A DEEP LEARNING APPROACH FOR PREDICTION OF IVF IMPLANTATION OUTCOME FROM DAY 3 AND DAY 5 TIME-LAPSE HUMAN EMBRYO IMAGE SEQUENCES

Mehryar Abbasi*† *Parvaneh Saeedi**‡ *Jason Au*$^{\text{II}\alpha}$ *Jon Havelock*$^{\text{II}\beta}$

* School Of Engineering Science
Simon Fraser University
Burnaby, BC, Canada
$^{\text{II}}$Pacific Centre for Reproductive Medicine
Burnaby, BC, Canada
Email:†mabbasib@sfu.ca, ‡psaeedi@sfu.ca, $^{\alpha}$jau@pacificfertility.ca, $^{\beta}$jhavelock@pacificfertility.ca

ABSTRACT

Various protocols have been developed to improve the success rate of In Vitro Fertilization (IVF). Earlier protocols were based on embryonic cell quality on embryos' third day. Newer protocols rely on the blastocyst quality (day-5 embryo). Artificial intelligence (AI) systems for automatic human embryo quality assessment seem to be the natural trend towards improving IVF's outcome. AI systems can potentially reveal hidden relationships between embryos' various attributes. To this date, most AI systems assess single blastocyst images. This paper proposes a novel approach that predicts the embryo implantation outcome from their time-lapse images. This approach consists of two models. One model evaluates each embryo based on its day-3 attributes, while the second model assesses the same embryo's day-5 image sequence. A Data Length Schedular (DLS) algorithm is developed addressing variations in blastocyst stage sequences' lengths. With an accuracy of 76.9%, the proposed system beats state of the art by 6%.

***Index Terms*—** IVF, Embryo, Deep learning, Artificial Intelligence, Implantation

1. INTRODUCTION

Reproductive diseases affect many couples around the globe [1, 2]. In-Vitro Fertilization (IVF) has been one of the most commonly used therapies for all causes of infertility. Even though substantial advancements have been made in IVF protocols, the success rate remains lower than desired [3]. In IVF, multiple OVAs retrieved from a patient's ovaries are fertilized in the lab environment and kept in controlled incubators for about five days before grading the embryos and transferring one or two embryos with the highest implantation potentials into the patient's uterus. Before substantial advances made in culture's quality and incubating technologies, many clinics preferred to transfer embryos on day three [4]. The rationale behind such a decision was that naturally, the women's uterus is the best incubator for an embryo to grow. Moreover, it seemed that often embryos would not continue to reach to day five stage [5]. So it would have been less risky and more cost-efficient to transfer embryos on day three [4]. However, often, patients seem to have a higher number of embryos on day three. Therefore, choosing an embryo on day three for transfer seemed to include some randomness. Because many of such embryos will not develop into blastocysts despite all clinical efforts [6], it has been observed that embryos that develop into blastocysts seem to have a higher chance of leading to a positive pregnancy [7]. Therefore, nowadays, most clinics use embryo quality assessments on blastocysts [8]. The most common method for embryo quality assessment is the morphological evaluation through visual attributes [9]. To quantify those attributes, multiple grading protocols have been developed. Evaluation methods based on day three and blastocyst have been the most popular methods for quality assessment. Not only do they coincide with the above transfer timing, but also they evolve to a substantially different stage of growth with relatively different yet distinctive morphological features [10, 7].

Today's primary strategy for optimizing IVF's outcome is to select and transfer embryos with the highest chance of success. The most common embryo selection technique is grading at the blastocyst stage (day five) [9], via assessing morphological attributes. Such morphological evaluation is done through visual assessment by experienced embryologists, making it time-consuming and subjective. Like many other medical image analysis applications, there has been a rising interest in using AI algorithms to analyze human embryo images. The following section describes these works.

1.1. Artificial Intelligent based human Embryo analysis

The works explored in this section investigate the practicality of AI-based models for automated analysis of human embryos in microscopic images. These works can be divided into two categories of (1) automatic grading and (2) implantation or live-birth outcome prediction.

The first category's work utilized AI-based systems to assign a grade to each embryo based on the morphological attributes of various components of it [11, 12, 13]. Approaches of the first group tried to address the central issue around morphological grading: subjectivity to the embryologist's knowledge and experience. The difference of opinions among expert embryologists could lead to grading inconsistencies [14] and lead to choosing embryos with lower implantation potentials. These methods try to minimize the grading inconsistencies by comparing their models' output against multiple experts' accumulated decisions. The grading methods based on morphological attributes of an embryo and the embryologists' experience help assess the quality of an embryo only to some extent; however, they are not wholly indicative of the outcome. For instance, multiple studies have shown that embryos graded as excellent only had a 50% success rate [13, 15, 16]. Indeed, there might be embryonic feature qualities that, for any reason, have not been observed, correlated, or included in the outcome prediction process yet. Therefore, it might be better to use an artificial system to directly use millions of features to predict the outcome.

The second category attempts to correlate between embryos physical attributes and their outcome after the embryos are transferred. The recorded outcome can be either in the form of implantation [17], or livebirth [18, 19, 20]. Such an actual outcome is used as the ground truth (GT), and AI models are created to predict the correct outcome through analyzing embryo images. One of the more recent developments in IVF technologies has been the introduction of time-lapse imaging systems within incubators, such as embryo scopes. These incubators have enabled continuous monitoring of embryos' development. Unfortunately, most AI-based developed systems only evaluate embryos quality based on a single shot of its blastocyst stage [18, 19, 11, 12]. Only a few methods have utilized time-lapse image sequences [13, 20]. Despite their efforts, these methods suffer from an oversight, which is the entire sequence's frames are assumed to have the same deciding attributes. The same processing unit is used to extract all embryo's image features, without any separation for the existing difference at various embryo stages. The fact that an embryo has different visual attributes at various development stages [21] is entirely overlooked. Multiple studies have shown that there is a direct correlation between the embryo's state at specific time-stamps and its viability/potentials [22, 23]. It is shown that day five transfers' success rate is correlated with the number of developed cells on day three [23]. [22] showed that timings of different stages are related to the embryo's final quality.

In this paper, we propose a system that exploits image sequence format and combines the morphological features of an embryo at two significantly different stages of its development to assess its quality.

1.2. Contributions

This paper presented an algorithm that utilizes an embryo's time-lapse image sequence to predict its implantation outcome. The presented algorithm consists of two CNN models. One model processes the day five embryo images, and the other evaluates the day three embryo frames. The two systems' results are combined to make a single decision regarding the embryo's quality (pregnancy outcome). When an algorithm focuses only on an embryo at one specific development stage, only the visual attributes existing at that stage will be processed, and perhaps the duration of the process will be dismissed. In [18, 19, 11, 12], embryos are evaluated only based on a single frame at their blastocyst stage. Other vital information, such as the length of time to reach the blastocyst stage, will be dismissed entirely. By processing an embryo during a window of time, we can embed the evolution speed into our model in addition to visual attributes. We proposed a Data Length Schedular (DLS) algorithm to regulate the training process by suppressing effects of a variable-length image sequence that naturally exists due to slow or fast-developing embryos.

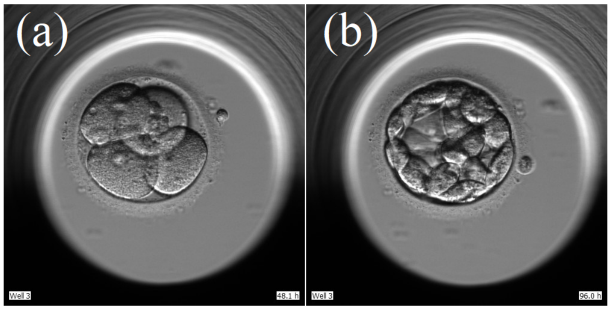

Fig. 1. Images of an embryo at the start of its (a) 3rd and (b) 5th days.

2. METHODOLOGY

2.1. Data

We used a dataset comprised of 130 time-lapse image sequences of individual embryos. Frames were captured at 15-minute time intervals by EmbryoScopeTM time-lapse incubator (Vitrolife). The image format is 8-bit 500×500 images. Each embryo day three sequence was comprised of the frames capture after the 48th hour but before the 72nd hour after fertilization. The day three sequences consisted of 96 consecutive images. The day five sequence were the frames captured after the 96th hour. The day five sequences ended when the embyros were transferred to the patients' uterus. Since different embryos' progress rate varies, the length of day five sequences was different (ranging between 67 and 96 frames). Two images of an embryo at its two different stages of day three and day five from one image sequence are shown in Fig. 1. The visual attributes of an embryo at these two times are significantly different. We automatically cropped and centered each image around each embryo and resized the image to 224×224 pixels. Out of our 130 sequences, 60 resulted in positive implantation, whereas the other 70 did not. The total number of images for day three and day five sequences were 12480 and 10097, respectively. We used 5-fold cross-validation for the training and testing of our model. We divided our dataset into five subsections of 26 samples. Four sets were chosen as the training set in each run, while the last set was used for testing. The algorithm was repeated five times with changing the test set in each round. The outcomes of all runs were averaged then.

2.2. Model architecture

Fig. 2 presents the CNN architecture proposed for this work. This architecture consists of two separate processing paths. The top path is designed to process day three embryo image sequences. The lower path is designated to evaluate day five image sequences. The output of each path is the implantation probability of the input embryo in the image sequence. Each convolutional layer of this structure is followed by a batch normalization layer [24] (momentum set to 0.01) and a ReLU activation function [25]. The dashed lines in Fig. 2 represent skip connections that use a convolutional layer with a 1×1 kernel size used for channel size matching.

2.3. Model Training

Each path in this model is trained separately. Relevant Images of each sequence are extracted and separated into two groups (day three, day five). This process is done based on the embryo's time

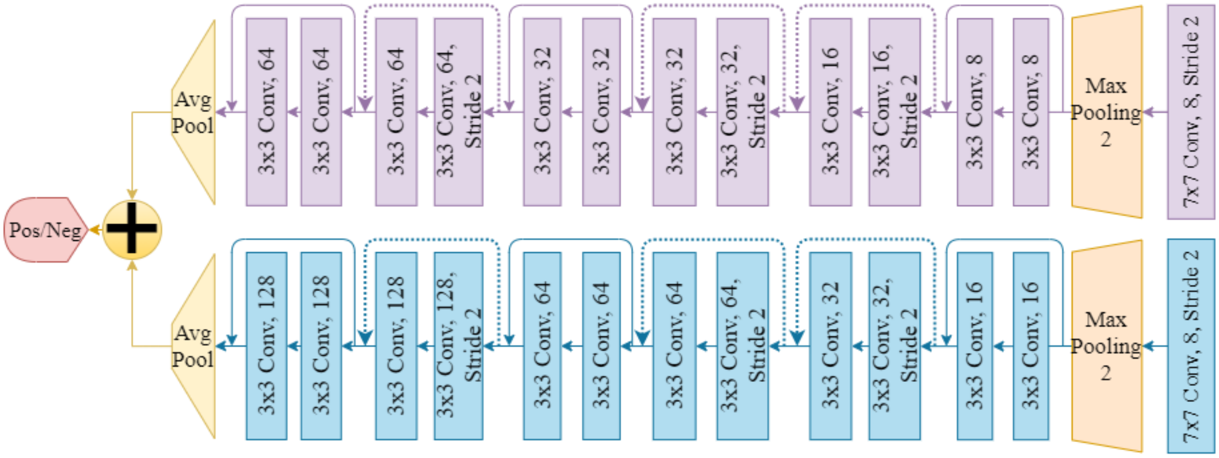

Fig. 2. Proposed CNN architecture for processing both day three (top) and day five (bottom) embryo image sequences.

displayed on the right bottom side of each frame (see Fig. 1). The displayed time area is cropped automatically and converted to text using Optical Character Recognition (OCR). When a sequence is called during training, one random image is selected from that sequence group and passed to the Data-batcher with that sequence label. If the embryo image sequence has a positive implantation outcome, the given label is positive (negative implantation outcome produces a negative label). In the consecutive times when the same sequence is called again, a new random frame is selected without having repeated selections. The no repeated selection scheme ensures that all frames are processed only once during the passage of one epoch. We have reached the choice of random selection by trial and error. Since the consecutive frames are too similar, if we train the model without a random selection of the frames, it will be quickly pushed toward an unwanted direction. Therefore it might steer away by overfitting or non-convergence. Training of each path is separate. When the day three path is being trained, the training data solely consists of images in the day three group. Consequently, during the day five model training, only the day five group images are used for training. The number of images for all of day three groups is the same and equals 96. However, the number of frames on day five section varies. Such image sequence length variation could cause performance degradation in the generality of the trained data. In a non-regulated training process, those sequences with a higher number of image frames could trigger data imbalance. Longer sequences usually have a lower progression rate, and therefore, the difference between consecutive frames could be small. Such a low variation from one frame to the next could be viewed as repetitive frames, triggering a bias imbalance in the training sequences. In order to suppress this potential issue, we propose Data Length Schedular (DLS) algorithm.

2.3.1. Data Length Schedular

DLS algorithm controls data from sequences in the training based on the training's progression; it spreads samples more evenly . The first DLS parameter is s. The training sequences are divided into s groups of $1/s$ percentile, $2/s$ percentile, $\ldots$, and s/s percentile in length. At the start of training, the first group is used. The next group replaces the current group if certain checkpoints are reached; the process continues until the last group is selected. There are two interchangeable modes for checkpoints: completion of every n epochs, the passage of p consecutive epochs without improvement in the validation loss. Here, we used the first mode with $s = 5$ and $n = 10$.

2.4. Model Testing

In order for the generated model to predict the implantation outcome for test embryo sequences, it has to go through the following steps:

1. The day three and day five images of the test sequence are extracted and separated into two groups. These images are extracted if their time puts them between 48-72th hour (day three) or after the 96th hour (day five). After applying OCR to each image's right bottom side, the displayed time would be converted to text, and the image time is retrieved.

2. Images in the day three group are passed to the day three model. For each image, a single number is produced. All of the numbers are averaged to create a single number representing the day three prediction (this number is used to evaluate the day three model performance).

3. The same operation is repeated; however, the day five image group is passed to the day five model.

4. The two outputs are then averaged with the manner shown in Fig. 2, outputting a single number.

5. The outputs are compared against the annotations of the origination embryo to calculate the model performance.

3. EXPERIMENTAL RESULTS

The presented model was trained and tested on Cedar cluster of Compute Canada [28]. The batch size for the training of both models was set to 256. The number of epochs for both operations was set to 100. The learning rate of both models was set to $1\text{e}{-4}$. The loss function used for training of both models was Binary cross-entropy loss. Adam optimizer [29] was used as the optimization algorithm.

The performance for both models were also calculated before their combinations. These calculations were done using the outputs in the second and third steps of the testing operation explained in Section 2.4. These results are shown in Table 1.

Table 1. Performance comparison of embyros' implantation outcome predictors.

Row No	Model	Percision [26]	Recall [26]	Jaccard-Index [27]	Accuracy [27]
1	Day three model	63.9	67.4	50.6	68.5
2	Day five model	70.6	69.0	52.6	69.2
3	Day five model + DLS	72.6	70.4	54.2	70.8
4	Combined Day three and Day five	72.6	72.3	56.7	72.3
5	Combined Day three and Day five + DLS	**79.6**	**76.4**	**61.8**	**76.9**
6	Image CNN classifer [17]	63.6	63.6	46.7	62.8
7	Image + Cell segmentation CNN classifer [17]	71.1	72.7	56	70.9
8	Handmade feature classifer [18]	61.5	60.5	44.0	62.0
9	Image + Morphological factors CNN [19]	70.2	71.4	55.3	74.3

Testing results for a day five model trained without the DLS algorithm are also shown in 3rd row of Table 1. In the model that used DLS, the s was set to 4. Furthermore, it was operated in its first mode with n set to 10. Comparing the results presented in the 2nd and 3rd row of Table 1 shows that using the DLS algorithm improves the model's performance. This improvement is intensified when the two-path model contains the RDS enabled day five model (row 7 against row 6 of Table 1).

Table 1 compares the performance of the proposed model against the state of the art. The third section (rows 8 and 9) of Table 1 displays related algorithms that predict the live-birth outcome. The first two sections (rows 1 to 7) contain algorithms that predict the implantation outcome. A positive implantation outcome refers to a positive pregnancy test. It may not be appropriate to compare these two types of algorithms against each other. However, they might be informative to the reader because they are in the same context.

4. CONCLUSION

This paper proposed a fully automated approach for predicting human embryo implantation outcomes from an embryo's time-lapse image sequence in IVF treatment. A CNN-based model was presented with two paths for separate analysis of day three and day five images. Each track produced a prediction for the outcome of the input image sequence. The final model combined these two predictions to make a more accurate prediction. Additionally, we introduced an algorithm (DLS) to tackle challenges caused by the length variation innate in an image sequence. These length variations are the effect of the slow-varying scene and could degrade the intelligent model's performance. The DLS algorithm reduced such degrading effects by controlling regulating the training data based on their lengths. Experimental results showed that the presented algorithm achieved the average implantation outcome accuracy of 76.9%. This model outperforms state of the art by 6% in accuracy when predicting the implantation outcome. Our model also outperforms the live-birth prediction method presented in [19] 6.5% in Jaccard-index and 2.6% in accuracy. These two measures are best means to quantify the performance of any classification system.

5. REFERENCES

[1] E Santos Filho, JA Noble, and D Wells, "A review on automatic analysis of human embryo microscope images," *The open Biomed. Eng. journal*, vol. 4, pp. 170, 2010.

[2] Marcia C Inhorn and Pasquale Patrizio, "Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century," *Human reproduction update*, vol. 21, no. 4, pp. 411–426, 2015.

[3] Silke Dyer, Georgina M Chambers, Jacques de Mouzon, Karl-Gösta Nygren, Fernando Zegers-Hochschild, Ragaa Mansour, Osamu Ishihara, Manish Banker, and Geoffrey David Adamson, "International committee for monitoring assisted reproductive technologies world report: assisted reproductive technology 2008, 2009 and 2010," *Human reproduction*, vol. 31, no. 7, pp. 1588–1609, 2016.

[4] Laura Rienzi, Filippo Ubaldi, Marcello Iacobelli, Susanna Ferrero, Maria Giulia Minasi, Francisco Martinez, Jan Tesarik, and Ermanno Greco, "Day 3 embryo transfer with combined evaluation at the pronuclear and cleavage stages compares favourably with day 5 blastocyst transfer," *Human Reproduction*, vol. 17, no. 7, pp. 1852–1855, 07 2002.

[5] Michael M Alper, Peter Brinsden, Robert Fischer, and Matts Wikland, "To blastocyst or not to blastocyst? that is the question," *Human Reproduction*, vol. 16, no. 4, pp. 617–619, 2001.

[6] PM Rijnders and CA Jansen, "The predictive value of day 3 embryo morphology regarding blastocyst formation, pregnancy and implantation rate after day 5 transfer following in-vitro fertilization or intracytoplasmic sperm injection.," *Human Reproduction (Oxford, England)*, vol. 13, no. 10, pp. 2869–2873, 1998.

[7] David K Gardner, Michelle Lane, John Stevens, Terry Schlenker, and William B Schoolcraft, "Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer," *Fertility and sterility*, vol. 73, no. 6, pp. 1155–1158, 2000.

[8] Abha Maheshwari, Mark Hamilton, and Siladitya Bhattacharya, "Should we be promoting embryo transfer at blastocyst stage?," *Reproductive BioMedicine Online*, vol. 32, no. 2, pp. 142–146, 2016.

[9] Claudio Manna, Loris Nanni, Alessandra Lumini, and Sebastiana Pappalardo, "Artificial intelligence techniques for embryo and oocyte classification," *Reproductive biomedicine online*, vol. 26, no. 1, pp. 42–49, 2013.

[10] Allison E Baxter Bendus, Jacob F Mayer, Sharon K Shipley, and William H Catherino, "Interobserver and intraobserver variation in day 3 embryo grading," *Fertility and sterility*, vol. 86, no. 6, pp. 1608–1615, 2006.

[11] Pegah Khosravi, Ehsan Kazemi, Qiansheng Zhan, Jonas E Malmsten, Marco Toschi, Pantelis Zisimopoulos, Alexandros Sigaras, Stuart Lavery, Lee AD Cooper, Cristina Hickman, 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.

[12] Tsung-Jui Chen, Wei-Lin Zheng, Chun-Hsin Liu, Ian Huang, Hsing-Hua Lai, and Mark Liu, "Using deep learning with large dataset of microscope images to develop an automated embryo grading system," *Fertility & Reproduction*, vol. 1, no. 01, pp. 51–56, 2019.

[13] Mikkel F Kragh, Jens Rimestad, Jørgen Berntsen, and Henrik Karstoft, "Automatic grading of human blastocysts from time-lapse imaging," *Computers in biology and medicine*, vol. 115, pp. 103494, 2019.

[14] Tatiana Puga-Torres, Xavier Blum-Rojas, and Medardo Blum-Narváez, "Blastocyst classification systems used in latin america: is a consensus possible?," *JBRA Assisted Reproduction*, vol. 21, no. 3, pp. 222, 2017.

[15] Yan-Yu Zhao, Yang Yu, and Xiao-Wei Zhang, "Overall blastocyst quality, trophectoderm grade, and inner cell mass grade predict pregnancy outcome in euploid blastocyst transfer cycles," *Chinese medical journal*, vol. 131, no. 11, pp. 1261, 2018.

[16] Shiping Chen, Hongzi Du, Jianqiao Liu, Haiying Liu, Lei Li, and Yuxia He, "Live birth rate and neonatal outcomes of different quantities and qualities of frozen transferred blastocyst in patients requiring whole embryo freezing stratified by age," *BMC Pregnancy and Childbirth*, vol. 20, no. 1, pp. 1–9, 2020.

[17] Reza Moradi Rad, Parvaneh Saeedi, Jason Au, and Jon Havelock, "Predicting human embryos' implantation outcome from a single blastocyst image," in *2019 41st Annu. Int. Conf. of the IEEE Eng. in Med. and Biol. Society*. IEEE, 2019, pp. 920–924.

[18] Alejandro Chavez-Badiola, Adolfo Flores-Saiffe Farias, Gerardo Mendizabal-Ruiz, Rodolfo Garcia-Sanchez, Andrew J Drakeley, and Juan Paulo Garcia-Sandoval, "predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning," *Scientific Reports*, vol. 10, no. 1, pp. 1–6, 2020.

[19] Yasunari Miyagi, Toshihiro Habara, Rei Hirata, and Nobuyoshi Hayashi, "Feasibility of predicting live birth by combining conventional embryo evaluation with artificial intelligence applied to a blastocyst image in patients classified by age," *Reproductive Medicine and Biology*, vol. 18, no. 4, pp. 344–356, 2019.

[20] D Tran, S Cooke, PJ Illingworth, and DK Gardner, "Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer," *Human Reproduction*, vol. 34, no. 6, pp. 1011–1018, 2019.

[21] Jacques Cohen and Kay Elder, *Human preimplantation embryo selection*, vol. 6, CRC Press, 2007.

[22] Karolina Fryc, Agnieszka Nowak, Barbara Kij, Joanna Kochan, Pawel M Bartlewski, and Maciej Murawski, "Timing of cleavage divisions determined with time-lapse imaging is linked to blastocyst formation rates and quality of in vitro-produced ovine embryos," *Theriogenology*, vol. 159, pp. 147–152, 2020.

[23] Catherine Racowsky, Katharine V Jackson, Natalie A Cekleniak, Janis H Fox, Mark D Hornstein, and Elizabeth S Ginsburg, "The number of eight-cell embryos is a key determinant for selecting day 3 or day 5 transfer," *Fertility and Sterility*, vol. 73, no. 3, pp. 558–564, 2000.

[24] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[25] Abien Fred Agarap, "Deep learning using rectified linear units(ReLU)," *arXiv preprint arXiv:1803.08375*, 2018.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] Lin Li, Yue Wu, and Mao Ye, "Experimental comparisons of multi-class classifiers," *Informatica*, vol. 39, no. 1, 2015.

[28] [Online]. Available: https://www.computecanada.ca/, ," .

[29] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.