Embryo Grading after In Vitro Fertilization using YOLO

Dewi Ananta Hakim[a1], Ade Jamal[a2], Anto Satriyo Nugroho[b3], Ali Akbar Septiandri[a4], Budi Wiweko[c5]

[a]Faculty of Science and Technology, University Al Azhar Indonesia, Indonesia
[1]dewi.ananta@if.uai.ac.id
[2]adja@uai.ac.id (Corresponding author)

[b]National Research and Innovation Agency, Indonesia
[3]anto.satriyo.nugroho@brin.go.id

[c]Indonesian Medical Education and Research Institution (IMERI), Faculty of Medicine, Indonesia University, Indonesia

Abstract

In vitro fertilization is an implementation of Assistive Reproductive Technology. This technology can produce embryos outside the mother's womb by manipulating gametes outside the human body. The success rate of in vitro fertilization is the selection of good-grading embryos. In this study, the authors used Yolo Version 3 to perform object detection objectively by introducing grades for each embryo image. The author uses an embryo image sourced from the Indonesian Medical Education and Research Institute with information on the quality of the embryo. In this study, the author separated the data into two schemes. The first scheme separates data into training data of 70%, 15% validation data, and 15% for testing data. The second scheme uses a Stratified K-Fold Cross-Validation with a fold value =3. In training, the writer configures the values of Max Batches=6000, Steps=4800,5400, Batch=64, and Subdivision=16 by doing image augmentation (saturation=1.5, exposure=1.5, hue=0.1, jitter=0.3, random=1). For each of the obtained mAP (Mean Average Precision) values for data separation schemes, one is 100.00% in the 6000th iteration, while for the two-data separation scheme, the highest mAP is 97.33%.% in the fold=3 and 5000th iteration. It means that both separation schemes are sufficient in terms of mAP.

Keywords: *In Vitro Fertilization, Image Processing, Object Detection, Computer Vision*

1. Introduction

The rapid development of technology certainly affects everyone to get fast and adequate information. With this change, object detection technology is often used in the medical field, including the detection of diseases, genetically engineered viruses, and other medical activities. Among many medical activities, there is a medical activity called Assistive Reproductive Technology (ART), a procedure for handling the combination of egg cells with sperm in the laboratory. These producing embryos, called In Vitro Fertilization (IVF) with embryo quality, are classified as Grades 1, 2, and 3. One way to identify these grades is to look visually at the shape of the embryo image captured using a laboratory microscope [1].

An alternative method has been studied that uses the image of the embryo to perform tasks from computer vision, including image classification. Septiandri used the fast.ai library by implementing several pre-trained convolution neural networks to classify embryos on the 3rd-day embryo image and cropping the data manually [2]. The image cropping results are divided into three grades based on the Veeck classification criteria. Grade 1 to grade 3 with the transferred learning model for feature extraction methods used includes Residual network with different types of depth (ResNet 18, ResNet34, ResNet50, and ResNet101) [3], Xception [4], MobileNetV2 [5], and densely connected convolutional networks [6]. This training aims to compare and find the highest learning rate. This study resulted in the highest accuracy value in the ResNet50 model, which was 91.79% for the 3rd-day embryo image [2].

In 2018, the Yolo (You Only Look Once) model presented a version update, Yolo Version 3. Yolo version 3 was here to improve upon previous versions of Yolo (Yolo version 2) [7]. Yolo version 3 is a model that simultaneously performs object recognition and regression with object detection output using a bounding box. Yolo Version 3 will use a multilabel classification approach for each bounding box [8]. Redmon's research [7] stated that Yolo version 3 was faster in Mean Average Precision (mAP) and Intersection over Union (IoU) values than other detection methods. Several previous studies used the Yolo version 3 model on an image: Norling conducted the study of the plant Silver Birch (Betula pendula) and Pinus Sylvestris using Yolo version 3. The researcher got an mAP value of 0.99 with data separation, 50% for training data, and 50% for data validation [9]. As a result, despite newer versions being available at the time this article is being written, we advocated using Yolo Version 3 for embryo grading in light of the positive outcomes from work previously discussed.

Further research was carried out by Han [10]. In this study, the dataset was obtained from the First Affiliated Hospital of Fujian Medical, foot care staff at the endocrinology department of the hospital. Special cameras in political studies total 2,688 data samples, all based on Asian skin color. The overall data is divided into four categories: grade 1, grade 2, grade 3, grade 4, and grade 5. Before the modeling was implemented in the dataset, the data was sized equations and converted into Data Annotation form as a preprocessing stage. The model was implemented based on the data preprocessed with Yolo V3. Then the performance calculation of the detection model was made using stacking tricks measurements on the dataset and compacted values. It gains above 1.36%; this needs to be underlined even though it reached 87% in this experiment. Several categories were degraded, such as grade 1 and grade 2. Similarity can lead to misjudgments or missing sections [10].

From the reference of previous research, the author focuses on how accurate the Yolo version 3 algorithm is for In Vitro Fertilization (IVF) embryo images and how the Yolo v3 model presentation results in recognizing each grade in In Vitro Fertilization (IVF) embryos. The limitation of this study is the data obtained from IMERI FKUI, which consists of Grade 1, Grade 2, and Grade 3, with the final output being a bounding box with a recognized embryo grade label.

2. Research Methods

2.1. Dataset

The dataset received from the laboratory of the Indonesia Medical Education and Research Institute (IMERI, FKUI) contains 325 images. Each image has various grades of embryos, indicating the embryo quality [1,2]. Of the 325 images identified:

Table 1. Data Distribution Based on Grade

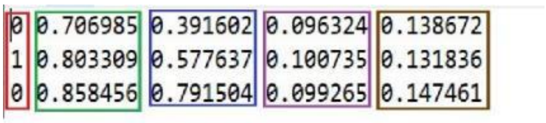
Class	Annotation	Total
Grade 1	0	169
Grade 2	1	342
Grade 3	2	47

2.2. In Vitro Fertilization (IVF) Embryo Description

The In Vitro Fertilization process outputs embryos that can be seen and captured using a microscope. Each embryo produced will represent the type or quality of the resulting embryo. In this study, we used In Vitro Fertilization (IVF) Embryo data from Grades 1, Grade 2, and Grade 3. Grade 1 is said to have the best quality embryo compared to other grades. Grade 2 is categorized as good enough quality because it still has implantation capacity, while Grade 2 and 3 is an embryo with a lower implantation rate than other grades [11].

2.3. Implementation Yolo V3

Yolo (You Only Look Once) is one of the algorithms for object detection tasks assisted by multi-class classification and regression algorithms to determine the labeling of the detected objects. As the name implies, Yolo V3 will predict objects at three scales, where feature extraction will be carried out at each scale. The Yolo model has a specific format for labeling data called Yolo Format. Where each labeling result will produce a value of:



```
0 0.706985 0.391602 0.096324 0.138672
1 0.803309 0.577637 0.100735 0.131836
0 0.858456 0.791504 0.099265 0.147461
```

Figure 1. Labeling Yolo Format, Represent: Annotation of class (1st column), x_center (2nd column), y_center (3rd column), Width(4th column), and Height (5th column)

Where the value is obtained from the following equation [7].

$$x\_center = \frac{\left(\frac{x\_\min + width_1}{2}\right)}{width_2} \quad (1)$$

$$y\_center = \frac{\left(\frac{y\_\min + height_1}{2}\right)}{height_2} \quad (2)$$

$$H = \frac{height_1}{height_2} \quad (3)$$

$$W = \frac{width_1}{width_2} \quad (4)$$

Description:
H = normalization of height
W = normalization of width
$width_2$ = overall image width
$width_1$ = bounding box width
$height_2$ = overall image height
$height_1$ = bounding box height
x_center= object Center X
y_center= object Center Y

In this study, the model training scheme for testing can be seen in the following figure:

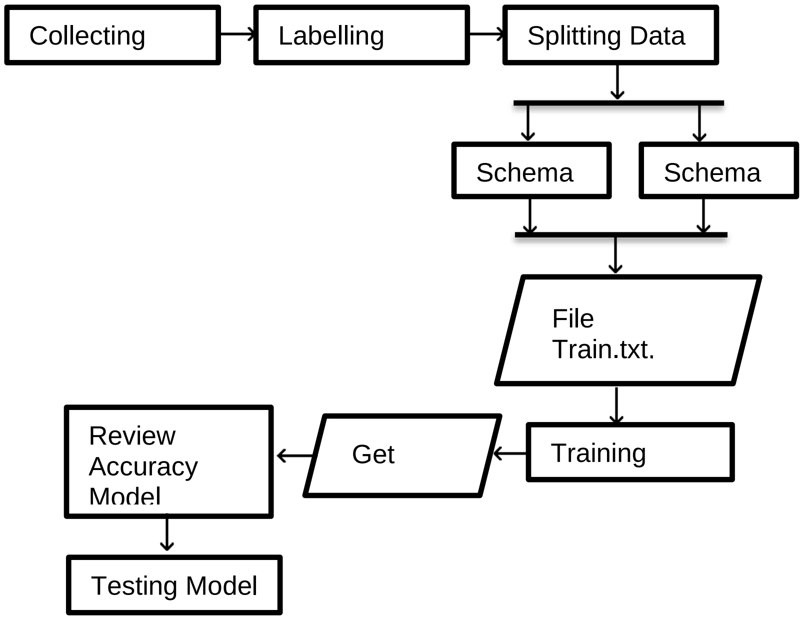


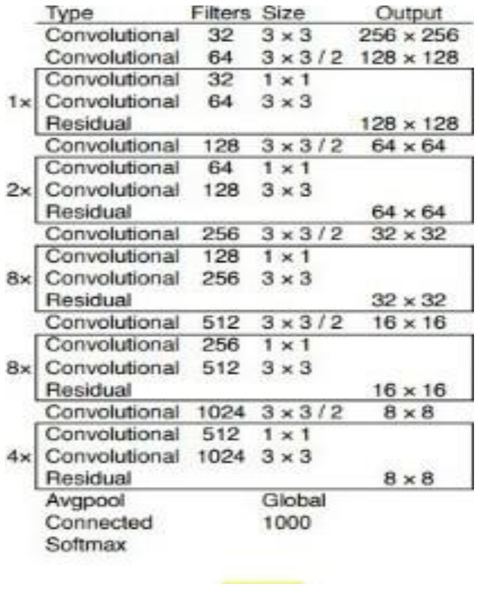
Figure 2. Methodology of the Training Process and Model Testing

2.3.1. Splitting Data

In this study, we experiment with 2 data separation schemes. First, data will be separated into 70% training data, 15% test data, and 15% validation data. Second, data will be separated using Stratified K-Fold Cross-Validation. Stratified K-Fold Cross Validation works to separate the data into training data and test data that refers to the Y variable as a control. With the target data, separation is the variable Y as the sampling class. So it can maintain the percentage of samples for each class. The value of K = 3 is determined based on the references of previous studies to perform the task of detecting objects with multi-class.

2.3.2. Darknet-53

Darknet-53 is the framework that forms Yolo in conducting model training. Built with C language and CUDA framework, it has 53 *convolutional layers* with few floating-point operations and higher speed [12], as shown in Figure 3.



	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 3. Architecture of Darknet-53

2.3.3. Model Performance Configuration

In this architectural study, we use parameter configurations and hyperparameters that can be customized where:

Table 2. Schema Configurasi Yolo Version 3 for Training

Parameter	Value
Batch	64
Max Batches	6000
Subdivision	16
Learning Rate	0.01
Filter	24
Steps	4800,5400
Saturation	1.5
Exposure	1.5
Hue	0.1
Jitter	0.3
Random	1

For Momentum values, learning rate, Batch, and Subdivision are set using practitioner rule values. That rule is a way to get good model accuracy. The rule of practice is to replace the values used in the same problem or find the best value by experimenting. Every value will be executed in the black box process [13].

2.3.4. Model Performance Parameters

The performance of the Yolo V3 model can be seen from the Mean Average Precision (mAP) and Intersection Over Union (IoU). The mAP is a standardization and accuracy matrix of object detection models for all existing classes where the value of mAP is 0 to 1. This value is obtained from the calculation between precision and recall [14].

$$precision = \frac{TP}{TP + FP} \quad (5)$$

$$recall = \frac{TP}{TP + FN} \quad (6)$$

Description:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The mAP value calculated by taking the average Average Precision (AP) from all classes, can be seen in the following equation:

$$AP = \frac{1}{11} \sum_{r \in \{0.0,\ldots,1.0\}} P_{interp} \quad (7)$$

$$P_{interp}(r) = \max_{\tilde{r} \geq r} p(\tilde{r}) \quad (8)$$

With r representing the set of recall values, p(r) is the precision value at a particular recall point. At the end of the detection, Yolo v3 displays a bounding box on the detected object, where this value represents how likely the box contains objects and how accurate the resulting bounding box is [15].

$$\text{box confidence score} = P_r(object) \times IoU \quad (9)$$

2.3.5. Testing Model

The testing model is used to see how well the model is built for detecting the test data. In testing the model, we conducted a study to do the test with the following scheme:

Table 3. Model Testing Experiments

Data Separation Scheme 1 and 2		
Original Image	Darkness (*gamma*=1.5)	Brightness (*gamma*=0.5)
Non-maximum suppression and *Without Non-maximum suppression*		

3. Result and Discussion

3.1. Training Model1

By using a GPU that has a capacity of 32 GB of RAM, the training provides graphics like the following:

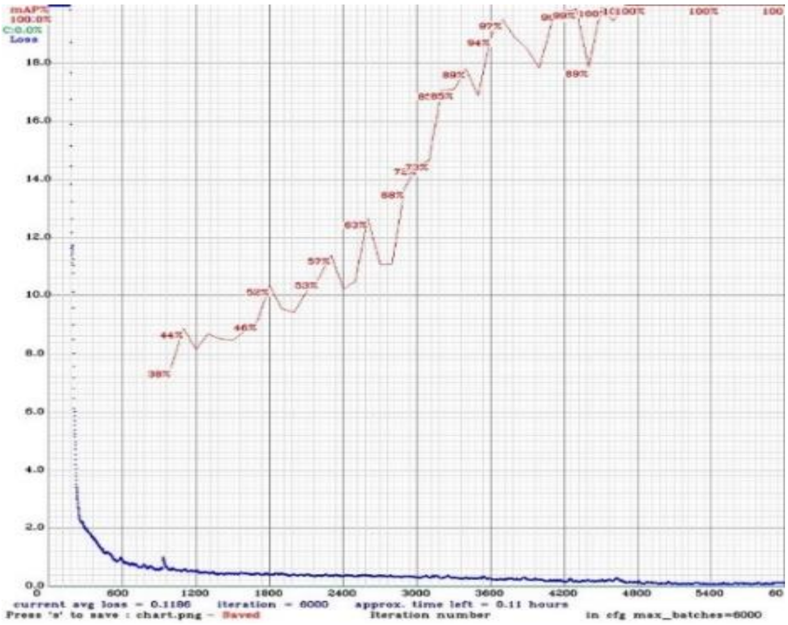


Figure 4. Curve mAP Data Separation Schema 1

Based on figure 4, the training model, by separating the data from the first scheme (training data, test data, and data validation), the average loss value from the first scheme is 0.118 with an average IoU of 88.68%. Note that the red curve (the upper one) denotes mAP values, and the blue (the bottom) denotes the average loss.

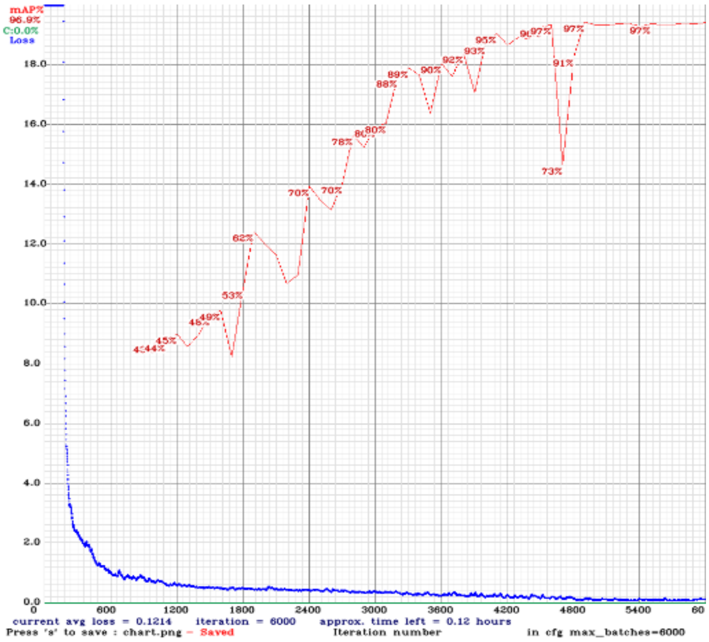


Figure 5. Curve mAP Data Separation Schema 2 Fold-1

Based on Fig 5 and a similar figure for fold 2 and 3, the model using the second data separation scheme (stratified k-fold cross-validation) with K-Fold=3 obtained that each loss value of K=1 is 0.121, K=2 is 0.143, and K=3 is 0.126.

Figures 4 and 5 show a graphical representation of the mAP validation (red dotted line) and training loss (blue dotted line) from the results of the model training process with a total of 6000 iterations. Based on the new darknet [7] documentation, the loss value obtained from the calculation results on the training data and mAP calculations uses validation data defined in the labeled data file.

3.2. Testing Model

Tests were carried out to see the model's success in detecting embryo objects using previously separated test data. In this test, the author will experiment with the addition of 2 different data separation schemes. The author wants to compare the original image and the image illuminated and darkened by non-maximum suppression. This test uses the weight of the largest mAP value from the results of 2 data separation schemes. The first schema data separation uses the results of the 6000th iteration, and the second schema data separation uses the results of the 5000th iteration and fold =3 With each of them applied, Threshold=0.5. As seen in the following pictures

3.2.1. Testing with First Schema of Data Splitting (Testing Image 1)

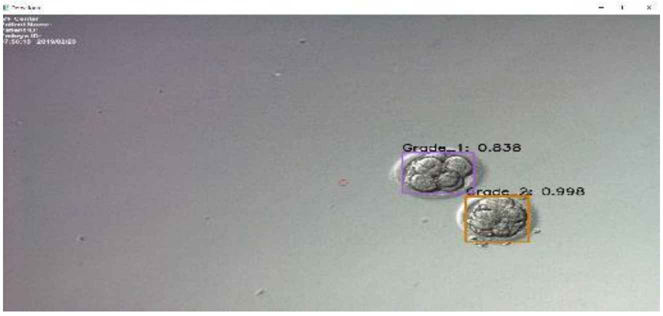


Figure 6. Original Image with Non-Maximum Suppression shows one bounding box Grade 1 and 1 bounding box Grade 2

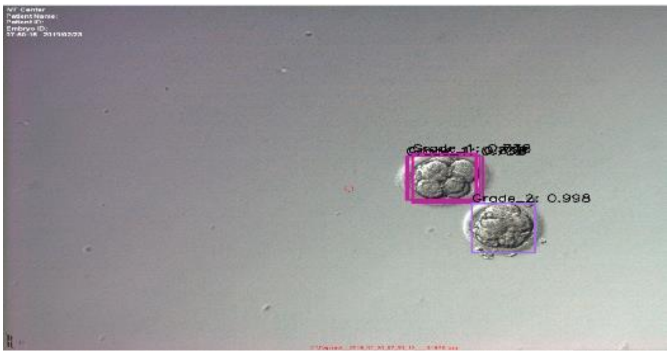


Figure 7. Original Image without Non-Maximum Suppression shows four bounding boxes Grade 1 and 1 bounding box Grade 2

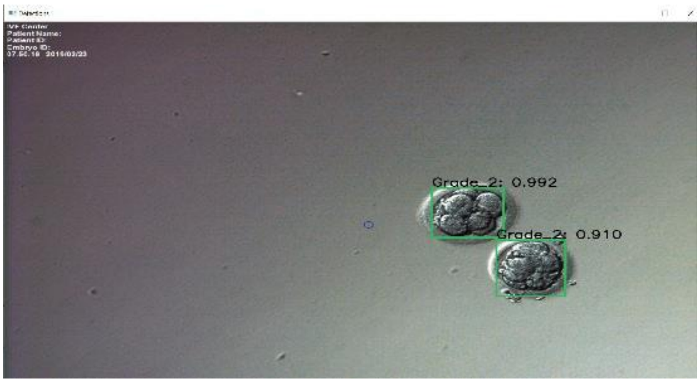


Figure 8. Original Image with Non-Maximum Suppression shows two bounding boxes Grade 2

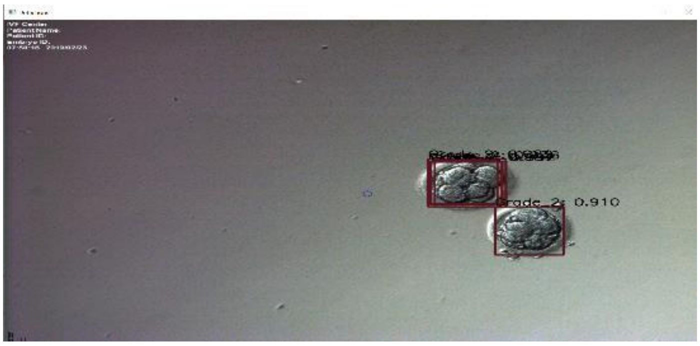


Figure 9. Original Image Without Non-Maximum Suppression shows four bounding boxes Grade 1 and 2 bounding boxes Grade 2

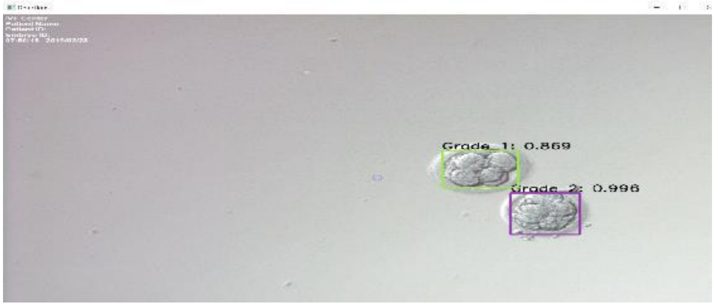


Figure 10. Brightness Image with Non-Maximum Suppression shows one bounding box Grade 1 and 1 bounding box Grade 2

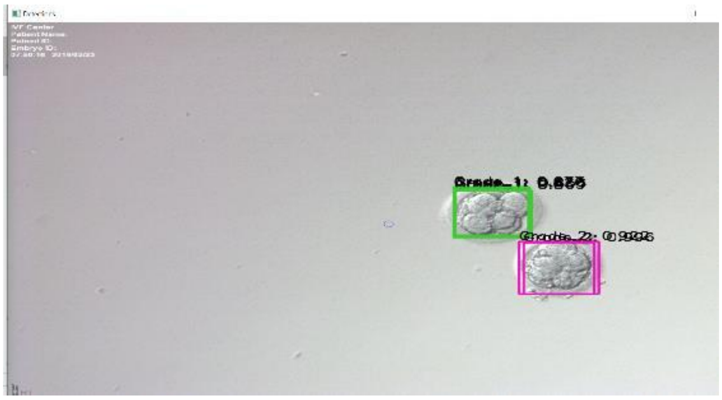


Figure 11. Brightness Image without Non-Maximum Suppression shows four bounding boxes Grade 1 and 2 bounding boxes Grade 2

3.2.2. Testing with Second Schema of Data Splitting (Testing Image 1)

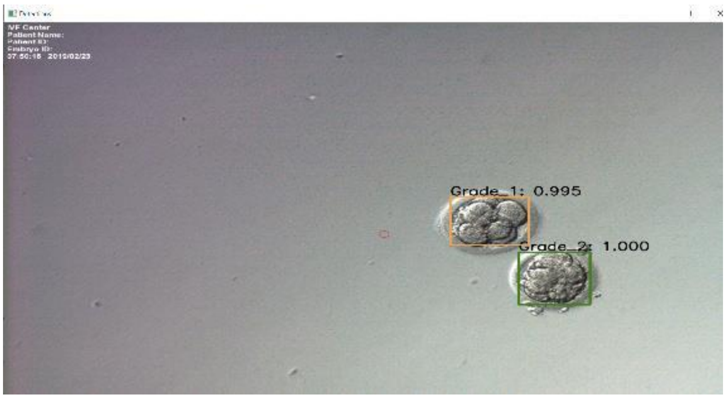


Figure 12. Original Image with Non-Maximum Suppression shows one bounding box Grade 1 and 1 bounding box Grade 2

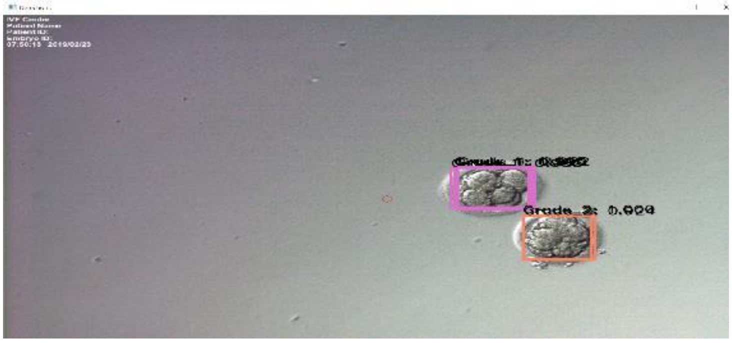


Figure 13. Original Image without Non-Maximum Suppression shows four bounding boxes Grade 1 and 1 bounding box Grade 2

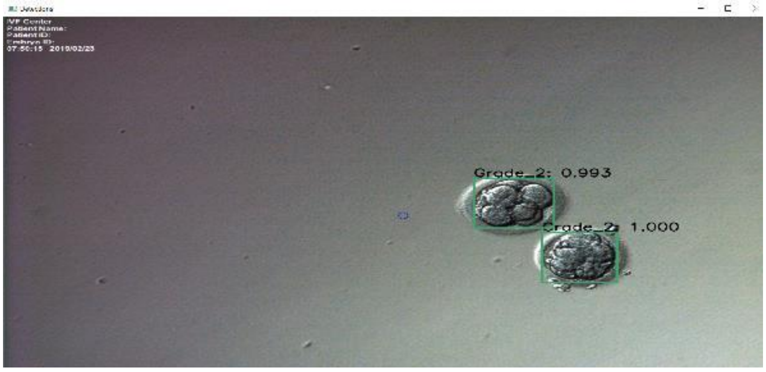


Figure 14. Darkness Image with Non-Maximum Suppression shows one bounding box Grade 1 and 1 bounding box Grade 2

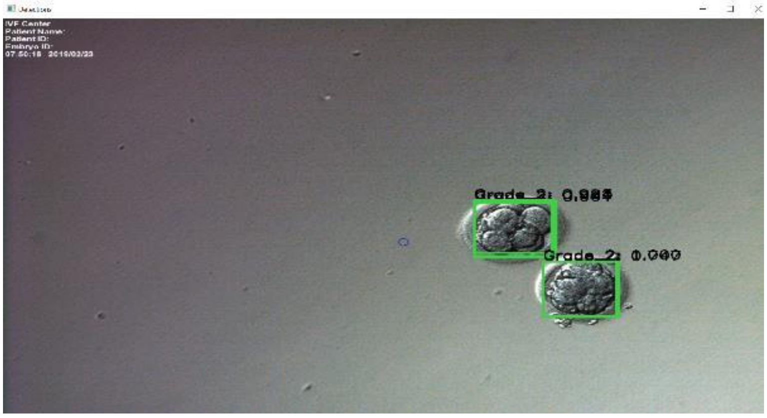


Figure 15. Darkness Image without Non-Maximum Suppression shows five bounding boxes Grade 1 and 2 bounding boxes Grade 2

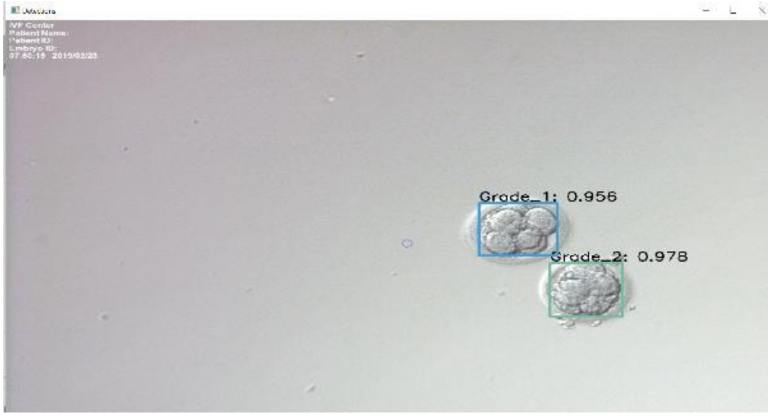


Figure 16. Brightness Image with Non-Maximum Suppression shows one bounding box Grade 1 and 1 bounding box Grade 2

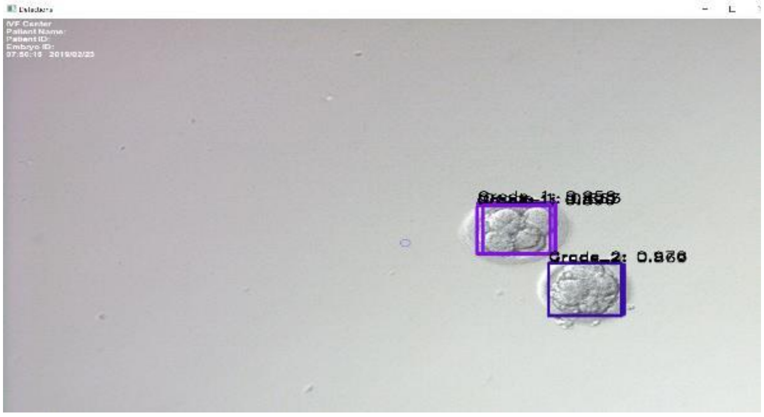

IVF Center
Patient Name:
Patient ID:
Embryo ID:
07:50:16 2019/02/23

Figure 17. Brigthness Image without Non-Maximum Suppression shows four bounding boxes Grade 1 and 2 bounding boxes Grade 2

Figures 6 to 17 show the results of embryo detection with two schemes, Figures 6 to 11 show the detection results with the first data separation scheme, and Figures 12 to 17 show the detection results with the second data separation scheme.

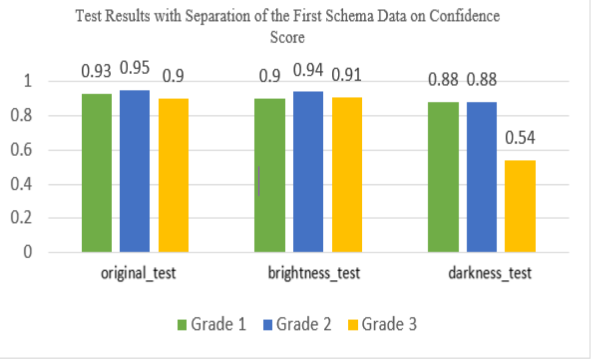


Figure 18. Comparison of Confidence score using the 1st data separation image

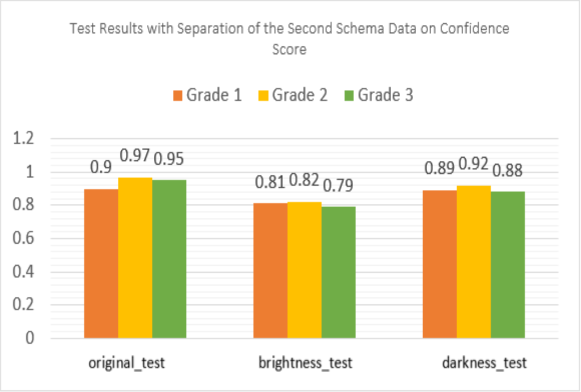


Figure 19. Comparison of Confidence score using the 2nd data separation image

From Fig.18 and 19, it can see that the highest confidence scores for grade 2 and grade 1 were in the original image test at 0.95 and 0.93, while for grade 3, the highest confidence score was in the illuminated image, which was 0.91. As for the one with the lowest confidence score, in the given image darkening, the average confidence score for grades one and grade 2 is 0.88, and grade 3 is 0.54.

For data separation with the second scheme, which refers to Fig.19, the highest confidence scores for grade 1, grade 2, and grade 3 are in the test with the original image of 0.90, 0.97, and 0.95. While the test using the described image has the lowest confidence score compared to the initial and darkened image test, namely for grade 1: 0.81, grade 2: 0.82, and grade 3: 0.79

So, from Fig. 18 and 19, it can be seen that the provision of images in decreasing and increasing lighting can affect the confidence score. Meanwhile, for the overall test image test results with the first separation scheme, the following results were obtained:

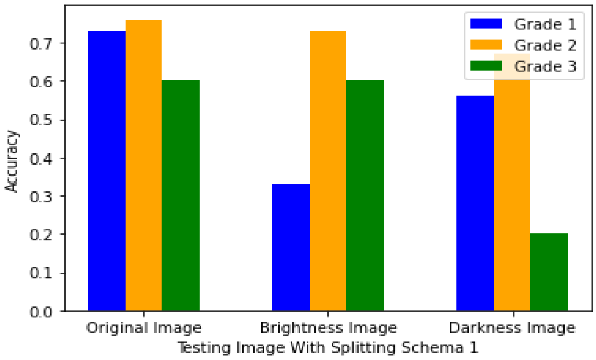


Figure 20. Comparison of Accuracy using the 1st data separation image

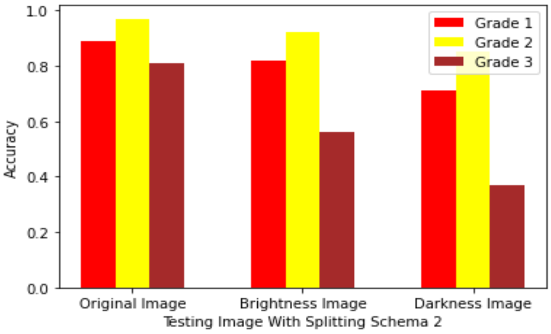


Figure 21. Comparison of Accuracy using the 2nd data separation image

Based on Figure 21, images with interference in the form of lighting and darkening have poorer Accuracy and F-1 Score results against testing with the original image. It is caused by various data factors, such as the image, supported by configuring image augmentation (light enhancement) at the pre-training stage, which is not the same as the test stage image. The first data separation scheme, which divides into validation, test, and training data, represents a bad model seen in the test results, which show values that are much different from the validation results based on Fig.4.

The previous discussion of the result implies that Yolo is sensitive to the light intensity of the input data image. However, using the original image from IMERI can still provide an excellent embryo grading solution.

4. Conclusion

The main idea of this study is to develop previous research and be able to detect the grade of IVF embryos. With the implementation of the Yolo version 3 model in this study, the Yolo version 3 model was able to get a good accuracy value. Based on the result, the Max Batches configuration or the number of iterations will affect the model's accuracy. It was found for the first schema data separation (Training, Test, and Validation Data) that the largest mAP value was at the 6000th iteration, which was 100.00%, while the smallest accuracy value was at iteration 1000, which was 37.56%. Our results are superior when compared to Han [10] for grading diabetic feet, where the maximum mAP is 91.5%, but the issue is not actually the same. As for the accuracy of the second data separation using Stratified K Fold Cross Validation, where the largest mAP value is in the 5000th iteration at fold=3, which is 97.33% with an average loss of 83.05%, with the smallest mAP being at the 1000th iteration of 42.98%. The separation of data using the second scheme can show a good model, while the first separation shows a bad model or is experiencing overfitting. The model cannot detect bounding boxes or correct labels for class grade x, which see from the high Specificity value, which is 0.86 in the original image, 0.77 in the darkened image, and 0.81 in the original and brightness image.

References

[1] M. F. Kragh, J. Rimestad, J. Berntsen, H. Karstoft, "Automatic Grading of Human Blastocyst from Time-Lapse Imaging," *Computers in Biology and Medicine*, vol.115, no. 103494, 2019.

[2] A. A. Septiandri, A. Jamal, P. A. Iffanolida, O. Riayati, and B. Wiweko, "Human Blastocyst Classification after In Vitro Fertilization using Deep Learning", in *The 7th International Conference on Advance Informatics: Concept, Theory, and Applications (ICAICTA)*, 2020, pp. 1-4.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning on Image Recognition", in *the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[4] F, Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions", in *the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251-1258.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. C. Chen, "Mobilenetv2: Inverted Residuals and Linear Bottlenecks", in *the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510-4520.

[6] G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Network", in *the IEEE Conference on Computer Vision and Pattern Recognition,* 2017, pp. 4700-4708.

[7] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement", *arXiv.org* 2018.arXiv:1804.02767, Available URL:https://doi.org/10.48550/arXiv.1804.02767.

[8] Z. Wang, "SEG-YOLO: Real-Time Instance Segmentation using YOLOv3 and Fully Convolutional Network", KTH Royal Institute of Technology Stockholm, School of Electrical Engineering and Computer Science (EECS), 2019.

[9] S. Norling, "Tree Species Classification with YOLOv3: Classification of Silver Birch (Betula Pendula) and Scots Pine (Pinus sylvestris)", KTH Royal Institute of Technology Stockholm, School of Electrical Engineering and Computer Science (EECS), 2019.

[10] A. Han, Y. Zhang, A. Li, C. Li, F. Zhao, Q. D. Liu, Y. Liu, X. Shen, S. Yan, and S. Zhou," Efficient Refinement on YOLOv3 for Real-Time Detection and Assessment of Diabetic Foot Wagner Grades", *arXiv.org*, arXiv:2006.02322, Available URL: https://doi.org/10.48550/arXiv:2006.02322

[11] M. Nomura, A. Iwase, K. Furui, T. Kitagawa, Y. Matsui, M. Yoshikawa, and F. Kikkawa, "Preferable Correlation to Blastocyst Development and Pregnancy Rates with a New Embryo Grading System Specific for Day 3 Embryos", *Journal of Assisted Reproduction and Genetics*, vol. 24, no. 1, p.23-28.

[12] A. Ammar, A. Koubaa, M. Ahmed, A. Saad and B. Benjdira, "Vehicle Detection from Aerial Images using Deep Learning: a Comparative Study", *Journal Electronics*, vol. 10. No. 7, p. 820, 2021.

[13] T. Yu and H. Zhu, "Hyper-Parameter Optimization: A Review of Algorithms and Applications", *arXiv.org*, 2020, arXiv.2003.05689, Available URL: https://doi.org/10.48550/arXiv.2003.05689.

[14] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview", arXiv.org, 2020, arXiv:2008.05756v1, Available URL: https://doi.org/10.48550/arXiv:2008.05756

[15] A. van Etten," Satellite Imagery Multiscale Rapid Detection with Windowed Network", 2018, arXiv:1809.09978v1, Available URL: https://arxiv.org/pdf/1809.09978.pdf