HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Accurate System for Assessment and Selection of Human Embryo After in Vitro Fertilization

Chae Young Kim[1] and Eun Jin Park[#]

[1]Posan High school, Republic of Korea
[#]Advisor

ABSTRACT

In Vitro Fertilization (IVF) is a complex medical procedure designed to assist individuals in overcoming infertility by facilitating the union of sperm and egg outside the human body. Not all embryos created during the IVF process have the same potential for successful implantation and development. Embryo grading allows embryologists to assess the quality of embryos based on specific criteria, such as cell division, symmetry, and the presence of fragmentation. By assigning grades, they can identify the most viable embryos with the highest likelihood of successful implantation. The conventional methods for embryo grading are time-consuming, labor-intensive, and prone to errors due to their subjective nature. Therefore, there is a pressing need for the development of an accurate and automated grading system to address these challenges. In this study, I propose a human embryo grading system based on machine learning. The proposed system utilizes embryo images as inputs and produces probabilities to predict the success or failure of each embryo. I utilize a convolutional neural network to develop the grading model and introduce a novel hierarchical feature extraction approach to enhance accuracy. Through extensive experiments, I demonstrate that the proposed method surpasses previous approaches with a significant performance margin.

Introduction

In recent decades, In Vitro Fertilization (IVF) has emerged as a groundbreaking medical procedure to offer hope to individuals and couples grappling with infertility. This assisted reproductive technology involves the union of sperm and egg outside the human body which provides a pathway to conception that was once elusive. Not all embryos created during the IVF process exhibit the same potential for successful outcomes. The success of IVF is contingent upon the identification and selection of high-quality embryos. This is a critical step that influences the likelihood of successful implantation and subsequent development.

Traditional methods of embryo grading, which evaluate criteria such as cell division, symmetry, and the presence of fragmentation, have long been the cornerstone of assessing embryo quality. These methods are fraught with challenges and stemming from their subjectivity, labor-intensive nature, and susceptibility to errors.

Addressing the limitations of current embryo grading methods is not merely an academic pursuit but a necessity in the field of reproductive medicine. The accurate identification of high-quality embryos is important in maximizing the success rates of IVF which can minimize costs and alleviate the emotional and financial burden on individuals and couples undergoing fertility treatments.

The main objective of this research is to develop and validate an accurate and automated grading system for human embryos after in vitro fertilization. The machine learning-based system aims to surpass the limitations of conventional grading methods which can provide a more efficient and reliable way of assessing embryo quality. I employ a convolutional neural network for the development of the embryo grading system. I

also introduce a novel hierarchical feature extraction approach to improve the system's accuracy. Through various experiments, it has been demonstrated that the proposed approach outperforms previous methods.

The following chapters are organized as follows: Chapter 2 provides background knowledge to facilitate a better understanding of the proposed method. Chapter 3 explains the detailed process of the proposed method, while Chapter 4 analyzes its performance through extensive experiments. Finally, Chapter 5 summarizes the research.

Related Work

In Vitro Fertilization

In vitro fertilization (IVF) is a complex and assisted reproductive technology procedure designed to help individuals or couples overcome infertility and achieve pregnancy. The term "in vitro" refers to the process taking place outside the body, specifically in a laboratory setting. The process typically begins with the administration of fertility medications to stimulate the ovaries to produce multiple eggs. This is important because multiple eggs increase the chances of successful fertilization.Once the eggs reach maturity, a minor surgical procedure known as egg retrieval is performed. A thin needle is inserted through the vaginal wall and into the ovaries to extract the eggs. On the same day as the egg retrieval, a sperm sample is collected from the male partner or a sperm donor. The sperm is then processed and prepared for fertilization.

The collected eggs and sperm are combined in a laboratory dish for fertilization. The fertilized eggs, now called embryos, are monitored for a few days as they undergo cell division. The embryos are cultured in a special incubator, and their development is carefully observed. In some cases, pre-implantation genetic testing may be performed to assess the genetic health of the embryos.

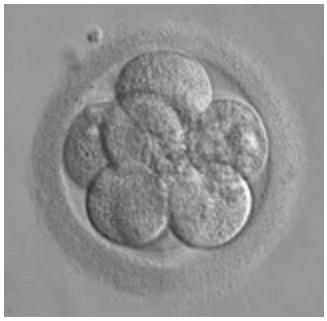

Figure 1. Human embryo image after in vitro fertilization

After a few days of culture, one or more healthy embryos are selected for transfer to the woman's uterus. This is typically done through the cervix using a thin catheter. If the embryo successfully implants into the uterine lining, pregnancy occurs. A pregnancy test is usually conducted about 10-14 days after the embryo transfer to determine if implantation has occurred.

Embryo grading helps identify embryos with the highest potential for successful implantation and subsequent development into a healthy pregnancy. Not all embryos created during the IVF process have the same likelihood of implanting in the uterus and progressing to a full-term pregnancy. Traditional methods of embryo grading involve visual inspection by embryologists who evaluate various characteristics, such as cell division patterns, symmetry, and the presence of fragmentation. These assessments help identify embryos with the highest potential for successful implantation and subsequent development into a healthy pregnancy.

To address these limitations, the integration of machine learning techniques in embryo grading has emerged as a promising solution. Machine learning algorithms, particularly convolutional neural networks, can be trained to analyze the embryo images and learn visual patterns associated with successful implantation.

Image Classification

Image classification is one of many computer vision tasks that assign predefined labels or categories to images based on their visual content. Image classification systems are frequently built using convolutional neural networks known for their consistent and high-performance capabilities across various computer vision challenges.

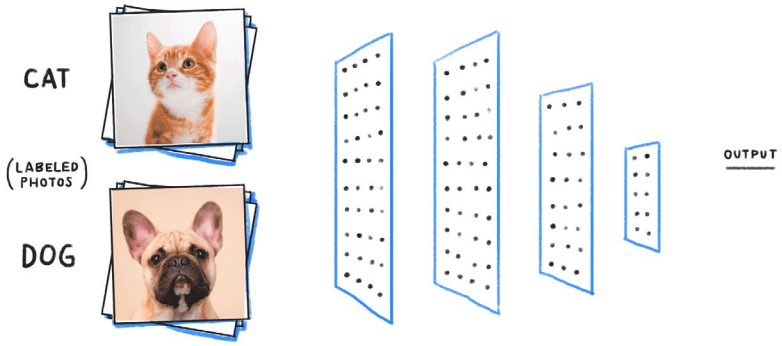

Figure 2. Architecture of the proposed lunar landscape segmentation method.

Figure 2 illustrates the architecture of an image classification system employing a convolutional neural network. The network takes images as input and extracts visual characteristics or feature maps that encapsulate crucial information for categorization. Then the neural network utilizes these feature maps to classify the final results. There are several well-established convolutional neural networks designed for the development of image classification systems such as VGG (Simonyan et al. 2014), ResNet (He et al. 2016), and HRNet (Wang et al. 2020).

In this research, I consider the human embryo grading problem as image classification. The system utilizes embryo images as input and generates binary classification outcomes distinguishing between success and failure. More detailed information about the proposed system is provided in Chapter 3.

Proposed Method

In this chapter, I present the detailed architecture and methodology of the proposed human embryo grading system.

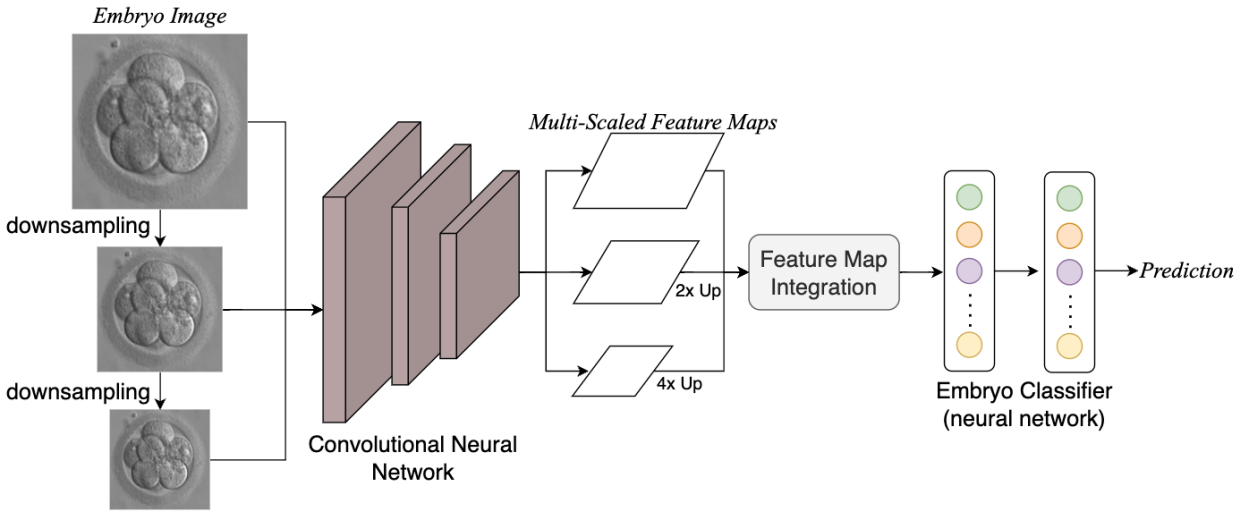

Figure 3. Architecture of the proposed human embryo grading system

Figure 3 provides an overview of the proposed method's overall architecture. The main idea of the proposed approach is the generation of pyramid images at three distinct levels from the input embryo image. These multi-scaled images introduce variations in dimensions which offer a comprehensive view of the embryo at different levels of granularity. This pyramid structure contributes to the model's capacity to capture both global and local features within the input image.

The multi-scaled images are then fed into a convolutional neural network which is well-suited for image analysis tasks. The convolutional neural network processes these images and generates feature maps that encapsulate rich and detailed features extracted from the multi-scaled representations. This approach proves advantageous, as the model gains the ability to discern and prioritize relevant features across different scales, enhancing its overall discriminative power. The feature maps obtained from the convolutional neural network encapsulate a rich representation which captures subtle details crucial for accurate embryo classification. This multi-scale feature extraction approach brings about an improvement in accuracy. This will be further investigated in Chapter 4.

To train the proposed network, I use the widely-used cross-entropy loss function which is a popular choice for training machine learning-based classification models. The cross-entropy loss function is calculated as Equation 1.

Equation 1: Cross-entropy loss function

$$L = -log_e \hat{y}$$

To develop the proposed system, I employed ResNet-50 which is a convolutional neural network architecture known for its commendable performance across various computer vision problems. While experimenting with different convolutional neural network architectures, ResNet-50 consistently delivered the most accurate results. For the training parameters, the model is trained for 80 epochs with a learning rate set to 0.0001. The batch size was configured to be 128, and all image patches were resized uniformly to dimensions of 512x512.

Experimental Results

Human Embryo Dataset

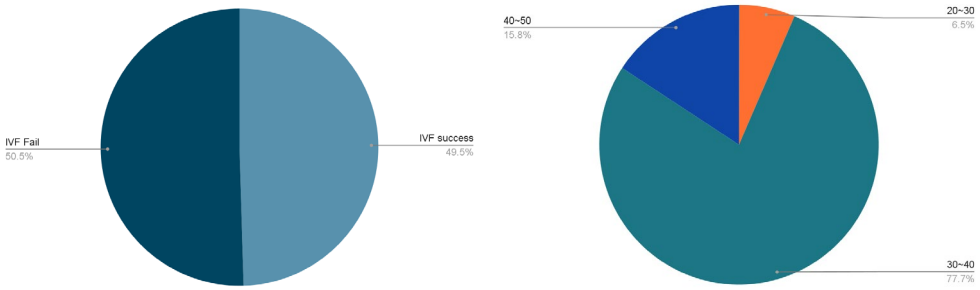

Figure 4. Dataset distribution (AI-Hub 2023) (left: distribution of IVF success of failure and right: distribution of patient ages)

In this chapter, I provide a comprehensive explanation of the dataset utilized in this research study. The dataset comprises a total of 14,989 embryo image samples and each categorized into one of two classes: IVF success or IVF failure. The dataset exhibits a balanced distribution between IVF success and failure samples, with 49.5% of the samples representing successful cases and 50.5% representing failures.

An analysis of the ages of individuals contributing to the dataset reveals a varied demographic profile. The majority of the participants, constituting 77.7% of the dataset, fall within the age range of 30 to 40 years. Additionally, 15.6% of participants are between 40 to 50 years old, and 6.5% are in the age group of 20 to 30 years.

Performance Evaluation

To assess the performance of the proposed method, I evaluated it using four commonly employed inference metrics in classification problems: accuracy, precision, recall, and F1-score. For the comparison methods, I selected several established methods such as VGG19 (Simonyan et al. 2014), MobileNetV2 (Sandler et al. 2018), Xception (Fran et al. 2017), HRNet-w32 (Wang et al. 2020) , ResNext-50 (Xie et al. 2017), and ResNet-50 (He et al. 2016) , known for their competitive performance across various computer vision tasks. These models were trained using identical parameters to ensure a fair and unbiased comparison.

Table 1. Inference metric comparison

Architecture	Accuracy	Precision	Recall	F1-Score
VGG19 (Simonyan et al. 2014)	0.9026 (± 0.0014)	0.8973 (± 0.0009)	0.8964 (± 0.0012)	0.9024 (± 0.0010)
MobileNetV2 (Sandler et al. 2018)	0.9034 (± 0.0006)	0.9088 (± 0.0008)	0.8995 (± 0.0010)	0.9041 (± 0.0009)
Xception (Fran et al. 2017)	0.9074 (± 0.0011)	0.9112 (± 0.0013)	0.8983 (± 0.0009)	0.9075 (± 0.0011)
HRNet-w32 (Wang et al. 2020)	0.9121 (± 0.0011)	0.9101 (± 0.0010)	0.9086 (± 0.0010)	0.9145 (± 0.0014)
ResNext-50	0.9164	0.9213	0.9119	0.9176

(Xie et al. 2017)	(± 0.0013)	(± 0.0011)	(± 0.0014)	(± 0.0012)
Resnet-50 (He et al. 2016)	0.9205 (± 0.0010)	0.9267 (± 0.0009)	0.9233 (± 0.0007)	0.9199 (± 0.0008)
Proposed Method	**0.9589** (± 0.0008)	**0.9651** (± 0.0010)	**0.9538** (± 0.0009)	**0.9594** (± 0.0010)

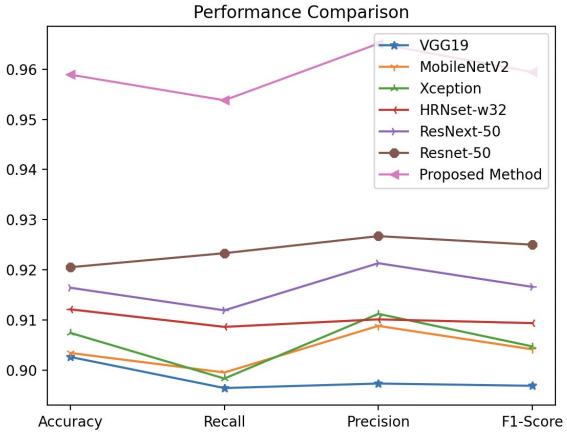

Figure 5. Inference metric comparison

Table 1 provides a comprehensive performance comparison between the proposed method and state-of-the-art classification approaches. Notably, VGG19, MobileNetV2, and Xception exhibit relatively lower performance due to their shallow network architectures which limit their ability to generate rich feature maps. Conversely, ResNext-50, ResNet-50, and HRNet-w32 demonstrate superior accuracy which benefits from their deeper layers that facilitate the extraction of more comprehensive and rich features. The proposed method surpasses all state-of-the-art methods by a significant margin. This comparison clearly demonstrates that the proposed approach enhances accuracy.

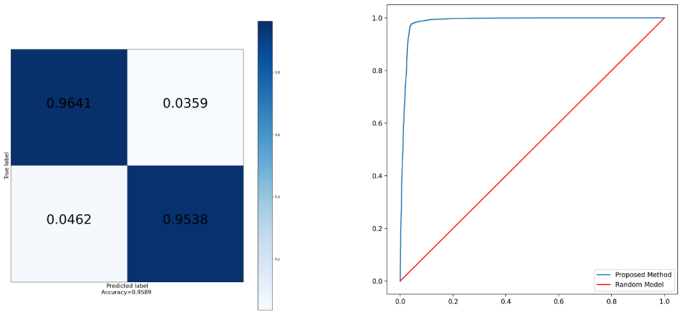

Figure 6. Dataset distribution (AI-Hub 2023)

Figure 6 illustrates the confusion matrix and ROC curve of the proposed method. The results showcase the robust performance of the proposed method in accurately classifying both negative and positive categories.

To investigate deeper into the proposed approach, I conducted an ablation study. Initially, I trained regular convolutional neural networks and I implemented the proposed hierarchical multi-scale feature extraction approach to each network. Both sets of networks were trained using identical parameters to ensure a fair comparison. The accuracy was then measured to evaluate and compare the results.

Table 2. Ablation study result

Architecture	Accuracy (Regular CNN Architecture)	Accuracy (Proposed Method)
VGG19	0.9026 (± 0.0014)	0.9328 (± 0.0008)
MobileNetV2	0.9034 (± 0.0006)	0.9305 (± 0.0011)
Xception	0.9074 (± 0.0011)	0.9398 (± 0.0007)
HRNet-w32	0.9121 (± 0.0011)	0.9435 (± 0.0015)
ResNext-50	0.9164 (± 0.0013)	0.9502 (± 0.0010)
Resnet-50	0.9205 (± 0.0010)	0.9589 (± 0.0008)

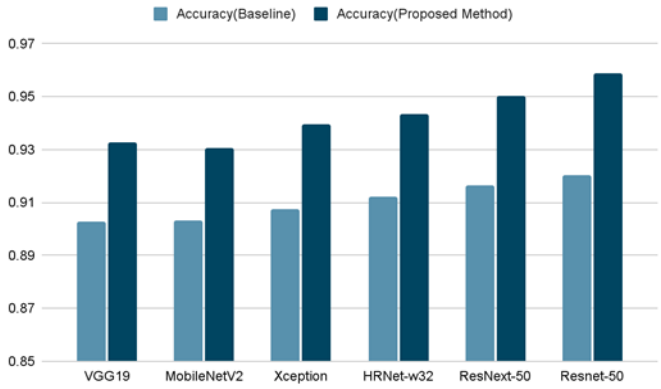

Figure 7: Ablation study result

Table 2 and Figure 7 provide a summary of the ablation study. The results demonstrate a consistent increase in accuracy when applying the proposed approach. Notably, for deeper network architectures like ResNet-50 or ResNext-50, the improvement in accuracy is more pronounced compared to others. This outcome substantiates the robust and consistent efficacy of the proposed approach in enhancing accuracy.

Finally, I conducted a data augmentation study to further enhance accuracy. Table 3 presents the specific augmentations applied and their corresponding results.

Table 3. Data augmentation study

Architecture	Accuracy (Resnet-50)
Baseline	0.9589
Color Jitter	0.9375 (-0.0214)
Gaussian Noise	0.9382 (-0.0207)
Grayscale	0.9547 (-0.0042)
Horizontal Flip	0.9601 (+0.0012)

Random Perspective	0.9597 (+0.0008)
Rotation	0.9612 (+0.0023)
Horizontal Flip + Random Perspective + Rotation	**0.9637 (+0.0048)**

The augmentation techniques that alter pixel intensity, such as color jitter, gaussian noise, or grayscale adjustments, did not contribute to increased accuracy. Conversely, those focusing on changing image geometry features proved effective in enhancing accuracy.

Conclusion

In this study, I have presented a comprehensive investigation into the development and evaluation of a novel human embryo grading system for in vitro fertilization (IVF). The proposed approach leverages advanced machine learning techniques, specifically a convolutional neural network for accurate and automated embryo classification. The experiments and analyses revealed promising results which showcased the superiority of the proposed method over established state-of-the-art models such as VGG19, MobileNetV2, Xception, HRNet-w32, ResNext-50, and ResNet-50. The proposed system consistently demonstrated robust performance. It excels in scenarios involving deeper network architectures. Exploring data augmentation strategies aimed asst refining accuracy revealed nuanced results. While pixel intensity changes through techniques like color jitter, gaussian noise, or grayscale adjustments did not yield significant improvements, augmentations focusing on image geometry features exhibited positive outcomes. In the future, I intend to deploy the proposed system in real-world scenarios to aid embryologists in their assessments.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

AI-Hub. (2023, Apr 1). *"Infertility treatment embryo image data"*: AI-Hub. https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=ty&dataSetSn=71367

Fran, C. (2017). Deep learning with depth wise separable convolutions. In IEEE conference on computer vision and pattern recognition (CVPR). https://doi.org/10.48550/arXiv.1610.02357

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). https://doi.org/10.48550/arXiv.1512.03385

Kaggle. (2022, Mar 26). "Image Classification using CNN": Kaggle. https://www.kaggle.com/code/arbazkhan971/image-classification-using-cnn-94-accuracy

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520). https://doi.org/10.48550/arXiv.1801.04381

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(10), 3349-3364. https://doi.org/10.48550/arXiv.1908.07919

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).
https://doi.org/10.48550/arXiv.1611.05431