Early prediction of the transferability of bovine embryos from videomicroscopy

Yasmine Hachani, Patrick Bouthemy, Elisa Fromont, Sylvie Ruffini, Ludivine Laffont, Alline de Paula Reis

EARLY PREDICTION OF THE TRANSFERABILITY OF BOVINE EMBRYOS FROM VIDEOMICROSCOPY

Y. Hachani[1], *P. Bouthemy*[1], *E. Fromont*[2]

[1]Inria, [2]Univ. Rennes, IUF, Inria, IRISA
France

S. Ruffini[4], *L. Laffont*[4], *A. De Paula Reis*[3,4]

[3]Ecole Nationale Vétérinaire d'Alfort
[4]University Paris-Saclay, UVSQ, INRAE, BREED
France

ABSTRACT

Videomicroscopy is a promising tool combined with machine learning for studying the early development of *in vitro* fertilized bovine embryos and assessing its transferability as soon as possible. We aim to predict the embryo transferability within four days at most, taking 2D time-lapse microscopy videos as input. We formulate this problem as a supervised binary classification problem for the classes *transferable* and *not transferable*. The challenges are three-fold: 1) poorly discriminating appearance and motion, 2) class ambiguity, 3) small amount of annotated data. We propose a 3D convolutional neural network involving three pathways, which makes it multi-scale in time and able to handle appearance and motion in different ways. For training, we retain the focal loss. Our model, named SFR, compares favorably to other methods. Experiments demonstrate its effectiveness and accuracy for our challenging biological task.

***Index Terms*—** video-microscopy, embryo, classification, CNN

1. INTRODUCTION

Most techniques used to study the mechanisms of embryonic development are incompatible with embryo survival. Video microscopy applied to bovine embryos produced by *in vitro* fertilization (IVF) (Fig.1), is a promising tool compatible, with survival, in association with the analysis power of machine learning techniques. It allows us to study the early development and to assess the transferability of *in vitro* fertilized embryos, i.e., the capacity to reach the blastocyst stage, suitable for transfer to a cow uterus. From the application point of view, having this ability to correctly predict on a large scale whether embryos can be transferred or not, is crucial for cattle breeding. With current methods, only 30% of transferred blastocysts actually result in pregnancy. This achievement should help reduce pregnancy failures and thus unnecessary inseminations.

However, two major problems arise: *i)* detailed analysis of each video is time-consuming for biologists and potentially limits a day-to-day practice, *ii)* to enable advanced biological studies on early mechanisms influencing embryo development, it is preferable to know the embryo transferability as soon as possible. Therefore, automating the transferability prediction by operating directly on videos of embryonic development is of key interest, while achieving the task the earliest possible.

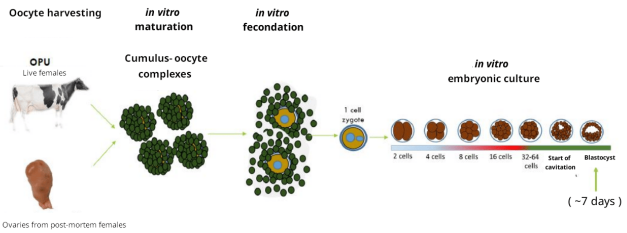

Fig. 1. The *in vitro* fertilization (IVF) process for bovine embryos.

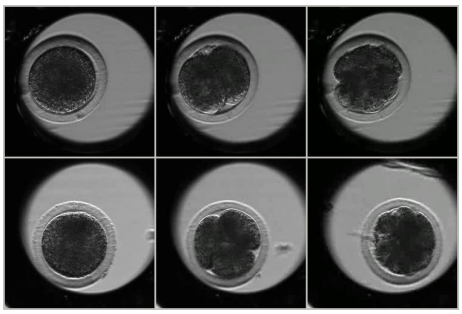

Fig. 2. Two sample videos of IVF bovine embryos, with three images taken at distant time instants (top row: an example from class transferable; bottom row: an example from class non transferable). The bovine embryo (dark grey), surrounded by the zona pellucida, is located in a micro-well (light grey) within the Petri dish (black). The embryo occupies a small part of the image.

Our overall objective is then to achieve a correct predic-

tion of the embryo transferability within four days at most, taking 2D time-lapse videos as input to be analyzed by 3D convolutional neural networks. This achievement will offer biologists new prospects such as better embryo sorting for further studies on embryonic genome activation (EGA) (at day 4) or morula formation (at day 5), decision-taking for earlier transfer into the uterus of a recipient female, which limits the duration of culture under sub-optimal conditions.

We formulate this problem as a two-class supervised classification one: the embryo is transferable (T class) or not transferable (NT class). This problem is challenging for three main reasons: 1) tricky embryo appearance and motion, 2) class ambiguity, and 3) small amount of annotated data. First, as illustrated in Fig.2, the microscopy videos display little contrast, involve a lot of noise and complex motion with transparency effects. The diverse embryos studied often exhibit poorly discriminating appearance between classes (see Fig.2 again), while the videos show complex morphological and temporal processes. Second, the intra-class variability is high, in the sense that the trajectories of the embryo development may substantially vary within a given class. Conversely, the inter-class distance is low. Indeed, observed development of two embryos from the two different classes may be fairly similar for the four first days. Third, because labeling is costly, there is only a limited database of videos labeled transferable or not transferable.

The rest of the paper is organized as follows. Section 2 describes related work. In Section 3, we present our 3D network with three pathways, and the associated training losses. Section 4 reports quantitative and comparative evaluations of the method. Section 5 contains concluding remarks.

2. RELATED WORK

Biologists have been working on embryonic development and phenotype formation for several decades. Thanks to IVF, significant advances have been achieved leading to major progress in medicine and breeding. Video-microscopy for bovine embryos has enabled biologists to observe different development trajectories leading to distinct phenotypes.

In human artificial reproductive medicine, the objective is to reduce the need for multiple pregnancies and losses during pregnancy. In [9], a first human embryo selection model was introduced based on manual annotation and a decision tree. Then, the authors of [20] developed an embryo selection model using a single static image captured by light microscopy. If the results of these studies were promising, the learning process was carried out *retrospectively*. Indeed, the pregnancy outcomes were predicted from embryos that were previously manually selected by an expert for transfer, leaving a large number of embryos outside the studies. These results may therefore include a certain level of over-fitting, as well as a number of biases related to embryo manipulation and transfer. Recently, a deep learning approach has been adopted in [1] to select human embryos from time-lapse image sequences acquired over five days, knowing that the video acquisition started around 24 hours after insemination. They used the Inflated 3D ConvNet (I3D) of [2] followed by a recurrent LTSM network. However, training is still carried out retrospectively. It leverages a very large dataset (about 100,000 videos).

Deep learning (DL) has also been used on human embryo videos to tackle different problems. For instance, in [5], the authors proposed a vector quantized variational autoencoder (VQ-VAE) to segment blastomere instance. Latest work mostly focuses on characterizing the different stages of embryo development, namely cell cleavage with intermediate stages defined by cell count (from 1-cell to 4-cell, sometimes up to 8-cell count), morula and blastocyst. In [7], the authors elaborated a development stage detector based on a 2D-CNN followed by a LSTM network classifier and added a synergic loss to learn embryo-independent features. In [10], the development stage classification was improved with EmbryosFormer, a three-headed model designed as an encoder-decoder deformable transformer inspired from Deformable DETR [23]. The authors of [15] adopted a different approach using the object detection technique YOLO v5 [11] and performing cell counting.

Bovine embryos are more difficult to study than their human counterparts, since their cells are darker, which makes, for example, cell counting quite difficult. Besides, biologists have reported [12] that the development trajectories may reflect different adaptation mechanisms and aptitudes for future gestation. In [12], the authors proceeded *prospectively*: they first characterized the trajectories and verified their biological interest, limiting the biases of the retrospective studies mentioned above. Observations based on the embryonic morphokinetic features have led to distinguish several families of trajectories distributed in transferable embryos (here, named T class) and non-transferable embryos (here, named NT-class). A prediction model was defined, based on random forest, leveraging many detailed annotations for each video.

Embryonic development could be seen as a form of action (in the computer vision terminology), and then, our classification problem could be seen as an action classification one. We therefore briefly review work on action recognition in videos since the advent of deep learning. The pioneering work [17] introduced a two-stream convolutional network taking both images and optical flow fields as input to leverage appearance and motion for action recognition. However, we experienced that optical flow was poorly estimated on bovine embryo videos. By considering the spatio-temporal video as a 3D volume, 3D convolutional networks have been extensively adopted since then for action recognition, as shown for instance in [2] with the inflated 3D convnet or in [19] with a 3D ResNet. In [3], the authors proposed a 3D network comprising two pathways, a Slow one devoted to appearance information with input video at a low frame rate, and a Fast one

with input video at high frame rate to better capture motion. This last model will be inspirational in our work.

To the best of our knowledge, the model we propose is the first DL-based model devoted to bovine embryos that are more difficult to handle. In addition, we focus on the transferability of IVF embryo in a prospective way, and the earliness of the prediction is an essential aspect of our work. Consequently, we consider a shorter period of embryonic development. Finally, we only leverage one annotation per video, that is, its class, transferable or not transferable, and a limited amount of annotated videos.

3. MODEL DESCRIPTION

As stated above, we address a two-class classification problem to predict the transferability of the IVF bovine embryos. The two classes are transferable embryos (T class - embryos with a potential of establishing a pregnancy, they can be transferred into a female recipient) *vs* non-transferable embryos (NT class - embryos with no or very low potential of establishing a pregnancy, they should not be transferred).

In practice, the expert biologist annotates the videos on a longer temporal basis than 4 days, to up to six or even eight days of the embryo development. This is nevertheless a light annotation, one label (the class) per video. In some cases, the expert biologist may need to understand the whole evolution of the embryo over the full video to decide on the class, as development trajectories may be similar up to a certain stage. On our side, we take these annotations for the same videos but restricted to four days of development. This explains why the inter-class distance may be small when considering videos of 4-day development only. Consequently, automatically predicting transferability at four days, i.e., carrying out our binary classification, is a complex video analysis task. In addition, only a small set of annotated videos is available, the images are noisy and poorly contrasted, subject to transparency effects, and motions in the video are not easy to identify.

3.1. Network architecture

We have designed a 3D network for our two-class classification problem in order to achieve early prediction of IVF bovine embryo transferability. We believe that a 3D convolutional network is more adapted to properly capture the spatio-temporal features characterizing the embryonic development, than for instance a recurrent neural network. Indeed, the morphokinetic features are quite intricate and an embryo development is not as smooth along time as a human action in a video. It is mainly specified by a few discrete events corresponding to the cell divisions, with rather random local motions in between.

Our 3D network presented in Fig.3 involves three pathways combined, with directed lateral connections. As in the SlowFast network [3], we have the *Slow* pathway taking the input video at a low frame rate and mainly dedicated to capture spatial features in images, the *Fast* pathway with input video at a high frame rate mainly devoted to the temporal features. The *Fast* pathway has a fraction β of channels and a temporal resolution α times higher than the *Slow* pathway. We call the third pathway *Regular*. It takes the input video at the same rate as *Fast* pathway, but it involves more channels in each layer. As motivated in the ablation study (Section 4.3), we use ResNet18 for the three pathways to build a light 3D network, which speeds up training and mitigates over-fitting. We replaced all ResNet batch normalization layers with group normalization layers [21], since group normalization is at least as good as batch normalization when trained with small or medium batch sizes, and it allows us to use Pytorch Lightning gradient accumulation technique efficiently.

We found the use of the three pathways beneficial, because appearance and motion are rather intertwined due to the transparency of cell membranes and the fact that we observe 2D projections, partly overlaid, of 3D cells. The three pathways bring complementary ways of handling appearance and motion to provide the right prediction. The outputs delivered by the last layer of each pathway are concatenated to feed the classifier. Directed lateral connections are included between pathways, We focused on two combinations of the lateral connections. The first one is illustrated in Fig.3 and comprises a connection from the *Regular* to the *Fast* pathways and from the *Fast* to the *Slow* pathways. The second one involves a fusion from the *Regular* to the *Slow* pathways and from the *Fast* to the *Slow* pathways; the *Regular* and the *Fast* pathways are not connected. We selected the first combination as explained in the ablation study (Section 4.3). We call our method SFR.

3.2. Loss function

We could adopt different loss functions. Since our data are unbalanced between the two classes *T* and *NT*, we have considered the focal loss [6], initially introduced for the object detection task. The focal loss can contribute to correct this imbalance, while focusing on the most difficult examples. The focal loss writes:

$$\mathcal{L}_f(\mathbf{v}, y) = -\sum_{c=1}^{2} \alpha_c (1 - \hat{p}(y_c|\mathbf{v}))^\gamma p(y_c|\mathbf{v}) \log \hat{p}(y_c|\mathbf{v}), \quad (1)$$

where $\mathbf{v}$ denotes the video input, y_c one of the two classes, $\hat{p}(y_c|\mathbf{v})$ the predicted probability of having class c given video $\mathbf{v}$, and $p(y_c|\mathbf{v})$ the true one, equal to 1 for the right class c regarding $\mathbf{v}$ since we are dealing with supervised classification. In addition, α_c is the weight for class c, γ the focusing parameter. The larger γ, the less importance is put to well-classified samples.

As reported in the ablation study, we also investigated the cross-entropy loss [22], defined for a binary classification by:

$$\mathcal{L}_{ce}(\mathbf{v}, y) = -p(y_c|\mathbf{v}) \log \hat{p}(y_c|\mathbf{v}) - (1 - p(y_c|\mathbf{v})) \log(1 - \hat{p}(y_c|\mathbf{v})). \quad (2)$$

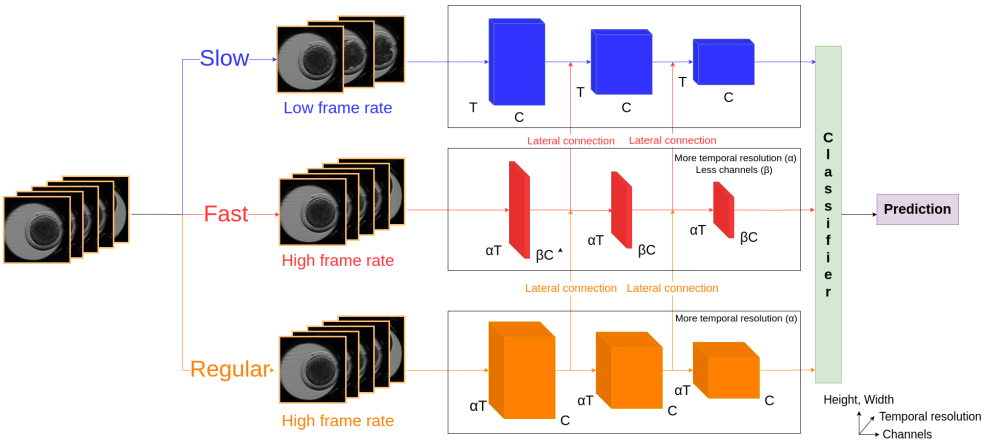

Fig. 3. Our 3D model combines three pathways, *Slow* and *Fast* and *Regular*, all implemented with 3D ResNet18 of different configurations. Input video is given at a lower temporal rate for the *Slow* pathway. The three outputs are combined before providing the prediction. Our SFR model includes directed lateral connections between the pathways as drawn in the figure.

3.3. Data augmentation

Various strategies can be adopted to deal with the lack of data. Data augmentation is a classical one [16]. Here, we can consider data augmentation applied to photometric, spatial, or temporal features of the video [14]. In practice, we only applied basic image manipulations to every frame of the videos: Gaussian noise addition, Gaussian blur, image flipping, image transpose, image cropping. All images in a given sequence are modified in the same way. We have thus multiplied the total number of videos in the training set by 20.

4. EXPERIMENTAL RESULTS

4.1. Video capture and video dataset

The videos are acquired as follows. The oocytes recovered from slaughterhouse ovaries and matured *in vitro* are brought into contact with frozen-thawed semen in a culture dish defining the starting point of the biological development of the embryos [12]. The embryos are put in microwells of Petri dishes around twenty-two hours after the *in vitro* fertilization. Each Petri dish contains sixteen microwells. It is placed into the PrimoVision system that comprises a simple transmission light microscope, and is filmed by the incubator camera.

The PrimoVision system takes a picture of the Petri dishes every fifteen minutes over eight days, and delivers 2D time-lapse video sequences that are subsequently divided into sixteen videos, one per embryo. Each video is annotated by a biologist with the *T*-label or the *NT*-label. For the purpose of our work, we have retained only the video footage of the first four days of embryo development. Since video acquisition begins only 22 hours after IVF, each processed video covers a period of 3 days and comprises around 300 images. The video dataset includes 947 videos, distributed into 763 for training and validation and 184 for test. Each set includes around 65% of non-transferable embryo videos.

4.2. Implementation details

Each model was trained using the AdamW optimizer [8], with a learning rate of 10^{-4} and the other parameters kept at their default values. We applied a cyclic learning rate scheduler as recommended in [18]. We trained the models using mini-batches of 32 samples artificially created thanks to the accumulate gradients technique, implemented in PyTorch Lightning, that accumulates gradients of small batches before performing a backward pass. We apply the stochastic weight averaging (SWA) [4] technique, which improves the generalization of our models by averaging the network weights at different, well-chosen epochs. We use early stopping to end training when the loss computed on the validation set increased ten epochs in a row. Then, we select the model at the epoch with the highest accuracy on the validation set (supervised training).

4.3. Ablation study

We have carried out an ablation study on the components of our model. Firstly, regarding the combination of the lateral connection, the first option (connection from the *Regular* to the *Fast* pathways and from the *Fast* to the *Slow* pathways) provided a better accuracy. Therefore, this is the one we will be using next.

We conducted an ablation experiment on the depth of the ResNet network to be used. We tested two possible depths, 18 layers and 50 layers, which would allow us to still have a rather light model. We trained our SFR model with ResNet18 and ResNet50 modules. For this experiment, we simply took the cross entropy loss, pending a decision on the γ parameter

SFR model	
Module	**Acc**
with 3D-Resnet18	72.9
with 3D-Resnet50	71.6

Table 1. Results on the binary classification in terms of accuracy (*Acc*) obtained by our SFR model with ResNet18 and ResNet50 modules using the cross entropy loss. We have carried out a dozen evaluations each time for different training seeds and folds, and we provide the average.

of the focal loss. Results on the binary classification (transferable *vs* non-transferable) are reported in Table 1. Since ResNet-18 results are slightly better and it is even lighter, we selected this architecture.

We omitted the directed lateral connections between the three pathways of our SFR model. We call SFR Late Fusion the model without any lateral connections. We use ResNet18 as motivated above. In the same time, we investigated both the cross entropy loss and the focal loss, thus combining these two ablation experiments. For the focal loss, we took $\gamma = 2$, as justified below. We set $\alpha_1 = 1.25$ and $\alpha_2 = 0.833$ according to the frequency of the two classes. Results are reported in Table 2. In all cases, the focal loss leads to increased performance. Furthermore, SFR performs better than SFR Late Fusion, which shows the importance of lateral connections.

SFR and SFR Late Fusion models	
Model and loss function	**Acc**
SFR(CE)	72.9
SFR(FL)	75.6
SFR Late Fusion(CE)	72.5
SFR Late Fusion(FL)	73.5

Table 2. Results in terms of accuracy (*Acc*) obtained by SFR and SFR Late Fusion with the cross entropy loss (CE) and focal loss (FL) for $\gamma = 2$. We have carried out a dozen evaluations each time for different training seeds and folds, and we provide the average.

SFR model	
Gamma value	**Acc**
SFR(FL) ($\gamma = 1$)	73.1
SFR(FL) ($\gamma = 2$)	75.6
SFR(FL) ($\gamma = 3$)	73.4

Table 3. Results on the binary classification in terms of accuracy (*Acc*) obtained by our model SFR (involving ResNet18) with focal loss and $\gamma \in \{1, 2, 3\}$. We have carried out a dozen evaluations each time for different training seeds and folds, and we provide the average.

Our last ablation experiment dealt with the setting of parameter γ of the focal loss function. We tested the focal loss for three values of parameter γ, $\gamma = 1, 2,$ and 3, knowing that the case $\gamma = 0$ is somehow equivalent to train the model with a weighted binary cross-entropy loss. We carried out the experiment on the binary classification with our SFR model. We report the results in Table 3. We can conclude that the best choice is $\gamma = 2$. By the way, the authors of [1] made the same choice.

4.4. Comparative experiments

We have carried out comprehensive comparative experiments on the early prediction of bovine embryo transferability. To evaluate the performance of all methods, we consider the following metrics: overall accuracy *Acc*, precision P_T (resp. P_{NT}) and recall R_T (resp. R_{NT}) for the T (resp. NT) class. We performed the binary classification with SlowFast [3] (the ResNet18 version of the code) and a classical 3D-RestNet18, using for both the focal loss, since this loss is more adapted to our problem as demonstrated in Table 2. We train the two methods on our training dataset. This yields a comparison between these two models and ours. In addition, we built a baseline model comprising a 2D convolutional neural network (CNN) followed by a recurrent neural network (RNN). The 2D CNN is implemented with a ResNet18 that learns spatial features on images. A recurrent GRU neural network handles the time dimension of the embryo video, taking as input the successive output of the 2D CNN. The output of the last cell of the GRU is sent to a fully connected layer to obtain the classification prediction.

Comparative results for all the tested models are collected in Table 4 with the use of the focal loss for all methods. As expected, the 2D network involving a recurrent neural network underperforms all the 3D networks. Our SFR method obtains the best accuracy rate, followed by ResNet18. In addition our model is much more stable than the others, on the different metrics, which is crucial. SlowFast does not perform as expected, probably due to the particular nature of the videos processed, very different from those videos considered in action recognition. In addition, our SFR method has the best precision score for the T class and the best recall score for the NT class, which is very important for the target application. Indeed, for cattle breeding, it is essential to predict transferable embryos correctly, in order to avoid unnecessary pregnancies by transferring non-transferable embryos.

We have also compared our method with others regarding the computation time to make a prediction, still in Table 4. For each model, we computed the average time for one inference by repeating 1000 predictions on a NVIDIA RTX A500. As expected, the 2D CNN with GRU is the fastest at inference, followed by SlowFast which is faster than a simple 3D-ResNet as shown in the original paper [3]. Next come 3D-ResNet18 and our SFR model, whose average time pre-

Model	Acc	P_T	R_T	P_{NT}	R_{NT}	Average Prediction Time
2D-Resnet18+GRU	67.6±4.1	54.6±5.6	48.4±19.9	74.2±5.9	78.3±8.6	0.037s
3D-Resnet18(FL)	74.6±4.0	64.1±6.5	67.4±12.4	82.0±4.6	78.6±7.8	0.205s
SlowFast(FL)	73.1±2.6	63.9±4.0	55.2±9.3	77.3±3.2	82.8±3.4	0.077s
Ours - SFR(FL)	**75.6±1.5**	**66.8±3.3**	62.5±3.4	80.1±1.1	**82.8±2.9**	0.294s

Table 4. Comparison of results obtained for the models 2D-Resnet18+GRU, 3D-Resnet18, SlowFast [3], and our model SFR, all models with the focal loss ($\gamma = 2$). We have carried out a dozen evaluations each time for different training seeds and folds, and we provide the mean and standard deviation for the accuracy, precision and recall for the T and NT classes.

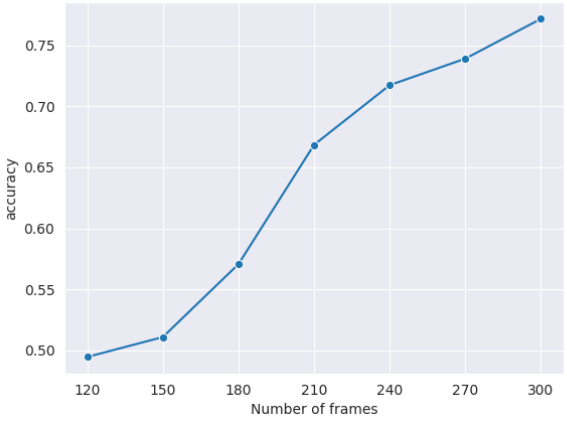

Fig. 4. Accuracy of early prediction for our SFR model with focal loss for the following range of video length: from 120 frames (about two days 1/4 of embryo development) to 300 frames (about four days).

diction is increased by approximately $89ms$ compared to the 3D-ResNet18, but this is not a problem for our application.

4.5. Earlier prediction

We wanted to check whether it is possible to make a prediction even before four days have elapsed. Therefore, we evaluate the performance of our SFR model with focal loss, when performing less-than-4-day Transferable *vs* Non transferable prediction. To do this, we test the model with increasingly shorter videos, removing the last thirty frames each time, which corresponds to removing information occurring during seven hours and a half. We stop testing at 120 frames, i.e., approximately two days 1/4 of embryo development. Results are plotted in Fig.4. We observe that the curve regularly climbs. Depending on the accuracy level acceptable for a given application, a usable prediction could be provided at an even earlier stage than the four days of the embryo development.

5. CONCLUSION

We have designed a 3D model to predict the embryo transferability within four days, taking 2D time-lapse microscopy videos as input. We state the problem as a supervised two-class classification one that remains however difficult. The three-pathway architecture makes our 3D model multi-scale in time regarding the input videos and able to manage appearance and motion in different ways. We successfully dealt with poorly discriminating embryo appearance and motion, in addition affected by transparency, and small inter-class distance. Experiments demonstrate the usefulness of directed lateral connections between pathways, and of the focal loss. We favorably compared our SFR model with other methods. SFR provides the best accuracy rates with much better stability than 3D-ResNet18 on all metrics. Thus, we are able to efficiently and accurately achieve the early prediction of bovine embryo transferability. Future work will be concerned with an explicit combination of classification earliness and accuracy in the training loss.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted using data available in our laboratory. No live animals or euthanised animals were used to create the original data. The semen was acquired from a commercial company and the cumulus oocyte complex were harvested from ovaries recovered *post-mortem* in a commercial slaughterhouse. Both these companies and our laboratory are based in France and state-approved. The necessary authorisations for the use of *post-mortem* biological material have been obtained from the responsible Ministry.

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge the collaboration of Dr. Véronique Duranthon and Brigitte Marquant-LeGuienne for the experimental protocol design for the embryo production. The authors declare no conflicts of interest. The production of the original embryo data was funded by CRB-Anim.

Yasmine Hachani's PhD grant is funded by Inria. Operation of the research project is also partly funded by the DIGIT-BIO program of INRAe.

8. REFERENCES

[1] J. Berntsen, J. Rimestad, J.T. Lassen, D. Tran, and M.F. Kragh. Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences. *PLOS One*, Feb. 2022.

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, June 2017.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K. He. SlowFast networks for video recognition. In *Int. Conf. on Computer Vision (ICCV)*, Seoul, Oct. 2019.

[4] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conf. on Uncertainty in Artificial Intelligence (UAI)*, August 2018.

[5] W.-D. Jang, D. Wei, X. Zhang, B. Leahy, H. Yang, J. Tompkin, D. Ben-Yosef, D. Needleman, and H. Pfister. Learning vector quantized shape code for amodal blastomere instance segmentation. In *Int. Symposium on Biomedical Imaging*, Cartagena de Indias, April 2023.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár Focal loss for dense object detection. In *Int. Conf. on Computer Vision (ICCV)*, Venice, June 2017.

[7] L. Lockhart, P. Saeedi, J. Au, and J. Havelock. Automating embryo development stage detection in time-lapse imaging with synergic loss and temporal learning. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Strasbourg, September 2021.

[8] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Int. Conf. on Learning Repesentations (ICLR)*, New Orleans, May 2019.

[9] M. Meseguer, J. Herrero, A. Tejera, K.M. Hilligsøe, N.B. Ramsing, and J. Remohí. The use of morphokinetics as a predictor of embryo implantation. *Human Reproduction*, 26(10):2658-2671, Oct. 2011.

[10] T.-P. Nguyen, T.-T. Pham, T. Nguyen, H. Le, D. Nguyen, H. Lam, P. Nguyen, J. Fowler, M.-T. Tran, and N. Le. EmbryosFormer: Deformable transformer and collaborative encoding-decoding for embryos stage development classification. In *Winter Conf. on Applications of Computer Vision (WACV)*, Waikoloa, January 2023.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, June 2016.

[12] A. De Paula Reis, M. Beghiti, S. Messoudi, B. Marquant-Le Guienne, L. Laffont, S. Ruffini, E. Canon, P. Adenot, N. Le Brusq, V. Duranthon, and A. Trubuil. Identification and mathematical prediction of different morphokinetic profiles of *in vitro* developed bovine embryos. In *34rd Meeting of the Association of Embryo Transfer in Europe*, hal-02737515, Nantes, Sep. 2018.

[13] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv:*1706.05098, 2017.

[14] M.C. Schiappa, Y.S. Rawat, M. Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1-37, July 2023.

[15] A. Sharma, M. H. Stensen, E. Delbarre, M. Siddiqui, T. B. Haugen, M. A. Riegler and H. L. Hammer . Detecting human embryo cleavage stages using YOLO V5 object detection algorithm. In *Nordic Artificial Intelligence Research and Development (NAIS)*, Oslo, June 2022.

[16] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:60, 2019.

[17] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, Montreal 2014.

[18] L. N. Smith. Cyclical learning rates for training neural networks. In *Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, March 2017.

[19] D. Tran, H. Wang, L. Torresani, J. Ray, Y. Le Cun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, June 2018.

[20] M VerMilyea et al. Development of an articial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. *Human Reproduction*, 35(4):770-784, April 2020.

[21] Y. Wu and K. He Group normalization. In *European Conference on Computer Vision (ECCV)*, Munich, September 2018.

[22] H. Yao, D.-L. Zhu, B. Jiang, and P. Yu. Negative log-likelihood ratio loss for deep neural network classification. In *Proc. of the Future Technologies Conference (FTC)*, AISC 1069, 2019.

[23] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, May 2021.