AI/ML Project – Embryo Analysis to Improve Success Rate of IVF

Anjana Padikkal Veetil
Department of Data Analytics for Business
St Clair College, Windsor, Ontario, Canada. N9A 5K3
Email: w0780343@myscc.ca

Nobin Ann Mathew
Department of Data Analytics for Business
St Clair College, Windsor, Ontario, Canada. N9A 5K3
Email : w0768877@myscc.ca

Venkata Bhagya Teja Maridu
Department of Data Analytics for Business
St Clair College, Windsor, Ontario, Canada. N9A 5K3
Email: w0788222@myscc.ca

Amal Mathew
Department of Data Analytics for Business
St Clair College, Windsor, Ontario, Canada. N9A 5K3
Email: w0773672@myscc.ca

Santosh Kumar Kantimahanti Lakshmi
Department of Data Analytics for Business
St Clair College, Windsor, Ontario, Canada. N9A 5K3
Email: w0780766@myscc.ca

Supervised by
Prof. Umair Durrani
St. Clair College

Contents

ABSTRACT

In-vitro fertilization (IVF) is the most efficient way to reduce infertility among human beings, which is increasing nowadays. Even with the high-end technologies available in this modern era the success rate of IVF is less. In the IVF process an embryologist manually grade the embryo, out of those embryos graded the best embryos are implanted in the women. There are chances of manual error during the human intervention while embryo grading, which can lead to multiple IVF cycles. In this project we classify the embryo images into two classes good embryo images and poor embryo images. To achieve this, a deep neural network method is leveraged as an artificial intelligence approach to predict the quality of the human embryo images. We used the framework known as STORK which has multiple pre trained models, out of which we used Google's Inception model algorithm. We achieved an accuracy of 96% while predicting the quality of embryo as good and poor.

1. INTRODUCTION

Approximately 16% which is 1 in 6 couples in Canada experience infertility (UCLA Health, 2020). This number has doubled since the 1980's. Globally, According to Reproductive Biological Endocrinology, 2015 48.5 million couples experience infertility. Statistics indicates that 6.1 million people in United states are affected by infertility and only half of them are undergoing related treatments.[1] Infertility is defined as a clinical condition of inability to conceive or get pregnant after one year or longer of unprotected sex.[2] IVF is a type of assistive reproductive technology (ART) for infertility treatment and surrogacy. Surrogacy is an arrangement where a woman agrees to deliver for another person where pregnancy is medically impossible[3].

Figure 1 explains the whole IVF process step by step. IVF involves many steps, and each cycle would take an average duration of 6 to 8 weeks (about 2 months). At first the individual would undergo an initial consultation and testing with an infertility specialist and would be prescribed medication for ovarian stimulation. Then the doctor retrieves the eggs from the woman's ovary. The retrieved eggs are fertilized with the sperm from the semen sample of a partner or a sperm donor in a culture medium in a laboratory. The fertilized eggs undergo embryo culture where the fertilized eggs are allowed to grow in an artificial medium (incubator) under supervision. After 3 – 5 days once the embryo reaches the blastocyst stage, the best embryos are selected by the embryologists based on the morphological attributes and are transferred into the woman's uterus. After two weeks of embryo transfer the couple undergoes a pregnancy test and the success of the IVF process is determined.[3]

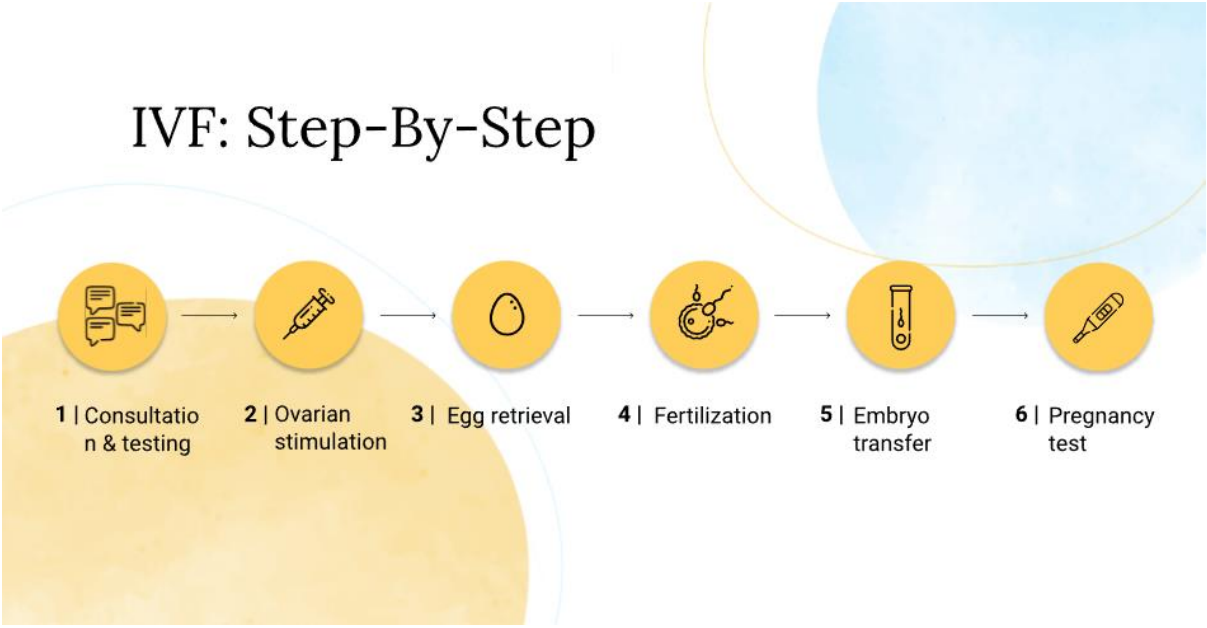

Figure 1 IVF PROCESS [2]

2. BACKGROUND

Standard morphological assessment of the embryo is an important objective of an embryologist, and it has always been the major tool for selecting the best embryo for transfer. A good embryo morphology remains one of the prime predictors for the success of the IVF process. Manual assessment may result in human error based on their experience, intuition, and expertise.[4] One cycle of an IVF process takes 6 to 8 weeks (about 2 months). Multiple rounds of IVF are common as the success rate depends on other factors such as age and past medical history. On average 3 IVF cycles are found to be clinically effective. The average cost of one round of IVF is estimated to be $15000, with a success rate of less than 50% per embryo transfer for women under the age of 35. Medication adds more expenditure to the process and the whole cost of the process may also vary based on other factors such as the clinic chosen for treatment, number of cycles administered and surrogacy in case. So, a couple experiencing infertility must invest lots of money in the IVF process to ensure a successful outcome based on the number of cycles required.

Multiple viable blastocysts were transferred to increase the chances of pregnancy, but this would result in multiple pregnancies and gestational complications in the mothers and babies.[1] Therefore, identifying the single best viable blastocyst is recommended which would reduce the number of cycles administered and eliminate multiple pregnancies and other related issues. Manual assessment of the embryo to determine the quality of the embryo can be automated using artificial machine learning algorithms which would classify the embryos into good or poor based on supervised machine learning classification model.

3. DATASET

3.1 Integrated Dataset Overview:

Training Images (220): Good 118, Poor 102
Test Images: 49

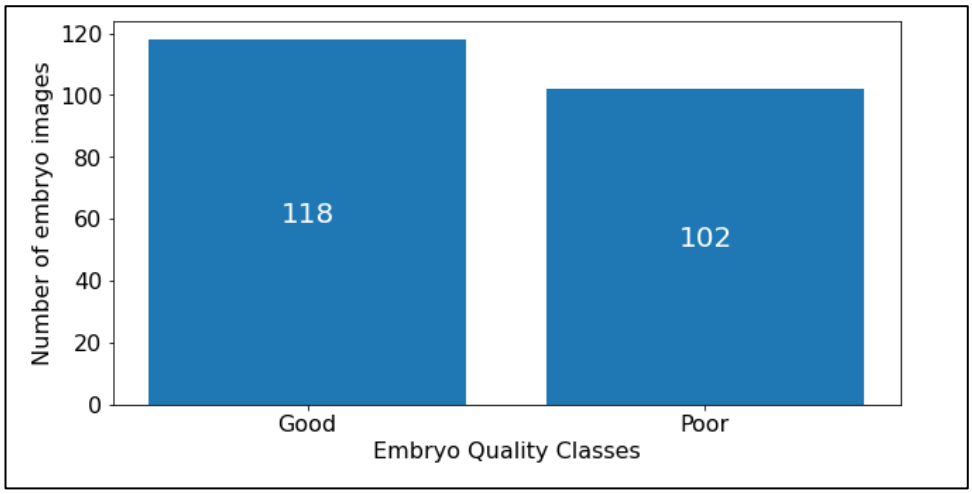

Figure 2: Bar chart for the distribution of good and poor-quality embryos.

As we have limited accessibility of embryo images, we integrated embryo images from multiple sources. Section 3.2 and 3.3 provides details on the data sources from where the images were collected. Figure 3 shows the image of a blastocyst and its four different compounds outer layer of zona pellucida, Trophectoderm, Inner cell mass and blastocoel. The embryologist evaluates these components and grade an embryo. According to the grade and the information of embryo implanted or not, the embryo images are classified as good and poor embryos.

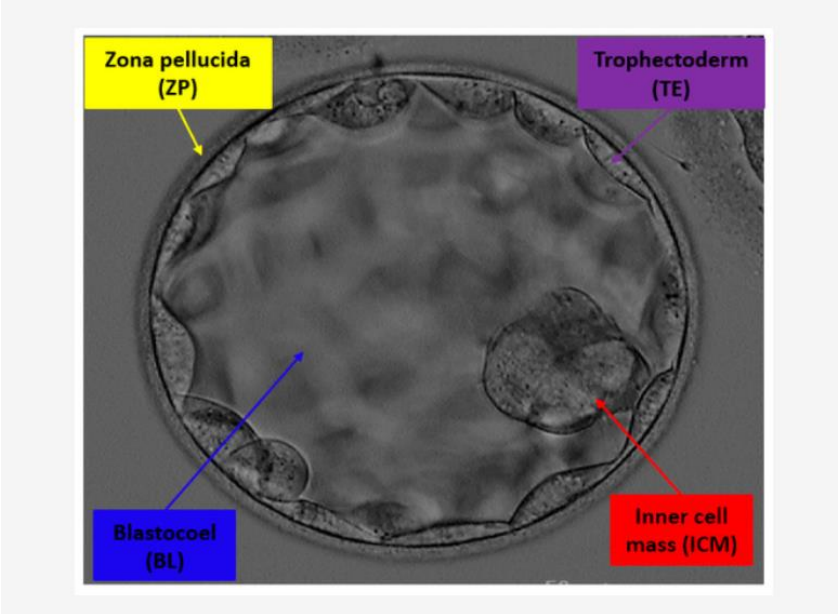

Figure 3: Blastocyst microscopic image with its components.[1]

3.2 Dataset 1 :

Source : Simon Fraser University.

Training Images : Good 76, Poor 60

Test Images: 35

This dataset is password protected and securely stored in Simon Fraser University vault and could be accessed only by authorized personal. These embryo images were collected from the patients who were admitted at the PCRM (Pacific Centre for Reproductive Medicine) Canada. There are total of 171 images that contains either good or poor embryo. The images were in .bmp format which was later changed into .jpg format as a part of image pre-processing.

3.3 Dataset 2 :

Source : STORK Framework.

Training Images : Good 42, Poor 42

Test Images: 14

This image dataset is publicly available and taken from the GitHub repository of STORK[5], these images of human embryos were obtained from the Centre for Reproductive Medicine at Weill Cornell Medicine. There are total of 98 images. This image dataset was already in .jpg format and did not need any image preprocessing.

4. METHODOLOGY

4.1 Data Preprocessing

4.1.1 Image file type modifications

The dataset from Simon Fraser University (SFU) was in bmp format, which we had to convert into jpg format for images to be used for training. Whereas the dataset from STORK was already in jpg format.

4.1.2 Image Data Labelling

The dataset from Stork was already labelled as good and poor images which is required for calculating the accuracy. Whereas the images from SFU was not labelled. Labelling embryo images from SFU has been performed based on the master list which was provided along with images. Report (Figure 4) is evaluated by an embryologist, and it has information such as the embryo grade and if the embryo was implanted into the woman or not which was used to label them as good or poor.

Information provided in report for each image:

1. **Grade:** The grade code, which contains three grades, indicates the embryo quality based on the pregnancy outcome as determined by statistical analysis of clinical data. The images of the blastocysts were labeled using the Veeck and Zaninovic grading system[6]
    I. Amount of blastocyst has expanded: Grades range from 1 to 6, with 6 being the most developed.[6]
    II. Quality of the Inner Cell Mass: Grades range from A to C, with A being the best quality.[6]
    III. Quality of the Trophectoderm: Grades range from A to C, with A being the best quality.[6]
2. **Outcome:** Provides information on whether embryo is implanted or not in the process of IVF. Figure 4 shows the outcome of each embryo image, if the outcome is shown as 0 it means that embryo was not implanted whereas if the outcome has 1 that means the embryo has implanted into the woman. The outcome also has a value 2 which we do

not include in the training or test dataset because those images with outcome 2 means we don't have information on that embryo if it was implanted or not. We then prefix the image file name with the class (Good or Poor).

File Name	GRADE	Expansion	ICM	TE	Outcome	Acquisition Year
	According to the Gardner's Method					
Blast_PCRM_R12-0137	4AA	4	A	A	1	2012
Blast_PCRM_R12-0160	4AA	4	A	A	1	2012
Blast_PCRM_R13-0006A	2AB	2	A	B	2	2013
Blast_PCRM_R12-0173a	3AB	3	A	B	0	2012
Blast_PCRM_R12-0173b	2AA	2	A	A	0	2012
Blast_PCRM_R12-0221a	3AA	3	A	A	1	2012
Blast_PCRM_R12-0221b	3AA	3	A	A	1	2012

Outcome:
0 = Not implanted
1 = Implanted
2 = Unknown

Figure 4: Embryo Grade Report for Dataset 1

4.1.3 Removing Unwanted Data

In SFU dataset 78 embryo images has recorded outcome value as 2 in the Grade Report (Figure 4). These images are not used for the training or testing as the value 2 refer that embryo is implanted or not is Unknown.

4.1.4 Merging two datasets

Dataset 1 (Sourced from Simon Fraser University) and Dataset 2 (Sourced from STORK framework) are merged based on the classes (good embryo and poor embryo) to increase the data for training the model, these images were divided in the ratio of 80:20 where 80% data was stored in train along with the STORK train data images and remaining 20% of the images stored in test directory along with the test images of STORK dataset.

4.2 Data Analysis

4.2.1 Software and Tools Used

Python 2.7 was used to carry out all the data analysis in this research.

TensorFlow version 1.15 and the TF-Slim Python library for defining, training, and evaluating models are used to implement STORK framework.

4.2.2 Data Modelling:

A Deep Neural Network (DNN) is used for embryo image analysis based on Google's Inception-V1 architecture. The STORK repository has multiple pre-trained models out of which Inception V1 is best suitable for image classification.

4.2.2.1 STORK Framework Introduction

We have used STORK framework to predict blastocyst quality to good or poor. STORK is a framework developed based on Google's Inception model, and we have trained the model for good-quality and poor-quality classification.

4.2.2.2 Training Model

This study included a total of 269 blastocyst images, and the data is divided into two parts, 80% of the images (220) are used for training the model and the remaining 20% images (49) are used for testing the trained model.

5. RESULTS

We have combined two embryo image datasets for achieving high accuracy in our classification model. There was a total of 220 human embryo images for training the model and 49 for testing. Out of 220 total embryo images 118 embryos were of good quality embryos and the remaining 102 were of poor-quality embryos.

Deep neural network has high accuracy in classification of human embryo images. We have used pre trained Inception V1 model algorithm to train the images. The basic inception V1 has four parallel layers of 1X1 convolution, 3X3 convolution, 5X5 convolution and 3X3 max pooling.[7]

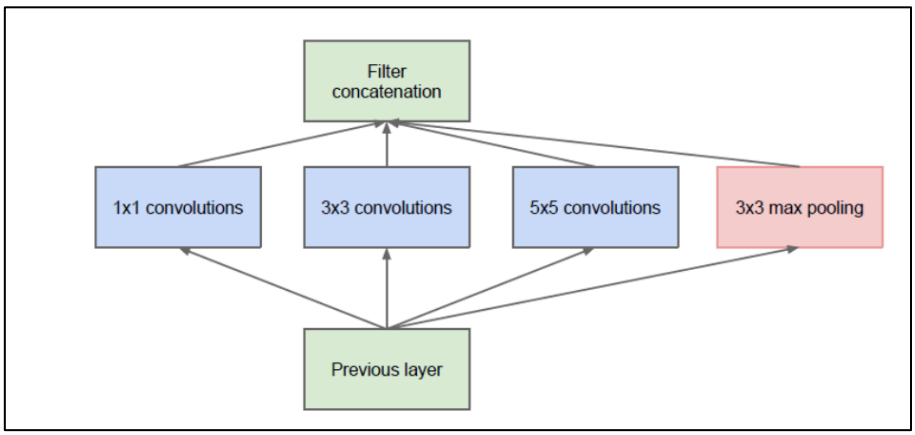

Figure 5: Inception V1 Naive Form.[7]

Figure 5 shows the basic Inception architecture of the four layers which is called Naive form. There are few drawbacks of this form as it requires high computational power and is time consuming.[7] The Inception V1 model is a transfer learning algorithm, where fine-tuning of the parameters is initially performed for all the layers.[5] We trained Inception V1 model with the two classes which is good-embryo and poor-embryo images. We used 5000 steps for training the deep neural network using STORK and used randomly selected test images of good and poor embryo images which were pre-labelled for calculating the accuracy of the test model. Our test model was able to predict the embryo quality with 96% accuracy. 47 embryo images were predicted accurately out of 49 images. The train accuracy of this model was 100% due to the limited availability of the dataset. We had 220 images of good and poor embryo images for training the model. We acquired a total precision of 0.95 for the good embryo to be predicted as good and the poor embryo to be predicted as poor i.e., Sum of true positive and true negative is 47 (c=47) out of 49(t=49) total embryo images. True positive stands for good embryo predicted as good and True negative represents poor embryo predicted as poor. The precision is calculated as c* 0.1 / t which gives 0.95 for our test images.

We had developed a data product by integrating the model with the web page where we can upload an embryo image and check the probability percentage of the image being good or poor. The web interface was developed using Bootstrap, HTML5, JavaScript, CSS, and Flask. The Data Product web page has two tabs, a home tab and a Report tab. Figure 6 shows the home page of the Data Product where we have provided a brief description of the In Vitro Fertilization.

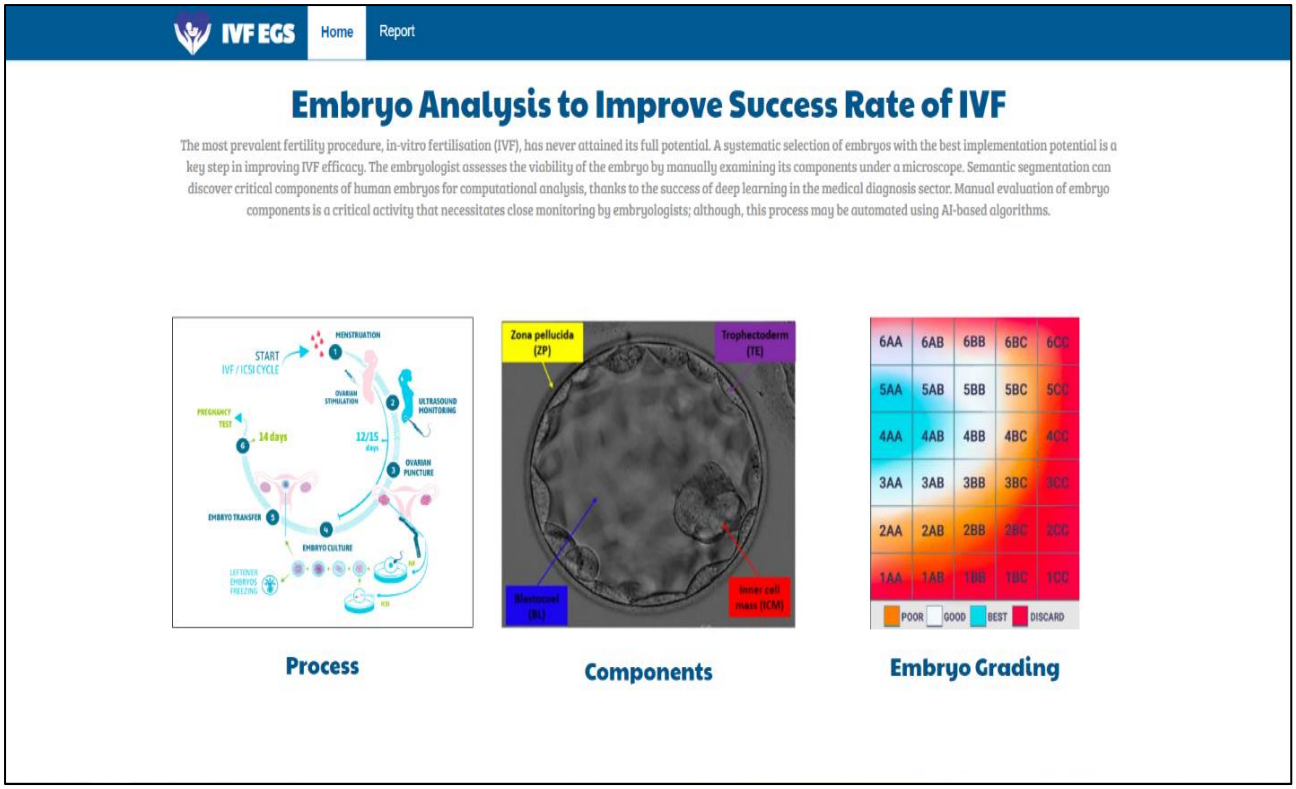

Figure 6: Data product Home Page

The Report tab shown in Figure 7 enables to upload an embryo image and there is a fetch report button which runs the predict model and predicts the probability of the uploaded embryo image being good and poor.

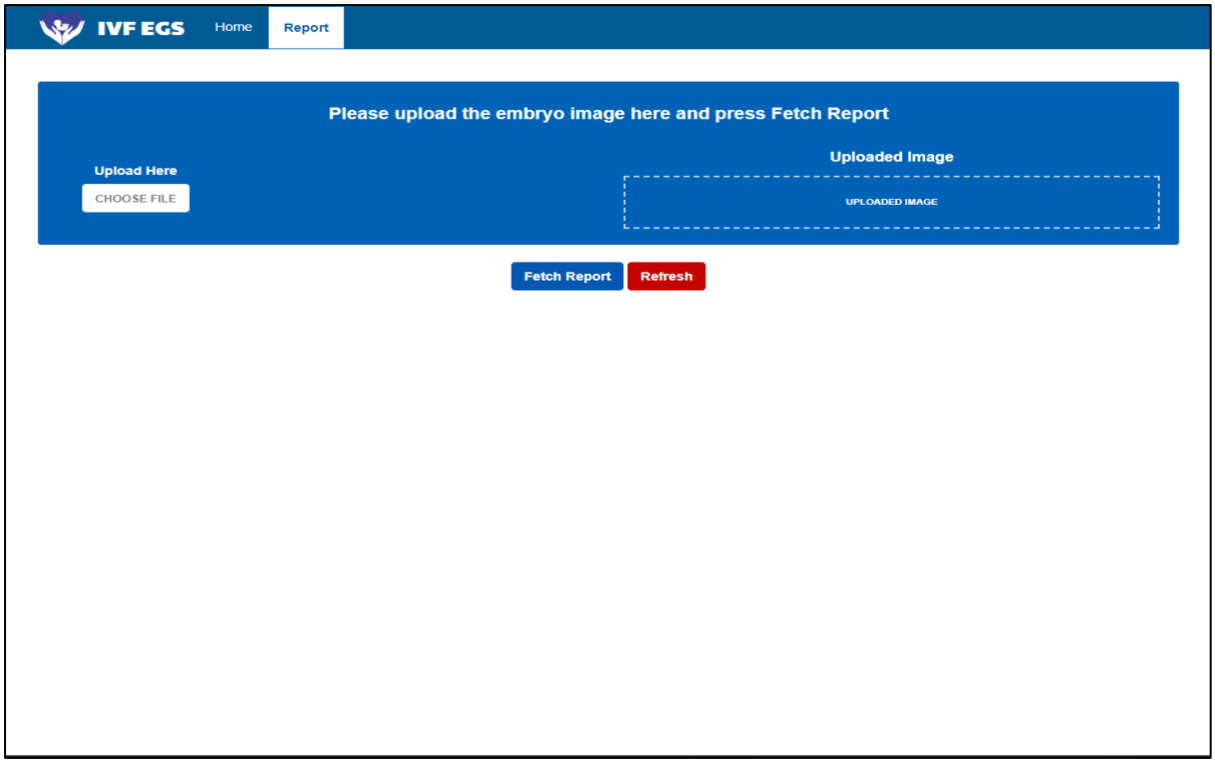

Figure 7: Report Fetch Page

6. DISCUSSION

Studies on evaluating human embryos are presently very few and focused primarily on morphological characteristics. Several research using traditional machine learning techniques have been published recently, such as support vector machine (SVM) and RF, and deep learning methods, such as CNN for either a grade assessment or a success forecast. To date, a variety of AI techniques have been applied to evaluate blastocysts. STORK outputs good and poor grades.

The benefit of our approach is that the entire image of the embryo is evaluated as opposed to just the predetermined, segmented features that embryologists are trained to analyse, allowing for quantification of all the available data. Convolution enables the AI to recognise morphological feature patterns that we are unable to evaluate. We have shown that deep learning techniques can produce precise quality evaluations.

The STORK architecture given here offers a technique that is simple to put into practise for a variety of applications, such as embryo grading. Significantly, the STORK framework is fully automated and does not require any manual augmentations or preprocessing on the input images. In effect, it offers embryologists an easy-to-use platform that doesn't require advanced computational understanding.

The result of the model (Figure 8) gives information regarding the quality of the embryo, classifying it as either Good or Bad. The classification is done based on the probability values of being Good or Bad for each image. Higher probability value for Good implies that the image is classified as Good, otherwise it is classified as Bad. A sample output has been shown in the Figure 8.

```
E:\STORK-master\STORK-master\Images\test/good_23765483_-15_3AA.jpg      0.9878179   0.012182053
E:\STORK-master\STORK-master\Images\test/good_23765483_-30_3AA.jpg      0.9999862   1.3875138e-05
E:\STORK-master\STORK-master\Images\test/good_23765483_-45_3AA.jpg      0.9999685   3.1443265e-05
E:\STORK-master\STORK-master\Images\test/good_23765483_0_3AA.jpg  0.9974751   0.0025248467
E:\STORK-master\STORK-master\Images\test/good_23765483_15_3AA.jpg 0.9994037   0.0005962936
E:\STORK-master\STORK-master\Images\test/good_23765483_30_3AA.jpg 0.9988055   0.0011944603
E:\STORK-master\STORK-master\Images\test/good_23765483_45_3AA.jpg 0.98130715  0.018692851
E:\STORK-master\STORK-master\Images\test/good_Blast_PCRM_R14-0326b.jpg  0.99946386  0.00053617754
E:\STORK-master\STORK-master\Images\test/good_Blast_PCRM_R14-0348a.jpg  1.0   4.89573e-08
```

Figure 8: Output.txt Result file

Figure 9 shows the probability of the quality of an embryo image being good or poor. If the probability of an embryo being poor is near to 1 and that of being good is near to 0 implies that the embryo is classified as poor embryo. While if its vice versa the embryo will be classified as good embryo. In the figure 9 Fair-Negative represents True Negative and Fair-Positive represents True Positive.

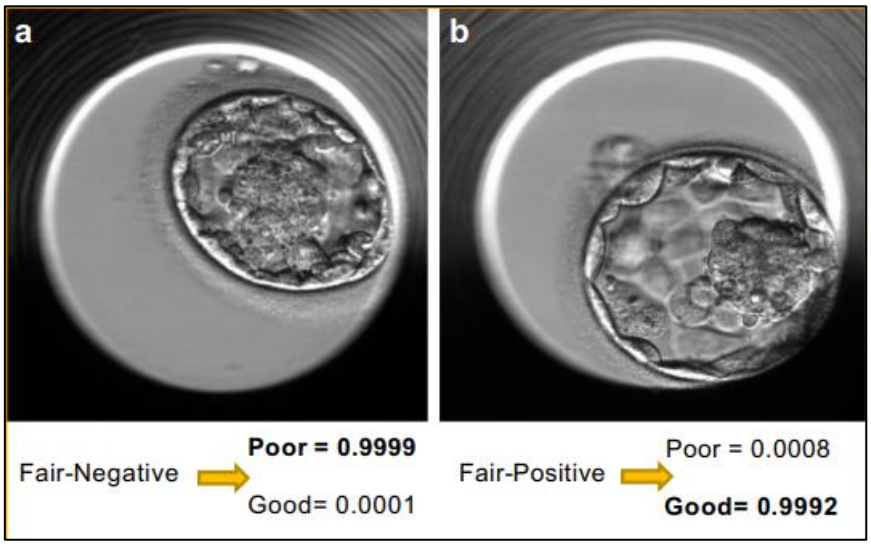

Figure 9: Probability of embryo quality

We used 2 metrics for measuring the performance of our model. First metric is “Precision”. It is used to measure the model's performance, which is the quality of a positive prediction made by the model. Precision refers to the number of true positives a divided by the total number of predictions (i.e., the number of true positives plus the number of false positives)[8]. To find the precision of the overall model, we used the following formula, and it has come around 95%.

$$Precision = C * 0.1/T$$

C – True Positive or True Negative
T – Sum of True Positives and False Positives or True Negative and false Negatives

Accuracy is another metric we used to measure the performance. Accuracy is the fraction of predictions our model predicted right [9]. When we calculate the accuracy for our framework, it is amounting to 96%. Formally, accuracy has the following definition: Accuracy = Number of correct predictions divided by Total number of predictions.

$$Accuracy = C * 0.1/T$$

C – Sum of True Positive and True Negative

T – Total number of images used for testing

However, it is mentioned that our approach has several limitations. For example, when the possibility of accurately predicting pregnancy using only embryo images that are labelled as "positive live birth" or "negative live birth,". The results demonstrated that based only on embryo morphology, the trained algorithm is unable to distinguish between positive and negative live births with accuracy.

The training process also requires loading a sizable number of medical images for training and validation, so even though STORK can run on conventional computer microprocessors (CPUs), ample system memory and graphics processing units (GPUs) make the training process faster (by at least an order of magnitude).

7. CONCLUSION

By employing our model, we could select the best viable embryo for transferring to the uterus. With the current IVF process, it is preferred to go for 3 cycles where 1 cycle costs up to $15,000 and 6-8 weeks. If our model could classify the embryos as good and poor there is a possibility of reducing the number of cycles in IVF.

So, to classify the embryo, we leveraged a deep neural network to identify the quality of each embryo and we also used a stork framework, to find the quality of the embryo images. We have classified these embryo images into two classes namely, good, and poor, by using the stork model. For testing the data, our model gives an accuracy of 96% on a total of 49 images, where 47 were predicted correctly. One of our limitations is that while training we got an accuracy of 100%, due to the limited number of datasets. Also, our model will go with the higher precision rate for true negative, where the poor embryo is predicted to be poor. We have developed a web page for the data product to Integrate the model, where an embryo image can be uploaded and the probability of that image being good, or poor can be fetched. The webpage can be used by a common person to check the quality (probability) of the uploaded embryo as good or poor.

8. GITHUB

This is the GitHub repository of our Project “AI-ML-Project-Embryo-Analysis-to-Improve-Success-Rate-of-IVF”

https://github.com/anjanapv/AI-ML-Project-Embryo-Analysis-to-Improve-Success-Rate-of-IVF

9. ACKNOWLEDGMENTS

This project is solely based on the GitHub repository called STORK (https://github.com/ih-lab/STORK). We acknowledge Simon Fraser University for providing access to the password protected dataset.

We would also like to thank Eric Tang (etang.mba2008@ivey.ca) for coordinating and providing dataset sources for this project.

We would like to extend our heartfelt gratitude for the continuous support and guidance of our professor Dr. Umair Durrani without whom this project wouldn't have been possible.

10. REFERENCES

[1] M. Arsalan and A. Haider, "Detecting Blastocyst Components by Artificial Intelligence for Human Embryological Analysis to Improve Success Rate of In Vitro Fertilization procedure which embryos are cultured the The (blastocyst transferred to the pat multiple were other is an embryo," *J. Pers. Med.*, 2022.

[2] Pacific Fertility Center, "A journey to parenthood with a team that cares," 2020. https://www.pfcla.com/services/in-vitro-fertilization/overview

[3] Wikipedia, "In vitro fertilisation," 2022. https://en.wikipedia.org/wiki/In_vitro_fertilisation

[4] A. Kirillova *et al.*, "Should we transfer poor quality embryos?," *Fertil. Res. Pract.*, vol. 6, no. 1, pp. 1–7, 2020, doi: 10.1186/s40738-020-00072-5.

[5] P. Khosravi *et al.*, "Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization," *npj Digit. Med.*, vol. 2, no. 1, pp. 1–9, 2019, doi: 10.1038/s41746-019-0096-y.

[6] Healthline, "All About IVF Embryo Grading," 2022. https://www.healthline.com/health/infertility/embryo-grading#day-5-embryo-grading

[7] Adith Narein T, "Inception V3 Model Architecture," 2022. https://iq.opengenus.org/inception-v3-model-architecture/#:~:text=Inception V1 When multiple deep layers of convolutions,filters of different sizes on the same level.

[8] C3 AI, "Precision." https://c3.ai/glossary/machine-learning/precision/#:~:text=Precision is one indicator of,the number of false positives).

[9] Data Robot, "Machine Learning Model Accuracy," 2022. https://www.datarobot.com/wiki/accuracy/#:~:text=Machine learning model accuracy is,input%2C or training%2C data