al*., 2015), ResNets utilize "skip connections" that permit information to bypass layers and flow directly through the network in addition to the regular convolutional layers. This direct pathway alleviates the vanishing gradient problem, allowing ResNets to attain markedly higher accuracy than conventional feedforward networks on the ImageNet benchmark. The triumph of ResNets ignited further exploration into residual architectures, with variants like Pre-activation ResNets investigating different design alternatives and achieving even superior performance. The fundamental principle of residual connections has become a cornerstone in numerous contemporary deep learning architectures, showcasing its enduring influence on the field.

InceptionV3 Model: InceptionV3, invented by Google (Szegedy *et al*., 2015), is a formidable CNN celebrated for its precision and efficiency in image classification. Drawing inspiration from the Inception architecture, it employs multiple parallel pathways with diverse filter sizes to capture a wide range of information within an image. This design tackles the issue of vanishing gradients, a limitation in training deeper networks, enabling InceptionV3 to attain substantial depth (48 layers) relative to its forerunners. Trained on the extensive ImageNet dataset, it showcases remarkable performance on various of tasks such as object detection and image recognition, achieving over 78% accuracy on the ImageNet classification benchmark. Despite its accomplishments, research is ongoing to explore enhancements in areas like interpretability and efficiency, for instance, employing pruning techniques to decrease model size without compromising accuracy. In summary, InceptionV3 serves a fundamental pillar in the progression of deep learning architectures, exemplifying the efficacy of meticulous design and efficient training strategies in addressing complex image analysis tasks.

Xception Model: Xception (Chollet, 2017) offers a distinctive CNN architecture engineered for enhanced efficiency while preserving high accuracy. Taking cues from InceptionV3, it substitutes conventional convolutional layers with depth-wise separable convolutions, decomposing the convolution operation into two stages: depth-wise convolution, which applies filters independently to each channel, and pointwise convolution, which amalgamates the features across channels using 1×1 convolutions. This strategy considerably diminishes computational expense while retaining information, enabling Xception to attain performance on par with InceptionV3 on image classification tasks with fewer parameters and reduced computational overhead. Even though it is not as prevalently employed as other CNN architectures like ResNet, Xception's inventive design is a noteworthy addition to efficient deep-learning architectures, showcasing the potential of depth-wise separable convolutions for future network design.

EfficientNet Model: EfficientNet, unveiled by Tan and Le (2019), is a collection of CNN architectures engineered to attain exceptional accuracy while preserving computational efficiency. This equilibrium is realized through an innovative compound scaling method that scales the depth, width, and resolution of the network uniformly, in contrast to traditional methods that scale them unevenly. This scaling strategy, coupled with the incorporation of squeeze-and-excite modules for channel-wise attention, enables EfficientNet models to achieve cutting-edge results on image classification tasks like ImageNet, all while requiring significantly fewer parameters and computational resources compared to their peers. For example, EfficientNet-B0, the smallest model in the series, achieves accuracy comparable to VGG16 but with only 5.3 million parameters, showcasing its substantial efficiency improvements. This emphasis on efficiency renders EfficientNet particularly suitable for applications where computational resources are constrained, such as mobile devices and embedded systems.

Embryo Dataset Collection and Labelling: We collect image and scoring information for each type of embryo in the Blastocyst stage at the Vajira IVF Clinic, Faculty of Medicine, Vajira Hospital, Navamindradhiraj University. According to the Istanbul Consensus (Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology 2011), the embryo assessment involves three primary criteria: State of Development, Inner Cell Mass Development, and Trophectoderm Development. Each criterion is assigned a score within a specific range. State of Development is evaluated within a range of 1 to 4. Inner Cell Mass Development measures the growth of the inner cell mass, scaled from 1 to 3. Trophectoderm Development indicates the development of the outer cell mass to become the placenta and other tissues, ranging from 1 to 3. For example, three numbers in the grading "1-2-3" represent scores for State of Development of 1, Inner Cell Mass Development of 2, and Trophectoderm Development of 3, respectively.

Besides stages from the Istanbul Consensus, three additional challenging stages are introduced by the Vajira IVF Clinic: Arrested, Early, and Morula. These three additional classes are challenging to classify since they have many common delicate features. Early and Morula scores are measured from features of Inner Cell Mass and Trophectoderm while the Arrested score indicates cessation of active cell division. In summary, there are a total of 19 embryo classes as follows: 2-1-1, 2-1-2, 2-1-3, 2-2-1, 2-2-2, 2-2-3, 2-3-3, 3-1-1, 3-1-2, 3-1-3, 3-2-1, 3-2-2, 3-2-3, 3-3-2, 3-3-3, 4-2-2, Arrested, Early, and Morula. Note that the missing classes are either merged with a similar class or neglected due to their seldom-found instances. Each Image collected from the laboratory comprises multiple embryos, captured at 60 hours post-fertilization within the culture dish, as depicted in Figure 1. To segment an image into individual embryos, the finetuned YOLOv7 model was employed to facilitate object detection. The outcome of this procedure is a collection of the individual embryo images, as illustrated in Figure 2. Once the individual embryo images have been extracted, each image is tagged with the grade

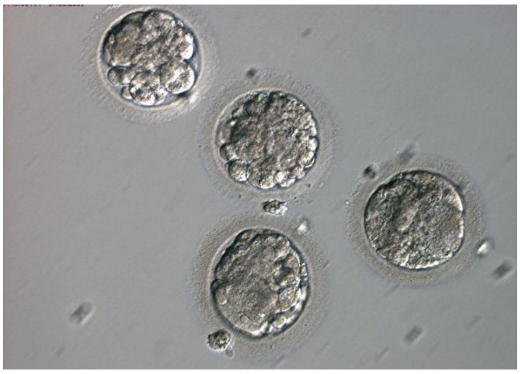

Figure 1. An image of multiple embryos in the Blastocyst stage

Figure 2. Individual embryos

Figure 3. Sample images of each class

recommended by the embryologist and stored in the treatment file. Figure 3 illustrates sample images of embryos in each of the 19 classes. The dataset comprises a total of 312 embryo images, distributed across 19 classes with varying images.

Data Augmentation: To increase the variety of images and conduct our experiments on a balanced dataset, data augmentation techniques - flipping, rotating, contrasting, exposure, and saturation adjusting - were applied to the original images, resulting in a total of 24,024 images. Afterward, the dataset was split into training, validation, and test sets in a 60:20:20 ratio. Consequently, the training set comprised 14,414 images, while both the validation and test sets contained 4,805 images each. Figure 4 depicts sample images from the augmentation process.

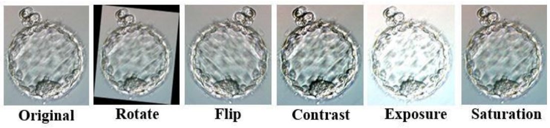

Figure 4. Augmented image samples

Model Creation: Seven well-known pretrained models - InceptionV3, Xception, VGG16, VGG19, ResNet152, ResNet101, and EfficientNet - are employed as candidates for the potential CNN backbone of the proposed model. We utilize each of the seven CNNs to extract features from embryo images without altering their pretrained weights.

Afterward, these features are passed to a sequence of batch normalization, dense (fully connected), and dropout layers (with a dropout probability of 0.5), as depicted in Figure 5. The upper diagram illustrates the overall process of our research. Each block represents a subsection in this paper. The lower diagram describes the proposed model for classifying the embryo dataset.

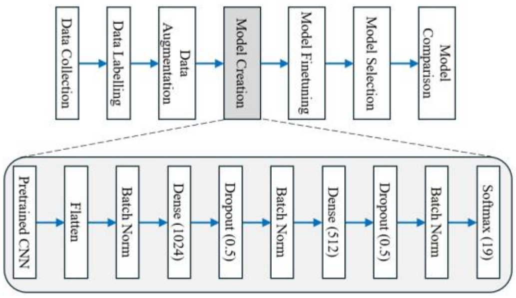

Figure 5. The overall process and the proposed model

Table 1. The number of parameters (x million) of each model

InceptionV3	Xception	VGG16	VGG19	ResNet152	ResNet101	EfficientNet
23.85	22.91	138.35	143.67	60.20	44.59	5.3

The proposed model utilizes the SoftMax activation function at the output layer to categorize the embryo images into 19 classes. Table 1 shows the number of parameters of all candidate models.

Results and Discussion

Model Finetuning and Selection: To identify the most efficient model for embryo grading, we train the proposed model on the training set using each of the seven pretrained CNNs as a backbone. We finetune all weights and biases of the proposed model that are not in the pretrained CNN. The training of each model is terminated at the point where the training loss and validation loss diverge. Due to space limitation, we illustrate samples of finetuning processes of InceptionV3 and Xception on Google Colaboratory with T4 GPU in Figure 6 and Figure 7, respectively.

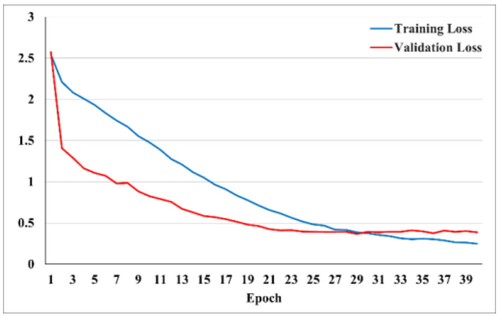

Figure 6. Finetuning for InceptionV3

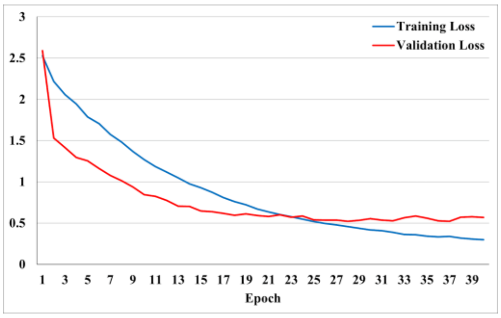

Figure 7. Finetuning for Xception

The performance of 7 models, with different optimizers (Adam and RMSprop), are shown in Figure 8 and Figure 9. Given that the models VGG16, ResNet101, and EfficientNet exhibit considerably lower metrics regarding accuracy, precision, recall, and F1-Score compared to other models, we have decided to exclude them from our candidate list. The remaining four CNN models - InceptionV3, Xception, VGG19, and ResNet152 - have competitively similar performance, with no more than 3% differences in all metrics. However, VGG19 and ResNet152 have substantially larger parameters than InceptionV3 and Xception, as shown in Table 1. These extra parameters increase the risk of overfitting the training data and consume too much space in the GPU's memory. Therefore, we discard them and keep only InceptionV3 and Xception as our candidates. Other hyperparameter settings for all models are given in Table 2.

Model Comparison: To find the best architecture from the remaining candidates, we compare the confusion matrix and the receiver operating characteristic (ROC) curve, including AUC, of both the InceptionV3 and Xception models in Figure 10 - Figure 11 and Figures 13 - Figure 14. Although ResNet152 (with RMSprop) is eliminated from our choice due to its enormous size, we still show its confusion matrix and ROC curve in Figure 12 and Figure 15 to show how well InceptionV3 and Xception compete with the larger model. Both InceptionV3 and Xception correctly predict the 19 classes at a high rate, as the intensity is shown along the diagonal line of the confusion matrices. However, three classes at the bottom right corners of the confusion matrices (Arrest, Early, and Morula) are difficult to predict. They are delicate and share many common characteristics, making it hard to distinguish between them.

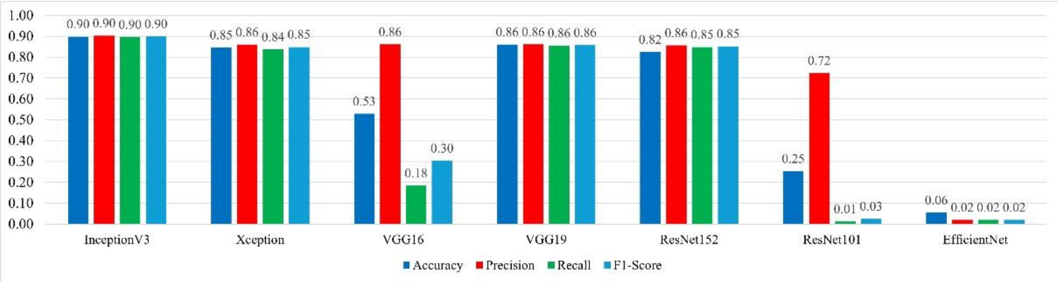

Figure 8. Test set accuracy with Adam optimization

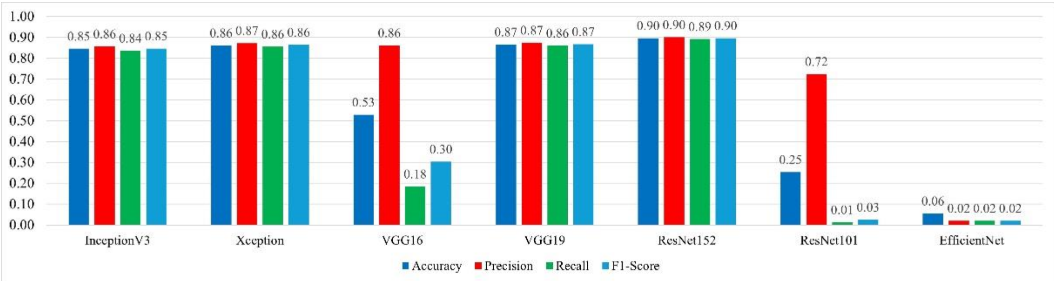

Figure 9. Test set accuracy with RMSprop optimization

Table 2. Hyperparameter settings

	Learning rate	Batch size	Early-stop epochs
InceptionV3	0.001	32	30
Xception	0.001	32	23
VGG16	0.001	32	15
VGG19	0.001	32	13
ResNet152	0.001	32	21
ResNet101	0.001	32	23
EfficientNet	0.001	32	49

It appears that ResNet152 struggles with predicting the Arrested, Early, and Morula stages, much like the confusion experienced by InceptionV3 and Xception. While ResNet152 exhibits slight improvement over the other two models in the remaining classes, its overall performance remains competitive with InceptionV3 and Xception (as shown in the AUCs in Figure 13 - Figure 15). If the ResNet152 model size is not significantly larger than the other two, it can be a potential model.

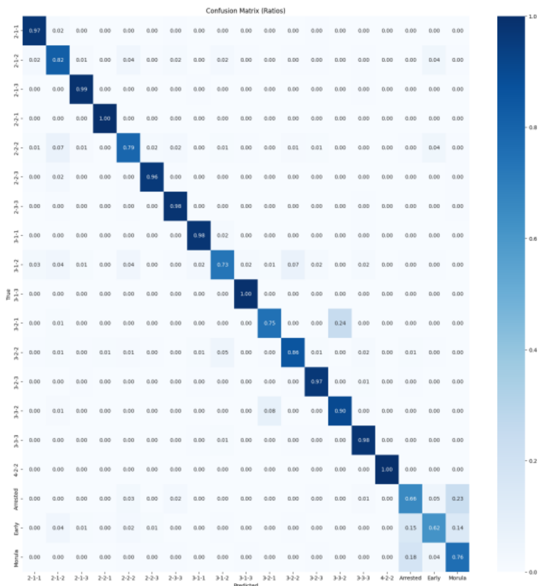

Figure 10. The confusion matrix of InceptionV3

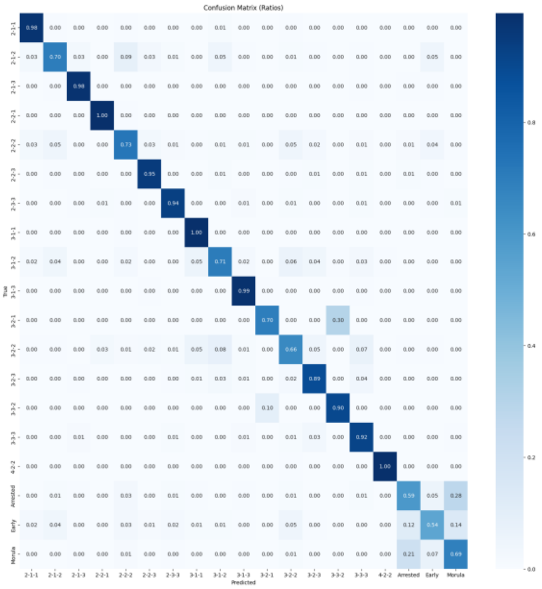

Figure 11. The confusion matrix of Xception

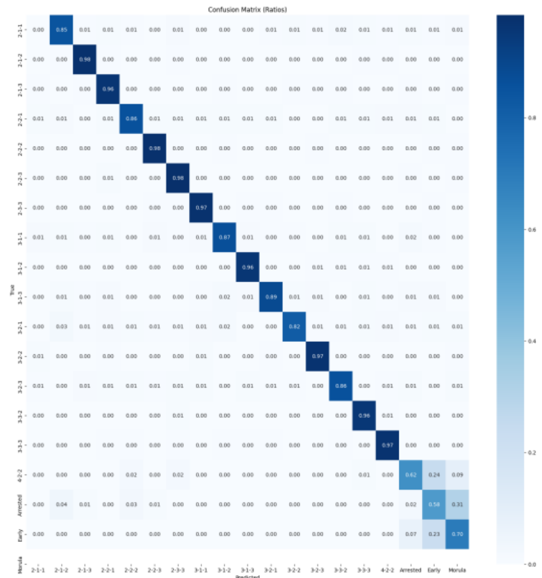

Figure 12. The confusion matrix of ResNet152

As Zeiler and Fergus (2014) introduced, we apply deconvolution to each convolutional layer by generating an input image that maximizes the activation of each filter. This process allows us to visualize how well each filter in CNN responds to the patterns. Figure 16 displays pattern visualizations from the first four filters of all candidates, given an original embryo image on the left. InceptionV3, Xception, and ResNet152 models efficiently extract essential features before transferring them to the dense layer, resulting in superior prediction performance. This experiment allows us to comprehend how better the InceptionV3, Xception, and ResNet152 models capture crucial features from the embryo images compared with other models. However, these filter patterns do not reflect the image region that the model pays attention to.

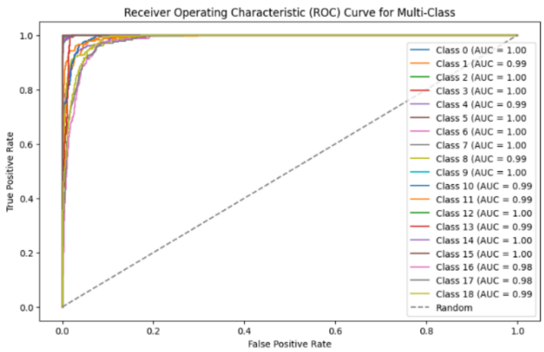

Figure 13. ROC of InceptionV3

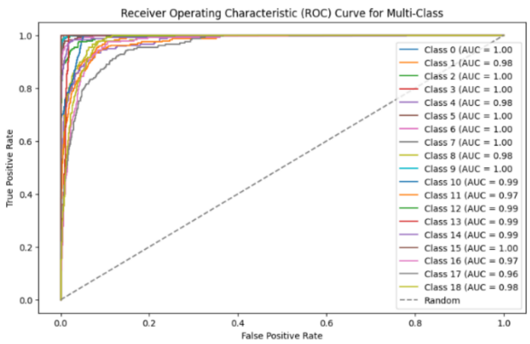

Figure 14. ROC of Xception

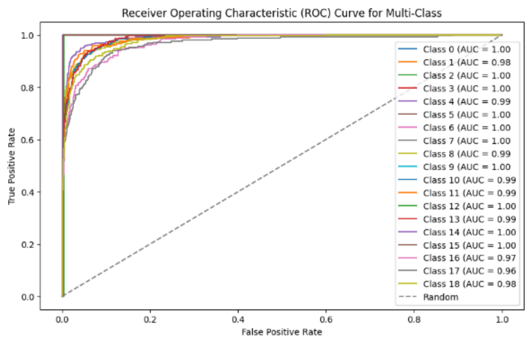

Figure 15. ROC of ResNet152

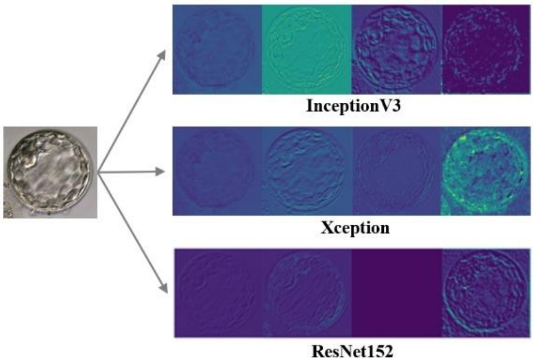

Figure 16. Pattern visualizations

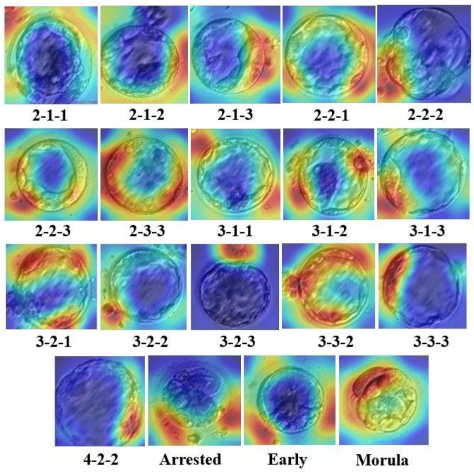

Figure 17. Class activation maps of InceptionV3

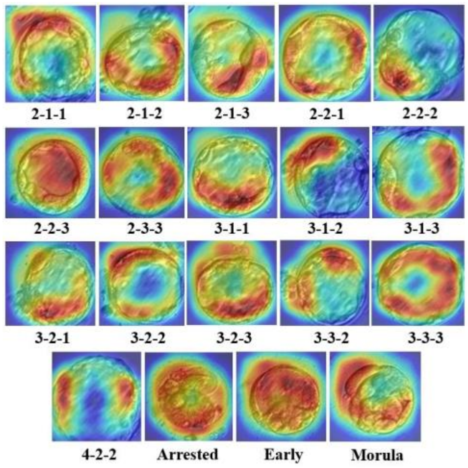

Figure 18. Class activation maps of Xception

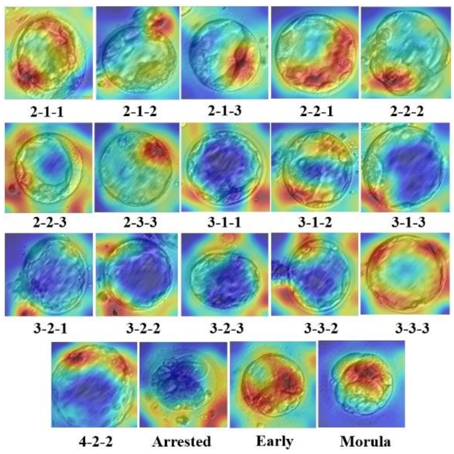

Figure 19. Class activation maps of ResNet152

In the final experiment, we aim to differentiate the performance of the InceptionV3 and Xception models by visualizing a class activation map (CAM) for each class in the embryo dataset. CAM is a graphical representation that maps out the regions within an image that CNN deems most pertinent for recognizing a particular category. It provides a quick visual summary across two dimensions, which helps us identify the most crucial or relevant features captured by each model.

Figures 17, 18 and 19 depict CAMs for the final CNN layers of InceptionV3, Xception, and ResNet152, respectively. Interestingly, InceptionV3 and ResNet152 concentrate on the external region of an embryo. Despite achieving high performance across all metrics, they sometimes focus on the unimportant features outside the embryo. Consequently, they may not accurately predict other unseen embryo images in the future. In contrast, the Xception model concentrates on the important features from the edge to an area inside the embryo. This focus aligns with what an embryologist considers when grading an embryo. As a result, the Xception model has greater potential for accurately predicting unseen embryo images and is our preferred choice for the CNN backbone.

It is worth discussing the characteristics of the image misclassified by most models. We discover that the embryo images from the Arrested, Early, and Morula classes are the most challenging to classify. Figure 20 shows sample images from three classes misclassified by most models. Images from the 3 classes share several characteristics. For instance, the amount of dense tissue inside the cell, the thickness of cell membranes, and the clear edges of internal dense tissue. These common characteristics often confuse most models and inexperienced human justification.

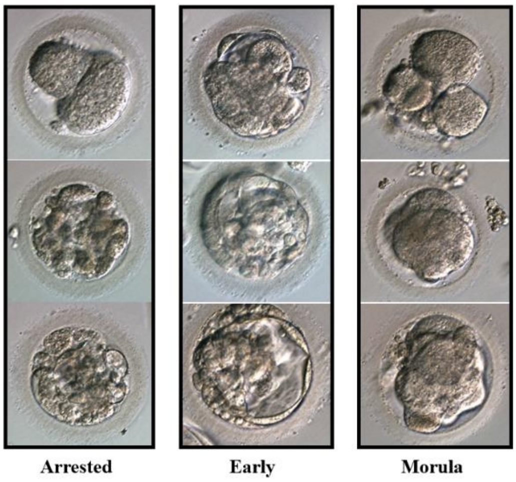

Figure 20. **Samples images misclassified by most models**

To summarize the experiment and model selection, we have identified two compact-sized candidates (InceptionV3 and Xception) as the most efficient models for creating an embryo quality scoring system. Since the model's size can be calculated by multiplying the models' parameter counts in Table 1 with the number of bytes (approximately 4 bytes) per parameter, comparing the parameter count also reflects the size (in bytes) ratio. Therefore, the size ratio of (InceptionV3: Xception: VGG19: ResNet152) is (23.85, 22.91, 143.67, 60.20), which is equivalent to the simplest ratio (1.04: 1.00: 6.27: 2.63). Regarding the model efficiency, both InceptionV3 and Xception demonstrate impressive performance. Their average AUCs (Area Under the Curve) are 0.9947 and 0.9915, respectively. However, based on their CAMs, we recommend the Xception model because it focuses more on the important features inside an embryo and has a high potential of accurately predicting unseen embryo images.

Conclusions

A deep-learning model has been proposed for classifying human embryo grading, utilizing a dataset from the IVF Clinic at Vajira Hospital in Thailand. This dataset is uniquely challenging, encompassing three additional embryo stages exclusive to the Vajira IVF Clinic. The model employs a pretrained Convolutional Neural Network (CNN) for the feature extraction. Among seven pretrained CNNs, InceptionV3 and Xception have exhibited exceptional performance. The model's Area Under the Curve (AUC) is notably high, and its confusion matrix surpasses the results of other models. The experiments conducted on the embryo dataset indicate encouraging results for the proposed model. The class activation map experiment recommends using the Xception model as the feature extractor for the proposed model, given its focus on the crucial features within the embryo. Although InceptionV3 and ResNet152 with RMSprop attain competitive results compared to Xception, both models focus on the wrong region when recognizing the embryo images. Moreover, their large number of parameters demands increased computations during feed-forward and backpropagation steps. Our study has the potential to make significant strides in the IVF field by minimizing human errors and addressing the need for skilled embryologists. It also holds promise for further development into a supportive tool for embryologists in evaluating embryo quality during the IVF process.

References

Arsalan, M., Haider, A., Choi, J., and Park, K.R. (2022). Detecting blastocyst components by artificial intelligence for human embryological analysis to improve success rate of in vitro fertilization. Journal of Personalized Medicine, 12(2):124. https://doi.org/10.3390/jpm12020124

Barnes, J., Brendel, M., Gao, V.R., Rajendran, S., Kim, J., Li, Q., Malmsten, J.E., Sierra, J.T., Zisimopoulos, P., Sigaras, A., Khosravi, P., Meseguer, M., Zhan, Q., Rosenwaks, Z., Elemento, O., Zaninovic, N., and Hajirasouliha, I. (2023). A non-invasive artificial intelligence approach for the prediction of human blastocyst ploidy: a retrospective model development and validation study. Lancet Digit Health, 5(1):e28-e40. https://doi.org/10.1016/S2589-7500(22)00213-8

Basari, I. and Gunawan, D. (2019). Automated detection of human blastocyst quality using convolutional neural network and edge detector. Proceedings of the 2019 1st International Conference on Cybernetics and Intelligent Systems (ICORIS); August 22-23, 2019; Denpasar, Indonesia, p. 181-184.

Berman, A., Anteby, R., Efros, O., Klang, E., and Soffer, S. (2023). Deep learning for embryo evaluation using time-lapse: a systematic review of diagnostic test accuracy. American Journal of Obstetrics & Gynecology, 229(5):490-501. https://doi.org/10.1016/j.ajog.2023.04.027

Berntsen, J., Rimestad, J., Lassen, J.T., Tran, D., and Kragh, M.F. (2022). Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences. PLoS One, 17(2):e0262661. https://doi.org/10.1371/journal.pone.0262661

Chavez-Badiola, A., Flores-Saiffe-Farías, A., Mendizabal-Ruiz, G., Drakeley, A.J., and Cohen, J. (2020). Embryo ranking intelligent classification algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reprod Biomed Online, 41(4):585-593. https://doi.org/10.1016/j.rbmo.2020.07.003

Chéles, D.S., Ferreira, A.S., de Jesus, I.S., Fernandez, E.I., Pinheiro, G.M., Dal Molin, E.A., Alves, W., de Souza, R.C.M., Bori, L., Meseguer, M., Rocha, J.C, and Nogueira, M.F.G. (2022). An image processing protocol to extract variables predictive of human embryo fitness for assisted reproduction. Applied Sciences, 12(7):3,531. https://doi.org/10.3390/app12073531

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI, USA, p. 1800-1807. https://doi.org/10.1109/CVPR.2017.195

Coward, K. and Wells, D. (2013). Textbook of Clinical Embryology. 1st ed. Cambridge University Press, New York, 342p. https://doi.org/10.1017/CBO9781139192736

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning Illustrated. MIT Press, Massachusetts, 775p.

Harun, M.Y., Rahman, M.A., Mellinger, J., Chang, W., Huang, T., Walker, B., Hori, K., Ohta, A.T., Harun, M.Y., Rahman, A., Mellinger, J., Chang, W., Huang, T., Walker, B., Hori, K., and Ohta, A. (2019). Image segmentation of zona-ablated human blastocysts. Proceedings of the IEEE International Conference on Nano/Molecular Medicine and Engineering (NANOMED); November 21-24, 2019; Gwangju, Republic of Korea, p. 208-213. https://doi.org/10.1109/NANOMED49242.2019.9130621

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV, USA, p. 770-778. https://doi.org/10.1109/CVPR.2016.90

Hu, J., Wang, H., Wang, J., Wang, Y., He, F., and Zhang, J. (2021). SA-Net: A scale-attention network for medical image segmentation. PLoS One, 16(4):e0247388. https://doi.org/10.1371/journal.pone.0247388

Kan-Tor, Y., Zabari, N., Erlich, I., Szeskin, A., Amitai, T., Richter, D., Or, Y., Shoham, Z., Hurwitz, A., Har-Vardi, I., Gavish, M., Ben-Meir, A., and Buxboim, A. (2020). Automated evaluation of human embryo blastulation and implantation potential using deep-learning. Advanced Intelligent Systems, 2(2):2000080. https://doi.org/10.1002/aisy.202000080

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Proceedings of the 2012 Advances in Neural Information Processing Systems 25 (NIPS 2012); December 3-8, 2012; Lake Tahoe, NV, USA, p. 84-90.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2,278-2,324. https://doi.org/10.1109/5.726791

Liao, Q., Zhang, Q., Feng, X., Huang, H., Xu, H., Tian, B., Liu, J., Yu, Q., Guo, N., Liu, Q., Huang, B., Ma, D., Ai, J., Xu, S., and Li, K. (2021). Development of deep learning algorithms for predicting blastocyst formation and quality by time-lapse monitoring. communications biology, 4(1):415. https://doi.org/10.1038/s42003-021-01937-1

Liu, H., Zhang, Z., Gu, Y., Dai, C., Shan, G., Song, H., Li, D., Chen, W., Lin, G., and Sun, Y. (2023). Development and evaluation of a live birth prediction model for evaluating human blastocysts from a retrospective study. eLife, 12:e83662. https://doi.org/10.7554/eLife.83662

Loewke, K., Cho, J.H., Brumar, C.D., Maeder-York, P., Barash, O., Malmsten, J.E., Zaninovic, N., Sakkas, D., Miller, K.A., Levy, M., and VerMilyea, M.D. (2022). Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. Fertility and Sterility, 117(3):528-535. https://doi.org/10.1016/j.fertnstert.2021.11.022

Milyea, M.V., Hall, J.M.M., Diakiw, S.M., Johnston, A., Nguyen, T., Perugini, D., Miller, A., Picou, A., Murphy, A.P., and Perugini, M. (2021). Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. Human Reproduction, 35(4):770-784. https://doi.org/10.1093/humrep/deaa013

Pan, S.J. and Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10):1,345-1,359. https://doi.org/10.1109/TKDE.2009.191

Rad, R.M., Saeedi, P., Au, J., and Havelock, J. (2019). BLAST-NET: Semantic segmentation of human blastocyst components via cascaded atrous pyramid and dense progressive upsampling. 2019 IEEE International Conference on Image Processing (ICIP); September 22-25, 2019; Taipei, Taiwan, p.1865-1869. https://doi.org/10.1109/ICIP.2019.8803139

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv, 1409.

Srivastava, R., Greff, K., and Schmidhuber, J. (2015). Highway networks. arXiv, 1505.00387.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV, USA, p. 2,818-2,826. https://doi.org/10.1109/CVPR.2016.308

Tan, M. and Le, Q.V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML 2019); June 9-15, 2019; Long Beach, CA, USA, p. 6105-6114.

Zeiler, M.D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision - ECCV 2014. LNCS, 8,689. https://doi.org/10.1007/978-3-319-10590-1_53