Apprentissage profond appliqué à la classification d'images microscopiques embryonnaires

Tristan Gomez, Harold Mouchère, Thomas Fréour, Magalie Feyeux

Apprentissage profond appliqué à la classification d'images microscopiques embryonnaires

T. Gomez[1], H. Mouchère[1] , T. Fréour [2 3], M. Feyeux[4],

[1] LS2N, UMR 6004, CNRS, Université de Nantes, France
[2] Service de médecine et biologie du développement et de la reproduction, CHU Nantes, France
[3] CRTI, Inserm, Université de Nantes, Nantes, France
[4] SFR Santé, MicroPICell core facility, CNRS, INSERM, Université de Nantes, Nantes, France

29 Juin 2020

Résumé

Dans le domaine de la fécondation in vitro (FIV), l'imagerie par time-lapse (ITL) est une technologie qui produit une vidéo montrant le développement de l'embryon pendant ses premiers jours. L'ITL permet d'annoter les instants auxquels commencent et finissent les différentes phases de développement de l'embryon. Cependant, le processus d'annotation prend du temps et nécessite des experts. Dans cet article, nous montrons que pour prédire les phases visibles dans les images issues de vidéos ITL, un ResNet-3D est meilleur qu'un équivalent 2D ou un ResNet-LSTM.

Mots-clés

apprentissage profond, fécondation in vitro, validation croisée, modèle temporel, convolution 3D, ResNet, LSTM, CNN, apprentissage machine

Abstract

In the field of in vitro fertilization (IVF), time-lapse imaging (TLI) is a technology that produces a video showing the development of the embryo during its first days. TLI makes it possible to annotate the moments at which the different stages of embryo development begin and end. However, the annotation process takes time and requires experts. In this article, we show that to predict the visible phases in images from TLI videos, a ResNet-3D is better than a 2D equivalent or a ResNet-LSTM.

Keywords

deep learning, in vitro fertilization, cross validation, time model, 3D convolution, ResNet, LSTM, CNN, machine learning

1 Introduction

L'infertilité est encore un problème mondial aujourd'hui puisqu'il touche 186 millions de personnes [1] et que le nombre de couples déclarant leur infertilité dans des centres en Europe augmente de $8-9\%$ par an [2]. La fécondation in vitro (FIV) est l'un des traitements les plus courants de l'infertilité. Elle implique une stimulation ovarienne suivie du prélèvement de plusieurs ovocytes, de la fécondation et de la culture d'embryons pendant 1 à 6 jours dans des conditions environnementales contrôlées. Les embryologistes sélectionnent ensuite un embryon à transférer en fonction de leur évaluation de la qualité des embryons.
L'efficacité de la FIV n'est pas encore optimale aujourd'hui [3]. Cette situation s'explique en partie par les limites actuelles des méthodes d'évaluation de la qualité des embryons. La méthode d'évaluation la plus courante est l'évaluation morphologique qui consiste à observer l'embryon au microscope régulièrement pendant ses premiers jours et à évaluer sa morphologie. Le croissance des embryons étant très lente, les biologistes n'obtiennent qu'une vision statique du développement. Cette méthode souffre toujours d'un manque de pouvoir prédictif et d'une grande variabilité inter et intra-biologiste [4].
L'Embryoscope© est un appareil qui permet aux embryons de se développer dans des conditions environnementales contrôlées et qui intègre une technologie d'ITL. L'ITL est une technologie récente qui consiste à prendre plusieurs photographies de l'embryon par heure pendant quelques jours et qui produit une vidéo time-lapse donnant une vision dynamique du développement. Elle permet d'observer en continu l'embryon sans le retirer des conditions contrôlées et stables de l'incubateur. Cette technologie permet une évaluation morpho-cinétique (MC) des embryons, c'est-à-dire que les biologistes ont une vision dynamique du développement de l'embryon et peuvent annoter certains évènements temporels (divisions des cellules, expansion du blastocyste, etc.).
Un travail récent a montré que la surveillance des vidéos time-lapse est associée à un taux de grossesse clinique et de naissance vivante significativement plus élevé, et à une perte précoce de grossesse significativement plus faible par rapport à la sélection morphologique [5] mais une autre étude a donné des résultats contradictoires [6]. En outre, plusieurs travaux [7, 8] ont proposé des modèles prédictifs de l'issue de la FIV (grossesse/non grossesse) basés sur des bases de données MC. Même si la variabilité a diminué grâce à l'ITL, elle reste un problème [9].
Cependant, l'annotation des paramètres MC est un processus qui prend du temps et qui nécessite des embryologistes

expérimentés. Dans cet article, nous évaluons rigoureusement plusieurs algorithmes pour prédire automatiquement les paramètres MC en utilisant des méthodes d'apprentissage profonds.

2 Description du problème

L'extraction automatique des paramètres-morphocinétique (MC) est une tâche de classification d'images qui consiste à classer chacune des images des vidéos en plusieurs classes, qui sont les différentes phases de développement auxquelles l'embryon peut se trouver. Les phases de développement étudiées dans les travaux cités ci-dessous sont choisies parmi les suivantes : $tPB2$, $tPNa$, $tPNf$, $t2$, $t3$, $t4$, $t5$, $t6$, $t7$, $t8$, $t9+$, tM, tSB, tB et enfin tEB. Le développement se déroule de la manière suivante : d'abord les pro-nucléi apparaissent puis disparaissent (phases $tPB2$, $tPNa$, $tPNf$), puis une cellule unique se divise jusqu'à qu'il y ait environ une dizaine de cellules (phases $t2$ à $t9+$), ensuite les cellules se compactent, le blastocyste apparaît et enfin ce dernier s'agrandit (phases tM, tSB, tB, tEB). Chacune de ces phases est définie en [10] et certaines sont illustrées en Figure 1. Ces phases sont ordonnées, il est très rare qu'un embryon revienne à une phase précédente durant son développement.

Chaque vidéo montre le développement d'un embryon particulier. Pour chaque vidéo, un biologiste expérimenté note l'heure de début et de fin de chaque phase de développement de l'embryon. Chaque image de chaque vidéo a donc une classe, qui est la phase à laquelle l'embryon se trouve dans cette image.

Il faut noter que certaines phases peuvent ne pas être visibles dans la vidéo. Par exemple un embryon peut contenir 2 cellules (phase $t2$) puis ces deux cellules peuvent se diviser chacune dans l'intervalle de temps qui sépare deux images consécutives (phase $t4$) sans qu'il n'y ait un moment dans la vidéo où l'embryon présente clairement 3 cellules (phase $t3$).

Chaque vidéo contient entre 300 et 600 images en nuance de gris et ont une résolution de 500×500 pixels. La durée réelle qui sépare deux images est de 10 à 20 minutes.

L'embryon est un objet en trois dimensions et il est parfois nécessaire d'ajuster le plan focal de l'Embryoscope afin de pouvoir observer correctement toutes les parties de l'embryon. Certains travaux extraient donc plusieurs vidéos par embryon, chacune enregistrée avec un plan focal différent, afin de donner une meilleure vision de l'embryon au modèle. Utiliser plusieurs plans focaux demande beaucoup de calculs, il est donc aussi possible d'extraire une seule vidéo en laissant l'Embryoscope choisir le meilleur plan focal à chaque image, c'est-à-dire le plan qui maximise la netteté de l'image.

Les différentes phases n'ont pas toutes la même durée donc les classes sont déséquilibrées. Par exemple la classe majoritaire est la classe $t9+$ et constitue 16% du total des images alors que la classe la moins fréquente ($t3$) constitue moins de 2% des images.

3 État de l'art

Nous présentons d'abord des travaux traitant de l'extraction automatique des paramètres MC et ensuite quelques travaux sur des tâches connexes : segmentation automatique de la masse cellulaire interne (MCI), du trophectoderme (TE) et enfin prédiction automatique de la qualité embryonnaire. Les approches classiques de classifications d'images et de segmentation de vidéos ont été largement utilisés.

3.1 Extractions des paramètres morpho-cinétiques

Plusieurs travaux ont tenté d'extraire automatiquement les informations morpho-cinétiques des vidéos en time-lapse brutes.

Certains travaux proposent une version simplifiée du problème avec moins de classes. Les auteurs de [11] et [12] se concentrent sur les phases entre 1 et 5 cellules et les auteurs de [13] proposent de ne modéliser que deux classes : blastocyste et non-blastocyste. Kanakasabapathy et al. [13] font valoir que la réduction du nombre de classes permet d'avoir des annotations plus nettes, car il y a beaucoup moins de variabilité intra et inter-opérateurs lorsque l'on ne modélise que ces deux classes. Des modèles comme AlexNet [11], un ensemble de U-Nets [14] dilatés résiduels [12] ou Xception [15, 13] sont utilisés comme classifieur. D'autres travaux [16, 17] proposent de travailler avec un nombre réduit de phases, à savoir les 6 premières phases de développement.

Le contexte obtenu en utilisant plusieurs trames voisines a également été exploité. La plupart des travaux proposent une fusion tardive en calculant les mêmes traits sur plusieurs trames successives, puis en agrégeant les informations pour produire une prédiction. Par exemple, il a été proposé de concaténer les vecteurs caractéristiques des images calculées par le même réseau de neurones convolutifs (CNN) [16], ou alors de les agréger à l'aide de max-pooling [17]. Il a aussi été montré qu'il est avantageux, lorsque plusieurs images sont utilisées, de prédire l'étiquette des images voisines en utilisant les caractéristiques de l'image actuelle [17]. Enfin, Lau et al. [18] utilisent un LSTM [19] pour prendre en compte les caractéristiques extraites à chaque pas de temps. Il faut noter que la fusion précoce d'informations a également été proposée dans [16] en concaténant les images sur la dimension du canal. Les différentes phases de développement sont strictement ordonnées : l'embryon peut parfois sauter une phase mais il est très rare qu'il revienne à une phase précédente. Cette contrainte est souvent exploitée par la programmation dynamique pour rendre la prédiction cohérente sur toute la vidéo [11, 18, 17, 16],

Plusieurs méthodes produisent des informations spatiales sur le blastocyste. Par exemple, il est possible d'apprendre à déduire la position du centroïde de chaque cellule de manière totalement supervisée [12]. Lau et al. utilisent l'apprentissage par renforcement pour recadrer l'image d'entrée sur une région d'intérêt avec un réseau de proposition

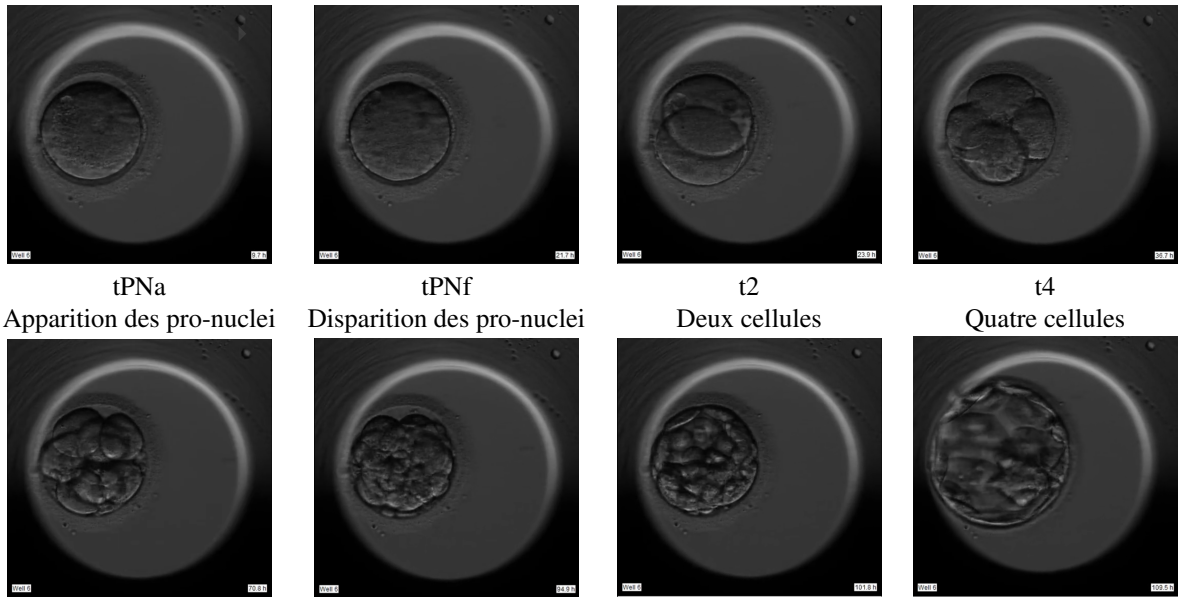


FIGURE 1 – Illustrations de 8 des 15 phases que nous utilisons

de région (RPN) [20] et à passer ensuite le résultat à un CNN [18]. L'image est divisée en une grille de 32×32 cellules et le RPN produit une distribution de probabilité sur celle-ci. Une cellule est choisie aléatoirement et la partie de l'image contenue dans la cellule est envoyée au classificateur. Ce dernier est entraîné en utilisant une entropie croisée et le RPN est entrainé en utilisant la méthode de "policy gradient" basé sur la performance du classificateur. Pour forcer le RPN à produire diverses distributions de probabilité, un terme d'entropie est ajouté à la fonction de coût. Ce détecteur apprend à localiser le blastocyste de manière faiblement supervisée.

3.2 Tâches liées

Les articles examinés ici couvrent les tâches suivantes : segmentation du TE, segmentation de la MCI et qualité des embryons. Pour extraire automatiquement la région du TE, une méthode *ad-hoc* basée sur l'algorithme Retinex a été développée [21]. Les auteurs l'utilisent pour nettoyer les régions à faible gradient, puis appliquent la méthode Canny Edge pour obtenir une estimation brute du TE. Enfin, ils utilisent l'algorithme des k-moyennes pour remplacer les composantes connexes par le centroïde auquel ils appartiennent, ce qui réduit le bruit dans le résultat. Pour segmenter la MCI, un FCN [22] et l'architecture U-Net [14] ont respectivement été proposés en [23] et [24]. Kheradmand et al. [23] utilisent un ensemble de données de 235 images augmentées de 36 rotations pour chaque image. Rad et al. [24] utilisent le même ensemble de données sans augmentation. Les auteurs ajoutent plusieurs convolutions dilatées dans la partie centrale de l'architecture afin d'augmenter le champ réceptif des neurones. Ils utilisent également un ensemble de quatre U-Nets différents, chacun ayant une résolution d'entrée différente (448×448, 384×384, 320×320, et 256×256), pour capturer un contexte plus large.

Les auteurs de [25] proposent des traits conçues à la main pour prédire la qualité des embryons bovins. Ils calculent d'abord la matrice de co-occurrence de niveau de gris (MCNG), la segmentation des blastocystes à l'aide de la méthode de transformée circulaire de Hough et la segmentation MCI à l'aide de la méthode des bassins versants. Ensuite, 36 variables scalaires sont calculées en utilisant les deux segmentations et le MCNG. Ces variables comprennent des traits comme le contraste, la corrélation, l'énergie, etc. Khosravi et al. [26] utilisent $10\,000$ images d'embryons, capturés à 110h post-insémination avec tous les plans focaux disponibles. Ils utilisent la méthode Veeck et Zaninovic [27] pour classer chaque embryon. Cette méthode manuelle attribue trois notes à l'embryon en fonction du développement du blastocyste, de la qualité de la MCI, et de la qualité du TE. Ces notes sont basées sur des critères morphologiques estimés par l'expert. Ensuite, ils classent les embryons en trois groupes : bonne qualité, qualité moyenne et mauvaise qualité (en fonction du taux de grossesse). Pour chaque embryon, 6000 images du groupe de bonne qualité et 6000 images du groupe de mauvaise qualité sont choisies au hasard. Un CNN [28] est entraîné pour faire la distinction entre les deux groupes. Enfin, ils étudient le lien entre la qualité de l'embryon et l 'âge du patient à l'aide d'un arbre de décision.

Les auteurs de [8] utilisent les paramètres morphocinétiques, les niveaux de fragmentation, la présence de multinucléation, l'uniformité des blastomères et l'âge des femmes passés comme entrée d'un NN en essayant de prédire s'il y aura ou non une implantation. La corrélation entre les caractéristiques de l'ACP et la grossesse a égale-

ment été étudiée. Milewski et al. [29] utilisent des images de 16 000 embryons, capturés entre 112 et 116 heures (jour 5) ou 136 à 140 heures (jour 6) de l'insémination. Ils entraînent un ResNet-50 [30] à prédire les notes données avec le système Veeck et Zahinovic [27], c'est-à-dire le développement des blastocystes, la qualité du TE et la qualité de la MCI.

4 Protocoles et modèles utilisés

Nous utilisons 953 vidéos d'embryons provenant de FIV réalisées entre 2011 et 2019 au laboratoire de biologie et de médecine de la reproduction du centre hospitalier universitaire de Nantes. Afin de réduire les temps de calculs, nous n'utilisons qu'un seul plan focal par embryon, qui est choisit automatiquement par l'Embryoscope pour chaque image de la vidéo. Nous utilisons les phases de développement évoquées en section 2, à savoir $tPB2$, $tPNa$, $tPNf$, $t2$, $t3$, $t4$, $t5$, $t6$, $t7$, $t8$, $t9+$, tM, tSB, tB et enfin tEB. Les modèles testés doivent donc résoudre un problème de classification à 15 classes.

Ce travail consiste à classer des images extraites de vidéos pour extraire la phase de développement correspondant à chaque image. Nous travaillons donc sur la même tâche que les travaux cités dans la section 3.1. Nous travaillons avec un plus grand nombre de phases que [13, 12, 11, 16, 17, 18] qui utilisent 6 phases ou moins. Nous profitons du fait que nous disposons déjà d'un grand nombre de vidéos où les 15 phases sont annotées et qu'il est possible que chacune de ces 15 phases soit importante pour prédire la qualité intrinsèque de l'embryon. Nous cherchons aussi à exploiter le contexte de chaque image, à savoir les images suivantes et précédentes, comme [16, 17, 18, 11]. Afin de simplifier le problème, nous ne cherchons pas à prédire la localisation de l'embryon ou des cellules, contrairement à [12, 18]. L'algorithme de Viterbi est utilisé pour rendre les prédictions du modèle cohérente tout au long de la vidéo, comme proposé dans [11, 18, 17, 16].

4.1 Modèles comparés

Plusieurs modèles ont été utilisés pour réaliser cette tâche de classification. Ils sont illustrés par la Figure 2 et détaillés ci-après.

Le modèle ResNet-18 [30] et plus généralement les modèles résiduels sont largement utilisés pour la classification d'image par exemple sur ImageNet [30]. Ce modèle est composé exclusivement de couches de convolutions (17 couches pour la version ResNet-18) et contient des connexions résiduelles toute les 2 couches. La résolution et le nombre de canaux des cartes de traits sont respectivement divisée et multiplié par deux toute les 4 couches. Après les convolutions, une couche de "average-pooling" permet d'obtenir un vecteur de traits, auquel on applique la couche "soft-max" finale pour faire la prédiction.

Le modèle ResNet-LSTM est la combinaison du modèle ResNet-18 avec un LSTM [19]. Le modèle LSTM a été conçu pour modéliser des séquences et a été appliqué avec succès dans des tâches telle que la reconnaissance de la parole [31]. Les pré-activations de l'avant-dernière couche sont utilisées comme vecteur de traits et sont transmises à un LSTM bidirectionnel à deux couches qui modélise l'évolution à travers les étapes du temps. La taille de chaque unité cachée est 1024. Une couche linéaire au-dessus du LSTM calcule les scores.

Le modèle ResNet3D [32] est une variante de ResNet conçue pour la classification de séquences d'images. Ce modèle permet de modéliser la séquence d'images en fusionnant les informations temporelles au fur et à mesure qu'on avance en profondeur dans le réseau. Ce modèle permet donc une fusion à la fois tardive et précoce de l'information. Pour cette application, les paramètres de max-pooling et de stride sont fixés à 1 dans la dimension temporelle. La suppression de l'agrégation temporelle est nécessaire car nous voulons une une prédiction pour chaque image. Nous utilisons la variante *r2plus1d_18'* proposée dans [32].

5 Expériences

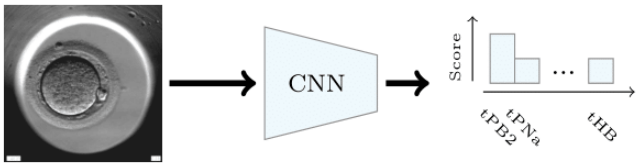
L'expérience principale de ce papier consiste à entraîner plusieurs architectures et à les évaluer à l'aide d'une validation croisée ($k = 10$). Les détails de l'expérience sont indiqués ci-dessous.

Méta-paramètres. Chaque batch est composé de 10 séquences de 4 images consécutives. La position de la séquence dans la vidéo est choisie aléatoirement dans la vidéo. La fonction de perte est optimisée avec la méthode de descente de gradient stochastique, avec un taux d'apprentissage constant de $0,001$ et un momentum de $0,9$. Nous appliquons du dropout [33] ($p = 0.50$) sur la dernière couche de chaque modèle pendant l'entraînement. Le modèle ResNet-18 traite chaque image de manière indépendante et lit donc les $10 \times 4 = 40$ images en parallèle, afin d'avoir un nombre d'images par batch égal pour les trois modèles et ainsi permettre une meilleure comparaison.

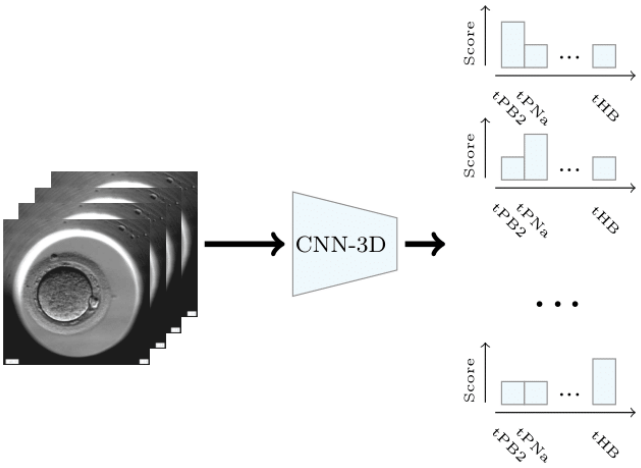
Lors de la validation, pour réduire l'occupation de la mémoire GPU, les modèles ne sont pas évalués sur l'ensemble de la vidéo en une seule fois, mais sur des séquences de 150 images. Comme chaque vidéo contient environ 500 images, quelques inférences suffisent pour analyser une vidéo entière.

Soit N le nombre total d'images d'entraînement et L le nombre d'images dans une séquence. Une époque se termine lorsque le modèle a vu N/L séquences. Pour sélectionner les séquences, nous utilisons des tirages aléatoires avec remise, c'est-à-dire que le modèle peut voir plusieurs fois la même image et peut ne pas voir certaines images au sein d'une époque.

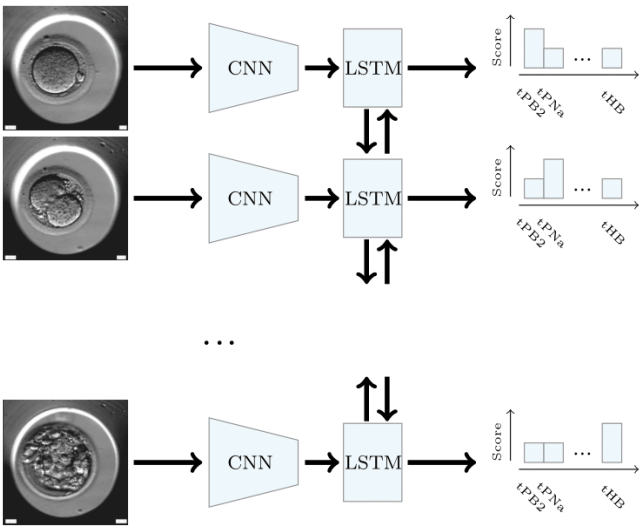
Nous utilisons 88% des vidéos pour l'entraînement, 6% pour la validation et 6% pour le test. Le modèle est entraîné jusqu'à ce que les performances sur l'ensemble de validation ne s'améliorent pas pendant 30 époques consécutives, avec un maximum de 80 époques. Le meilleur modèle de l'ensemble de validation est alors restauré et évalué sur l'ensemble de test.



(a) Le modèle ResNet-18



(b) Le modèle ResNet-3D



(c) Le modèle ResNet-LSTM

FIGURE 2 – Les différents modèles utilisés.

Pré-traitement et sélections des vidéos. Aucune augmentation de données n'est utilisée. Les images sont réduites de 500×500 à 224×224 pour réduire l'occupation de mémoire GPU.

Certains embryons se développent mal et leur croissance peut être interrompue par un biologiste. Lorsque cela se produit, étant donné que ces embryons ne seront pas transférés les biologistes ne les annotent pas aussi soigneusement que quand il s'agit d'embryons avec un développement normal. Afin de réduire le bruit dans les annotations, nous choisissons de ne garder que les vidéos montrant au moins 6 phases de développement distinctes. Le nombre de vidéos final est de 788.

Initialisation des poids. Les poids du ResNet-18 et du ResNet3D sont respectivement pré-entrainés sur ImageNet [34] et Kinetics [35]. Les poids du LSTM sont initialisés de manière aléatoire.

Matériels et logiciels utilisés Les modèles ont été implémenté avec la bibliothèque Pytorch 1.4.0 [36]. Une partie des modèles a été entrainés sur deux cartes Tesla P100 (32 Go de mémoire graphique au total) et, pour des raisons de temps, le reste des modèles a été entrainé sur CPU.

5.1 Métriques

Plusieurs métriques sont utilisées pour évaluer les modèles. La précision p est une des métriques les plus utilisées en classification d'images et se calcule par la proportion d'images correctement étiquetées par le modèle. Nous calculons également cette métrique en utilisant l'algorithme de Viterbi au préalable afin que les prédictions soient cohérentes tout au long de la vidéo, comme il est souvent fait dans la littérature [11, 18, 17, 16]. Cette deuxième métrique est notée p_v. La précision temporelle p_t est la proportion moyenne des transitions de phase qui sont prédites suffisamment proches de la transition réelle correspondante. Par "assez proche", nous entendons que le temps de la transition prédite ne se situe pas à plus d'une heure du temps de la transition réelle. Elle exige donc que les prédictions soient rendues cohérentes en utilisant Viterbi. Une métrique similaire est utilisé dans [37]. Enfin, la corrélation linéaire c entre le temps de transition prédit et le temps réel de la transition correspondante. Elle nécessite donc également Viterbi.

5.2 Résultats

Dans le Tableau 1 sont indiquées les performances moyennes avec écart type après 10 validations croisées et dans le Tableau 2 sont indiqués les p-valeurs montrant la signifiance statistique des écarts entre les performances moyennes. Les valeurs moyennes de la corrélation linéaire sont très proche de 1 probablement parce que l'algorithme de Viterbi impose d'avoir des prédictions cohérentes et donc une transition entre deux phases tardives sera toujours prédite après une transition entre deux phases au début du développement. Cela implique que les instants auxquels les modèles prédisent une transition sera toujours une fonction monotone croissante du temps, ce qui biaise la corrélation vers 1.

L'architecture ResNet-3D permet un gain significatif de performance par rapport à ResNet-18 et ResNet-LSTM. Le fait que le modèle ResNet-LSTM ait une performance qui ne soit pas significativement supérieure à celle de ResNet indique qu'il faudrait peut-être augmenter la capacité de la partie LSTM de ce modèle (augmentation du nombre de couches, de la taille des unités cachées) ou la longueur des séquences traitées. Il est intéressant de noter que même sur des séquences très courtes (4 images), ResNet-3D arrive à obtenir des performances significativement meilleures que ResNet-18. Cela indique que le gain de performance pourrait éventuellement être encore meilleur sur des séquences

Modèle	p	p_v	c	p_t
ResNet	0.7 ± 0.03	0.72 ± 0.03	0.9917 ± 0.0018	0.51 ± 0.04
ResNet-3D	$\mathbf{0.74 \pm 0.03}$	$\mathbf{0.77 \pm 0.03}$	$\mathbf{0.9934 \pm 0.0011}$	$\mathbf{0.61 \pm 0.02}$
ResNet-LSTM	0.71 ± 0.04	0.72 ± 0.04	0.9913 ± 0.0023	0.52 ± 0.05

TABLE 1 – Les performances des modèles évalués. Les valeurs indiquées sont les moyennes accompagnées des écart-types.

Métrique	Modèle	ResNet	ResNet-3D
p	ResNet-3D	**0.01**	
	ResNet-LSTM	0.43	0.1
p_v	ResNet-3D	**0.01**	
	ResNet-LSTM	0.8	**0.02**
c	ResNet-3D	**0.02**	
	ResNet-LSTM	0.67	**0.02**
p_t	ResNet-3D	$\mathbf{5e-06}$	
	ResNet-LSTM	0.63	**0.00013**

TABLE 2 – p-valeurs obtenues suite à la validation croisée. Le modèle ResNet-3D est significativement meilleur que ResNet-18 selon toutes les métriques et que ResNet-LSTM selon trois métriques. Le gain apporté par ResNet-LSTM par rapport à ResNet-18 n'est pas significatif.

plus longues. Les métriques p et p_v ont le défaut de pénaliser autant les modèles qui proposent des transitions de phases loin de vraies transitions que ceux qui se trompent de quelques images seulement. La métrique de précision temporelle prend en compte cet aspect : un modèle qui prédit un changement de phase près du changement de phase réel est favorisé par rapport à un modèle qui est loin de la vérité. Cette métrique montre que la performance du ResNet-3D est largement supérieure à celle de ResNet ou ResNet-LSTM (61% contre 51% ou 52%). Le gain de performance de ResNet-3D est donc sous-estimé si l'on se réfère seulement aux métriques de précision ou précision temporelle.

La figure 3 illustre bien en quoi ResNet-3D est meilleur que ResNet-18. Ce dernier produit des prédictions qui sont instables alors que ResNet-3D parvient à faire des prédictions plus cohérentes, probablement parce qu'il agrège les images au niveau temporel.

6 Travaux futurs

Les expériences faites dans ce travail n'utilise qu'un seul plan focal par image, ce qui empêche d'avoir une bonne vision de certaines parties du puits dans lequel se trouve l'embryon (i.e les zones situées au dessus et en dessous du plan focal). Les Embryoscopes capturent chaque image en utilisant tous les plans focaux ce qui produit des images 3D. Utiliser les images provenant de tous les plans focaux permettrait probablement d'améliorer la performance. Il serait aussi intéressant de forcer le modèle à se concentrer sur une partie de l'image afin de rendre sa décision plus interprétable comme par exemple avec de l'attention dure dans [18] ou avec de l'attention douce dans [38]. L'attention douce ne réduit pas la taille de l'image traitée mais permet de mettre en valeur n'importe quel détail dans l'image, peu importe sa forme, ce qui améliore l'interprétabilité. Dans le futur nous souhaitons exploiter de plus longues séquences d'images et donc mieux tirer partie des systèmes avec LSTM, mais il faut résoudre le problème d'occupation mémoire GPU que cela pourrait entraîner. Enfin, nous allons utiliser les paramètres MC prédits par notre modèle ainsi que d'autres variables biologiques pour prédire le taux de grossesse.

7 Remerciement

Cette étude est financée par le projet NExT (I-SITE) ANR-16-IDEX-0007.

Références

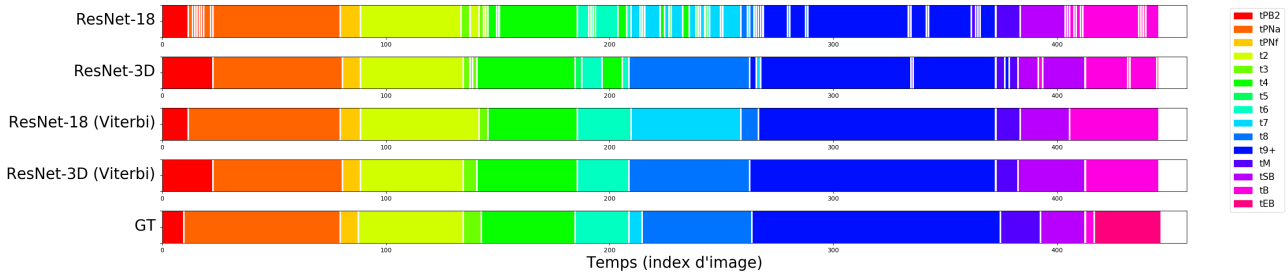
[1] Marcia C. Inhorn and Pasquale Patrizio. Infertility around the globe : new thinking on gender, reproductive technologies and global movements in the 21st century. *Human Reproduction Update*, 21(4) :411–426, 03 2015.

[2] A. P. Ferraretti, V. Goossens, M. Kupka, S. Bhattacharya, J. de Mouzon, J. A. Castilla, K. Erb, V. Korsak, A. Nyboe Andersen, and European IVF-Monitoring (EIM) Consortium for the European Society of Human Reproduction and Embryology (ESHRE). Assisted reproductive technology in Europe, 2009 : results generated from European registers by ESHRE. *Human Reproduction (Oxford, England)*, 28(9) :2318–2331, September 2013.

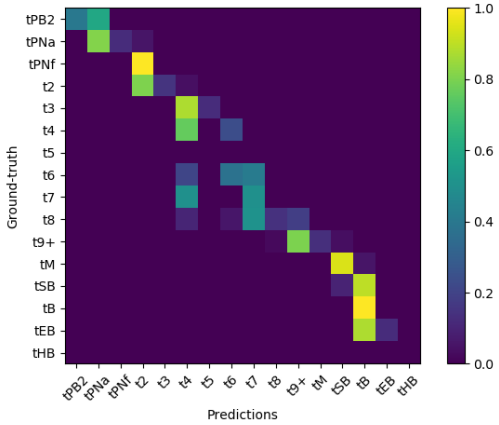
[3] S. Dyer, G.M. Chambers, J. de Mouzon, K.G. Nygren, F. Zegers-Hochschild, R. Mansour, O. Ishihara, M. Banker, and G.D. Adamson. International Committee for Monitoring Assisted Reproductive Technologies world report : Assisted Reproductive Technology 2008, 2009 and 2010†. *Human Reproduction*, 31(7) :1588–1609, 05 2016.

[4] Joe Conaghan, Alice A. Chen, Susan P. Willman, Kristen Ivani, Philip E. Chenette, Robert Boostanfar, Valerie L. Baker, G. David Adamson, Mary E. Abusief, Marina Gvakharia, Kevin E. Loewke, and Shehua Shen. Improving embryo selection using a computer-automated time-lapse image analysis test plus day 3 morphology : results from a prospective multicenter trial. *Fertility and Sterility*, 100(2) :412–419.e5, Aug 2013.

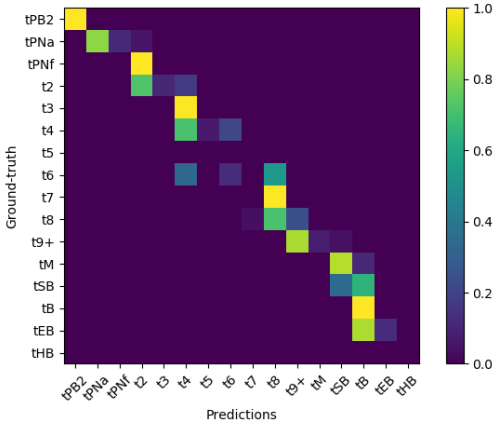
[5] pubmeddev and Pribenszky C. al, et. Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces



(a)



(b)



(c)

FIGURE 3 – Visualisation des prédictions faites par un modèle ResNet-18 et un modèle ResNet-3D sur une vidéo. Les modèles ont eu le même jeu d'entraînement, de validation et de test. La vidéo analysée ici fait partie du jeu de test. 3a) Une frise montrant les prédictions faites les deux modèles. Les deux lignes du haut montrent les prédictions brutes du modèles, les deux lignes suivantes montrent les prédictions obtenues après avoir appliqué l'algorithme Viterbi sur les prédictions brutes. La dernière ligne montre la vérité terrain et chaque couleur représente une phase de développement. Sans utiliser Viterbi, le modèle ResNet-3D donne des prédictions assez stables d'une image à l'autre alors que ResNet-18 change régulièrement de prédiction. La figure est mieux visible en couleur. En 3b) et 3c) sont montrées respectivement les matrices de confusion de ResNet-18 et ResNet-3D

early pregnancy loss : a meta-analysis. - PubMed - NCBI.

[6] Minghao Chen, Shiyou Wei, Junyan Hu, Jing Yuan, and Fenghua Liu. Does time-lapse imaging have favorable results for embryo incubation and selection compared with conventional methods in clinical in vitro fertilization ? A meta-analysis and systematic review of randomized controlled trials. *PloS One*, 12(6) :e0178720, 2017.

[7] Bjørn Molt Petersen, Mikkel Boel, Markus Montag, and David K. Gardner. Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3. *Human Reproduction (Oxford, England)*, 31(10) :2231–2244, 2016.

[8] Robert Milewski, Agnieszka Kuczyńska, Bożena Stankiewicz, and Waldemar Kuczyński. How much information about embryo implantation potential is included in morphokinetic data ? a prediction model based on artificial neural networks and principal component analysis. *Advances in Medical Sciences*, 62(1) :202 – 206, 2017.

[9] Linda Sundvall, Hans Ingerslev, Ulla Knudsen, and Kirstine Kirkegaard. Inter- and intra-observer variability of time-lapse annotations. *Human reproduction (Oxford, England)*, 28, 09 2013.

[10] H. Nadir Ciray, Alison Campbell, Inge Errebo Agerholm, Jesús Aguilar, Sandrine Chamayou, Marga Esbert, and Shabana Sayed. Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. *Human Reproduction*, 29(12) :2650–2660, December 2014.

[11] Aisha Khan, Stephen Gould, and Mathieu Salzmann. Deep convolutional neural networks for human embryonic cell counting. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 339–348, Cham, 2016. Springer International Publishing.

[12] R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Blastomere cell counting and centroid localization in microscopic images of human embryo. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, Aug 2018.

[13] Manoj Kumar Kanakasabapathy, Prudhvi Thirumalaraju, Charles L. Bormann, Hemanth Kandula, Irene Dimitriadis, Irene Souter, Vinish Yogesh, Sandeep Kota Sai Pavan, Divyank Yarravarapu, Raghav Gupta, Rohan Pooniwala, and Hadi Shafiee. Development and evaluation of inexpensive automated deep learning-based imaging systems for embryology. *Lab Chip*, 19 :4139–4145, 2019.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net : Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[15] François Chollet. Xception : Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[16] Nathan H Ng, Julian McAuley, Julian A Gingold, Nina Desai, and Zachary C Lipton. Predicting embryo morphokinetics in videos with late fusion nets |& dynamic decoders, 2018.

[17] Z. Liu, B. Huang, Y. Cui, Y. Xu, B. Zhang, L. Zhu, Y. Wang, L. Jin, and D. Wu. Multi-task deep learning with dynamic programming for embryo early development stage classification from time-lapse videos. *IEEE Access*, 7 :122153–122163, 2019.

[18] Tingfung Lau, Nathan Ng, Julian Gingold, Nina Desai, Julian J. McAuley, and Zachary C. Lipton. Embryo staging with weakly-supervised region selection and dynamically-decoded predictions. *CoRR*, abs/1904.04419, 2019.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8) :1735–1780, November 1997.

[20] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN : towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[21] A. Singh, J. Au, P. Saeedi, and J. Havelock. Automatic segmentation of trophectoderm in microscopic images of human blastocysts. *IEEE Transactions on Biomedical Engineering*, 62(1) :382–393, Jan 2015.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.

[23] S. Kheradmand, A. Singh, P. Saeedi, J. Au, and J. Havelock. Inner cell mass segmentation in human hmc embryo images using fully convolutional network. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1752–1756, Sep. 2017.

[24] R. M. Rad, P. Saeedi, J. Au, and J. Havelock. Multi-resolutional ensemble of stacked dilated u-net for inner cell mass segmentation in human embryonic images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3518–3522, Oct 2018.

[25] José Celso Rocha, Felipe José Passalia, Felipe Delestro Matos, Maria Beatriz Takahashi, Marc Peter Maserati Jr, Mayra Fernanda Alves, Tamie Guibu de Almeida, Bruna Lopes Cardoso, Andrea Cristina Basso, and Marcelo Fábio Gouveia Nogueira. Automatized image processing of bovine blastocysts produced in

vitro for quantitative variable determination. *Scientific Data*, 4 :170192 EP –, Dec 2017. Data Descriptor.

[26] Pegah Khosravi, Ehsan Kazemi, Qiansheng Zhan, Jonas E. Malmsten, Marco Toschi, Pantelis Zisimopoulos, Alexandros Sigaras, Stuart Lavery, Lee A. D. Cooper, Cristina Hickman, Marcos Meseguer, Zev Rosenwaks, Olivier Elemento, Nikica Zaninovic, and Iman Hajirasouliha. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *npj Digital Medicine*, 2(1) :21, 2019.

[27] Lucinda L. Veeck and Nikica Zaninovic. An Atlas of Human Blastocysts. Library Catalog : www.crcpress.com.

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[29] 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*, 01(01) :51–56, 2019.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[31] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278, Dec 2013.

[32] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. *CoRR*, abs/1708.07632, 2017.

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15 :1929–1958, 2014.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3) :211–252, 2015.

[35] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[36] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[37] M Feyeux, Arnaud Reignier, M Mocaer, J Lammers, Dimitri Meistermann, P Barrière, Perrine Paul, Laurent David, and Thomas Fréour. Development of automated annotation software for human embryo morphokinetics. *Human reproduction (Oxford, England)*, 03 2020.

[38] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better : Recurrent attention convolutional neural network for fine-grained image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.