

BACHELOR THESIS
COMPUTING SCIENCE



RADBOUD UNIVERSITY

Embryo Classification using Neural Networks

Author:
Ilse Arwert
s4592778

First supervisor/assessor:
Prof. Tom Heskes
t.heskes@science.ru.nl

Second supervisor:
Martine van Miltenburg
m.vanmiltenburg
@mckinderwens.nl

Second assessor:
MSc Twan van Laarhoven
tvanlaarhoven@cs.ru.nl

August 24, 2020

Abstract

Infertility is an issue that affects roughly 186 million people worldwide. This issue is partially overcome using fertility treatments that often produce in vitro embryos, which undergo a visual morphology assessment to determine embryo quality. The potential of using neural networks to aid this subjective process has barely been examined. In this thesis, we researched the potential of using STORK, Inception V1 and Inception V3 as tools to aid in the qualitative classification of embryos produced in vitro at the MCK Fertility Center. We did this by analyzing how accurately these neural networks could predict whether embryos had been selected for transfer into the uterine cavity. The analysis shows that Inception V1 and Inception V3 had an accuracy of 68% and 78% respectively, while STORK lagged behind at roughly 50%. In conclusion, neural networks are a potentially viable addition to the embryo quality assessment process, although further research is required.

Contents

1	Introduction	2
2	Preliminary knowledge	3
2.1	Embryo quality	3
2.2	Neural Networks	4
3	Related Work	7
4	Tools	9
4.1	STORK	9
4.2	Inception V1	10
4.3	Inception V3	12
5	Method	14
5.1	Compiling the Dataset	14
5.2	STORK	15
5.3	Standalone V1 and V3	16
6	Results	17
6.1	STORK	17
6.2	Inception V1	19
6.3	Inception V3	20
6.4	Combined observations	21
7	Discussion	23
8	Conclusions	25
A	Appendix	28
A.1	Inception V1 Architecture	29
A.2	Inception V3 Architecture	30
A.3	Results of V1 and V3	31

Chapter 1

Introduction

The MCK Fertility Center (Medisch Centrum Kinderwens) is a fertility center located in Leiderdorp. People who desire to have children but are unable to do so without medical intervention can go there for help with conceiving a child. Several techniques are possible, such as the well-known IVF (in vitro fertilization) treatment or gamete donation. Many of these treatments result in *in vitro* embryos: embryos created outside the womb. They are kept in an incubator to monitor their first days. From the multiple embryos created for a couple, an embryologist selects the embryo most likely to survive. This choice is based on a visual morphological assessment, taking into account several criteria such as cell division rate and symmetry. The selected embryo is then transferred back into the mother, hopefully resulting in a successful pregnancy.

This thesis will examine the possibility of automating the selection of the most viable embryo. Specifically, it raises the question: is classifying embryos using convolutional neural networks a viable option for the MCK Fertility Center?

The dataset provided by the MCK Fertility Center is structured quite unconventionally compared to the datasets that STORK is intended for. Therefore, this thesis will not only test the accuracy of STORK. Instead, it will compare different strategies for classifying the images provided by the MCK Fertility Center. Firstly, we evaluate the accuracy of classifying the embryos using STORK. Then, we show the accuracy obtained by retraining the final layer of Inception V1, the neural network inside STORK. Then, we retrain the final layer of another neural network, Inception V3, which is known for higher accuracy compared to Inception V1.

This thesis contains explanations of the architectures of each strategy, the complete methodology of the research, and its results. Then, we discuss the meaning of those results and their limitations. This is followed by a discussion of related work, our conclusions, and recommendations for potential future research. Finally, the appendix contains auxiliary information.

Chapter 2

Preliminary knowledge

This chapter contains preliminary knowledge to help provide context for the following chapters. In particular, we have provided a short explanation of the process of in vitro fertilization, explaining why embryo quality assessment is vital, what “embryo quality” actually means and which characteristics embryologists usually use to determine it. In addition to this, we present some background knowledge regarding the structure and functionality of neural networks.

2.1 Embryo quality

The first “IVF baby” was born in 1978 [16]. The mother was infertile due to blocked Fallopian tubes, which prevented sperm from reaching her oocytes. An oocyte was retrieved from one of her ovaries using a needle and subsequently fertilized in a laboratory using her husband’s sperm. Two days later, the embryo was transferred into the uterine cavity. The procedure turned out to be successful: the embryo implanted, the woman became pregnant and carried the pregnancy to term.

Nowadays, IVF treatments are used for couples suffering from subfertility due to a range of causes such as male subfertility, endometriosis, or tubal problems. A second technique called ICSI (intra-cytoplasmic sperm injection) was introduced in the early 1990s. ICSI is identical to IVF until the laboratory phase; for ICSI, one motile sperm cell is injected into a mature oocyte with a small glass needle. This technique can aid heterosexual couples in which the man has extremely poor sperm quality [13]. Worldwide, an estimated 6+ million children have been born after a successful IVF or ICSI treatment. In the Netherlands, this number is approximately 80.000. Every year, more than 13.000 IVF and ICSI cycles are performed in the Netherlands alone [1].

The first IVF treatments were performed in a natural menstrual cycle. This means that only one oocyte could be retrieved: generally, in

a natural cycle, only one oocyte matures at a time. This is why fraternal twins are rare. To increase the success rate of IVF/ICSI treatments, follicle-stimulating hormone (FSH) is used to induce ovarian hyperstimulation. Daily injections of FSH result in multi-follicular growth, which allows a higher number of oocytes to be retrieved at once [11]. More oocytes routinely lead to the development of more embryos, which introduces a new question: how does one choose which embryos to transfer, or to freeze?

Several morphological aspects of developing embryos factor into this selection. These include the number of cells, the division rate, the variability of sizes in cells, and the amount of fragmentation (small particles resulting from the degeneration of a cell during the process of division).

After 30 years of research and fertility treatments, not a single morphological marker has been identified that can predict the future success of an embryo with real certainty. The recent development of time-lapse microscopy has made permanent record-keeping a reality, but reliable embryo selection using a single parameter or algorithm applicable to all patients remains elusive [6]. Recently, researchers have started to examine the possibility of using automated image classification to determine embryo quality, although this technique is still in its infancy.

2.2 Neural Networks

“...a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs”

Dr. Robert Hecht-Nielsen, as quoted in [4]

The simplest definition of artificial neural networks, shown above, was coined in 1989 by Robert Hecht-Nielsen, the neuro- and computer scientist who authored one of the first textbooks on the subject. It cuts to the core of what makes neural networks different from traditional computing. Where traditionally, computing was done by a central processing unit with access to all data and instructions, neural networks consist of many small processing elements. Each performs a simple computation and passes on its output to the next processing element. These networks are modeled on mammalian brains, although they are much less complex; large neural networks consist of perhaps thousands of neurons, human brains of 86 billion [8].

The main feature of neural networks is that they comprise several interconnected layers. Neural networks typically consist of an input layer, several ‘hidden layers,’ and an output layer. Of these, the hidden layers are where the actual processing happens. See Figure 2.1 below for a simplified example.

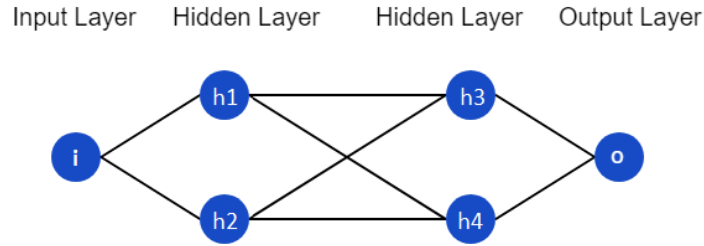


Figure 2.1: A simplified neural network.

In the figure above, the layers consist of multiple interconnected components, shown in blue. These represent the layers' processing elements, called *neurons*. Generally, the input layer of a neural network distributes the input to each neuron in the first hidden layer. Each new hidden layer then receives input from the previous layer's neurons, which is processed by the current layer's neurons to create an output, which it forwards to the next layer. Eventually, the different outputs are merged into one conclusion by the output layer.

The way the neurons in the hidden layers process the input they receive and the ways different neurons connect both influence what happens to the input received by that layer. This is how different layers can have different effects on the data. The neural networks that this thesis focuses on consist of several types of layers, the most important of which are explained below.

A key layer in image classification is the convolutional layer. Convolutional layers are used to scan images for certain features. These layers examine input images in small patches to determine the probability that that patch contains a specific feature. The output consists of a feature map with the respective probabilities of each patch containing that feature. Patches may overlap. Convolutional layers are usually named after the size of their feature filters. For example, a 5x5 convolutional layer will examine patches of five by five pixels.

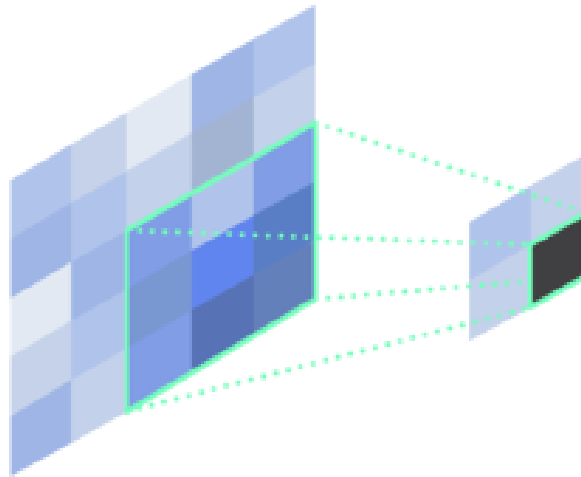


Figure 2.2: Input and output of a 3x3 convolutional layer. From [14].

As shown in the figure above, convolutional layers reduce the dimensions of the data, which reduces the number of operations necessary later in the network. So too do pooling layers, another type of layer often used in image classification. Pooling layers reduce the amount of information contained in the feature maps they receive from convolution layers, keeping the most important information. This lowers the resolution of the feature map to reduce the dimensions of the data and the computational cost. There are two types of pooling: average pooling and max pooling. Average pooling outputs the average value of each patch on the input feature map, and max pooling calculates the maximum value of each patch.

The third type of layer found in any neural network is the fully connected layer, in which every neuron in a hidden layer connects to every neuron in the next hidden layer (as in Figure 2.1). A drawback of these is that they have a high computational cost. Therefore, these are used for specific purposes such as “flattening” the output of convolution and pooling layers. The final layer of a neural network is typically a fully connected layer which outputs a single vector containing the respective probabilities that the image belongs to each available class.

Chapter 3

Related Work

The idea of embryo classification using neural network has some basis in the existing literature, although it is a small and relatively new research field. While a few different teams have researched the possibilities of automating embryo classification, there is limited literature in which neural networks are used for this challenge. Most of the works below, therefore, are related due to similar subject matter rather than methodology.

The first article on the subject was published in 2012 by Filho et al [7]. It detailed their work in assessing the viability of automating embryo grading, based on one of the visual criteria used by embryologists. Their tactic was to calculate the number of cells based on different characteristics extracted from an image of the embryo, as cell number is an indicator of blastocyst quality. They could calculate the number of cells with a reasonable 67-92% accuracy. This is not good enough to be a heavyweight factor in embryo grading, but it did open the door to using image analysis as a potential tool for embryo analysis.

In recent years, it has been shown repeatedly that convolutional neural networks, or CNNs, are an effective method for solving different medical imaging problems. Examples are diagnosing breast cancer [15], detecting diabetic retinopathy [2], and many others.

In 2019, before the start of this thesis, one major article was published on the subject, detailing the development of the STORK framework for embryo classification. STORK has an embryo classification accuracy of over 97%. It consists of a collection of scripts built around GoogLeNet's Inception V1 neural network. As described previously, STORK classifies several images for each embryo, each image with a different focal layer. STORK then averages out the grades for the different layers to conclude the embryo's final grade. It has been tested for robustness using images of embryos from three fertility centers: two in the United States and one in Spain, and is robust for those datasets [10].

In addition to this, a study published in April 2020 found a convolu-

tional neural network to be significantly more consistent than a group of embryologists. Both the neural networks and the embryologists were tasked with grading a set of embryos on a scale of 1 (poor) to 5 (excellent). Unbeknownst to the embryologists, each embryo was also rotated 90, 180 or 270° and presented along with the original images. Consistency is then defined as the percentage of cases where the given grade did not depend on rotation. The researchers observed 83.92% consistency for the neural network, compared with 52.14-57.68% for the embryologists (likely due to the subjective nature of visual morphological embryo grading). This shows that generally, automated grading of embryo quality is viable as an addition to manual grading, if not a complete replacement of it [3].

Chapter 4

Tools

In this chapter, we provide more in-depth information regarding STORK, and discuss the architectures of Inception V1 and Inception V3.

4.1 STORK

The STORK framework was developed in 2018 by Khosravi, Kazemi, Zhan et al. Their article *Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization* covers the development, structure, and results of the STORK framework [10].

In this article, they outline the design and developmental process of STORK, and examine its accuracy in embryo quality classification.

STORK’s design followed several distinct phases. Firstly, embryologists used the Veeck and Zaninovic grading system to grade a set of images of human embryos, provided by an embryology lab. This system assigns embryos different grades based on several morphological features [22]. The researchers converted these grades into good-quality, fair-quality, or poor-quality labels for each image. The good- and poor-quality embryos were compiled into a training set and used to fully re-train Inception V1, which is the convolutional neural network the research team decided to integrate into STORK. Once Inception V1 was re-trained, it was tasked with classifying the ‘fair-quality’ embryos, now as either good- or poor-quality.

Inception V1 was trained with 12001 black-and-white images of 500x500 pixels: 6000 images of 877 good-quality embryos, and 6001 images of 887 poor-quality embryos, captured at precisely 110 hours post insemination. Each embryo was photographed with seven different focal depths, or ‘layers’.

Testing the resulting neural network, now called STORK, using a randomly selected test set of good- and poor-quality embryos showed that STORK could predict the quality of those embryos with an accuracy of 96.94% per image. This increased to 97.53% after averaging the predictions of multiple layers of the same embryo.

Attempting to classify the fair-quality embryos resulted in a skewed divide: 82% of the fair-quality embryo images were labeled good-quality, while only 18% were classified as poor-quality. Although skewed towards good-quality, several factors do suggest that this divide is correct. For example, the average patient age in each group fits: fair-quality images reclassified as “good” came from younger patients, on average, than fair-quality images reclassified as “poor”.

The project code and notes on how to implement and execute STORK are available on the STORK GitHub [17].

4.2 Inception V1

Inception V1 was the name of a neural network developed in 2014 for the ILSVRC14 (ImageNet Large-Scale Visual Recognition Challenge 2014) by Szegedy et al. It is a convolutional neural network that is 22 layers deep. The designers’ main goal was to attain a high accuracy while keeping the computational cost constant. It is also known as GoogLeNet.

The main distinctive feature of Inception V1 is known as the “Inception module”. These modules consist of six convolutional layers and a concatenation layer to merge their outputs. Inception modules are designed to extract features of different sizes on the same input image. To facilitate this, they include 1x1, 3x3, and 5x5 convolutional layers, all receiving the same input. The outputs of these convolutional layers are then stacked with the output of a parallel pooling path, which the designers decided to add due to the essential nature of pooling operations in neural networks [18]. The architecture of the first iteration of these models is shown in Figure 3.1.

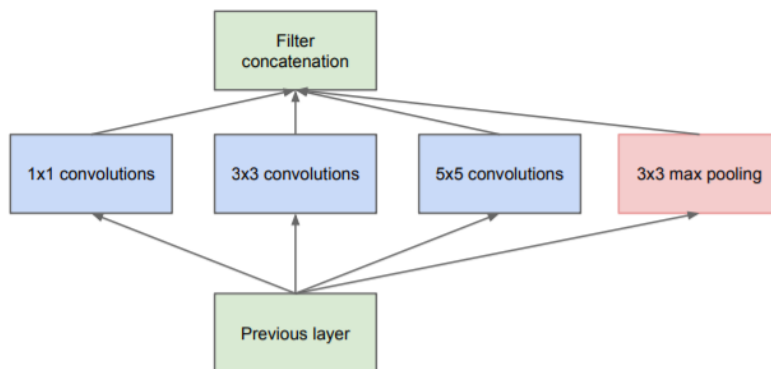


Figure 4.1: The naive structure of the V1 Inception modules. From [18].

The issue with this design is that merging the output of the pooling layer directly with those of the convolutional layers leads to a computational cost that was higher than acceptable. The number of outputs would increase

between modules, leading to computational blow-up. Therefore, the designers added 1x1 convolutional layers to reduce the dimensions of the data. These were inserted before the expensive 3x3 and 5x5 convolutional layers, as well as after the max-pooling layer. This choice kept computational costs in check. The final V1 Inception module design is shown in Figure 3.2.

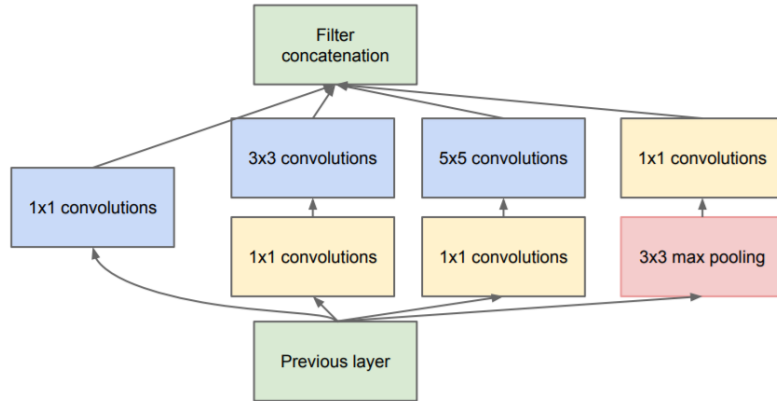


Figure 4.2: The new structure of the V1 Inception modules. From [18].

The architecture of Inception V1 consists of nine of these Inception modules in sequence, surrounded by some pre-processing and output processing layers. There are also some max-pooling layers in between the Inception modules to halve the dimensions of the data. There are two branches outside this sequence, which is shown in Figure 3.3. These branches are only active during the training phase, not during testing. They calculate the prediction error, commonly called the loss. The lower the loss, the better the network’s current performance. The weights of the inputs into the neurons are adjusted during training to obtain better performance by minimizing the calculated loss.

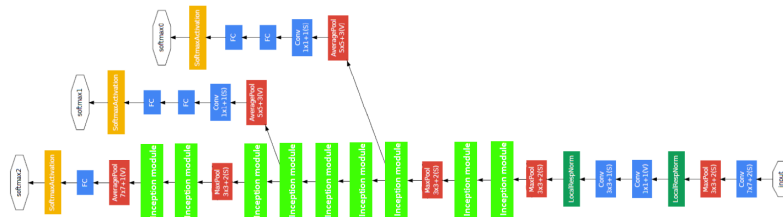


Figure 4.3: The architecture of Inception V1. Adapted from [21].

A full-size version of this figure is available in the Appendix.

4.3 Inception V3

Inception V1 was well-designed in terms of both speed and accuracy at the time of its invention. However, improvements were made over the years, and new insights reached. As a result, several iterations of Inception have been released over time. One of these is Inception V3, which is known for having a much lower computational cost than Inception V1. Inception V3 was presented in the same paper as Inception V2. The changes between V1 and V3 will, therefore, be presented without considering V2.

In Inception V3, the 5×5 convolutions inside several of the Inception modules were replaced by two 3×3 convolutions. In addition to this, in several Inception modules, the 3×3 convolutions were replaced by 3×1 and 1×3 convolutions in sequence. Both of these design changes were made to reduce computational cost: a 5×5 convolution is 2.78 times more expensive in terms of computational cost than a 3×3 convolution, so even two 3×3 convolutions are cheaper than a single 5×5 convolution. In addition to this, factorizing a 3×3 convolution into 1×3 and 3×1 convolutions is roughly 33% cheaper to compute [19].

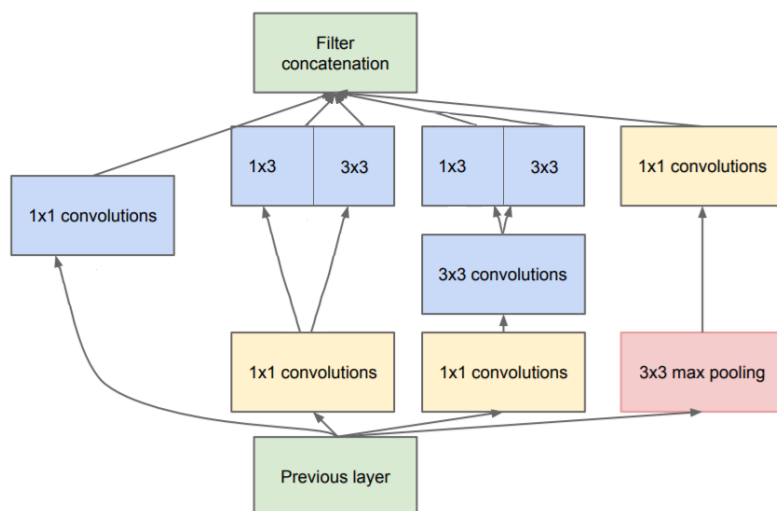


Figure 4.4: The structure of one of the V3 Inception modules. Adapted from [18].

As illustrated in Figure 3.4, the new Inception modules were made wider to avoid adding more depth; the 1×3 and 3×1 convolutions are positioned in parallel rather than in sequence. This avoids reducing the dimensions of the data too far, which would result in loss of information and lower accuracy. It ensured that the accuracy of the new Inception V3 was as high as that of V1, even with much fewer computations.

The architecture of Inception V3 consists of eleven Inception modules and only a single auxiliary branch to calculate the loss. It has 42 layers,

mostly due to the extra convolution layers in the Inception modules. Figure 3.5 shows a compressed schematic diagram of its architecture.

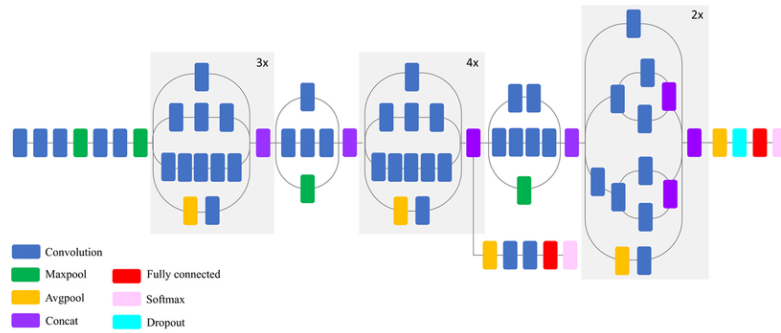


Figure 4.5: The architecture of Inception V3. From [12].

A full-size version of this figure is available in the Appendix.

Chapter 5

Method

5.1 Compiling the Dataset

For this research, a total of 64 images of 40 different human embryos were provided by the MCK Fertility Center. The dataset consisted of fully anonymous data. The images were captured using Geri®embryo incubators, which create time-lapse videos of embryo development using an integrated embryoscope. The images used were captured at 65 hours post-insemination in three focal depths. Each image was labeled with “selected” or “discarded” by embryologists employed at the MCK Fertility Center.

A note on reproducibility: the dataset analyzed in this thesis is not and will not be publicly available. Due to privacy and security concerns, the data will not be redistributed. However, the methods discussed below are not specific to this dataset and can be applied to any embryo imaging data.

The dataset provided by the MCK Fertility Center is structured quite differently to the datasets for which STORK is intended. One difference is that during this research, the MCK transferred embryos after three days of in vitro incubation, rather than the more common five days. The datasets STORK was trained with and tested on all contained five-day old blastocysts. The MCK did switch to transferring embryos after five days of incubation in May 2020, although the dataset only contained images of three-day-old embryos for the sake of consistency.

A second notable difference was that in related literature, the quality-classification datasets were divided into “good”, “fair” and “poor” quality embryos. The dataset provided by the MCK Fertility Center was divided into the more ambiguous “selected” (the embryos which were transferred into the mother’s uterine cavity) and “discarded” (the embryos which were not transferred). Sometimes, a couple produces several high-quality embryos, only one of which will be selected (unless the couple decides to freeze the others, so that they can use those in the future). Similarly, if there are only poor-quality but potentially still viable embryos, the best one will still be

transferred to the mother. Therefore, the “selected” dataset may contain relatively poor quality embryos, and the “discarded” subset may contain high-quality embryos.

In addition to this, the average age of the clientele of the MCK Fertility Center is quite high compared to other Fertility Centers, which can be seen from their relatively low success rate [1]. Age is a contributing factor to embryo quality, meaning that the MCK Fertility Center’s patients do, on average, produce lower quality embryos (which result in a successful pregnancy less often). Therefore, there is a chance for a poor-quality embryo to be transferred, introducing a degree of ambiguity to this split.

The images were cropped at 500x500 pixels, centered on the embryos. The dataset contained a total of 64 images of 40 embryos, split equally into good-quality and poor-quality.

5.2 STORK

Instructions on how to set up and work with STORK are available on the GitHub of the development team behind STORK [17]. The framework includes scripts for the classification itself, training and different accuracy measures. In order to work with STORK as intended by its creators, these were used throughout the stage of research where we worked with STORK without modification.

The framework came with a checkpoint of the Inception V1 network that the creators used, and the instruction to train this network using the provided training data ($N = 90$). The network was trained for a total of 500 iterations.

We first used STORK to classify the MCK Fertility Center embryos after only training the neural network inside the framework with the STORK training data. This was done in order to gain insight in the basic accuracy of STORK as well as its compatibility with the dataset provided by the MCK Fertility Center.

After this yielded a baseline (see Results for more information), we moved to re-training the final layer of STORK with the data provided by the MCK Fertility Center, to help account for the aforementioned differences between the MCK dataset and the datasets STORK was trained with. Instead of doing this within the script provided in the STORK framework, we decided to re-train the final layer of Inception V1 as a standalone network, which allowed for more control over the training process. In addition to this, it would enable us to more accurately compare the results of Inception V1 to those of Inception V3. Comparing Inception V3 as a standalone module to Inception V1 through the lens of STORK would be skewed, since the Inception V1 model used in STORK has been fully re-trained by the team behind STORK.

5.3 Standalone V1 and V3

For the purpose of testing the accuracy of both Inception V1 and Inception V3, a script was created based on the `retrain.py` script available via the Tensorflow GitHub [20]. It is available on the GitHub of the author of this thesis [9].

The script was modified specifically to facilitate easy K-fold cross validation. This was necessary due to the small size of the dataset; using 15% of the dataset for testing means a testing set of ten images. If even a single embryo in that set is anomalous or simply ambiguous in quality, it can have a significant impact on testing accuracy. The original `retrain.py` script randomly selects a number of images in each class to be used as a testing set; that was changed so that the original set is instead divided randomly into ten folds. Retraining the final layer is then done using one of those ten folds as a testing set, and the other nine as a training set. This can be repeated for each fold, resulting in ten accuracy scores. Especially with such small testing sets, these scores may vary wildly between folds. Therefore, averaging those provides a more reliable measure of algorithm accuracy than only executing the algorithm once, for a single randomly selected test set.

We continued training for 500 iterations each time, logging training and testing accuracy every five iterations. 500 iterations is quite low in the world of machine learning, but this already resulted in overfitting, and training for any longer would not increase accuracy.

Chapter 6

Results

The aim of this thesis is to evaluate the potential of using neural networks to qualitatively classify embryos for the MCK Fertility Center. In order to do so, we evaluated three variants for accuracy in embryo prediction: the STORK framework, which utilises a modified Inception V1; Inception V1 as a standalone module; and Inception V3.

6.1 STORK

Testing the accuracy of STORK’s pre-trained Inception V1 without re-training the final layer yielded the following results when tested on all the data supplied by the MCK Fertility Center (see Figure 6.1). The accuracy of using the trained model to predict STORK’s own test set has also been included for contrast. The accuracy of classifying the MCK Fertility Center’s dataset fluctuated around 0.5, no better than random. There was too much discrepancy between the dataset used to train STORK and the dataset provided by the MCK.

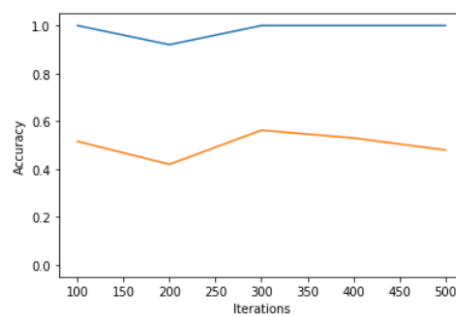


Figure 6.1: The accuracy of using STORK to classify STORK’s testing dataset (orange) and the MCK’s dataset (blue) after 100, 200, 300, 400 and 500 training iterations.

Figure 6.2, below, shows the confusion matrices corresponding to the results of attempting to predict the classes of the MCK Fertility Center's embryos. Interestingly, the majority of the incorrectly labeled embryos were poor-quality (*not transferred*) embryos labeled as 'good' (*transferred*).

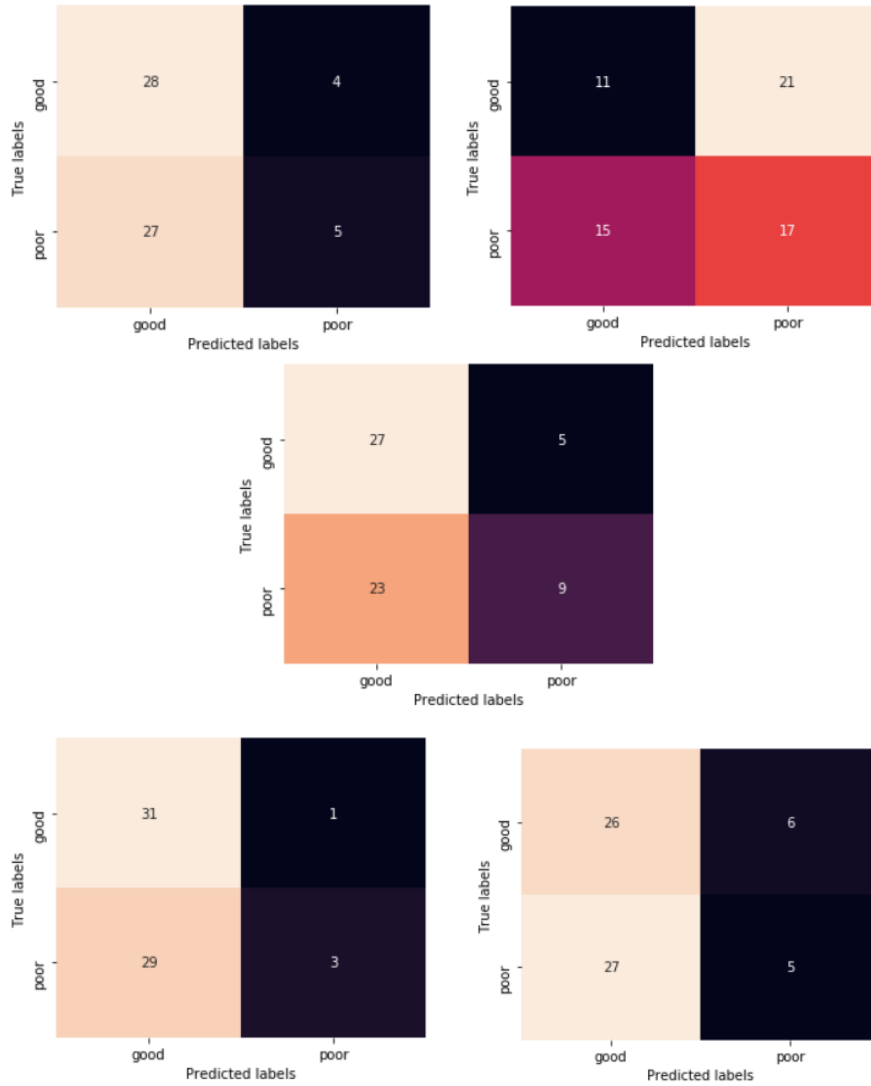


Figure 6.2: The confusion matrices showing the results of predicting the labels for the MCK dataset after training with STORK's training data. From top to bottom, left to right, the confusion matrices are shown in order for every 100th iteration.

6.2 Inception V1

Inception V1 as a standalone network was found to be both more accurate and faster than STORK; re-training the final layer of the Inception V1 for 500 iterations took minutes rather than two hours.

Using 10-fold validation to evaluate the accuracy of the model yielded accuracy scores which varied widely, as shown in Figure 5.1 below. The average accuracy of the model was 68%.

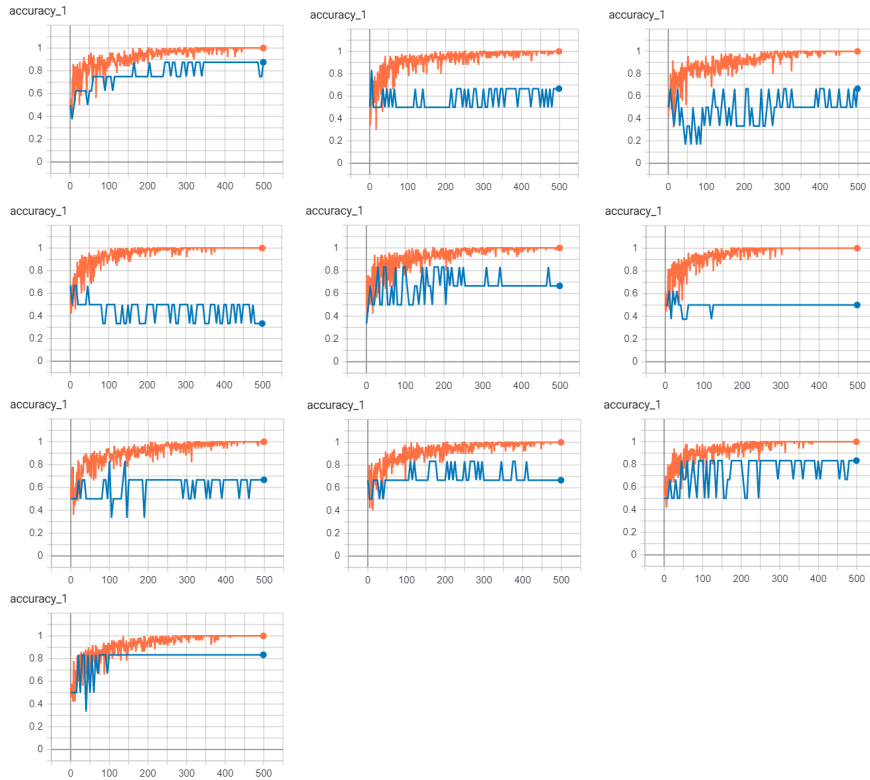


Figure 6.3: The accuracy of each fold, with the number of training steps on the x-axis and the accuracy on the y-axis. The orange lines represent training accuracy at different stages of the training process, the blue lines represent testing accuracy.

6.3 Inception V3

Inception V3 was more accurate than both STORK and Inception V1 as a standalone network, although it was slightly slower than Inception V1.

Using 10-fold validation to evaluate the true accuracy of the model yielded accuracy scores that varied widely, as shown in Figure 5.2 below. The average accuracy of the model was 78%.

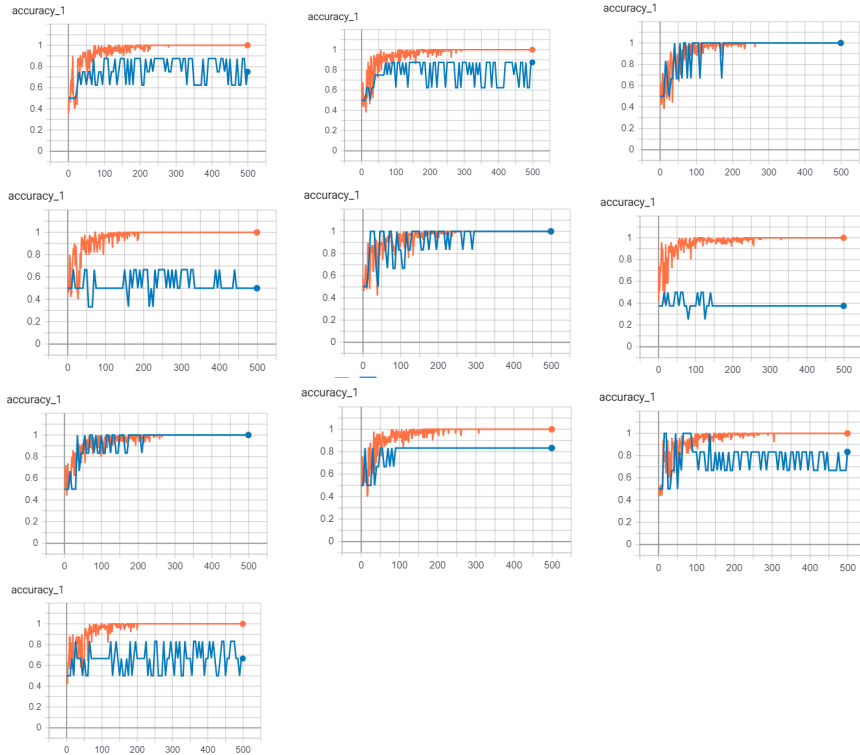


Figure 6.4: The accuracy of each fold, with the number of training steps on the x-axis and the accuracy on the y-axis. The orange lines represent training accuracy at different stages of the training process, the blue lines represent testing accuracy.

6.4 Combined observations

We recorded the results of a second run of 10-fold cross validation, using the same folds for both classifiers (see section A.3 of the Appendix). The information obtained in this way was reorganized into several different tables (see below), to show a number of statistical differences between Inception V1 and Inception V3’s performances.

To examine whether the difference in accuracy between Inception V1 and Inception V3 was statistically significant, we performed McNemar’s test, using the contingency table below (see Table 6.1).

	V1 Correct	V1 Incorrect	Total
V3 Correct	35	14	49
V3 Incorrect	8	7	15
Total	43	21	64

Table 6.1: A contingency table showing the results of using Inception V1 and Inception V3 to predict whether embryos were transferred or discarded.

The McNemar test statistic (Chi-square statistic value, with 1 degree of freedom) is used most often to calculate the statistical significance of the difference in performance between the two classifiers. However, if the sum total of occasions where the classifiers disagree is less than 25, this is inaccurate. Here, this sum total is $8+14=22$. Therefore, we instead used McNemar’s Chi-squared test with continuity correction, as proposed by Edwards [5]. With a significance threshold α of 0.05, this test revealed that the classifiers make errors in similar proportions, only on different instances of the test set ($\chi^2(1, N = 64) = 1.14, p > \alpha$).

However, the neural networks did respond differently to ‘poor-quality’ (not transferred) embryos, as shown in Table 6.2.

	V1 Correct	V3 Correct
good	21	22
poor	22	27

Table 6.2: A table showing the number of correctly classified images by both Inception V1 and Inception V3, per class.

Inception V1 and Inception V3 were similarly good at identifying a ‘good-quality’ (transferred) embryo correctly (66% and 69% accuracy, respectively), but Inception V3 was better at identifying ‘poor-quality’ (not transferred) embryos correctly, with 85% accuracy where Inception V1 had 67%.

Overall, both networks created a slightly skewed divide. Inception V3 classified a total of 27 images as ‘good’, and 37 as ‘poor’. Inception V1 divided the embryos more evenly, with 31 being classified as ‘good’ and 33 as ‘poor’.

As shown in the contingency table (Table 6.1), Inception V1 and Inception V3 were both incorrect in seven cases. These seven embryos are pictured in Figures 6.5 and 6.6. Notably, nearly half of these incorrectly classified embryos were located at the edge of their petri dishes, resulting in dark corners. This only occurred in 10.9% of the images in the dataset (7 out of 64).

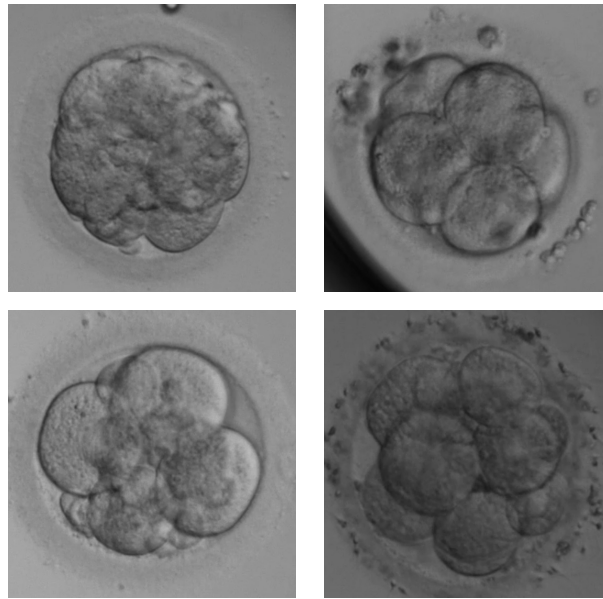


Figure 6.5: The four embryos incorrectly labeled ‘poor’ by both Inception V1 and Inception V3.

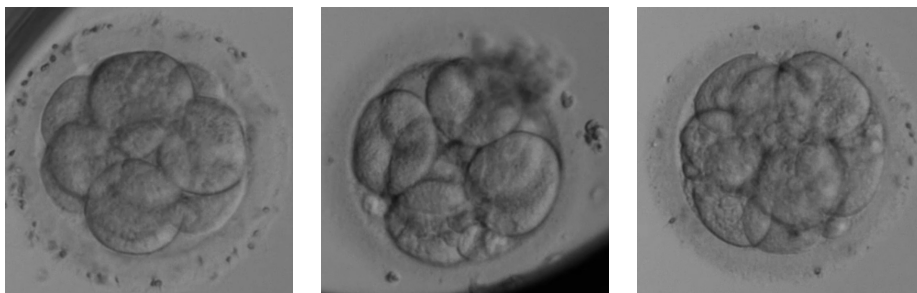


Figure 6.6: The three embryos incorrectly labeled ‘good’ by both Inception V1 and Inception V3.

Chapter 7

Discussion

Both Inception V1 and Inception V3 outperformed STORK, which was expected. STORK had been trained for a dataset containing images of five-day-old embryos divided into ‘good-quality’ and ‘poor-quality’, where the embryos in our dataset were labeled more ambiguously (‘transferred’ and ‘not transferred’) and were only three days old. The datasets were simply too dissimilar for STORK to make accurate predictions. However, in May 2020, the MCK Fertility Center switched from transferring embryos back to the intended mother after three days of incubation, to transferring them after five days. As the STORK framework was intended for classifying five-day-old embryos, and the dataset used for this thesis only contained three-day-old embryos, it is possible that the accuracy of STORK will increase when testing it on the embryos created after this switch. More research is needed to evaluate the effects of this change on the performance of all three of the networks used in this thesis.

Inception V3 had a higher accuracy than Inception V1, with 78% and 68%, respectively. However, they made similar mistakes, as shown by McNemar’s test, meaning that this difference is not statistically relevant. Therefore, their results will be discussed jointly.

Both Inception V1 and Inception V3 classified more embryos as ‘poor’ than ‘good’. This produces a skewed divide, although the original dataset had an equal divide. However, this skewed result may not be as incorrect as it seems at first glance. As mentioned before, the MCK Fertility Center’s clientele is older than average, which suggests that the embryos they produce will, on average, be lower-quality. Therefore, the likelihood of poor-quality embryos being transferred is higher than that of high-quality embryos not being transferred. It is possible that the high number of embryos being classified as poor is because they really are relatively poor in quality. However, more research is required to investigate this possibility, potentially including the patients’ age and/or whether the embryo transfer resulted in a successful pregnancy.

Keeping in mind that neural networks usually require several hundreds of images when transfer training, like we did, and several thousands of images when training from scratch, it can be safely said that our 64-image dataset was very, very small to attempt this kind of research with. This means that the results of our research have limited generalizability, which could be improved with further research by including a larger dataset.

Although the 78% accuracy we obtained with Inception V3 is not as high as the original developers of STORK obtained with their framework, it is surprisingly good when taking the aforementioned difficulties into consideration. Therefore, we conclude that using neural networks to classify embryos is *most likely* a viable option for the MCK Fertility Center. Future research is needed to observe the impact of having more data available, although we expect accuracy to improve quickly if more training data is available.

Chapter 8

Conclusions

The aim of this thesis was to examine the viability of using neural networks to classify embryos based on quality for the MCK Fertility Center. Based on the evaluation of the performance of three different structures (the STORK framework, Inception V1 and Inception V3), it can be concluded that neural networks are a potentially viable addition to the selection procedure.

The results indicate that Inception V3 is the most accurate, with an accuracy of 78%. Although this is not accurate enough to be included in the selection procedure as-is, obtaining such a high accuracy even with a small dataset shows that there is the potential for better results. This research raises the question of how accurate embryo classification using neural networks could be, if a larger dataset were available.

To better understand the implication of this research, future studies could address the comparative accuracy of these neural networks now that the MCK Fertility Center has switched to using five-day-old embryos for transfer. Especially STORK might be reconsidered, as that used five-day-old embryos during its training phase.

The field of embryonic quality assessment using neural networks is still in its infancy; there is little existing literature in this niche field of expertise. This thesis is the first study to examine the potential of using neural networks in this area in the Netherlands, although it will certainly not be the last.

Bibliography

- [1] Ivf cijfers 2018 per kliniek. <https://www.degynaecoloog.nl/wp-content/uploads/2019/12/IVF-cijfers-2018-per-kliniek.pdf>, December 2019.
- [2] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C. Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology & Visual Science*, 57(13):5200, October 2016.
- [3] Charles L. Bormann, Prudhvi Thirumalaraju, Manoj Kumar Kanakasabapathy, Hemanth Kandula, Irene Souter, Irene Dimitriadis, Raghav Gupta, Rohan Pooniwala, and Hadi Shafiee. Consistency and objectivity of automated embryo assessments using deep neural networks. *Fertility and Sterility*, 113(4):781–787.e1, Apr 2020.
- [4] Maureen Caudill. Neural networks primer, part i. *AI Expert*, 2(12):46–52, December 1987.
- [5] Allen L. Edwards. Note on the “correction for continuity” in testing the significance of the difference between correlated proportions. *Psychometrika*, 13(3):185–187, September 1948.
- [6] Craig Niederberger et al. Forty years of IVF. *Fertility and Sterility*, 110(2):185–324.e5, July 2018.
- [7] E. Santos Filho, J.A. Noble, M. Poli, T. Griffiths, G. Emerson, and D. Wells. A method for semi-automatic grading of human blastocyst microscope images. *Human Reproduction*, 27(9):2641–2648, June 2012.
- [8] Suzana Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3, 2009.
- [9] Ilse Arwert GitHub. <https://github.com/ilse-arwert/thesis>.
- [10] P. Khosravi, E. Kazemi, Q. Zhan, et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *npj Digital Medicine*, 2(1), Apr 2019.

- [11] Nick S. Macklon, Richard L. Stouffer, Linda C. Giudice, and Bart C. J. M. Fauser. The science behind 25 years of ovarian stimulation for in vitro fertilization. *Endocrine Reviews*, 27(2):170–207, April 2006.
- [12] Masoud Mahdianpari, Bahram Salehi, Mohammad Rezaee, Fariba Mohammadimanesh, and Yun Zhang. Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery. *Remote Sensing*, 10:1119, 07 2018.
- [13] G PALERMO. Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte. *The Lancet*, 340(8810):17–18, July 1992.
- [14] PerceptiLabs. <https://blog.perceptilabs.com/>.
- [15] Dina Ragab, Maha Sharkas, Stephen Marshall, and Jinchang Ren. Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ*, 7:e6201, 01 2019.
- [16] P.C. Steptoe and R.G. Edwards. BIRTH AFTER THE REIMPLANTATION OF A HUMAN EMBRYO. *The Lancet*, 312(8085):366, August 1978.
- [17] STORK GitHub. <https://github.com/ih-lab/STORK>.
- [18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. 2014.
- [19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.
- [20] TensorFlow GitHub. <https://github.com/tensorflow/tensorflow>.
- [21] Sik-Ho Tsang. Review: Googlenet (inception v1)- winner of ilsvrc 2014 (image classification), Jun 2019.
- [22] L.L. Veeck and Nikica Zaninovic. Grading criteria for human blastocysts. *An Atlas of Human Blastocysts*, 118, 01 2003.

Appendix A

Appendix

A.1 Inception V1 Architecture

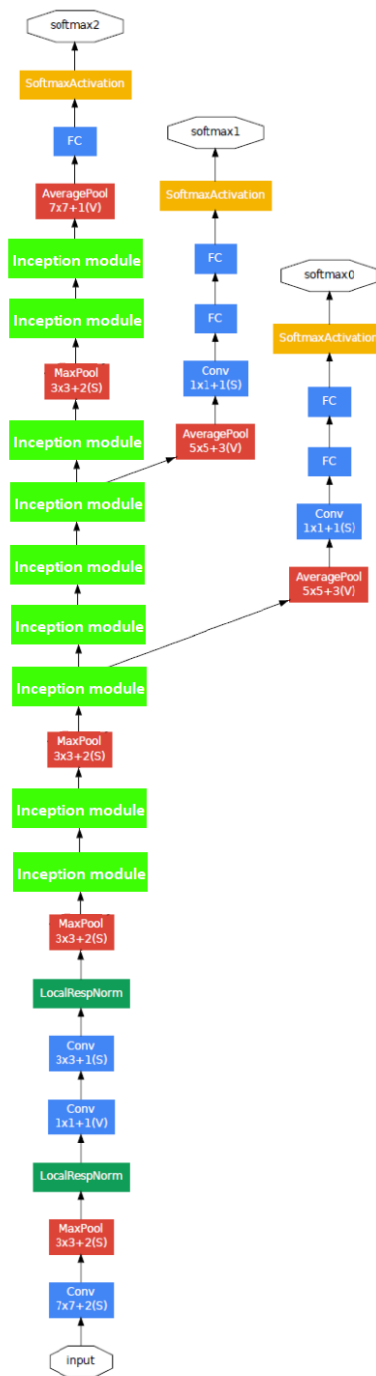


Figure A.1: The architecture of Inception V1. Adapted from [21].

A.2 Inception V3 Architecture

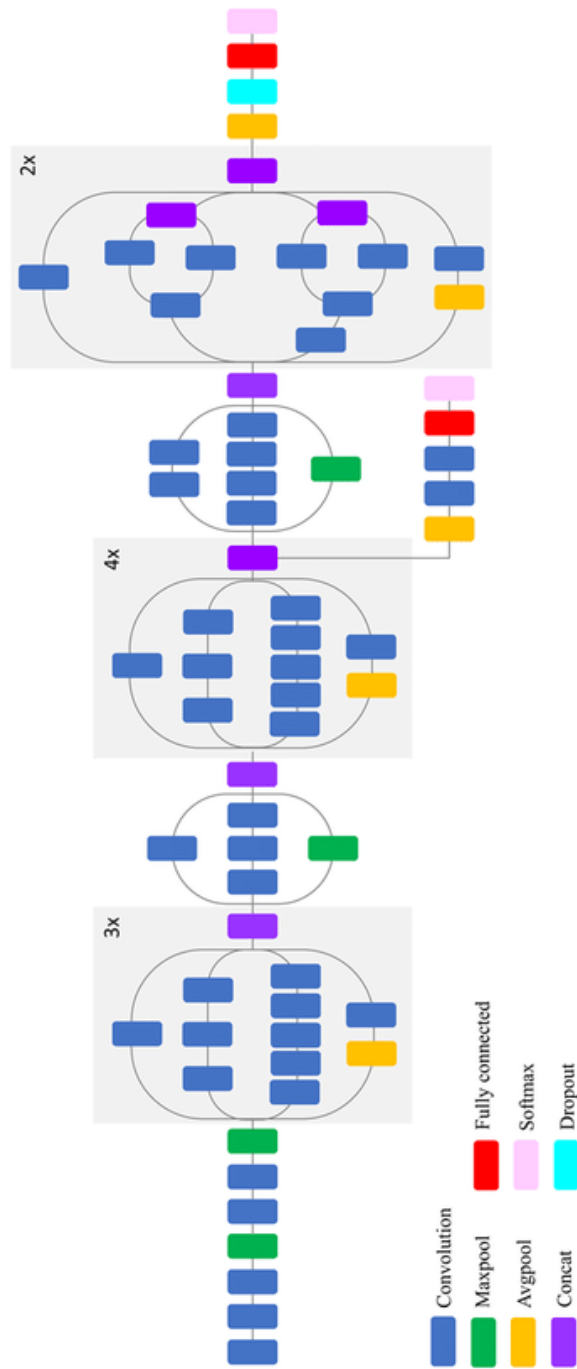


Figure A.2: The architecture of Inception V3. From [12].

A.3 Results of V1 and V3

‘+’ indicates that the image was classified correctly, and ‘-’ the opposite.

Image ID	V1	V3	Image ID	V1	V3
goodA	+	+	poorA	+	+
goodB	+	+	poorB	-	-
goodC	+	+	poorC	+	-
goodD	-	-	poorD	-	+
goodF	+	+	poorE	+	+
goodG	+	-	poorF	+	+
goodH	+	+	poorG	+	+
goodI	-	-	poorI	+	-
goodJ	+	-	poorJ	-	+
goodK	+	+	poorK	+	+
goodL	+	+	poorL	-	-
goodM	-	+	poorM	-	+
goodN	+	+	poorN	-	+
goodO	-	-	poorO	+	+
goodP	+	-	poorP	+	+
goodQ	+	+	poorQ	+	+
goodR	+	+	poorR	+	+
goodS	+	+	poorS	+	+
goodT	+	-	poorT	+	+
goodU	+	+	poorU	+	+
goodV	+	-	poorV	+	+
goodW	-	+	poorW	+	+
goodX	-	+	poorX	+	+
goodY	-	+	poorY	+	+
goodZ	+	+	poorZ	-	-
good1	+	+	poor1	-	+
good2	-	-	poor2	+	+
good3	+	+	poor3	+	+
good4	+	-	poor4	+	+
good5	-	+	poor5	-	+
good6	-	+	poor6	+	+
good7	-	+	poor7	-	+