Prédiction précoce de la transférabilité d'embryons bovins par vidéomicroscopie

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

Prédiction précoce de la transférabilité d'embryons bovins par vidéomicroscopie

Yasmine Hachani[1] Patrick Bouthemy[1] Sylvie Ruffini[3] Ludivine Laffont[3]
Elisa Fromont[4] Alline De Paula Reis[3,5]

[1]Inria Rennes
[4]Université de Rennes, IUF, Inria, IRISA
[5]Ecole Nationale Vétérinaire d'Alfort
[3]University Paris-Saclay, UVSQ, INRAE, BREED

yasmine.hachani@inria.fr

Résumé

La vidéomicroscopie combinée à l'apprentissage automatique est un outil en passe de devenir indispensable pour étudier le développement des embryons bovins fécondés in vitro *et évaluer leur transférabilité de manière la plus précoce possible. Notre objectif est de pouvoir prédire la transférabilité de l'embryon dans un délai maximal de quatre jours de développement, en prenant comme données d'entrée des vidéos de microscopie 2D. Nous formulons ce problème comme un problème de classification binaire supervisée pour les classes* transférable *et* non transférable. *Les difficultés sont toutefois importantes et de trois ordres : 1) apparence et mouvement complexes, imbriqués et peu discriminants, 2) distance entre classes réduite, 3) faible volume de données annotées. Nous proposons un réseau neuronal convolutionnel 3D à trois voies, ce qui le rend multi-échelle dans le temps et capable de traiter l'apparence et le mouvement de différentes manières. Pour l'entraînement, nous retenons la fonction de perte focale. Notre modèle, appelé SFR, se compare favorablement à d'autres méthodes à travers les diverses expériences que nous avons menées et qui démontrent son efficacité et sa validité pour cette tâche biologique difficile.*

Mots Clef

vidéo-microscopie, embryon, classification, réseau neuronal convolutif

Abstract

Video-microscopy is a valuable 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 of embryonic development 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. *However, important challenges arise and are threefold : 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, the focal loss is the best choice. Our model, named SFR, compares favorably to other methods. Experiments demonstrate its effectiveness and accuracy for our challenging biological task.*

1 Introduction

Les techniques utilisées pour étudier les mécanismes du développement embryonnaire sont souvent incompatibles avec la survie de l'embryon. La vidéomicroscopie appliquée aux embryons bovins produits par fécondation *in vitro* (FIV) fournit un grand nombre d'informations morphocinétiques du développement et est compatible avec la survie de l'embryon. Cette technique est donc prometteuse pour son utilisation à la fois en recherche et en élevage, en particulier associée à la puissance d'analyse des techniques d'apprentissage automatique. Ensemble, elles permettent d'étudier le développement embryonnaire initial et d'évaluer la transférabilité d'un embryon fécondé *in vitro*, c'est-à-dire la capacité d'atteindre le stade de blastocyste, et ainsi d'être apte à être transféré dans un utérus de vache. La capacité de prédire correctement, précocement et à grande échelle si les embryons peuvent être transférés ou non est cruciale pour la recherche sur le développement des embryons bovins et pour l'élevage bovin.

Deux enjeux majeurs se posent dans la pratique actuelle : *i)* l'analyse visuelle de chaque vidéo prend beaucoup de temps aux biologistes et limite potentiellement une utilisation à grande échelle, *ii)* l'analyse de transférabilité par les biologistes nécessite en général l'observation sur une durée de l'ordre de sept jours de développement, car il peut y avoir peu de différences entre les trajectoires de développement ; or, pour permettre des études biologiques avancées sur les mécanismes influençant le développement initial de

l'embryon, il est préférable de connaître la transférabilité de l'embryon le plus tôt possible. Par conséquent, l'automatisation de la prédiction de la transférabilité opérant sur les vidéos et une prédiction le plus tôt possible dans le développement embryonnaire représentent des intérêts majeurs à la fois pour la recherche en biologie du développement embryonnaire bovin et pour l'utilisation de la fécondation *in vitro* en élevage bovin.

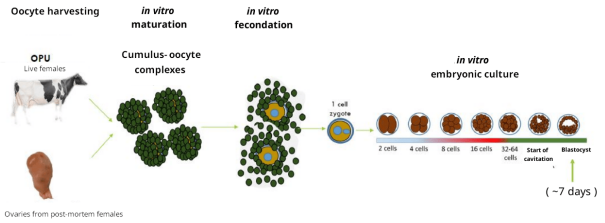

FIGURE 1 – Processus de fécondation *in vitro* (FIV) de l'embryon bovin.

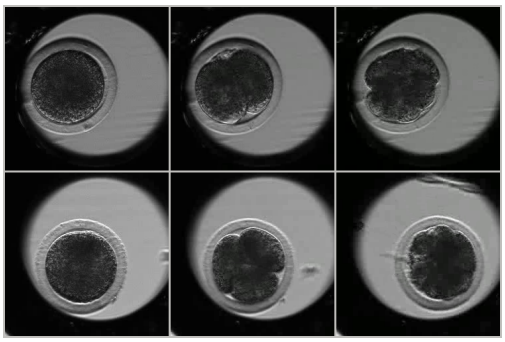

FIGURE 2 – Deux exemples de vidéos d'embryons bovins produits *in vitro* (PIV), avec trois images prises à des instants différents (rangée du haut : un exemple de la classe transférable ; rangée du bas : un exemple de la classe non-transférable). L'embryon bovin (gris foncé), entouré de la zone pellucide, est situé dans un micropuits (gris clair) à l'intérieur de la boîte de Pétri (noir). L'embryon occupe en fait une petite partie de l'image.

Ainsi, notre objectif principal est de parvenir à une prédiction correcte de la transférabilité de l'embryon dans un délai maximum de quatre jours, en prenant des vidéos 2D accélérées (ou time-lapse en anglais) comme données d'entrée à analyser par des réseaux neuronaux convolutifs. Le choix de quatre jours est motivé par les aspects biologiques suivants : en plus de l'apport pour l'élevage bovin, cette avancée pourra offrir aux biologistes de nouvelles perspectives de recherche. Outre la prise de décision précoce pour un transfert dans l'utérus d'une femelle receveuse, elle permettra aussi un meilleur tri des embryons pour des études sur l'activation du génome embryonnaire (EGA) (au jour 4 chez le bovin) ou sur la formation de la morula (au jour 5). Dans ce contexte, nous posons ce problème de transférabilité comme une classification supervisée à deux classes : l'embryon est transférable (classe T) ou non transférable (classe NT). Ce problème reste difficile pour trois raisons principales : 1) l'apparence et le mouvement de l'embryon sont souvent difficilement discernables, 2) la variabilité intra-classe est forte et la distance inter-classe faible, et 3) le faible volume de données annotées disponibles. En effet, comme illustré en figure 2, les vidéos de microscopie présentent peu de contraste, comportent beaucoup de bruit et des mouvements complexes avec des effets de transparence. Les divers embryons étudiés affichent souvent une apparence peu discriminante entre les classes (voir à nouveau figure 2), alors que les vidéos montrent des processus morphologiques et temporels complexes. Deuxièmement, la variabilité intra-classe est élevée en ce sens que les trajectoires de développement de l'embryon peuvent varier considérablement au sein d'une classe donnée. À l'inverse, la distance inter-classe est faible, car le développement observé de deux embryons appartenant à deux classes différentes peut être relativement similaire pendant les quatre premiers jours. Troisièmement, la production d'embryons et l'annotation étant coûteuses, il n'existe qu'une base de données limitée de vidéos globalement étiquetées transférables ou non transférables.

Le reste du papier est organisé comme suit. La section 2 décrit l'état de l'art. Dans la section 3, nous présentons notre réseau convolutif 3D à trois voies, que nous nommerons SFR, ainsi que les fonctions de perte étudiées. La section 4 reporte les évaluations quantitatives et comparatives de la méthode. La section 5 contient les commentaires de conclusion.

2 État de l'art

Depuis plusieurs décennies, les biologistes s'intéressent au développement embryonnaire et la formation des phénotypes. Grâce à la FIV, des avancées significatives ont été réalisées et ont permis des progrès importants en médecine et en élevage. La vidéomicroscopie des embryons bovins a permis aux biologistes d'observer différentes trajectoires de développement conduisant à des phénotypes distincts.

Dans le domaine de la procréation médicalement assistée, l'objectif est de réduire le nombre de grossesses multiples et de pertes de grossesse. Dans [9], un premier modèle de sélection d'embryons humains a été introduit, basé sur une annotation manuelle et un arbre de décision. Ensuite, les auteurs de [20] ont développé un modèle de sélection d'embryons en utilisant une seule image statique capturée par microscopie optique. Si les résultats de ces études étaient prometteurs, le processus d'apprentissage a été réalisé *rétrospectivement*. En effet, les résultats des grossesses ont été prédits à partir d'embryons préalablement sélectionnés manuellement par un expert pour le transfert, ce qui a laissé un grand nombre d'embryons en dehors des études. Ces résultats peuvent donc inclure un certain niveau de surapprentissage ainsi que des biais liés à la manipulation et au transfert des embryons. Récemment, une approche par apprentissage profond a été adoptée par [1] pour sélection-

ner des embryons humains à partir de séquences de vidéos "time-lapse" acquises sur cinq jours, sachant que l'acquisition vidéo a commencé environ 24 heures après l'insémination. Les auteurs ont utilisé le Inflated ConvNet 3D (I3D) de [2] suivi d'un réseau récurrent LTSM. Cependant, l'apprentissage est toujours effectué rétrospectivement. Ils tirent par contre parti d'un très grand ensemble de données (environ 100 000 vidéos).

L'apprentissage profond a également été utilisé sur des vidéos d'embryons humains pour résoudre différents problèmes. Par exemple, dans [5], les auteurs ont proposé un autoencodeur variationnel à registre de vecteurs (VQ-VAE) pour segmenter des blastomères. Les travaux les plus récents se concentrent principalement sur la caractérisation des différents stades de développement de l'embryon, à savoir la division cellulaire avec des stades intermédiaires définis par le nombre de cellules (de 1 cellule à 4 cellules, parfois jusqu'à 8 cellules), la morula et le blastocyste. Dans [7], les auteurs ont élaboré un détecteur de stade de développement exploitant un réseau neuronal convolutif 2D suivi d'un LSTM comme classifieur et ont ajouté une fonction de perte synergique pour apprendre des caractéristiques indépendantes de l'embryon considéré. Dans [10], la classification du stade de développement a été améliorée avec EmbryosFormer, un modèle à trois têtes conçu comme un transformer codeur-décodeur déformable inspiré de Deformable DETR [23]. Les auteurs de [15] ont eux adopté une approche différente en utilisant la technique de détection d'objets YOLO v5 [11] et en comptant les cellules.

Les embryons bovins sont plus difficiles à étudier que leurs homologues humains, car leurs cellules sont plus sombres, ce qui rend, par exemple, le comptage des cellules très difficile. Par ailleurs, des biologistes ont observé [12] que les trajectoires de développement peuvent refléter différents mécanismes d'adaptation et aptitudes pour la gestation future. Dans [12], les auteurs ont procédé de manière *prospective* : ils ont d'abord caractérisé les trajectoires et ensuite vérifié leur intérêt biologique, limitant ainsi les biais des études rétrospectives mentionnées plus haut. Les observations s'appuyant sur des caractéristiques morphocinétiques embryonnaires ont permis de distinguer plusieurs familles de trajectoires entre les embryons transférables et les embryons non transférables. Un modèle de prédiction a été défini, impliquant un modèle de forêts aléatoires et s'appuyant sur de nombreuses annotations détaillées et standardisées pour chaque vidéo.

Le développement embryonnaire peut aussi être considéré comme une forme d'action (dans la terminologie de la vision par ordinateur), et notre problème de classification pourrait donc être considéré comme un problème de classification d'action. Nous passons donc brièvement en revue des travaux sur la reconnaissance d'actions dans les vidéos depuis l'avènement de l'apprentissage profond. Le travail pionnier [17] a introduit un réseau convolutionnel à deux flux prenant à la fois des images et des champs de flots optiques en entrée afin d'exploiter l'apparence et le mouvement pour la reconnaissance d'actions. Cependant, nous avons constaté que le flot optique était mal estimé sur les vidéos d'embryons bovins. En considérant la vidéo spatio-temporelle comme un volume 3D, les réseaux convolutifs 3D ont été largement adoptés depuis lors pour la reconnaissance d'actions, comme le montrent par exemple [2] avec le Inflated ConvNet 3D ou [19] avec un ResNet 3D. Dans [3], les auteurs ont proposé un réseau 3D à deux voies, une voie *Slow* consacrée à l'information de l'apparence avec une vidéo d'entrée à faible fréquence d'images, et une voie *Fast* avec une vidéo d'entrée à haute fréquence d'images afin de mieux capturer le mouvement. Ce dernier modèle sera une source d'inspiration dans notre travail.

À notre connaissance, le modèle que nous proposons est le premier modèle basé sur l'apprentissage profond consacré aux embryons bovins. En outre, nous nous concentrons sur le potentiel de transférabilité des embryons produits *in vitro* (PIV) de manière prospective, et sur la précocité de la prédiction est un aspect essentiel de notre travail. En conséquence, nous considérons une période plus courte de développement embryonnaire. Pour ce faire, nous n'utilisons qu'une seule annotation par vidéo, c'est-à-dire sa classe, transférable ou non transférable, et un nombre limité de vidéos annotées.

3 Description du modèle

Comme formulé plus haut, nous considérons un problème de classification binaire pour prédire la transférabilité des embryons bovins issus de FIV. Les deux classes sont :

- la classe notée T des embryons transférables, à savoir les embryons au potentiel d'aboutissement à une grossesse, ils peuvent être transférés dans une receveuse femelle ;
- la classe notée NT des embryons non transférables, embryons au potentiel nul, ou très faible, de mener à une gestation, ils ne doivent pas être transférés.

En pratique, le biologiste expert annote les vidéos sur une période plus longue que quatre jours, jusqu'à sept (voire huit) jours de développement de l'embryon. Avec ce temps d'observation allongé, une annotation très fiable peut être effectuée, notamment pour les cas difficiles où il peut s'avérer nécessaire de comprendre l'évolution complète de l'embryon sur ce temps long pour décider de la classe. en effet, les trajectoires de développement peuvent être proches jusqu'à un certain stade. De notre côté, nous prenons ces annotations pour les mêmes vidéos, mais limitées à quatre jours de développement.

La prédiction automatique de la transférabilité à seulement quatre jours, c'est-à-dire notre classification binaire, est une tâche d'analyse vidéo complexe. En outre, seul un petit ensemble de vidéos annotées est disponible, les images sont bruitées et peu contrastées, sujettes à des effets de transparence, et les mouvements dans la vidéo ne sont pas faciles à identifier.

3.1 Architecture du réseau

Nous avons conçu un réseau convolutif 3D pour notre problème de classification à deux classes afin d'obtenir une prédiction précoce de la transférabilité des embryons bovins produits *in vitro*. Nous pensons qu'un réseau convolutif 3D est mieux adapté pour capturer correctement les caractéristiques spatio-temporelles d'apparence et de mouvement du développement embryonnaire observé. En effet, les caractéristiques morphocinétiques sont assez complexes et imbriquées, le développement d'un embryon n'étant pas aussi lisse dans le temps qu'une action humaine dans une vidéo. Il est principalement caractérisé par quelques événements discrets correspondant aux divisions cellulaires, avec des mouvements locaux plutôt aléatoires entre les deux.

Notre réseau 3D présenté à la figure 3 comprend trois voies qui peuvent être combinées de deux manières : avec ou sans connexions latérales directes entre les voies. Comme dans le réseau SlowFast [3], nous avons la voie *Slow* qui prend en entrée une version de la vidéo à faible fréquence d'images et qui est de fait principalement dévolue à la capture des caractéristiques spatiales des images ; puis, la voie *Fast* avec en entrée la vidéo à la fréquence d'acquisition initiale qui porte, elle, principalement sur les caractéristiques temporelles de la vidéo. La voie *Fast* a une fraction β des canaux et une résolution temporelle α fois plus élevée que la voie *Slow*. Nous introuduisons une troisième voie *Regular*. Elle prend en entrée la vidéo à la même fréquence que celle de la voie *Fast*, mais elle implique plus de canaux dans chaque couche.

Comme motivé dans l'étude sur les versions alternatives (ou options) des composantes de notre modèle, décrite à la section 4.3, nous utilisons des ResNet18 pour les trois voies afin de construire un réseau 3D léger, ce qui accélère l'apprentissage et tend à éviter des effets de sur-apprentissage. Nous avons remplacé toutes les couches de Batch Normalization de ResNet par des couches de Group Normalization [21]. En effet, la Group Normalization est au moins aussi performante que la Batch Normalization lorsqu'elle est entraînée avec des batchs de petite ou moyenne taille, et elle nous permet d'utiliser efficacement la technique d'accumulation du gradient de Pytorch Lightning.

La troisième voie, dite *Regular*, s'avère pertinente, car l'apparence et le mouvement apparent sont étroitement liés en raison de la transparence des membranes cellulaires et du fait que nous observons des projections 2D, partiellement superposées, de cellules 3D. Les trois voies apportent ainsi des manières complémentaires de traiter et combiner apparence et mouvement dans la vidéo, afin de fournir au final la bonne prédiction. Les sorties des dernières couches de chaque voie sont concaténées avant d'être envoyées au classifieur. De plus, des connexions latérales orientées sont incluses entre les voies. Nous appelons notre modèle SFR. Nous avons envisagé deux combinaisons de connexions latérales. La première est illustrée à la figure 3 et comprend une connexion de la voie *Regular* à la voie *Fast* et de la voie *Fast* à la voie *Slow*. La seconde implique une fusion de la voie *Regular* à *Slow* et de la voie *Fast* à *Slow*, les voies *Regular* et *Fast* n'étant alors pas connectées. Nous avons choisi la première combinaison comme expliqué en section 4.3.

3.2 Fonction de perte

Nous pouvons envisager différentes fonctions de perte. Comme nos ensembles de données sont déséquilibrés entre les deux classes *T* et *NT*, nous avons opté pour la fonction de perte focale [6], initialement introduite pour la tâche de détection d'objets. La fonction de perte focale peut contribuer à corriger ce déséquilibre, tout en se concentrant sur les exemples les plus difficiles. Elle s'écrit :

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

où v représente la vidéo en entrée, y_c l'une des deux classes, $\hat{p}(y_c|v)$ la probabilité prédite d'appartenir à la classe c compte tenu de la vidéo v, et $p(y_c|v)$ la probabilité réelle, égale en fait à 1 pour la bonne classe c compte tenu de v, puisqu'il s'agit d'une classification supervisée. De plus, α_c est le poids de la classe c, γ le paramètre de focalisation. Plus γ est grand, plus les exemples difficiles à classer auront un impact dans l'entraînement.

Nous avons aussi testé comme alternative dans la section 4.3 la fonction de perte classique pour les tâches de classification, à savoir la fonction de perte d'entropie croisée [22], définie pour une classification binaire par :

$$\begin{aligned}\mathcal{L}_{ce}(v, y) = &- p(y_c|v) \log \hat{p}(y_c|v) \\ &- (1 - p(y_c|v)) \log(1 - \hat{p}(y_c|v)).\end{aligned} \tag{2}$$

3.3 Augmentation de données

Différentes stratégies peuvent être adoptées pour faire face au manque de données. L'augmentation des données est une stratégie classique [16]. Ici, nous pouvons considérer l'augmentation des données appliquée aux caractéristiques photométriques, spatiales ou temporelles de la vidéo [14]. En pratique, nous n'avons appliqué que des manipulations d'image de base à chaque image des vidéos : ajout de bruit gaussien, flou gaussien, retournement d'image, transposition d'image, recadrage d'image. Toutes les images d'une séquence donnée sont modifiées de la même manière. Nous avons ainsi augmenté dans un rapport de 20 le volume de vidéos pour l'entraînement.

4 Résultats expérimentaux

4.1 Acquisition des vidéos et jeu de données

Les vidéos ont été acquises à l'INRAE de la manière suivante. Les ovocytes récupérés sur les ovaires d'abattoir et maturés *in vitro* sont mis en contact avec de la semence congelée-décongelée dans une boîte de culture définissant le point de départ du développement biologique des embryons [12]. Les zygotes présumés sont placés dans des

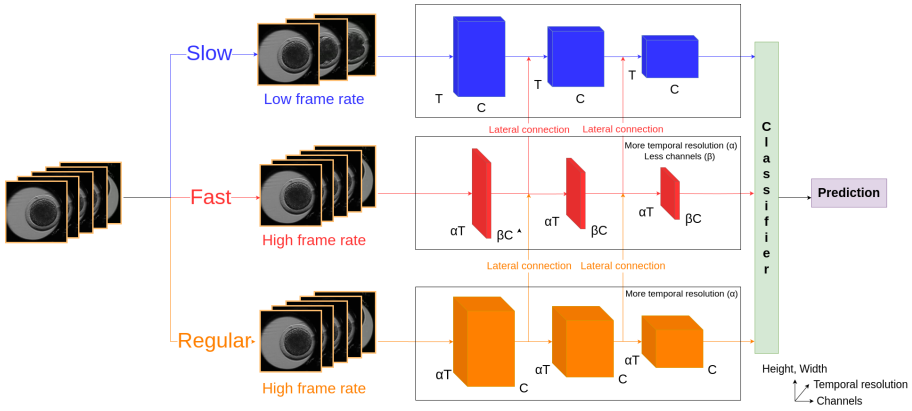

FIGURE 3 – Notre modèle 3D SFR combine trois voies, *Slow* et *Fast* comme dans [3], ainsi que *Regular*, toutes implémentées avec des ResNet18 3D. La vidéo d'entrée est donnée à une fréquence temporelle plus faible pour la voie *Slow*. Les trois sorties sont combinées avant de fournir la prédiction. De plus, des connexions latérales orientées sont incluses entre les voies, comme le montre la figure.

micro-puits de boîtes de Petri environ vingt-deux heures après la fécondation *in vitro*. Chaque boîte de Petri contient seize micro-puits. Chaque boîte de culture est placée sur un équipement permettant de réaliser les vidéos time-lapses avec un microscope à lumière transmise (système Primovision), dans un incubateur avec atmosphère tri-gaz humidifiée adaptée à la culture embryonnaire.

Le système PrimoVision prend une photo des boîtes de Petri toutes les quinze minutes pendant la durée de la culture embryonnaire à partir desquelles sont constituées les séquences vidéo 2D time-lapse. Celles-ci sont ensuite divisées en seize vidéos, une par embryon. Chaque vidéo est annotée par un biologiste avec l'étiquette T ou l'étiquette NT. Pour les besoins de notre travail, nous n'avons retenu que les séquences vidéos des quatre premiers jours de développement de l'embryon. Étant donné que l'acquisition des vidéos commence seulement 22 heures après la FIV, chaque vidéo traitée couvre une période de trois jours et comprend environ 300 images. L'ensemble de données vidéo est ainsi formé de 947 vidéos, réparties en 763 pour l'entraînement et la validation et 184 pour le test. Chaque ensemble contient environ 65% de vidéos d'embryons non transférables.

4.2 Détails d'implémentation

Chaque modèle a été entraîné à l'aide de l'optimiseur AdamW [8], avec un gain d'apprentissage (learning rate) de 10^{-4}, les autres paramètres étant laissés à leurs valeurs par défaut. Nous avons appliqué un planificateur de gain d'apprentissage cyclique comme recommandé dans [18]. Nous avons entraîné les modèles en utilisant des mini-batchs équivalents à 32 exemples grâce à la technique d'accumulation des gradients mise en œuvre dans PyTorch Lightning. Cette dernière cumule les gradients des petits batchs avant la rétropropagation du gradient. Nous appliquons la technique de moyenne stochastique des poids (SWA) [4] qui améliore la généralisation de nos modèles en moyennant les poids du réseau à différentes epochs bien choisies. Nous utilisons l'arrêt anticipé pour mettre fin à l'entraînement, lorsque la fonction de perte calculée sur l'ensemble de validation a crû pendant dix epochs consécutives. Ensuite, nous sélectionnons le modèle correspondant à l'epoch ayant le plus grand taux de bonne classification sur l'ensemble de validation (sachant que l'on met en œuvre un entraînement supervisé).

4.3 Etude des composantes de notre modèle

Nous avons étudié le choix des différentes composantes (ou options principales) de notre modèle. Cette étude est habituellement désignée sous le terme de "ablation study" en anglais.

Tout d'abord, en ce qui concerne la combinaison des connexions latérales, la première option (connexion de la voie *Regular* à la voie *Fast* et de la voie *Fast* à la voie *Slow*) a fourni un meilleur taux de bonne classification que la seconde option. C'est donc celle qui sera retenue par la suite.

Modèle SFR	
Modules	**Acc**
avec 3D-Resnet18	72.9
avec 3D-Resnet50	71.6

TABLE 1 – Résultats de la classification binaire (Acc pour taux de bonne classification) obtenus par notre modèle SFR avec les modules ResNet18 et ResNet50 (et la fonction de perte d'entropie croisée). Nous avons effectué une dizaine d'évaluations à chaque fois pour différentes "seeds" et "folds" d'entraînement, et nous fournissons la moyenne.

Nous avons testé deux variantes d'architecture de ResNet, 18 couches et 50 couches, car cela nous permet dans tous les cas d'avoir un modèle assez léger, sachant que nous ne disposons que d'un ensemble réduit de vidéos d'entraîne-

ment. Nous avons donc entraîné notre modèle SFR avec les modules ResNet18 et ResNet50. Pour cette expérience, nous avons simplement utilisé la fonction de perte d'entropie croisée en attendant de décider du paramètre γ de la fonction de perte focale. Les résultats de la classification binaire (transférable *vs* non transférable) sont fournis dans le tableau 1. ResNet18 apparaît plus approprié que ResNet50, comme le taux de bonne classification est sensiblement supérieur avec ResNet18 qui est de plus une architecture plus légère.

Nous avons testé la fonction de perte d'entropie croisée vis à vis de la fonction de perte focale pour notre modèle SFR incluant des connexions latérales dirigées entre les trois voies. Nous avons aussi évalué la configuration (appelée SFR Late Fusion) sans aucune connexion latérale et ce pour les deux fonctions de perte. Ces tests ont été menés avec ResNet18 comme préconisé plus haut. Pour la fonction de perte focale, nous avons pris $\gamma = 2$, comme justifié plus loin. Nous avons fixé $\alpha_1 = 1.25$ et $\alpha_2 = 0.833$ en fonction de la fréquence des deux classes. Les résultats obtenus pour notre modèle SFR et son alternative SFR Late Fusion avec les deux fonctions de perte sont consignés dans le tableau 2. Pour les deux modèles, la fonction de perte focale donne de meilleurs résultats. De plus, SFR est plus performant que SFR Late Fusion, ce qui montre l'importance des connexions latérales.

Modèles SFR et SFR Late Fusion	
Modèles et fonctions de perte	**Acc**
SFR(CE)	72.9
SFR(FL)	75.6
SFR Late Fusion(CE)	72.5
SFR Late Fusion(FL)	73.5

TABLE 2 – Résultats (Acc pour taux de bonne classification) obtenus par nos deux modèles SFR et SFR Late Fusion avec la fonction de perte d'entropie croisée (CE) et la fonction de perte focale (FL) avec $\gamma = 2$. ResNet18 est utilisé pour tous les modèles. Nous avons effectué une dizaine d'évaluations à chaque fois pour différentes "seeds" et "folds" d'entraînement, et nous fournissons la moyenne.

Modèle SFR	
Valeur de gamma	**Acc**
SFR(FL) ($\gamma = 1$)	73.1
SFR(FL) ($\gamma = 2$)	75.6
SFR(FL) ($\gamma = 3$)	73.4

TABLE 3 – Résultats de la classification binaire (Acc pour taux de bonne classification) obtenus par notre modèle SFR (module ResNet18) entraîné avec la fonction de perte focale pour $\gamma \in \{1, 2, 3\}$. Nous avons effectué une dizaine d'évaluations à chaque fois pour différentes "seeds" et "folds" d'entraînement, et nous fournissons la moyenne.

Notre dernière expérience porte sur la spécification du paramètre γ de la fonction de perte focale. Nous avons testé la fonction de perte focale pour trois valeurs du paramètre γ, $\gamma = 1, 2,$ et 3, sachant que le cas $\gamma = 0$ équivaut en quelque sorte à entraîner le modèle avec une fonction de perte d'entropie croisée pondérée. Nous avons réalisé l'expérience sur la classification binaire avec notre modèle SFR. Les résultats sont présentés dans le tableau 3. Nous retenons la valeur $\gamma = 2$ qui fournit le meilleur taux de bonne classification. Au demeurant, les auteurs de [1] ont également choisi la valeur $\gamma = 2$.

4.4 Expériences comparatives

Nous avons réalisé des expériences comparatives sur la prédiction à quatre jours de la transférabilité des embryons bovins. Pour évaluer et comparer les performances de toutes les méthodes, nous considérons les métriques suivantes : taux de bonne classification Acc, précision P_T (resp. P_{NT}) et rappel R_T (resp. R_{NT}) pour la classe T (resp. NT).

Nous avons effectué la classification binaire avec SlowFast [3] (la version ResNet18) et un 3D-RestNet18 classique, en utilisant toutefois pour ces deux modèles la fonction de perte focale, comme cette fonction est plus adaptée à notre problème (voir tableau 4). Nous entraînons les deux méthodes sur notre ensemble de données d'entraînement. En outre, nous avons construit un modèle de base comprenant un réseau neuronal convolutionnel 2D (CNN) suivi d'un réseau neuronal récurrent (RNN). Le CNN 2D est implémenté avec un ResNet18 qui apprend les caractéristiques spatiales des images. Un réseau neuronal récurrent type GRU traite la dimension temporelle de la vidéo de l'embryon, en prenant comme entrée les sorties successives du CNN 2D. La sortie de la dernière cellule du GRU est envoyée à une couche entièrement connectée pour obtenir la prédiction de classification.

Les résultats comparatifs pour tous les modèles testés sont rassemblés dans le tableau 4 avec l'utilisation de la fonction de perte focale ($\gamma = 2$) pour toutes les méthodes. Comme attendu, le réseau 2D avec un réseau neuronal récurrent est moins performant que tous les réseaux 3D.

Notre méthode SFR obtient le meilleur taux de bonne classification, secondé par ResNet18, mais notre modèle est beaucoup plus stable que ResNet18, y compris sur les autres métriques, ce qui est crucial. SlowFast n'a pas le comportement espéré, sans doute à cause de la nature particulière des vidéos traitées, très différentes des vidéos considérées en reconnaissance d'actions. De plus, notre méthode SFR a le meilleur score de précision pour la classe T et le meilleur score de rappel pour la classe NT, ce qui est très important pour l'application visée. En effet, pour l'élevage bovin, il est primordial de bien prédire les embryons transférables afin d'éviter des grossesses superflues en transférant des embryons non transférables.

Nous avons également comparé notre méthode avec d'autres en ce qui concerne le temps de calcul pour effectuer une prédiction, toujours dans le tableau 4. Pour chaque modèle, nous avons calculé le temps moyen pour une infé-

Modèle	Acc	P_T	R_T	P_{NT}	R_{NT}	Temps Moyen de Prédiction
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 – Comparaison des résultats obtenus pour les modèles 2D-Resnet18+GRU, 3D-Resnet18, SlowFast [3], et notre modèle SFR, tous les modèles étant entraînés avec la fonction de perte focale ($\gamma = 2$). Nous avons effectué une dizaine d'évaluations à chaque fois pour différentes "seeds" et "folds" d'entraînement, et nous fournissons la moyenne et l'écart-type.

rence en répétant 1000 prédictions sur un GPU NVIDIA RTX A500. Comme prévu, le CNN 2D avec GRU est le plus rapide pour l'inférence, suivi par SlowFast qui est plus rapide qu'un simple 3D-ResNet comme le montre l'article original [3]. Viennent ensuite 3D-ResNet18 et notre modèle SFR, dont le temps moyen de prédiction augmente d'environ $89ms$ par rapport à 3D-ResNet18, mais ce n'est pas un problème pour notre application.

4.5 Prédiction précoce

Nous voulions vérifier, expérimentalement à ce stade de l'étude, s'il était possible de faire une prédiction avant même que les quatre jours de développement ne se soient écoulés. Par conséquent, nous évaluons les performances de notre meilleur modèle, SFR avec fonction de perte focale, lorsque nous effectuons une prédiction Transférable (T) ou Non Transférable (NT) à moins de quatre jours. Pour ce faire, nous testons le modèle avec des vidéos de plus en plus courtes, en supprimant à chaque fois les trente dernières images, ce qui correspond à la suppression des informations survenues pendant sept heures et demie. Nous arrêtons les tests à 120 images, soit environ 2 jours 1/4 de développement de l'embryon. Les résultats sont représentés en figure 4. Nous observons que la courbe monte régulièrement. Le modèle permet de classer les embryons avec une précision supérieure à 70% dès trois jours et demi du développement (deux jours et demi de vidéo), c'est-à-dire environ une demi-journée avant la mise en route du génome embryonnaire, ce qui peut s'avérer également intéressant. Ainsi, il est possible de sélectionner tôt les embryons pour des études réalisées sur les étapes intermédiaires du développement (16 cellules, morula et jeune blastocyste). En ce qui concerne l'élevage, une décision à trois jours et demi du développement permettrait aux biologistes de gagner encore une demi-journée supplémentaire pour préparer la logistique des transferts embryonnaires.

5 Conclusion

Nous avons conçu un modèle convolutif 3D pour prédire la transferabilité de l'embryon bovin produit par FIV dans les quatre premiers jours de développement, en prenant comme données d'entrée des vidéos de microscopie 2D. Nous avons formulé le problème comme une classification supervisée à deux classes qui reste cependant difficile pour plusieurs raisons. L'architecture à trois voies rend

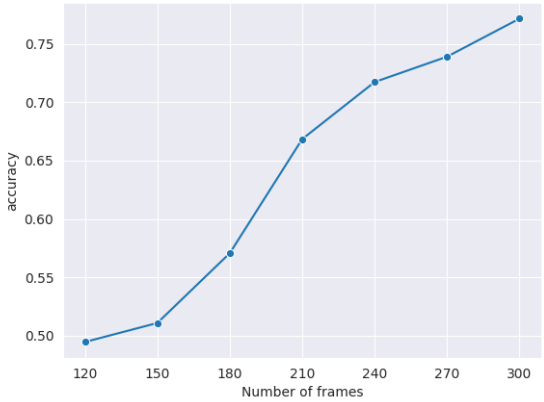

FIGURE 4 – Taux de bonne classification de la prédiction précoce pour notre modèle SFR avec fonction de perte focale pour la plage de longueur vidéo suivante : de 120 images (environ deux jours un quart du développement de l'embryon) à 300 images (environ quatre jours)

notre modèle 3D multi-échelle dans le temps en ce qui concerne les vidéos d'entrée et capable de combiner apparence et mouvement de différentes manières. L'apparence et le mouvement dans les vidéos sont difficiles à appréhender, étant affectés par la transparence des embryons et leur aspect sombre. De plus, la distance inter-classe est faible et la variabilité intra-classe importante.
Les expériences ont montré la pertinence des connexions latérales dirigées entre les voies de notre modèle SFR et du choix de la fonction de perte focale. Nous avons favorablement comparé notre modèle SFR à d'autres méthodes. SFR fournit les meilleurs taux de bonne classification avec une bien meilleure stabilité que ResNet18 sur toutes les métriques. Nous sommes donc en mesure de prédire avec efficacité et une précision suffisante la transférabilité précoce des embryons bovins. Les travaux futurs chercheront à combiner formellement précocité et précision de classification dans le modèle.

6 Respect des normes éthiques

Cette étude a été réalisée à partir des données disponibles dans le laboratoire BREED de l'INRAE. Aucun animal vivant ou euthanasié n'a été utilisé pour créer les données

originales. La semence a été acquise auprès d'une société commerciale et les complexes cumulus-ovocytes ont été récoltés sur des ovaires récupérés *post-mortem* dans un abattoir commercial. Ces sociétés et notre laboratoire sont basés en France et agréés par l'État. Les autorisations nécessaires à l'utilisation de matériel biologique *post-mortem* ont été obtenues auprès du ministère compétent.

7 Remerciements

Les auteurs souhaitent remercier Véronique Duranthon et Brigitte Marquant-LeGuienne pour leur collaboration à la conception du protocole expérimental de production d'embryons. Les auteurs ne déclarent aucun conflit d'intérêt. La recherche liée à la production des données a été financée par CRB-Anim et APIS-GENE.

Références

[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 *Conference on Uncertainty in Artificial Intelligence (UAI)*, Monterey, August 2018.

[5] W.-D. Jang et al. 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 et al. The use of morphokinetics as a predictor of embryo implantation. *Human Reproduction*, 26(10) :2658-2671, Oct. 2011.

[10] T.-P. Nguyen et al. EmbryosFormer : Deformable transformer and collaborative encoding-decoding for embryos stage development classification. In *Winter Conference 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 et al. 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 et al. 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.