Prédiction de l'âge cérébral chez le sujet sain en IRM anatomique par le deep learning

Paul Herent

UNIVERSITÉ PARIS DESCARTES

AVERTISSEMENT

Cette thèse d'exercice est le fruit d'un travail approuvé par le jury de soutenance et réalisé dans le but d'obtenir le diplôme d'Etat de docteur en médecine. Ce document est mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt toute poursuite pénale.

Code de la Propriété Intellectuelle. Articles L 122.4

Code de la Propriété Intellectuelle. Articles L 335.2-L 335.10

USPC
Université Sorbonne Paris Cité

UNIVERSITÉ PARIS DESCARTES

Faculté de Médecine PARIS DESCARTES

Année 2019 N° 73

THÈSE

POUR LE DIPLÔME D'ÉTAT

DE

DOCTEUR EN MÉDECINE

Prédiction de l'âge cérébral chez le sujet sain en IRM anatomique par le *deep learning*

Présentée et soutenue publiquement

le 22 mai 2019

Par

Paul HERENT

Né le 17 juin 1988 à Mont Saint Aignan (76)

Dirigée par M. Le Docteur Julien Savatovsky

Jury :

Mme La Professeure Laure Fournier, PU-PH …….. Présidente

Mme La Professeure Nathalie Lassau, PU-PH

M. Le Professeur Alain Luciani, PU-PH

RESUME EN FRANÇAIS

Nombre de mots : 545 (< 700)

Objectifs

Définir un processus explicite de prétraitement utilisable en clinique pour les données IRM
Prédire l'âge du cerveau à l'aide de divers algorithmes d'apprentissage automatique et d'apprentissage en profondeur
Définir les pièges courants inhérent à la méthodologie du Machine learning.
Vérifier par des méthodes d'interprétabilité si les connaissances apprises par les algorithmes sont celles en rapport avec le mécanisme du vieillissement cérébral.
Tester la validité du modèle entraîné sur une cohorte indépendante

Matériel et méthodes

Nous avons utilisé 1597 IRM pondérées en T1 en accès libre issues de 24 hôpitaux.
Le prétraitement a consisté à appliquer : la correction du champ de biais N4, l'enregistrement dans l'espace MNI152, la normalisation de l'intensité de la matière grise et blanche, le stripping du crâne et la segmentation du tissu cérébral. La
prédiction de l'âge du cerveau a été réalisée avec la complexité croissante des données saisies (histogrammes, données) et des modèles pour l'entraînement (modèles linéaires, modèles non linéaires tels que Gradient Boosting over decision tree, et enfin des réseaux de neurones convolutionnels (Convolutional Neuronal networks, CNN) 2D et 3D).
Les travaux sur l'interprétabilité des modèles consistaient (i) à procéder à la visualisation de données, telles que des cartes de corrélation entre l'âge et la valeur des voxels, et à générer (ii) des cartes montrant les régions cérébrales d'intérêt pour l'apprentissage.
Enfin, nous avons testé l'applicabilité de ce biomarqueur sur une cohorte indépendante de sujets sains et Alzheimer.

Résultats
Le temps de prétraitement des images était de 5 min pour une IRM 3D T1.
Nous avons trouvé une corrélation significative entre l'âge connu et le volume de matière grise avec une corrélation r = -0,74. Notre meilleur modèle prédisait l'âge cérébral avec une erreur absolue moyenne (EAM) de 3,60 ans, avec un réseau de neurones à convolutions (CNN) fine-tuned et pré-entraîné sur ImageNet (méthode de Transfer learning). L'EAM était à 5.5 ans en corrigeant notre modèle de l'effet centre.
Nos travaux sur l'interprétabilité ont révélé comme importantes pour les prédictions des régions cérébrales connues pour être impliquées dans le vieillissement (substance grise, en particulier l'insulta et les thalami, les ventricules, la substance blanche périventriculaires)

Conclusions

La prédiction de l'âge cérébral via le deep learning pourrait définir un biomarqueur du vieillissement cérébral, utilisable dans la pratique neuroradiologique quotidienne, Notre travail sur l'interprétabilité montre que les modèles apprennent le plus dans des régions cérébrales connues pour être affectées par le vieillissement (substance grise, siège de l'atrophie, substance blanche, affectée par la leucoaraiose, et la dilatation des ventricules avec l'âge), ce qui permet d'avoir plus confiance dans leur applicabilité par rapport à un modèle jusque là considéré comme une « boîte noire ».

Enfin, l'utilisation du modèle sur une autre cohorte de patients a montré (i) sa capacité de généralisation sur des données indépendantes, et d'autre part son application dans des maladies auxquelles il n'a pas été entraîné : l'âge estimé par l'algorithme est plus élevé chez des patients Alzheimer. Ce qui constitute une piste pour le dépistage plus précoce de cette maladie, via des méthodes d'imagerie. L'utilisation de méthodes similaires en IRM multimodale, incluant notamment l'IRM fonctionnelle, pourrait constituer une piste intéressante.

Ce biomarqueur pourrait être utilisé en substitution de la volumétrie dans des études visant à améliorer les performances de l'imagerie morphologique dans le diagnostic des pathologies dégénératives et dans les essais cliniques de traitements de ces maladies.

Mots clés : IRM, intelligence artificielle, deep learning, apprentissage profond, machine learning, apprentissage statistique, vieillissement cérébral, cerveau, démence, Alzheimer

ABSTRACT

Count word : 476

Objectives

To define a clinically usable preprocessing pipeline for MRI data
To verify by interpretability methods whether knowledge is learned by algorithms are those that are related to the mechanism of cerebral aging.
Test the validity of the model on an independent cohort

Data & Methods

We used 1597 open-access T1 weighted MRI from 24 hospitals.
Preprocessing consisted in applying : N4 bias field correction, registration to MNI152 space, white and grey stripe intensity normalization, skull stripping and brain tissue segmentation
Prediction of brain age was done with growing complexity of data input (histograms, grey matter from segmented MRI, raw data) and models for training (linear models, non linear model such as gradient boosting over decision trees, and 2D and 3D convolutional neural networks).
Work on the interpretability of models involved (i) visualizing, displaying maps, maps to corrode, displaying values, and (ii) generating heat maps which permitted to identify regions of interest used by the algorithm in its learning.
Finally, we tested the validity of this biomarker on an independent cohort of healthy subjects and Alzheimer's patients.

Results

Processing time seemed feasible in a radiological workflow : 5 min for one 3D T1 MRI.
We found a significant correlation between age and gray matter volume with a correlation r = -0.74. Our best model obtained a mean absolute error (MAE) of 3.60 years, with fine tuned convolution neural network (CNN) pretrained on ImageNet. The MAE was of 5.5 years with a method correcting the center effect.

Our work on interpretability on simpler models permitted to observe heterogeneity of prediction depending on brain regions known for being involved in ageing (grey matter, ventricles). Occlusion method of CNN showed the importance of Insula and deep grey matter (thalami, caudate nuclei) in predictions.

Conclusions

Predicting the brain age using deep learning could define a biomarker of cerebral aging, usable in daily neuroradiological practice. Our work on interpretability shows that models learn the most in brain regions known to be affected by aging. (grey matter, seat of atrophy, white matter, affected by leukoaraiosis, and ventricles, which tend to dilate with ageing). This methods give more confidence in their applicability of CNN in clinical practice, previously considered as a " Black Box ".

Finally, the use of the model on another cohort of patients showed (i) its ability to generalize on independent data, and on the other hand its application in diseases to which it was not trained: the estimated age by the algorithm is higher in Alzheimer's patients. This constitutes a pathway for earlier detection of this disease, via imaging methods. The use of similar methods in multimodal MRI, including functional MRI, could be an interesting avenue.

This biomarker could be used as a surrogate for Grey matter volumetry in studies aimed at improving the performance of morphological imaging in the diagnosis of degenerative diseases and in the clinical trials of treatments for these diseases.

Key words : MRI, artificial intelligence, deep learning, machine learning, brain aging, dementia, Alzheimer

REMERCIEMENTS

Mme la professeure Laure Fournier, vous me faites l'honneur de présider ce jury et d'évaluer ma thèse, je vous en remercie. Vous qui par vos travaux de recherche et votre implication pédagogique suivez de près l'évolution des radiomiques.

Mme la professeure Nathalie Lassau, vous me faites l'honneur de juger mon travail, je vous prie de trouver en ces mots l'expression de toute ma reconnaissance et mon respect, pour votre implication, notamment dans le challenge d'intelligence artificielle des journées françaises de radiologie, excellente façon de faire travailler ensemble médecin et ingénieurs sur ce sujet.

Mr le professeur Alain Luciani, vous me faites également l'honneur de juger mon travail, je vous remercie pour votre intérêt et vos lumières concernant les questionnements éthiques, stratégiques et politiques que représentent l'arrivée de l'intelligence artificielle dans le monde de la radiologie, pour vos réflexions au sein du groupe SFR IA.

A mon directeur de thèse, le docteur Julien Savatovsky, merci pour ton soutien dans ce travail, ton ouverture d'esprit et ton aide dans mes questionnements sur la voie à tracer dans ce domaine innovant de la radiologie.

Du côté d'Owkin…Merci à
Simon Jegou, pour ton mentorat en machine learning, et cette amitié qui est née de ce travail
Thomas Clozel et Gilles Wainrib, pour votre accueil à Owkin, très bienveillant

Milles merci à vous également :
Valentin et Sylvain Amé Toldo, pour votre aide sur les belles figures et la conception du blogpost,
Toute l'équipe Owkin, pour l'excellent travail d'équipe que nous avons fait (et je l'espère que nous ferons encore) entre Paris et New York: Anna Huyghues Despointes, Anna I. Bondarenko, Pierre Courtiol, Derek T. Russell-Kraft, Cedric Whitney, Meriem Sefta, Vincent Lepage , Adrian Gonzalez, Maxime SE, Paul Jehanno, Raphaël Léger, Alicia Simion, Eric Tramel, Mikhail Zaslavskiy, Pierre Manceron, Chloé Simpson, Paul Mabillot, Valentin Amé, Mathieu Galtier, Camille Marini, Sylvain Toldo, Sylvain Toldo, Charlie Saillard, Olivier Dehaene, Olivier Moindrot, May Kang Yu, et aux tous nouveaux arrivants que je ne connais pas encore.

Pascal Roux, pour ton soutien, ton aide, tes conseils, ta bienveillance dans le tout début de mes choix professionnels.

Mes parents, mes frères, ma famille, vous qui m'avez soutenu depuis le début de ces longues études.

Axelle, pour ta patience, ton soutien, et tout le reste
Mes amis : Luis, Laetitia, Lucciano, Sarah, Simon, Vanessa, Karim, Kim, Arnaud, Manon, Antoine, Arnaud, Kelly, Thomas, Francesca, Philippe, Emma, Aurelien, Maxime, Adrien, Michael, Clement,

Barnabé, Charles, Eri, Alejandro, Jeremy, Antoine, Thibaud, Delphine, David, et tous ceux que j'oublie (ne m'en voulez pas) qui m'accompagnent dans la vie en dehors de la médecine. Merci à vous d'être là.

UN MOT SUR OWKIN

Cette thèse est le fruit d'un travail effectué en collaboration avec les membres du laboratoire de recherche et développement de la société Owkin

-À propos d'Owkin:

OWKIN a été co-fondé en 2016 par le Dr Thomas Clozel, hématologue, et le Dr Gilles Wainrib, pionnier dans le domaine de l'intelligence artificielle appliquée en biologie. OWKIN a passé la phase de validation du concept et fournit maintenant ses algorithmes innovants d'Intelligence artificielle à plusieurs des plus grands centres anticancéreux et sociétés pharmaceutiques en Europe et aux États-Unis. Avec des bureaux à New York, Paris, Londres, Nantes nous sommes fiers de créer une culture d'entreprise axée sur la transparence, la collaboration, les défis, l'optimisme et le plaisir.

L'équipe d'Owkin est internationale, multidisciplinaire et est constituée d'une équipe de *data scientists* experimentés, médecins et d'un board stratégique axé sur la *business intelligence*. Leurs data scientists sont parmi les meilleurs au monde, avec plusieurs Maîtres Kaggle (au sein du top 100), un des plus performants du DREAM Challenge et des publications dans les journaux de référence en machine learning tels que ICML, NIPS et autres.

Coordonnées :
https://owkin.com/
Owkin France,
75 rue de Turbigo,
75003 Paris

UN MOT SUR LE BLOG :

Ce travail académique est prolongé par un travail pédagogique accessible en ligne. Nous avons en effet co-rédigé avec Simon Jegou, data scientist chez Owkin, un post de blog visant à démystifier le *Machine learning* pour les médecins. Nous vous invitons à le lire sur le lien suivant :

https://medium.com/owkin/a-machine-learning-survival-kit-for-doctors-97982d69a375

Vous pouvez également photographier avec votre smartphone le QR code ci-dessous :

figure 0 : notre QR code pour aller voir notre article de blog sur le sujet

Table des matières

I INTRODUCTION:

I Qu’appelle-t-on « intelligence artificielle » en 2019 ?

Certaines confusions peuvent être faites dans les définitions: aujourd'hui, ce que l'on appelle «l'intelligence artificielle» en médecine concerne principalement des tâches d'apprentissage machine supervisé. Le deep learning est un sous-type de techniques d'apprentissage automatique appelé «deep» (profond) en raison de l'architecture numérique utilisant un nombre important de couches de neurones artificiels (1). La majorité des articles sur l'apprentissage approfondi utilisent des réseaux de neurones convolutifs (Convolutionals neuronal networks, CNN), secondaire à son succès observé dans le cadre de compétitions d’algorithmes de Vision par ordinateur (2), où ces techniques ont dépassé de façon significative les performances de traitement d’information visuelle.

I Quelles sont les limites de la radiologie actuelle ?

L’interprétation radiologique aujourd’hui

Une partie du travail d'un radiologue aujourd'hui peut être décrite comme une tâche perceptuelle et cognitive : en tant que tâche perceptuelle, le médecin doit reconnaître les caractéristiques considérées comme normales ou non. En tant que tâche cognitive, il doit décrire les caractéristiques pertinentes dans un rapport radiologique écrit et enfin conclure en fournissant au patient et à ses médecins quelle est son interprétation globale : un ensemble d’hypothèses diagnostiques, des critères pronostics, et le suivi d’éléments objectifs pour guider la décision thérapeutique de l’équipe de soins.

Certaines interprétations de caractéristiques sont très connues et classées de manière consensuelle, corrélées à un risque connu. Par exemple, dans l’imagerie du cancer du sein, l’American College of Radiologist (ACR) définit la classification de Bi-Rads (Système de

compte rendu et de données d'imagerie du sein) (3), qui vise à classer les caractéristiques radiologiques caractéristiques en 6 classes (de 0 à 5) correspondant à 6 probabilités différentes de malignité. Ainsi, le Bi-Rads fournit un outil de décision basé sur des preuves pour le médecin, indexé sur un risque probabiliste de malignité (pratiquer un examen de suivi, faire une biopsie…).
Cependant, obtenir des résultats reproductibles n'est pas une tâche aisée, pour plusieurs raisons.

Argument de la crise de la reproductibilité dans la recherche médicale

Les données de la littérature médicale sont publiées dans des revues à comité de lecture, dont le rôle est notamment de garantir une rigueur d'application de la méthodologie scientifique. Or ce n'est pas suffisamment le cas en pratique.
En 2016, une enquête publiée dans Nature met en lumière le problème de la reproductibilité en sciences (4). Dans cette enquête, 1576 chercheurs ont été interrogés sur un problème de réplication via un bref questionnaire. Plus de 70% d'entre eux ont essayé de reproduire les expériences d'un autre scientifique et plus de la moitié n'ont pas réussi à reproduire leurs propres expériences. Concernant la recherche en médecine, le même schéma de réponses est apparu : les répondants ont rapporté entre 55 et 75% d'échecs lors de la reproduction d'une expérience. Sur la base de ce fait, il semble difficile d'appeler «Evidence Based Medicine» une médecine basée sur des expériences non reproductibles, malgré l'existence de méthodes d'évaluation de la qualité des preuves, comme par exemple les directives GRADE (5). Il existes de nombreuses causes pour expliquer cela, parmi lesquelles la pression de publication, le biais de publications d'études positives, l'impossibilité d'exécuter de façon indépendante les analyses statistiques par limite d'accès aux données, le manque de puissance statistique du fait de volumes de données insuffisants, le manque de coopération entre les équipes de recherche, le manque de temps des médecins chercheurs.

Argument des erreurs médicales :

Les erreurs en médecine ont une prévalence significative (le taux de diagnostics manqués, incorrects ou retardés est estimé à 10% à 15%) (6).
Plus spécifiquement, en radiologie, le taux d'erreur rétrospectif rapporté parmi les examens radiologiques peut monter jusqu'à 30% (7). Si certaines erreurs sont sans conséquences, certaines ont un impact sur la prise en charge du patient, et représentent de plus un coût élevé pour les systèmes de santé, estimées à plus de 38 milliards de dollars par an aux États-Unis.

Dans la revue sus-citée (6), les trois causes d'erreurs liées les plus fréquentes concernaient le défaut de détection d'une anomalie (42% des erreurs recensées), la satisfaction de la recherche (une autre anomalie est manquée à la découverte d'une première anomalie ; 22% des erreurs) et un raisonnement erroné (c.-à-d. une erreur de caractérisation ; chiffre élevé à 9%). Parmi les causes d'erreurs, on peut citer l'implication de la fatigue (8), le manque d'ergonomie des outils employés, les interruptions du radiologue lors de son interprétation, les problèmes d'organisation dus à l'augmentation du volume d'images médicales demandées dans la pratique médicale, l'utilisation encore insuffisamment systématique de scores qualitatifs … Ces erreurs pourraient également expliquer la grande variabilité intra-observateur observée dans l'interprétation radiologique (9,10), Cette variabilité pourrait être notamment être limitée par l'utilisation plus systématique de score qualitatifs de

Quel avenir souhaitable pour l'imagerie médicale ?

Au-delà de la perception visuelle humaine par des outils de quantification informatiques : la place de la radiomique

Des données pertinentes sont présentes dans les données d'imagerie qui sont humainement difficiles à quantifier. Les données radiologiques peuvent être implantées dans des modèles incluant d'autres données (données cliniques, données génomiques, données histologiques). Une nouvelle discipline, appelée Radiomique a été créée pour répondre à ces défis. Il peut être défini comme une «extraction à haut débit d'éléments d'image quantitatifs issus de l'imagerie médicale standard de soins qui permet d'extraire des données et de les appliquer au sein de systèmes d'aide à la décision clinique afin d'améliorer la précision diagnostique, pronostique et prédictive» (pour une plus grande précision). Pour une revue, voir (11).

De manière plus synthétique, on pourrait définir la radiomique comme l'analyse computationnelle des images médicales.

Certains outils de radiomique se sont déjà montrés efficaces. Un bon exemple est l'analyse de texture. Il est humainement difficile de quantifier le degré d'hétérogénéité d'une tumeur, par exemple, contrairement aux algorithmes d'analyse de texture. De plus, certaines publications ont montré que cette méthode peut refléter à l'échelle tissulaire une mutation génomique. Cela pourrait avoir un impact important sur l'accélération de la médecine de précision et la réduction des coûts de séquençage des tissus. Certaines preuves récentes ont montré l'intérêt de cette technique pour la prédiction de la mutation de l'EGFR dans le scanner d'un adénocarcinome pulmonaire (12) ou de la mutation IDH1 dans l'IRM d'un gliome de bas grade (13).

Un changement de paradigme au sein de la radiomique avec l'arrivée du deep learning.

2012 est l'année où une nouvelle famille d'algorithme, basée sur des architectures profondes de réseaux de neurones artificiels (*deep learning*), s'est montrée pertinente dans des tâches de *computer vision.* Après cette date, de nombreux cas d'usage en médecine ont été publiés : en histologie (pour une revue, voir (14), pour un article récent publié par notre équipe, voir (15), en dermatologie (16), ophtalmologie (17), radiothérapie (pour une revue, voir (18) …
Le champ d'applications de l'apprentissage en profondeur en radiologie est potentiellement très vaste et pourrait révolutionner chaque étape de traitement des images médicale: reconstruction d'images (19), segmentation de lésions, diagnostic (caractérisation de la lésion, prédiction pronostique,…), pronostic et suivi sous traitement, mais aussi corrélations histologie-radiologie, génétique-radiologie… Pour des revues listant ces applications, voir (20), (21).

Comme le mentionnent les livres blanc de la Société canadienne ou Française de radiologie axé sur l'intelligence artificielle (22),(23), la combinaison de l'amélioration de la disponibilité de grands ensembles de données et de l'augmentation de la puissance de calcul permet d'arriver plus vite à des cas d'usages cliniques. Pour cela, une collaboration étroite entre médecins et ingénieurs est nécessaire, permettant une synergie entre experts des données médicales et experts des sciences des données. Ce travail est le fruit de cette synergie.

Une boîte noire ou une intelligence artificielle explicable ?

Pour pouvoir être mis en œuvre en pratique clinique, les algorithmes doivent être compris.
Il existe un compromis bien connu dans la communauté Machine learning entre Performance et interprétabilité de ces outils mathématiques. Un algorithme simple est plus interprétable, mais moins précis. Un algorithme plus complexe peut produire des hautes performances (le deep learning en est l'archétype actuel), mais on lui reproche d'être une boîte noire, c'est à dire qu'il est impossible d'expliciter directement son fonctionnement et donc le mécanisme amenant à la prédiction.
De plus, les systèmes d'IA sont très sensibles aux biais, et leur prédiction peut être le résultat de la capture de facteurs de confusion au lieu de répondre à la question posée.

Dans ce travail nous aborderons le problème de la sensibilité aux biais en abordant l'exemple de l'effet centre. Chaque hôpital possède une IRM construite par un constructeur donné, avec des réglages de séquences donnés, un traitement de l'image brute donnée, et une population donnée (âge, sex, ethnicité….). Ainsi, il faut se hâter de ne pas conclure lors

des premières prédictions obtenues par de tels algorithme, en s'assurant que la performance donnée correspond à la tâche demandée (ici, prédire l'âge cérébral ; et non pas prédire l'hôpital d'où provient l'image)
Tous ces facteurs de variabilité se traduisent dans l'image finale et constituent des biais qui pourront être « captés » par l'algorithme. Nous expliciterons comment pallier à ces effets dans la partie méthode.

Cette sensibilité aux biais peut définir le manque de "sens commun" de la machine, c'est à dire une captation de biais qu'un humain n'aurait jamais fait. Pour surmonter ce manque de sens commun et pour une meilleure acceptabilité en pratique clinique, un modèle doit être aussi précis et explicable que possible. Le vrai défi est ici. Il existe tout un champ de recherche en mathématiques appliqués interessant à l'interprétabilité des modèles.

Parmi les méthode d'interprétabilité les plus connues, on peut citer la méthode LIME (*Local Interpretable Model Agnostic Explanations*), (24) qui comme son nom l'indique est un modèle qui peut s'appliquer indépendamment du modèle d'entraînement utilisé pour les prédictions (d'où son caractère « agnostique »), consistant à rendre « représentable » un vecteur issu d'une abstraction algorithmique (un vecteur dans un espace latent). Pour le cas des images, cela consiste en la génération d'une carte de « température » (heat map) montrant des régions d'importance dans les prédictions, permettant alors de faire le lien avec les caractéristiques de la région de l'image donnée et de juger de la pertinence de cette région.
Une autre méthode, celle que nous avons employé dans ce travail, consiste à occlure une région de l'image donnée à un algorithme entraîné et d'en étudier l'impact sur les prédictions (25). Cette méthode a l'avantage elle aussi d'être agnostique de l'architecture du modèle.
Pour un aperçu des différentes techniques d'interprétabilité utilisables en médecine, voir (26).

La prédiction de l'âge cérébral.

La Physiologie du vieillissement cérébral est classiquement étudiée de deux manières : structurelle et fonctionnelle.

Etude structurelle du vieillissement cérébral

De manière structurelle, les études sur le vieillissement cérébral ont porté sur les caractéristiques de vieillissement de la matière grise, de la substance blanche, et la résultante de la perte de substance cérébrale traduite par une augmentation de volume de

liquide céphalo-rachidien (LCR). Pour la matière grise, des techniques morphométriques telles que l'épaisseur corticale à base de voxel (VBCT) et la morphométrie à base de voxel (VBM) ont permis de quantifier l'atrophie corticale dans différentes régions du cerveau (27). Par exemple, avec la technique VBM, un processus d'atrophie accélérée a pu être mesuré dans des régions telles que l'insula, les gyri pariétaux supérieurs, les sillons centraux et les sillons cingulaires, alors qu'aucun effet de vieillissement n'a été observé dans l'amygdale, l'hippocampe et le cortex entorhinal, montrant l'hétérogénéité du processus d'atrophie cérébrale (28).

Concernant l'étude de la substance blanche, la leucoaraïose s'est révélée être un bon prédicteur du vieillissement physiologique. Ce terme vient du grec *leukos* : « blanc » et *araios* : « raréfaction ». En IRM, elle se traduit par des hypersignaux FLAIR, (le terme « white matter hyperintensities » est le terme usuel dans la littérature). En pondération T1, elle se traduit par un hyposignal. D'un point de vue histologique, elle correspond à des altérations vasculaires (micro-angiosclérose) avec épaississement fibrohyalin de la substance blanche. Des corrélations ont été trouvées entre la leucoaraïose et le déclin cognitif : une atteinte de la substance blanche est associée à de moins bonnes performances cognitives (29). La physiopathologie de l'atteinte des petits vaisseaux est multifactorielle, avec un rôle d'inflammation, de perturbation de la barrière hémato-encéphalique, de facteurs génétiques, de facteurs ischémiques (30).

Etude fonctionnelle du vieillissement cérébral

Le vieillissement cérébral a également été étudié avec par techniques d'IRM fonctionnelle. La méthode de connectivité fonctionnelle, permettant d'étudier les relations entre différentes régions du cerveau en fonctionnement, a permis d'identifier les régions du cerveau associées à la réserve cognitive (définie par la divergence entre les symptômes cliniques et les effets du vieillissement dans la maladie d'Alzheimer). Par exemple, des régions telles que les régions temporales médiales et le cortex cingulaire antérieur ou postérieur étaient associées à la réserve neurale et à la compensation neurale (stratégie cognitive alternative en cas de dysfonctionnement cognitif) dans les régions frontales et le réseau d'attention dorsale (31).

Le vieillissement cérébral en radiologie

Aujourd'hui, en routine radiologique, l'évaluation du vieillissement physiologique consiste principalement en 6 tâches, telles que :

- (i) l'évaluation catégorielle de l'atrophie (par exemple: le score de l'atrophie du lobe temporal médial (MTA / score de Scheltens), visant à distinguer une atrophie normale de l'atrophie pathologique (32),
- (ii) évaluation de la dilatation des ventricules (la tâche consiste à identifier l'atrophie de l'hydrocéphalie) (33)
- (iii) l'évaluation des hyperintensités de la substance blanche (White matter Hyperintensities, WMH) sur les séquences IRM T2 et FLAIR. Cette tâche consiste tout d'abord à identifier le profil de leucoaraïose par rapport à d'autres modèles de WHM (par exemple: inflammatoire, comme dans la sclérose en plaques), et en second lieu, le niveau de leucoaraïose avec l'échelle de Fazekas (34).
- (iv) la détection, le décompte et l'évaluation de la distribution spatiale des microbleeds cérébraux (Cerebral microbleeds, CMB), permettant de mieux caractériser les sous types de maladie des petits vaisseaux : liés à l'hypertension, les CMB on une répartition spatiale profonde et sont associés à un risque cardiovasculaire élevé. Liés à l'angiopathie amyloïde, ils sont de distribution superficielle, lobaire. Pour une revue récente, voir (35).
- (v) l'évaluation de séquelles d'AVC
- (vi) l'évaluation des espaces de Virchow Robin

Schématiquement, 3 séquences IRM sont utilisées pour ces tâches: IRM pondérée en T1 en echo de gradient, en T2 (séquence FLAIR) pour détecter les anomalies de la substance blanche (et le LCR), T2 * ou SWI pour les évaluations de CMB.

La prédiction de l'âge du cerveau : un biomarqueur potentiel dans les maladies neurologiques

La prédiction de l'âge du cerveau en IRM avec des techniques de machine Learning consiste à essayer de prédire Y (l'âge du cerveau) à partir de X (les données d'IRM) avec un modèle calculant une fonction f (X) pour prédire Y.

Un article récemment publié traitait précisément de ce sujet (36) et nous fournissait un résultat "state-of-the-art" pour cette tâche: L'erreur de prédiction était de 4,65 ans (Erreur absolue moyenne, MAE) avec des réseaux de neurones convolutifs (CNN) 3D formés à l'IRM cérébrale brute pondérée en T1 (c'est-à-dire au prétraitement minimal) avec un ensemble de données de 2001 sujets.

La prédiction de l'âge du cerveau en machine learning a été étudiée comme biomarqueur du vieillissement accéléré de certaines maladies telles que la schizophrénie (37), la sclérose en plaques (38), les lésions cérébrales traumatiques (39), les *Mild cognitive impairment* (déclin cognitif léger) (40), la maladie d'Alzheimer (41).
La prédiction de l'âge du cerveau a également été évaluée dans la sclérose en plaques, montrant la différence entre les personnes en bonne santé et les patients (38) : le cerveau

d'un patient atteint de SEP avait en moyenne un âge estimé de 10 ans de plus qu'un sujet sain).

L'âge estimé est corrélé à d'autres biomarqueurs connus du vieillissement : motricité (force de préhension plus faible), souffle (fonction pulmonaire moins performante), cognition (vitesse de marche plus lente, moindre flexibilité cognitive), stress chronique (charge allostatique plus élevée) et mortalité (42).

Certains articles ont mis l'accent sur l'interprétabilité des CNN dans des travaux connexes tels que la tâche de classification de la maladie d'Alzheimer sur l'IRM structurelle (43) ont montré que les régions corticales (en particulier dans l'hippocampe) et les ventricules étaient importantes pour la prédiction. Cependant, à notre connaissance, aucun travail sur l'interprétabilité avec une méthode similaire n'a été utilisé sur des sujets sains.

II NOTRE ETUDE

Objectifs de cette étude:

- Définir les principales étapes préliminaires d'un projet d'apprentissage automatique: accès aux données, nettoyage des données, analyse des données.
- Définir un pipeline de prétraitement utilisable en clinique pour la prédiction de l'âge cérébral.
- Reproduire les résultats précédents (prévision de l'âge cérébral avec 3D CNN) et essayez de les dépasser.

- Comparer les performances de CNN avec différents modes de représentation des données (histogramme des intensités, masque de segmentation, images traitées), et des des modèles de complexité croissante.
- Révéler certains pièges inhérents à la méthodologie du machine learning.
- Employer des méthodes d'explicabilité / interprétabilité des CNN.
- Explorer la validité du modèle sur une cohorte indépendante de sujets sains et de patients Alzheimer.

MÉTHODE

Obtention des données

Après téléchargement et nettoyage des données, nous avons pu constituer une cohorte de 1597 IRM pondérées en T1, séquence en echo de gradient, issues de cohortes publiques de sujets sains : Jeu de données 1 issue de la base d' IXI)) (563 sujets) et jeu de données 2 du Functional Connectome Project (1054 sujets).
Sur le jeu de données 1, les IRM ont été effectuées sur 1,5 tesla et 3 tesla (IRM Philips 3T et 1.5T et une IRM GE 1.5T). Sur le jeu de données 2, la majorité des IRM ont été réalisées sur des IRM à 3 T. Les descriptions exhaustives des paramètres des séquences et puissances respectives des IRM par centres sont disponibles via ce lien : https://www.nitrc.org/docman/?group_id=296.
Pour une analyse ancillaire, nous avons également téléchargé des données publiques de patients Alzheimer, issus de la cohorte publique ADNI (489 sujets).

Toutes les précisions sur les séquences sont présentes sur ce lien : http://adni.loni.usc.edu/methods/mri-tool/mri-analysis/.

Analyse des données

La figure 1 est un résumé de l'analyse des données sur les deux jeux de données que nous avons réalisée (les données issues de la cohorte ADNI ne sont pas détaillées dans cette figure). La figure 1A montre les écarts de nombre de patients en fonction du centre et la taille relative des deux jeux de données. La figure 1B montre l'hétérogénéité de la distribution par âge selon les centres.

Sur la figure 1C, nous avons effectué une tâche d'apprentissage automatique non supervisée, appelée t-SNE embedding, afin d'identifier des groupes de patients (« clusters ») présentes dans le jeu de données.
La méthode de t-SNE est une méthode de réduction de dimension couramment utilisée par les ingénieurs en intelligence artificielle, car elle permet de rapidement visualiser un espace de grande dimension (ici les 1597 matrices 3D correspondant aux IRM) dans une matrice 2D ou 3D (ici, chaque IRM était résumée en un point, sur un plan 2D).
Cette méthode a permis d'identifier un fort biais par rapport au centre où ont été acquis l'IRM, tels que le nombre de sujets par centres (hôpitaux), la répartition de l'âge, le sexe, mais principalement en raison de l'acquisition d'images : type d'IRM utilisé, taille du champ de vue (FOV) pour l'acquisition du cerveau, méthode de « *defacing* » employée pour anonymiser l'ensemble de données, variations du rapport signal sur bruit en raison du réglage des paramètres utilisés pour la séquence d'IRM pondérée en T1, la puissance du champ magnétique employé…
Cette étape permet de montrer l'intérêt de la standardisation : nous ne voulons pas que nos modèles d'apprentissage automatique capturent les biais par rapport au centre de prédiction de l'âge (par exemple, il existe des centres avec des valeurs aberrantes (un centre avec beaucoup de sujets très âgés), si notre modèle identifie des caractéristiques relatives à ce centre (ex : contraste spécifique à l'IRM utilisée), elle l'aidera à prévoir l'âge avancé, au lieu de capturer les « features » d'intérêt : c'est à dire les signes radiologiques du vieillissement cérébral (atrophie, leucoaraïose, dilatation ventriculaire).

Pour cela, de multiples stratégies peuvent être adoptées pour contrôler ces facteurs de confusions :

- augmenter la taille de l'échantillon (lorsque c'est possible) avec plus de centres;
- augmentation des données (data augmentation) : le but de cette technique est de montrer au modèle les mêmes images en phase d'apprentissage, mais de manière différente (effectuer certaines opérations de base telles que des rotations, des

translations d'images, différentes échelles d'intensité, etc.)

- méthode de validation croisée (cross validation) adaptée entre données d'entraînement et données de test.

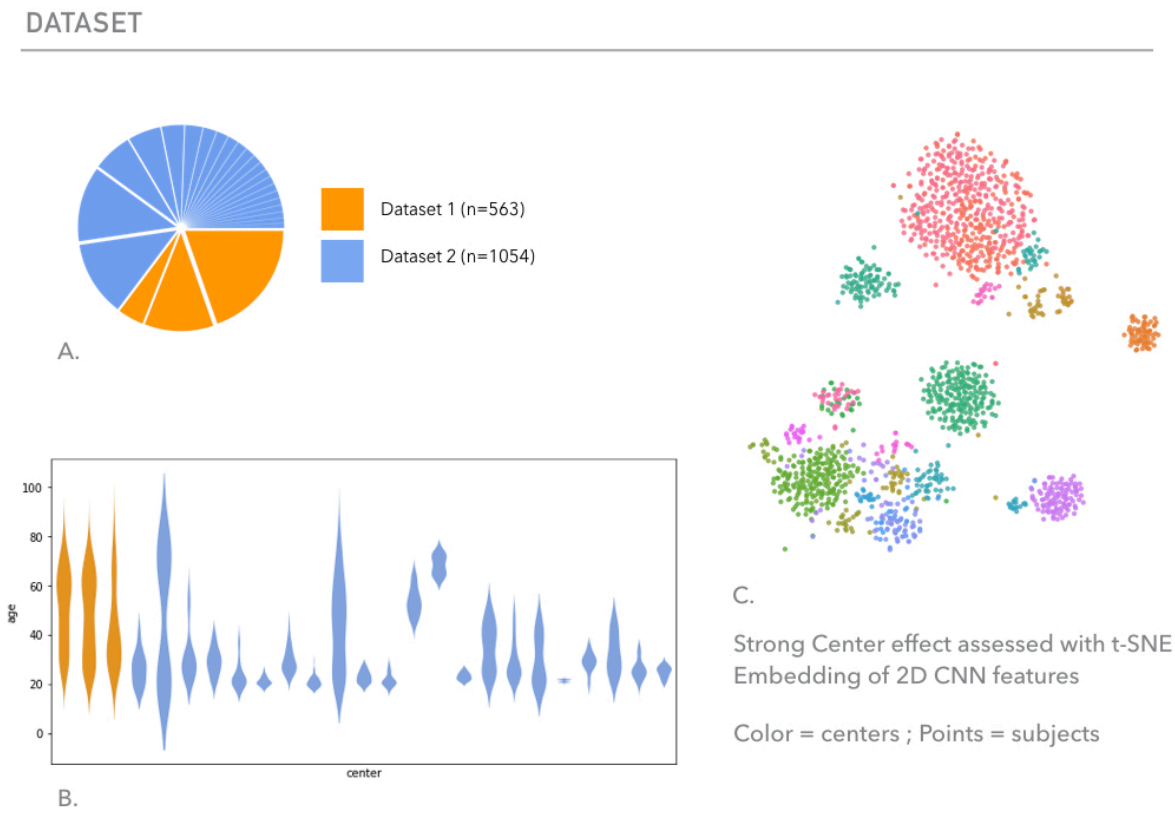

Figure 1: Description du jeu de données des sujets sains (hors cohorte ADNI).

1A : boîte à camembert illustrant des jeux de données (dataset) et leurs importances respectives. La couleur bleue représente le dataset 1, la couleur orange le dataset 2. Les parts du camembert correspondent aux centres (hôpitaux / centres de recherche) d'où proviennent les données.
1B : Violin Plot correspondant aux répartitions des âges suivant les centres. En orange ceux issus du dataset 1, en bleu ceux du dataset 2.
1C : méthode de clustering (t-SNE embedding).

Prétraitement des données :

Un premier point clé est qu'avant d'entraîner le modèle, une étape de prétraitement importante est nécessaire. Cela a pris pour nous la majorité du temps étudié pour ce projet (environ 3/4 du temps sur un projet de 6 mois).
Nous avons proposé ici un pipeline de prétraitement minimal, afin de contrôler les biais communs observés dans les données brutes. Ces outils sont open-source et assez faciles à mettre en œuvre.
Un résumé schématique présente ce processus dans la figure 2.

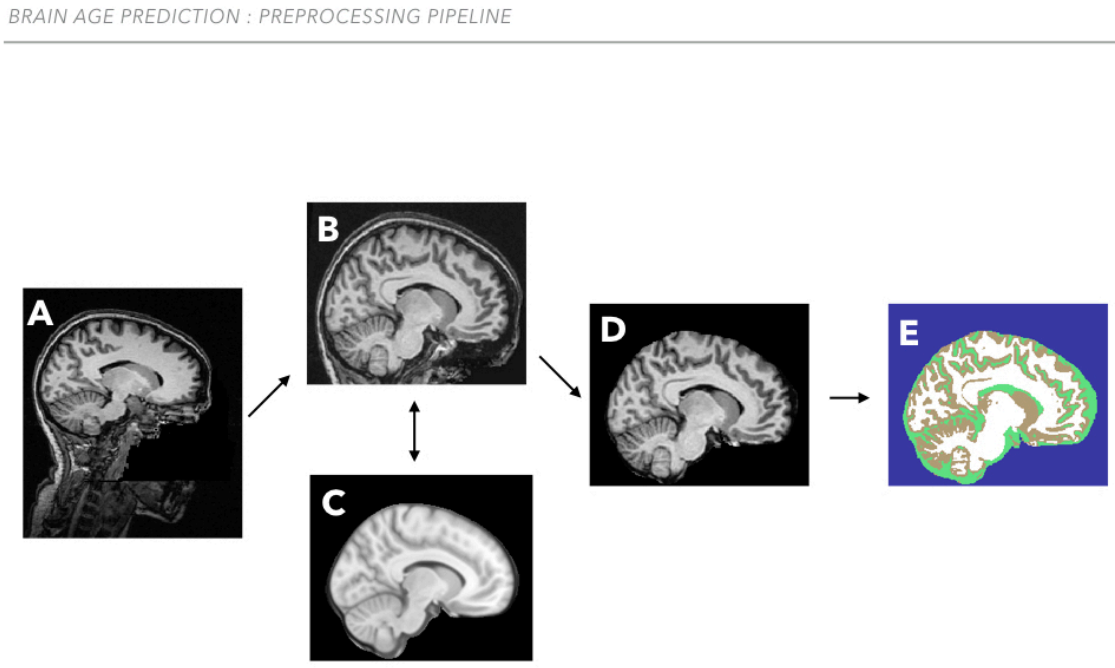

Figure 2: vue schématique du pipeline de prétraitement. A. Données brutes. B. Cropping. C. Coregistration et transformations affines avec l'atlas MNI 152. D. Skull stripping (qui consiste à enlever de l'image le crâne (peau, os, muscles principalement). E. Segmentation de la matière grise / matière blanche et Liquide céphalorachidien.

Recadrage:

Figure 2.A, on peut observer données brutes. Dans le plan sagittal, nous pouvons voir qu'un méthode d'anonymisation a été effectuée (lien pour la méthode employée ici), en enlevant les voxels relatifs aux visages (méthode de « defacing »). Le cou est visible sur cette image, mais n'était pas toujours visible, en fonction du champ de vue utilisé. Un recadrage standardisé centré autour du cerveau était nécessaire.

Coregistration:

Dans la Figure 2. B, afin d'automatiser ce recadrage, nous avons utilisé des outils de coregistration. Nous avons utilisé le modèle MNI152 pour cela (le modèle est présenté à la figure 2.C). Le modèle MNI (Institut national de Montréal) est un atlas cérébral qui représente la moyenne de 152 IRM de sujets sains, utilisant une transformation affine à 9 paramètres ((44) Evans et al. 1993).
Les transformations spatiales 3D sont de deux types : linéaire (transformation rigide et affine) et non linéaire (le plus utilisé: transformation difféomorphique). La registration dans la matrice du MNI permet de normaliser la taille relative de chaque cerveau sans perdre d'informations sur la proportion relative d'atrophie. C'était aussi un moyen de régulariser les singularités de genre, les cerveaux des femmes ayant tendance à être plus petits que ceux des hommes ((45)Ruigrok et al. 2014).

Correction du Biais N4 :

Ce biais inhérent à toute séquence d'IRM consiste à corriger les intensités de voxels dues à la présence d'une non-uniformité d'intensité aux basse fréquence (pour plus de détail, voir ((46)Tustison et al. 2010) (cf. figure 3). Cette étape a été réalisée à l'aide du logiciel Ants (toolbox open-source de normalisation d'image IRM)

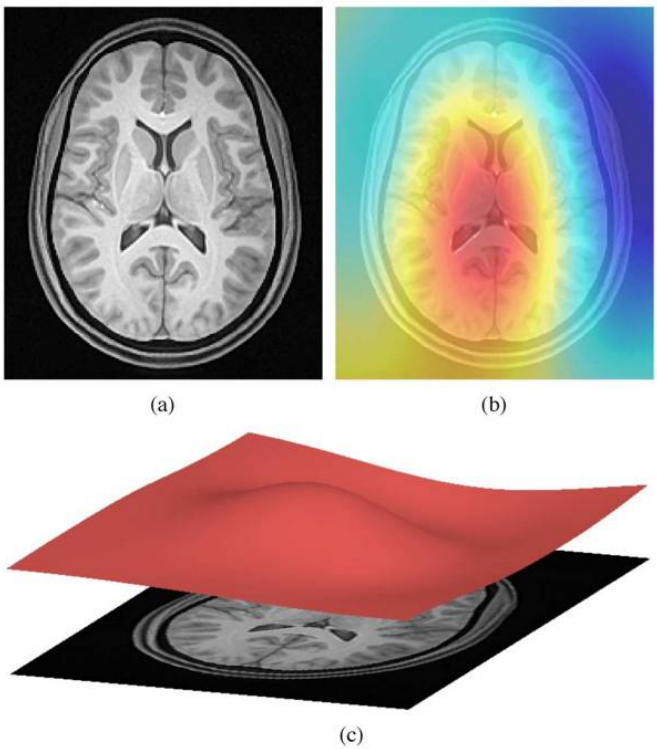

Figure 3: Carte de chaleur correspondant aux inhomogénéités des intensités de pixels avant correction du biais N4. Image issue de Tustison et al. 2010 (46)

Skull stripping

Le "décapage du crâne" a consisté à retirer le tissu non cérébral de l'image (cf. Figure 2.D). Cette étape a été réalisée afin d'évaluer l'effet de la présence de tissu non cérébral (muscle, peau, yeux, os) sur la prédiction de l'âge. Afin de réaliser le skull stripping, nous avons effectué une transformation affine donnant à chaque cerveau des dimensions identiques calquées sur le MNI. Le skull stripping pouvait alors être fait, en utilisant comme masque de soustraction le masque MNI, permettant alors d'enlever toutes les structures en dehors du masque. L'information liée à l'hétérogénéité de l'atrophie a pu ainsi être conservée.

Normalisation d'intensité

En IRM, il n'y a pas de valeur absolue pour les intensités de pixels, contrairement au scanner (où les intensités de pixels sont corrélées à la densité du tissu, en unités Hounsfield). Le contraste relatif est évalué entre les tissus : dans le cerveau, sur les séquences pondérées en T1, où la matière grise est plus sombre que la substance blanche (cf. figure 4), autrement dit, la valeur du voxel de la substance grise est inférieure à celle de la substance blanche. La conséquence en est une énorme variabilité entre les valeurs en fonction de l'hôpital où les images ont été acquises. Un moyen simple de contrer cet effet consiste à normaliser les intensités centrées sur un maximum local, correspondant à un type de signal dans le cerveau.

Premièrement, nous avons centré autour de 1 la valeur des pixels sur un maximum local correspondant à la substance blanche, hyperintense dans une image pondérée en T1. Pour cela, nous avons détecté le pic de substance blanche dans l'image avec un algorithme de détection de maximum) et divisé la valeur des intensités de pixels par la valeur de ce pic (cf figure 5)
Dans un deuxième temps, nous avons effectué une amélioration de la normalisation de l'intensité, ce qui a permis d'avoir un point constant pour le maximum local correspondant à la matière grise (cf. figure 6). Pour cela nous avons détecté le pic de matière grise et blanche à partir des masques de segmentation. Nous avons normalisé le pic de matière blanche (même méthode qu'auparavant) et interpolé en second lieu la valeur du pixel entre 0 et 1 afin d'obtenir le pic de gris à la valeur de 0,75. Nous avons ensuite évalué cette normalisation plus avancée sur les performances.

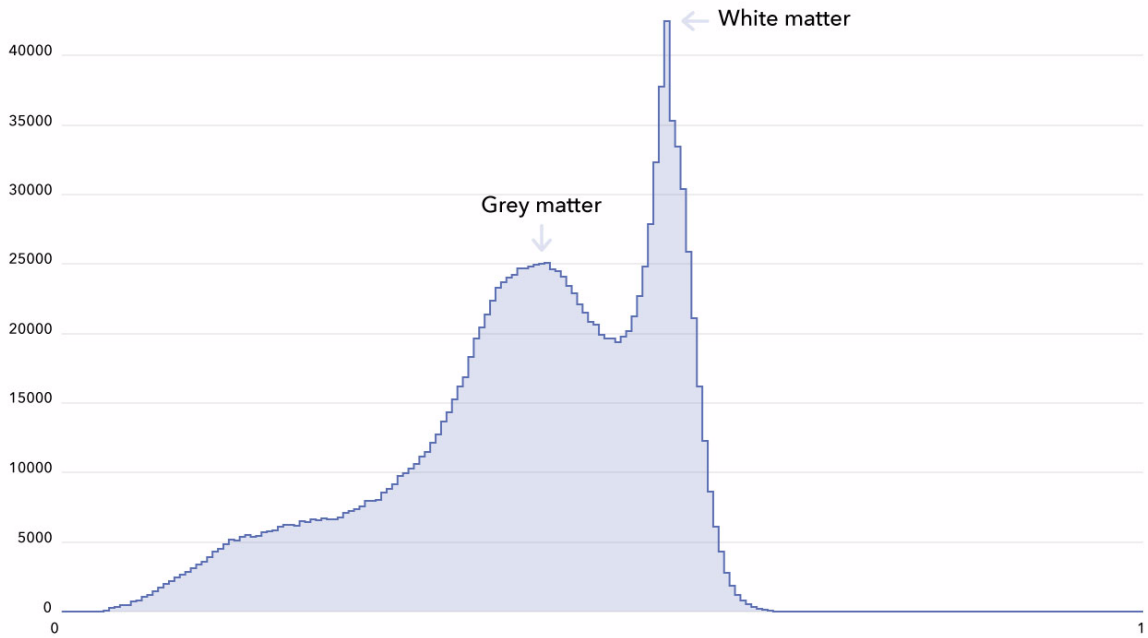

Figure 4. Histogramme des intensités dans les données de l'ensemble de l'IRM: La normalisation d'intensité a été effectuée autour du pic d'intensité de la matière blanche, visible à droite de l'image.

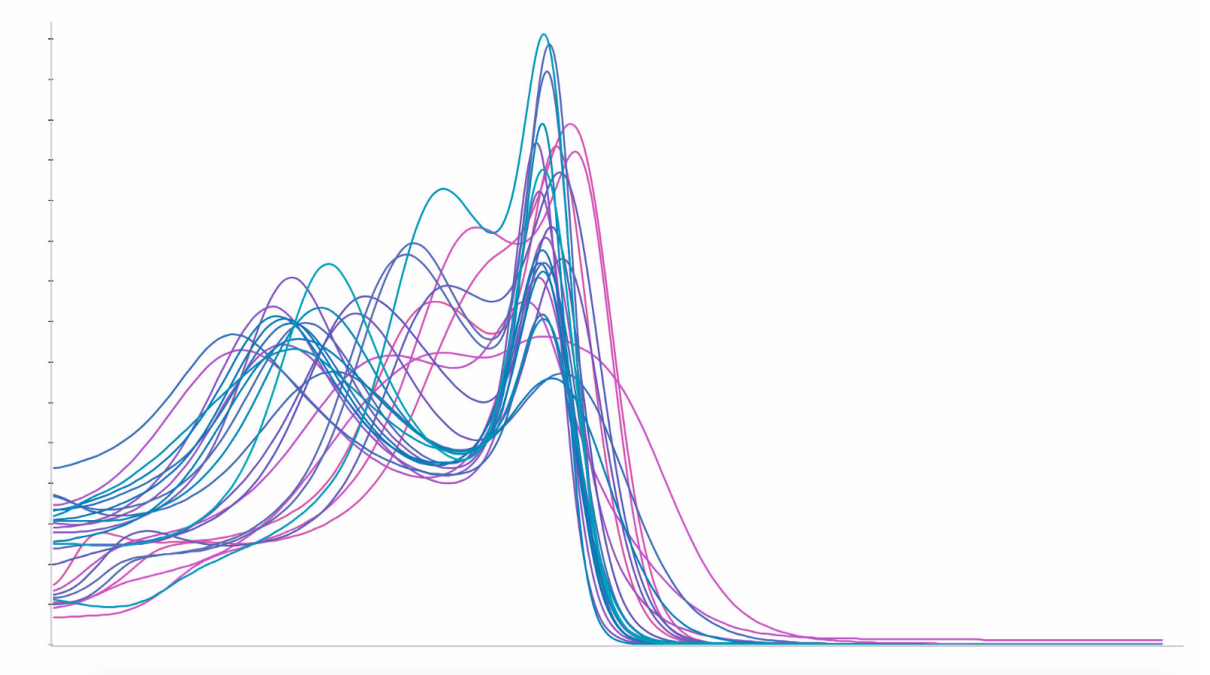

Figure 5. Figure représentant les courbes issues des histogrammes moyens issus de chaque centre (v1) après normalisation de l'intensité sur le maximum local correspondant à la substance blanche.

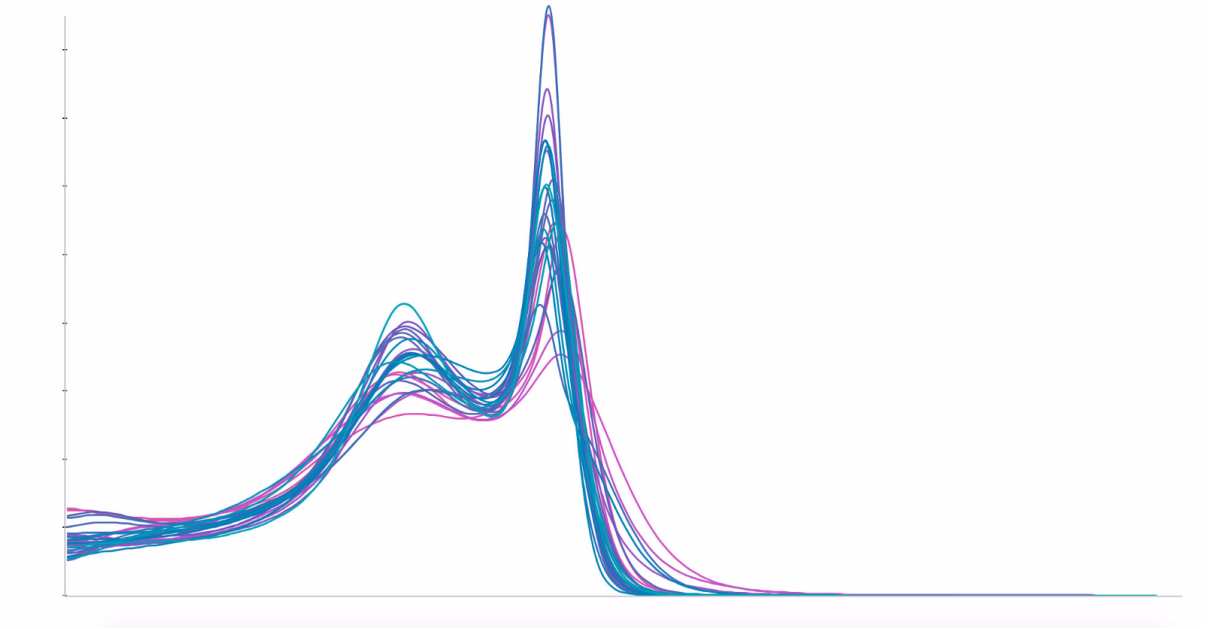

Figure 6: Histogramme (v2) des intensités après normalisation des intensités maximales de la substance grise et blanche.

Segmentation du tissu cérébral:

Comme vous pouvez le voir dans Figure 2.E, cette dernière étape consistait à séparer la matière grise en substance blanche et le LCR.

Comme pour les histogrammes, la segmentation de différents tissus est un moyen de réduire la dimensionnalité des données, c'est-à-dire, la complexité. A partir de valeurs de signal continues entre la substance grise, la substance blanche et le liquide céphalorachidien (LCR), nous réduisons les intensités de pixels en valeurs catégorielles : 0 pour le fond, 1 pour le LCR, 2 pour la substance grise, 3 pour la substance blanche.
Avec cette méthode, nous avons conservé les corrélations spatiales entre les différents éléments du tissu cérébral, mais nous avons perdu les variations subtiles du signal dans chaque type de tissu. Nous avons d'abord utilisé la bibliothèque Ants mais nous n'étions pas satisfaits du résultat. Nous avons finalement utilisé la boîte à outils FSL pour cette étape (47).
On sait que la quantité de matière grise décroit avec l'âge (28). Nous avons calculé les corrélations de Pearson entre l'âge et le volume de la matière grise à partir de la segmentation, c'est-à-dire la somme des voxels correspondant à la matière grise (cf figure 7).
Comme prévu, une forte corrélation négative a été observée entre l'âge (axe des Y) et le volume de matière grise (axe des X) ($r = -0,74$, $p < 0,05$). Les corrélations étaient plus faibles entre l'âge et le volume de LCR ($r = 0,40$) et les masques de segmentation de la substance blanche ($r = 0,13$).

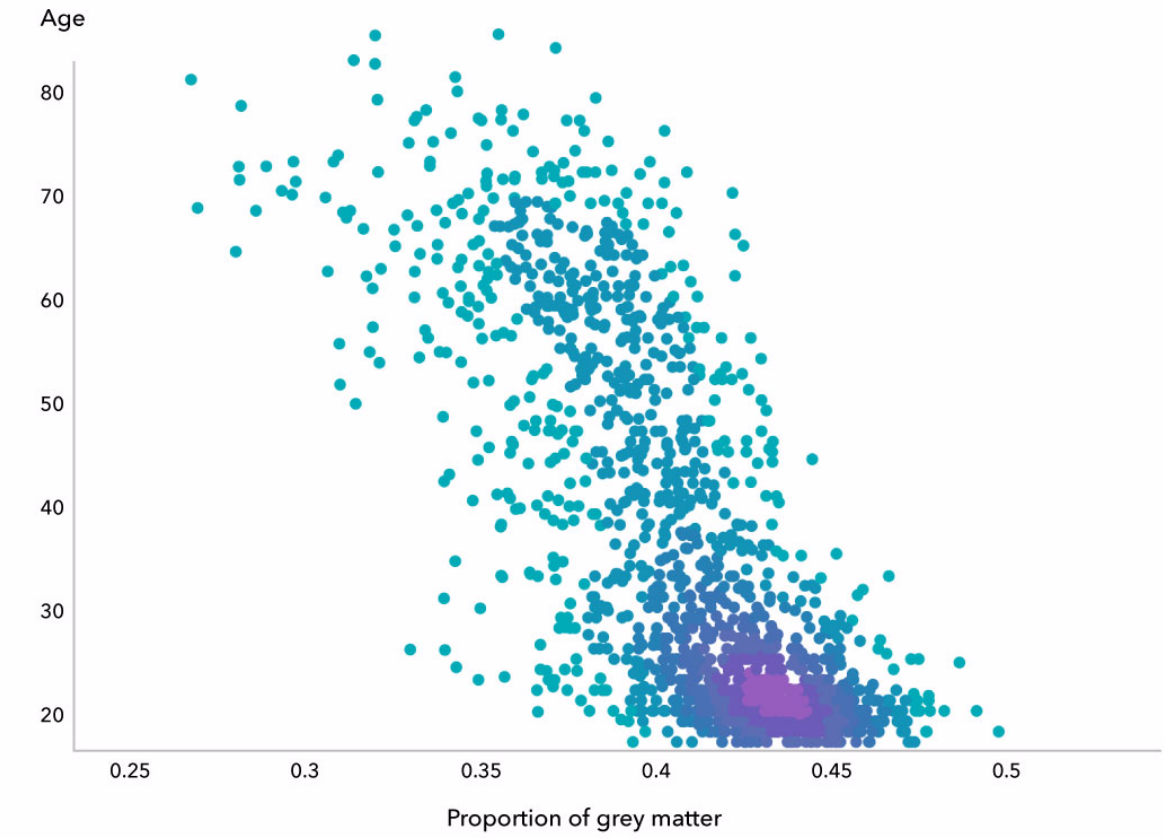

Figure 7: corrélation entre l'âge et le volume de matière grise à partir de masques de segmentation. Les couleurs représentent la densité des points.

Modèles d'apprentissage automatique : Prédire Y à partir de X, avec différentes architectures, et l'intérêt de la validation croisée.

À cette étape, nous avons un jeu de données vierge et des données prétraitées de différentes manières.
Nous avons réduit la dimensionnalité des données et extrait les entités de différentes manières afin d'avoir différentes données en entrée : histogrammes d'intensités des voxels, masques segmentés puis données brutes après « skull stripping ». Afin d'évaluer l'effet de l'extraction du crâne sur les prévisions, nous avons également entraîné des modèles avec des données brutes.

A propos des métriques employées, et comment comparer les résultats.

L'âge médian (réel) était de 36 ans. Nous avons calculé l'erreur absolue moyenne (MAE) à partir de l'âge médian des sujets (MAE à 13,73 années avec un random split) afin de comparer les prédictions.
Ainsi, si un algorithme calcule une erreur de prédiction supérieure à l'erreur absolue moyenne calculée à partir de l'âge médian (autrement, dit s'il n'est pas capable de prédire mieux qu'une valeur constante), cela signifie que sa prédiction ne fait pas mieux qu'une prédiction faite au hasard.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

Le calcul de la MAE suit la formule suivante :
Ou n correspond à la taille de l'échantillon, y(j) correspond à l'âge du sujet j et y^(j) correspond à l'âge prédit du sujet j.
L'âge prédit est l'âge estimé par notre algorithme, issue de la phase d'apprentissage.
La somme des différences entre âge réel et âge prédit divisé par la taille de l'échantillon définit donc la MAE dans notre étude.

Nous avons employé l'erreur absolue moyenne (MAE) comme métrique d'évaluation d'erreur pour nos algorithmes. Il est possible de choisir de calculer la MAE ou la RMSE, deux métriques assez proches pour évaluer la précision des prédictions de variables continues. Nous n'en avons choisi qu'une en raison du temps d'entraînement des modèles les plus complexes.
Plus l'erreur de prédiction est grande, moins la performance de l'algorithme est bonne.

L'intérêt des méthodes "baselines"

Nous avons formé des modèles linéaires (régression linéaire, régression Ridge) avec différentes entrées de données : histogrammes des intensités de l'IRM entière, masque de segmentation de matière grise et données brutes. Nous avons également entraîné un modèle non linéaire moins complexe que les CNN, avec une méthode appelée Gradient boosting : CatBoost (48).
Le boosting est un méta-algorithme (c'est-à-dire un algorithme qui s'applique à un autre algorithme) basé sur l'idée d'agréger progressivement de nombreux algorithmes simples, appelés apprenants faibles (weak learners), pour obtenir un apprenant final fort (strong learner) (49). Plus spécifiquement, chaque apprenant faible est optimisé pour minimiser l'erreur sur les données de formation en utilisant la somme des apprenants faibles précédents comme entrée supplémentaire.
Alors que les méthodes de deep learning se sont révélées extrêmement efficaces pour les données «structurées» (images, sons, textes, séries chronologiques, etc.), les méthodes de Boosting constituent l'algorithme non linéaire "par défaut" dans la communauté du machine learning pour tout autre type de données, et sont donc classiquement utilisés pour avoir une idée des performances, et ce d'autant qu'ils sont assez rapide à entraîner comparé aux CNN.

Modèle non linéaire complexe: réseaux de neurones convolutifs (CNN)

Ce qu'on appelle Deep Learning est une famille de modèles non linéaires appelés réseaux de neurones, qui partagent une architecture entièrement différenciable et structurée en couches. Les méthodes d'apprentissage profond ont récemment été mises à l'honneur pour leurs performances impressionnantes en vision par ordinateur (50), en reconnaissance vocale (51), en traitement du langage naturel, et l'apprentissage par renforcement (53), et ces succès ont été attribués à trois facteurs principaux: la disponibilité de jeux de données plus volumineux, d'une plus grande puissance de calcul et des algorithmes plus sophistiqués.

D'un point de vue mathématique, l'un des ingrédients clés qui rend les réseaux de neurones aussi efficaces est la capacité à intégrer le calcul différentiel, bien adapté à la structure des données. En vision par ordinateur en particulier, l'utilisation de réseaux de neurones convolutif (CNN) constitue un outil prometteur. Dans ce travail, nous avons utilisé certains des méthodologies avancées de deep learning, tels que les residual connections qui permet de garder un apprentissage efficace tout en tirant parti de la profondeur du réseau (plus de profondeur équivaut grossièrement à capturer plus de complexité) (54) et la *batch norm*, méthode qui permet notamment un entraînement plus rapide, et de limiter le manque de généralisation par surinterprétation (overfitting), (55). Tous les modèles ont été formés sur un seul GPU (NVidia Titan Xp ou Geforce GTX 1080) et ont été écrits en python avec la librairire Keras.

CNN 2D

Nous avons utilisé des modèles de CNN 2D avec deux approches. La première consistait à entraîner des modèles from scratch, c'est à dire sans pré-entraînement préalable, la seconde à utiliser l'apprentissage par transfert (transfer learning) avec un réseau pré-entraîné (Resnet50), puis après une data augmentation.
L'architecture du CNN 2D entraîné from scratch contenait 4 blocs répétés de convolutions (3x3), une couche batch norm 2D, une rectified linear unit (Relu), une couche d'average pooling 2D. Le nombre de feature maps a été fixé à 16 dans le premier bloc et a été doublé après chaque pooling layer pour en déduire une représentation suffisamment riche du cerveau.
La prédiction finale de l'âge a été obtenue en utilisant une dernière couche appellée fully connected layer, qui permettait de sortir une valeur du réseau correspondant à l'âge prédit.
Dans chaque application, les poids de réseau ont été formés en minimisant l'erreur moyenne absolue (MAE) à l'aide d'une RMS propagation et d'un Nesterov momentum optimizer (Nadam).
Le principe de l'apprentissage par transfert (utilisant un réseau de neurones pré-entraîné tel que Resnet 50 (54) est décrit comme suit : il est possible de transférer vers une nouvelle tâche la manière dont un algorithme extrait des entités de millions d'images vers un nouvel ensemble de données (dans le cas présent : Imagenet , un ensemble de données entraîné sur 1,4 million d'images). Dans ce travail, nous avons retiré les deux dernières couches d'Imagenet et ajouté deux couches de convolution, afin de les affiner avec la spécificité de notre jeu de données.

La data augmentation a consisté à appliquer des opérations sur les histogrammes (netteté, contraste et luminosité), une rotation (+/- 5 °), un décalage en largeur et en hauteur de +/- 5 pixels), un zoom (facteur de 0,01). L'intérêt de la data augmentation est de montrer artificiellement plus de données à l'algorithme. Plus vous disposez de données, plus vous obtiendrez théoriquement les meilleures performances (c'est-à-

dire une meilleure généralisation). Nous nous sommes notamment concentrés sur l'augmentation de la variété des valeurs d'histogrammes afin de limiter les biais dus à cette hétérogénéité provenant principalement des acquisitions d'images.

CNN 3D

Pour l'utilisation des réseaux CNN 3D, nous avons fait en sorte que les images aient toutes les même dimensions (182x218x182 voxels). Nous avons construit une architecture identique à celle du papier de référence (36) afin de reproduire les résultats (pour une représentation schématique, voir la figure 8). Celui-ci contenait 5 blocs de couche de convolution (3x3x3), une couche ReLU, une couche de convolution (3x3x3), une couche de batch norm 3D, une couche ReLu et enfin une couche max pooling (2x2x2). Le nombre de feature maps a été défini à huit dans le premier bloc et doublé après chaque couche de pooling. La prédiction finale de l'âge a été obtenue à l'aide d'une couche fully connected layer, qui mappait la sortie du dernier bloc sur une valeur de sortie unique, de façon similaire à celle employée dans les réseaux 2D.
Dans chaque application, les poids de réseau ont été formés en minimisant la MAE à l'aide de différents optimiseurs: descente de gradient stochastique (SGD); Root Mean Square propagation and gradient momentum (ADAM) et une variante de ADAM (ADAMAX). Nous avons conservé les paramètres de réglages par défaut proposés par Keras pour chaque expérience.

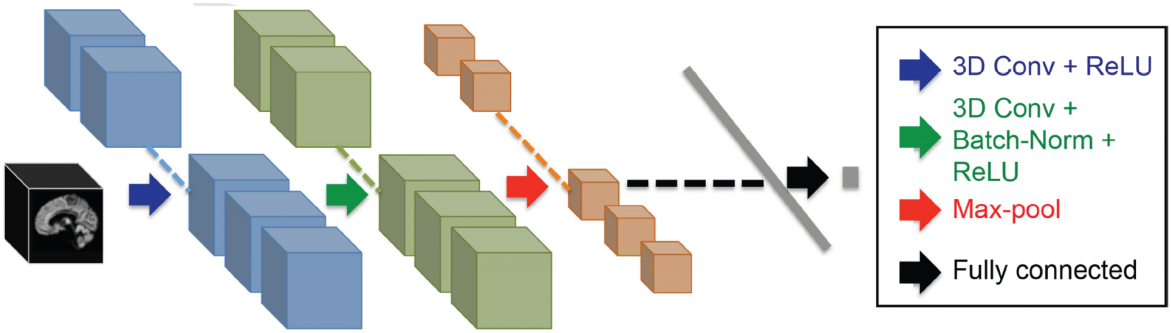

Figure 8: Représentation schématique de du réseaux CNN 3D issu du papier de Cole et al. 2017 (36).

Intérêt de l'utilisation de différentes méthodes de validation croisée:

La cross validation consiste à séparer le dataset en deux parties (spliting), une partie pour la phase d'entraînement, et une partie pour la phase de test. L'algorithme est ainsi testé sur des données qu'il n'a pas "vu" lors de la phase d'entraînement.

Le random split consiste en une validation croisée K-fold avec (k = 5). 4 folds sont utilisés pour former le modèle (80% de l'ensemble de données), 1 fold étant dédié à la phase de test.
L'inconvénient du *random split* est que l'algorithme est entraîné dans la phase d'entraînement à des données issues d'un centre qui sera également présent dans le data set de test. L'algorithme peut alors apprendre à reconnaître le centre plus que l'âge du patient. Ce qui constitue un biais, bien connu dans les études multicentriques : l'effet centre. Une façon d'estimer la taille de l'effet centre est de faire une cross validation non plus aléatoire, mais par centre. Pour cela, nous avons divisé le jeu de données en 5 fold, défini par leur centre (c.-à-d. que les données d'un hôpital donné ne pouvaient pas être à la fois en set d'entraînement et de test).

Travail sur l'interprétabilité / explicabilité des modèles :

Dans ce travail, nous nous sommes concentrés sur l'interprétabilité de modèles simples. Nous avons d'abord étudié les voxels corrélés avec l'âge sur différents masques de segmentation. Deuxièmement, nous avons généré des heat maps afin de visualiser les régions d'intérêt pour les prévisions effectuées avec un modèle linéaire simple (Ridge regression). Enfin, notre travail sur l'interprétabilité de CNN a consisté à occlure une partie des images à l'algorithme et de mesurer de son impact sur les prédictions pour une région occluse donnée, sur la base de la méthode de (25).

RÉSULTATS

Prédictions des différents modèles

Prédictions basées sur les histogrammes des intensités

Algorithme	Méthode de validation croisée	
	Random split	Split par centre
Prédiction aléatoire à partir de l'âge médian	13,73 +/- 13,49	15,12 +/- 11,87
Prédiction aléatoire de l'âge médian par hôpital	8,66 +/- 9,66	X
Histogramme v1 - Regression linéaire sur l' Histogramme des intensités	9.08 +/- 11.19	14.22 +/- 9,65
Histogramme v1 - Gradient boosting	5,74 +/- 5,73	11,52 +/- 8,67
Histogramme v2 - Régression linéaire	7,35 +/- 6,41	9,51 +/- 7,35
Histogramme v2 - Gradient boosting	**5,86 +/- 5,94**	**8,67 +/- 7,39**

Tableau 1: prédictions de l'âge du cerveau effectuées sur des histogrammes des intensités issus des images. (Erreur absolue moyenne (Mean absolute error, MAE) +/- écart type).

Dans le tableau 1, nous pouvons observer que la régression linéaire est moins précise que la méthode non linéaire par gradient boosting. De plus, la méthode de validation croisée a une forte influence sur les résultats. Les performances sont meilleures en utilisant un random split, car cette méthode ne prend pas en compte l'effet centre.

Nous pouvons observer qu'une normalisation plus fine des histogrammes (donnant une valeur de pixel constante aux pics correspondant à matière grise et matière blanche) améliore les performances à la fois pour le random split que le split par centre. De plus, les différences relatives d'erreurs de prédiction entre ces deux méthodes de cross validations étaient moins importantes avec cette méthode. (pour les prédictions du gradient boosting par exemple : histogramme v1 : 11.52-5.74 = 5.78 ans et histogramme v2 : 8.67-5.86 = 2.81 ans).
Le principal résultat à retenir en utilisant les histogrammes des intensités en tant que données d'entrée est le suivant : la meilleure performance est obtenue par une méthode de Gradient boosting (MAE de 5.86 ans dans une cross validation aléatoire, avec une normalisation avancée des pics d'intensité). Cette performance est à nuancer par l'effet centre, facteur de confusion pour les prédictions. En contrôlant l'effet centre par une cross validation adaptée, la meilleure MAE est de 8.67 ans.

Prédictions basées sur les masques de segmentation de matière grise

Algorithme	Méthode de validation croisée	
	Random split	split par centre
Histogramme V2 - Gradient boosting	5,86 +/- 5,94	8,67 +/- 7,39
IRM segmentée - Régression linéaire	**4,70 +/- 3,84**	**6,17 +/- 4,75**
IRM segmentée - Gradient boosting	4.92 +/- 4,38	7,09 +/- 5,50

Tableau 2: Prédiction de l'âge du cerveau réalisée sur un masque segmenté de matière grise (erreur absolue moyenne +/- écart type).

Dans le tableau 2, nous montrons les prédictions effectuées avec des masques de segmentation de la matière grise en tant que données d'entrées (input) des modèles. Nous pouvons voir que les prédictions sont plus précises que celles faites avec des histogrammes (ie une erreur de prédiction plus faible). Les performances sont équivalentes au modèle simple, et aucun gain de performance n'a été observé avec une méthode non linéaire telle que le Gradient Boosting.

Avec la validation croisée effectuée par centre, les performances sont plus proches entre les deux méthodes de cross validation (par exemple: 4,70 ans contre 6,17 pour la régression linéaire), ce qui suggère un effet centre plus faible avec cette méthode. Nous pouvons en déduire que la binarisation des valeurs de pixels au sein du masque de segmentation limite l'effet des biais présents dans le signal (exemple: rapport signal sur bruit spécifique inhérent aux différentes IRM).

Prédictions sur les images prétraitées avec des réseaux de neurones convolutifs (CNN) 2D

Algorithme	Méthode de validation croisée	
	Random split	split par centre
Histogramme V2 - Gradient boosting	5,86 +/- 5,94	8,67 +/- 7,39
Masque de segmentation- Régression linéaire	4,7 +/- 3,84	6,17 +/- 4,75
2D CNN from scratch	4,57 +/- 4,72	6,94 +/- 5,76
2D CNN - data augmentation (DA)	4,29 +/- 4,47	6,14 +/- 5,47
CNN pré-entraîné avec DA	3,60 +/- 3,67	6,57 +/- 5,11
Moyenne des CNN pré-entraînés avec DA et régression linéaire	**3,92 +/- 3,55**	**5,50 +/- 4,30**

Tableau 3: prévision de l'âge cérébral des données issues du pipeline de prétraitement avec des CNN 2D (erreur absolue moyenne +/- écart type).

Dans le tableau 3, on peut voir les résultats obtenus avec des modèles entraînés sur images prétraitées (rappel : normalisation sur la substance grise et blanche (sans segmentation), la correction du biais N4, co-registration dans l'espace MNI et le skull stripping). Nous pouvons observer que les CNN 2D from scratch sont moins précis que ceux pré-entraînés (méthode de transfer learning). La data augmentation a

permis une amélioration des performances indépendamment de la méthode de validation croisée. Le meilleur résultat a été observé avec un CNN 2D pré-entraîné sur random split (MAE: 3,60 ans), dépassant les performances publiées (4.6 ans dans le papier de J. Cole (36)). Toutefois, ce gain de performance n'a pas été observé avec une cross validation par centre (MAE: 6,57 ans).
En faisant une moyenne de performances de différents modèles (pre-trained CNN avec data augmentation et régression linéaire) on obtient des résultats plus robustes vis a vis de l'effet centre (on passe de 6.57 ans d'erreur à 5.5 ans en split par centre), avec des performance similaires en random split (3.92 vs 3.60 ans)

Prédiction sur les images prétraitées avec les CNN 3D

Lab	**Ensemble de données (nombre de sujets)**	***skull strippin g***	**Data Augmenta tion**	***Optimizer***	**validation croisée**
					Random split
J Coles, Neuroi mage, 2017	2001	Non	Oui	SGD	4.65
Owkin Lab	1597	Oui	Non	SGD	7.10
		Oui	Non	ADAM	5.02
		Oui	Non	ADAMAX	4.78

Tableau 4: Prédiction de l'âge du cerveau effectuée sur des données brutes sans crâne avec les 3D CNN (MAE).

Avec les 3D CNN formés from scratch sur des images prétraitées, nous avons obtenu des résultats similaires aux performances publiées sur le sujet (4,78 ans vs 4,65 ans dans l'article de original (36), sans Data Augmentation, avec un jeu de données plus petit et en utilisant un optimizer différent (ADAMAX, version dérivée d'Adam, au lieu de la méthode Stochastic Gradient Descent (SGD).

Travaux sur l'interprétabilité des modèles

Cartes de corrélations sur masques de segmentation :

Dans la figure 9, nous avons étudié les corrélations locales entre les voxels du masque de segmentation et l'âge du cerveau. A gauche, seuls les voxels segmentant la matière grises sont représentés. Sur le côté droit, seul le LCR. Les voxels bleus représentent des corrélations négatives et les voxels rouges des corrélations positives. Comme prévu, les aires corticales et les noyaux gris centraux étaient négativement corrélées avec l'âge sur le masque de segmentation de la matière grise (flèche continue) et positivement avec l'espace sous-arachnoïdien et les ventricules dans le masque de segmentation du LCR (flèche avec les cercles en pointillé), compatibles avec les features connues du vieillissement (atrophie et dilatation ventriculaire secondaire). Cependant, des corrélations positives ont été trouvées sur le masque de segmentation de la matière grise (flèche en pointillé «carré»). Cela était dû à une "mauvaise" segmentation de la matière grise. La leucoaraïose est en hyposignal sur les séquences T1, comme la matière grise. L'algorithme de segmentation a interprété cette caractéristique comme faisant partie de la matière grise.

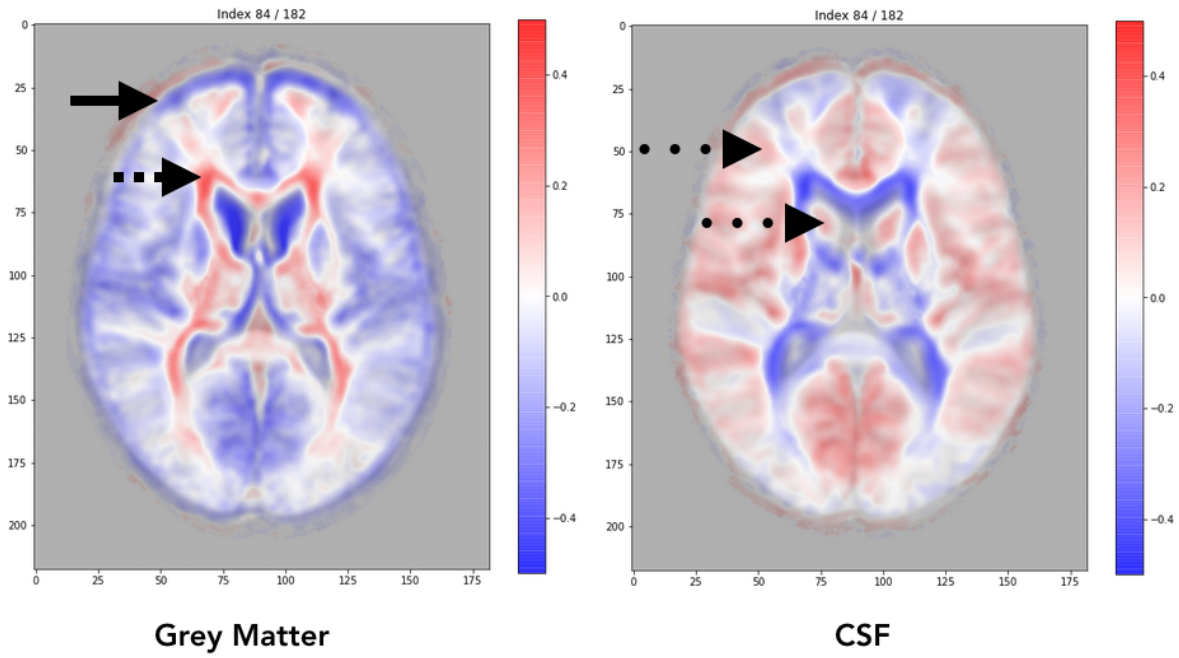

Figure 9: cartes de corrélations obtenues sur un masque de segmentation. En rouge les corrélations positives, en bleu les corrélations négatives. Plus les voxels sont bleus, moins ils sont présents lorsque l'âge augmente, et inversement pour les voxels rouges. A gauche, corrélations Age-Voxels issues des masques de segmentation des régions en hyposignal T1 (censé segmenter la matière grise). En réalité, l'algorithme

de segmentation a également segmenté les hyposignaux T1 en rapport avec la maladie des petits vaisseaux (leucoaraiose), corrélée positivement avec l'âge. A droite, carte de corrélations issue des segmentations du LCR (liquide céphalorachidien, ou cerebrospinal fluid, CSF).

Possibilité d'interprétation de modèles simples:

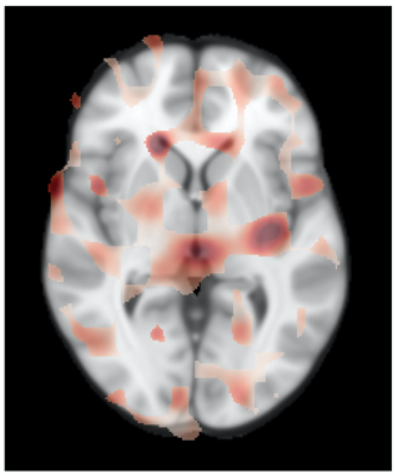

Figure 10. Cartes de pondération (weight map) à partir d'un modèle de régression de crête (Ridge regression) effectuée sur les masques de segmentation**.** Les zones rouges représentent les zones où les poids du modèle (ie les valeurs de paramètres du modèle, en valeur absolue) sont les plus corrélés à l'âge.

À la figure 10, Les zones en rouge représentent les zones où la matière grise était la plus corrélée avec l'âge selon une ridge regression. Les zones les plus corrélées semblaient être la matière grise et les ventricules (corne antérieure des ventricules latéraux et le troisième ventricule). Cette observation est compatible avec la physiologie du vieillissement, combinant une atrophie associée à une dilatation secondaire et progressive des ventricules.

Méthode d'interprétation des CNN:

Sur la figure 11, nous avons occlus avec de petites régions carrées d'images et observé leur incidence sur la prédiction sur CNN 2D pré-entraîné. Pour une case cachée donnée, une erreur de prédiction donnée a été calculée. Plus le carré est rose, plus l'erreur de prédiction est élevée. À gauche, nous avons appliqué cette technique aux sujets les plus jeunes de l'ensemble de données (âgés de moins de 30 ans). À droite, sur les sujets les plus âgés (plus de 60 ans). Nous pouvons observer, en plus des ventricules et des régions périventriculaires, l'importance de l'insula et des noyaux gris pour la prédiction de l'âge.

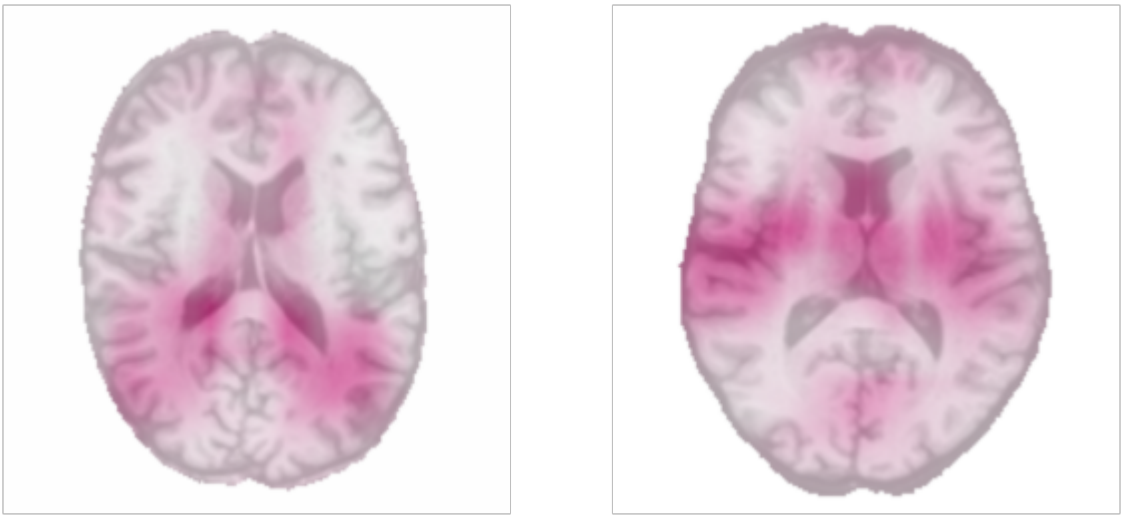

Figure 11: Interprétabilité des CNN 2D: «Cartes d'occlusion» avec la méthode de Zeiler et Fergus (25). À gauche, sujets les plus jeunes (moins de 30 ans); à droite, les plus âgés (plus de 60 ans).

Effet du skull stripping :

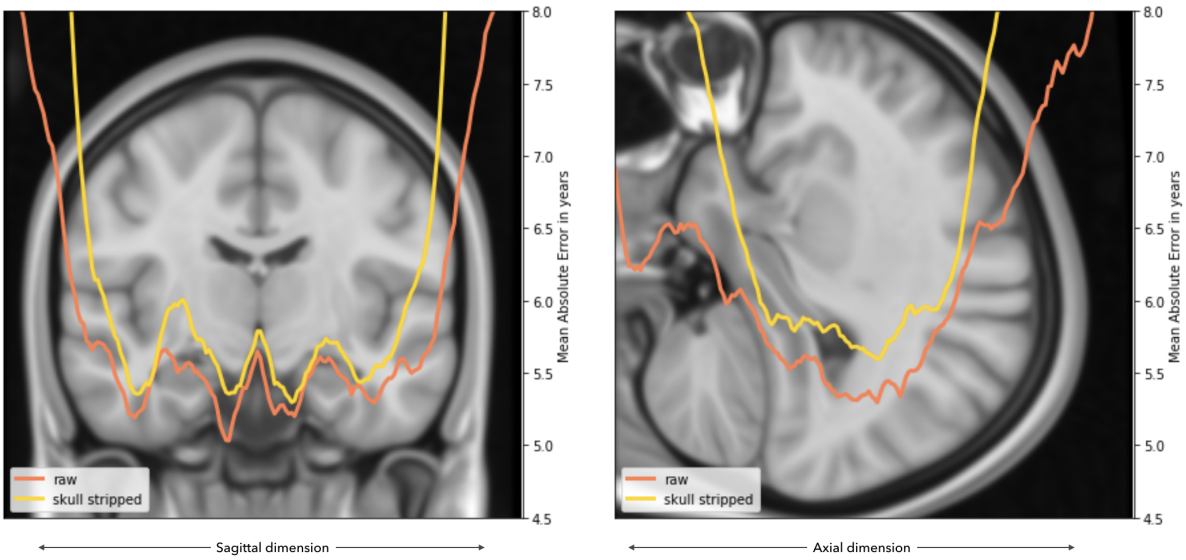

Figure 12: Effet du skull stripping sur les prédictions. Nous avons commencé par extraire des features avec Resnet sur des coupes 2D sur un plan sagittal ou axial, sur des données sans crâne et des données brutes. Nous avons utilisé ces fonctionnalités comme entrées dans un modèle de régression logistique pour prédire l'âge. Nous avons ensuite cartographié la valeur de MAE à partir des images avec skull stripping (courbe jaune) vs données brutes (courbes orange) sur la tranche correspondante dans un plan orthogonal.
Sur la figure 12, à gauche : nous voyons dans cette vue coronale la «projection» de l'erreur prédite sur chaque tranche sagittale (c'est-à-dire que chaque colonne de pixel de l'image est associée à sa prédiction moyenne issue de la coupe sagittale

correspondante). La prédiction était plus difficile sur les bords du cerveau où des données moins informatives étaient disponibles. Le même procédé a été appliqué sur l'image de droite (chaque colonne de pixel de l'image est associé à sa prédiction issue de la coupe coronale correspondante.).
Le principal message de ces figures est le suivant : l'algorithme exploite des informations qui ne sont pas issues du cerveau lors qu'il prédit l'âge sur l'irm sans skull stripping, et cela est mis en évidence (i) par une plus faible erreur de prédiction globale (la courbe orange est toujours plus basses que la jaune) et (ii) par des plus faibles erreurs de prédictions dans les régions excentrées (plus grand écart des valeurs entre la courbe orange et la courbe jaune dans les régions correspondant à la peau, les os, muscles, yeux…).

Application d'une potentielle application de ce biomarqueur du vieillissement : le diagnostic de la maladie d'Alzheimer

Jusqu'à présent, nous avons utilisé la prédiction de l'âge à partir d'analyses cérébrales comme prétexte pour vous présenter les bases de l'apprentissage automatique. Cependant, l'estimation de l'âge physiologique du cerveau pourrait être utile pour mieux comprendre les maladies neurodégénératives telles que la maladie d'Alzheimer. En guise d'expérience finale, nous avons appliqué nos modèles à 489 sujets de la Base de données ADNI.

Ces sujets sont répartis en deux catégories: sujets contrôle (269 sujets) et patients atteints de la maladie d'Alzheimer (220 sujets).
Nous avons téléchargé les données, les avons nettoyées, les avons passées dans notre pipeline de prétraitement et avons appliqué les modèles de régression linéaire et CNN entraînés. Les patients de cette base de données étaient beaucoup plus âgés (en moyenne de 75 ans) que dans l'ensemble de données que nous avons utilisé pour la formation (en moyenne de 35 ans) et, comme on pouvait s'y attendre, nos modèles ont systématiquement sous-estimé - l'âge des sujets en bonne santé. Cet échec met en évidence une limitation importante des modèles d'apprentissage automatique: s'ils ne sont pas formés sur un échantillon représentatif de la population, ils peuvent très mal fonctionner sur des sujets non vus. Ce manque de transférabilité d'un domaine à un autre est peut-être l'un des défauts majeurs de l'application du machine learning en santé.
Cependant, il est très intéressant de noter que lorsque nous avons tracé (Figure 13) la distribution des différences entre l'âge du sujet rapporté et l'âge du cerveau prédit avec l'algorithme final (régression linéaire + CNN), nous avons trouvé une différence moyenne de 6 ans entre les patients atteints de la maladie d'Alzheimer et les sujets en bonne santé. Bien que le modèle n'ait jamais été exposé à l'IRM d'un

sujet atteint de la maladie d'Alzheimer pendant son entraînement, le résultat du modèle peut distinguer les sujets atteints de la maladie d'Alzheimer avec un ROC-AUC score de 76%. Cela tend à confirmer que les cerveaux des sujets atteints de la maladie d'Alzheimer ont en quelque sorte des caractéristiques corrélées à un vieillissement cérébral accéléré (40), corroborant l'hypothèse selon laquelle l'âge prédit du cerveau pourrait être un nouveau biomarqueur des maladies neurodégénératives (38), (37).

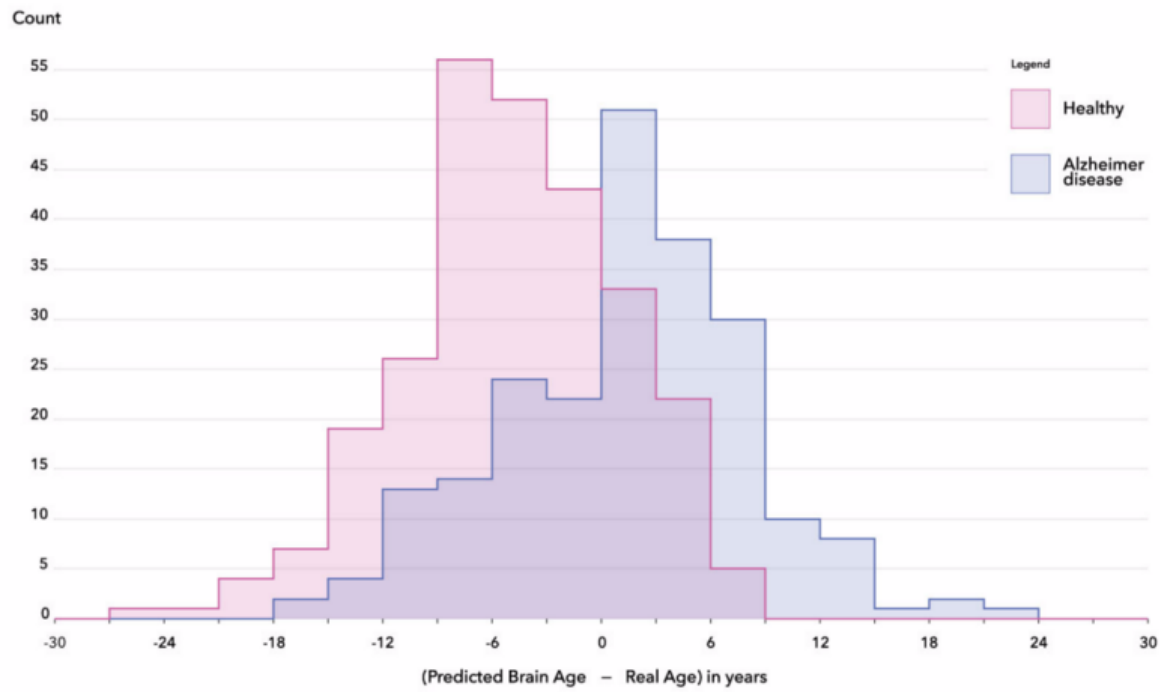

Figure 13 : Distribution de la différence entre l'âge prédit et celui rapporté chez les sujets Alzheimer (bleu) et sains (rose). On note une différence de 6 ans en moyenne entre les témoins normaux et les malades d'Alzheimer sur les 489 patients de la base de données ADNI que nous avons collectés.

DISCUSSION :

Notre pipeline de prétraitement des données est utilisable en routine radiologique

Les étapes consistant à définir comment nettoyer les données pour créer un pipeline de prétraitement efficace sont longues et ont pris la majeure partie du temps (3/4 du

temps sur un projet de 6 mois). Une fois celui-ci défini, son utilisation est rapide et est donc utilisable en routine, car il prenait environ 5 min par IRM, sous réserve d'accès à une puissance de calcul suffisante (nos ordinateurs avaient plus de 128 Go de Ram, plusieurs dizaines de CPU, et un GPU avec une dizaine de Giga de RAM).

L'utilisation de données d'entrées croissantes en complexité améliore les performances

À partir de ce pipeline de prétraitement, nous avons pu évaluer les prédictions sur différents niveaux de représentation du cerveau, par ordre croissant de complexité : des histogrammes aux images brutes en passant par les masques de segmentation de la matière grise. Même le mode de représentation le plus simple (histogramme des intensités) conservait des informations sur le vieillissement cérébral et permettait d'obtenir des résultats bien meilleurs que la prédiction aléatoire (meilleures performances avec régression linéaire): MAE de 5,86 ans contre 13,73 ans avec une prédiction aléatoire. Une normalisation plus précise des pics correspondants aux pics de matière grise et blanche était un moyen de limiter la perte de performance liée à l'effet centre.

Les meilleurs résultats du papier original (36) ont été obtenus en utilisant un masque de segmentation 3D de matière grise en tant que données d'entrée et un CNN 3D. La MAE était de 4,16 ans, contre 4,65 sur les données brutes (c'est à dire l'IRM, pas le masque de segmentation). L'auteur a affirmé que l'utilisation de données brutes pourrait permettre d'éviter de longues étapes de prétraitement et de rendre ce type d'algorithme utilisable dans un flux de travail radiologique sans prétraitement. En pratique, la génération d'un masque de segmentation peut être longue (pouvant se compter en heures de calculs), en fonction des outils utilisés et de la puissance de calcul disponible.
L'intérêt de pratiquer le skull stripping reste encore flou et doit être étudié dans les cas pathologiques. D'une part, nous pouvons considérer que la prédiction de l'âge à partir de données brutes (c'est-à-dire le cerveau et les tissus non cérébraux) constitue un biais : prédire l'âge de la tête au lieu de l'âge du cerveau. Cela pourrait ne pas être pertinent pour évaluer une maladie neurologique où une altération des tissus est observée uniquement sur le cerveau. Par contre, les différences de prédiction entre les données skull stripped vs images brutes sont très similaires (différence de MAE: moins de 0,5 an, cf. figure 12) et constituent probablement une différence acceptable en pratique clinique. Cependant, il n'est pas impossible que le tissu non cérébral soit informatif pour de meilleures prédictions. Il existe en effet des corrélations entre le volume de graisse cutanée et le volume de substance blanche : un volume de graisse sous cutané élevé est associé à une volume de substance blanche sous corticale moindre (56).

L'utilisation de modèles croissants en complexité améliore les performances, en 2D seulement.

Les modèles linéaires ont été dépassés par les modèles non linéaires dans la prédiction de l'âge du cerveau sur les histogrammes, mais cette tendance n'a pas été observée avec les masques de segmentation de la matière grise. Les modèles simples et non linéaires sont efficaces et donnent des résultats similaires à ceux d'architectures plus complexes telles que les CNN (Gradient boosting : MAE de 4,70 années sur les masque de segmentation de matière grise, contre 3,60 ans avec CNN 2D pré-formé sur des données brutes). Cependant les CNN ont dépassé les modèles non linéaires simples (méthode de gradient boosting).

Le CNN 2D pré-entraîné était le meilleur modèle que nous ayons entraîné (MAE: 3,60 ans), mais il était toujours sensible à l'effet centre (MAE: 6,57 ans avec split par centre). Cette perte de performance a été surprenante pour nous. En comparaison, le CNN 2D formé from scratch était plus «résilient» à l'effet centre (MAE: 4,29 ans avec random split, contre 6,14 ans avec split par centre). Nous avons observé lors de nos phases d'entraînement que la convergence de la Loss (fonction de coût calculée pour estimer la MAE) était obtenue plus rapidement avec un CNN pré-entraîné en 2D. Nous pouvons émettre l'hypothèse que cette convergence «trop rapide» a entraîné un overfitting et une mauvaise généralisation, entraînant une perte de performance sur le jeu de données de test.

Notre CNN 3D a reproduit des performances similaires à celles publiées (notre meilleur modèle: MAE de 4,78 ans vs 4,65 ans sur du papier d'origine (36) avec moins de données (1597 vs 2001 dans le papier original), mais était moins performant que le CNN 2D dans notre étude. Cela peut s'expliquer par le coût de calcul plus élevé lié à l'utilisation de ce modèle. Les informations pertinentes étaient probablement suffisantes sur les tranches 2D. L'intérêt d'utiliser un CNN 3D dépend probablement du problème à résoudre. Par exemple, une meilleure performance a été observée sur une tâche de détection de microbleeds avec des CNN 3D (Dou et al. 2016), en raison de la nécessité d'utiliser des informations 3D pour différencier les microbleeds des microbleeds mimics (par exemple: un vaisseau) sur des coupes 2D.

À propos de l'interprétabilité des modèles:

Nous avons souligné l'importance de l'interprétabilité des modèles, ce qui a permis de comparer nos résultats avec le profil biologique connu du vieillissement. Pour les modèles moins complexes, nous avons souligné l'importance de la dilatation ventriculaire et des régions où la matière grise est située dans le cerveau. De plus,

ces techniques permettent de montrer de manière plus compréhensible l'hétérogénéité du vieillissement cérébral.

La méthode d'interprétabilité employée sur les CNN a permis de faire émerger un motif plus fin que sur les autres méthodes. Les cartes d'occlusion selon (25) sur les modèles CNN ont révélé de façon "*data driven*" l'importance de l'insula et des noyaux gris centraux en tant que régions pertinentes pour les prédictions. Cela converge avec les résultats de (28), utilisant une autre méthode (morphométrie volumétrique par voxel (VBM, Voxel Based Morphometry)), où le vieillissement cérébral a été décrit comme accéléré dans l'insula et avec les travaux (57), montrant une accélération plus marquée de l'atrophie dans les noyaux gris centraux (thalami et capsule interne).
Cette méthode présente un avantage majeur : indépendamment de la façon dont est construite le modèle, il est possible de « voir ce que l'algorithme voit » en l'interrogeant de façon post-Hoc.
D'autres méthodes d'interprétabilité existent et mériteraient d'être étudiées dans l'interprétation des prédictions des modèles : certaines permettent de générer des cartes de chaleur sur les régions d'intérêt pour les prédictions, de façon similaire à celles employées dans cette étude (on citera les méthodes SHAP ou LIME), ou encore la méthode ATLAS Activation, permettant directement de visualiser des cartes d'activations à chaque couche des réseaux de neurones, publiée sur le blog de Google distill par Carter et al. (58) (Carter, et al., "Activation Atlas", Distill, 2019).

L'âge prédit du cerveau en tant que biomarqueur neuroradiologique?

Une étude utilisant des IRM de 36891 sujets (59) a montré l'intérêt de prédire l'âge dans les troubles cérébraux, en particulier dans les pathologies comme la démence, les troubles cognitifs légers (Mild cognitive impairment) ou la schizophrénie, où un «écart» a été observé entre les distributions de prédiction, par rapport aux sujets sains, de façon similaire à notre travail fait sur la base ADNI, ce qui suggère une reproductibilité et une fiabilité de ce biomarqueur.

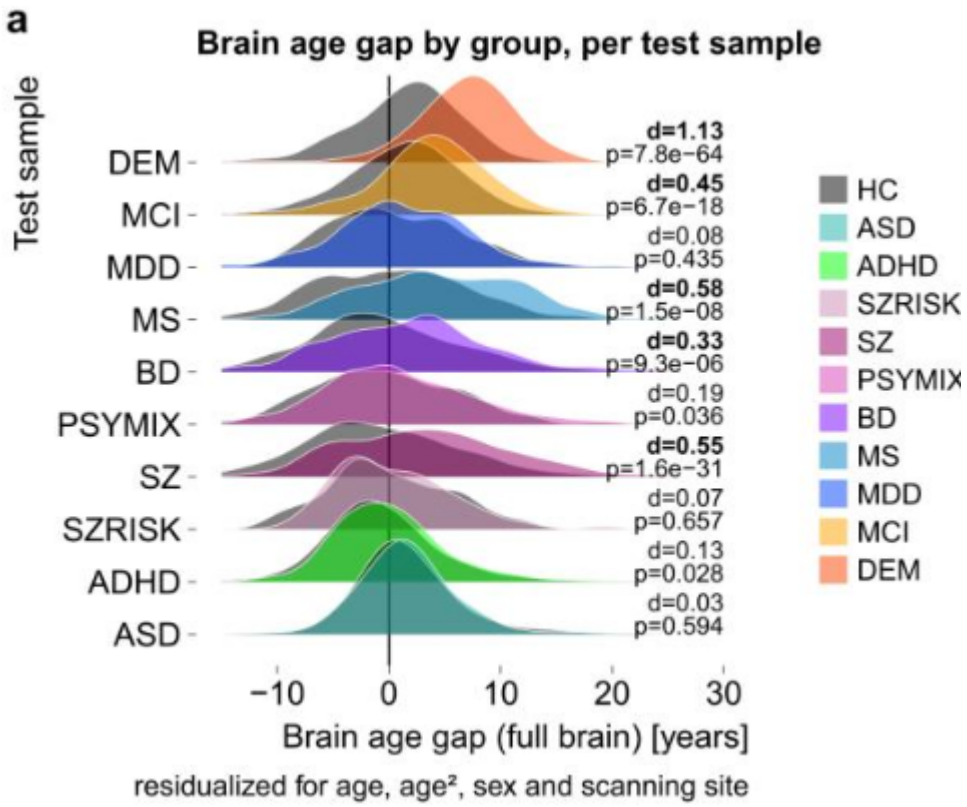

Figure 14 : issue de l'article de (59), employant une méthode similaire à la nôtre (ici la différence entre âge prédit et âge réel) sur des patients déments (DEM) mais également d'autres pathologies comme les Mild cognitive impairment, la sclérose en plaque (MS), les troubles bipolaires (BD), les troubles psychotiques (PSYMIX), la Schizophrénie (SZ), patients avec prodromes des schizophrénie (SZRISK), les troubles de l'attention et hyperactivité (ADHD), troubles du spectre autistique (ASD). le "d" correspond au Cohen's d, métrique qui permet de mesurer la séparabilité entre deux distributions.

Dans un autre article (60), les auteurs ont proposé une métrique dérivée de la MAE, qui vise à quantifier l'incertitude de l'erreur de prédiction (en calculant une covariance). Cette métrique permettait une meilleure précision de la classification que le MAE pour une tâche de classification telle que les mild cognitive impairment (AUC de 0,81 avec covariance, 0,68 avec MAE), la maladie d'Alzheimer (0,77 avec covariance, 0,68 avec MAE), l'autisme (AUC de 0,60 avec covariance, 0,53 avec MAE). Cependant, ces résultats n'ont pas été suffisamment performants pour constituer un outil de diagnostic pertinent sur le plan clinique et est probablement dû à l'insuffisance d'information présente dans l'unimodalité de l'IRM dans ces travaux.

Certains auteurs ont proposé l'IRM multimodale pour ce type de tâche de classification. Par exemple, les auteurs de l'étude (61) ont utilisé une IRM structurelle et fonctionnelle (pondérée en T1, tenseur de diffusion et IRMf resting state) et ont montré des performances bien supérieures (meilleure AUC de 0,922) pour classifier

démence frontotemporale vs sujet sain. Cette approche multimodale n'était cependant pas concluante pour la classification Alzheimer vs sujet sain.
Le modèle utilisé utilisait un modèle linéaire (elastic net regularized regression). Cependant, cet article a été utilisé pour un petit ensemble de données (37 patients avec une maladie d'Alzheimer probable et 28 patients avec une démence frontotemporale, par rapport à 35 sujets témoins). D'autres études seront nécessaires pour mieux appréhender l'apport diagnostique de la multimodalité de l'IRM dans les modèles prédictifs.

Limites et perspectives

À propos de la séquence IRM utilisée dans cette étude:

Nous pouvons révéler certaines limites de la prédiction de l'âge cérébral dans notre travail :

- (i) l'utilisation de l' IRM unimodale (seule l'IRM pondérée en T1 a été étudiée ici, alors qu'en routine radiologique, un protocole classique d'IRM cérébrale est multimodal (plusieurs séquences : au moins T1, FLAIR, DWI, T2 * et TOF) ;
- (ii) Aucune évaluation spécifique de la leucoaraïose n'a été réalisée (toutefois, ces données ont été capturées par segmentation de la substance grise, cf. figure 9)
- (iii) l'hétérogénéité des paramètres de séquences IRM et les différences d'intensité de champ magnétique (1,5 et 3 Tesla principalement), responsables d'un effet centre, prix à payer pour constituer un *dataset* de taille suffisante.

À propos des limites de CNN:

Une des limites de CNN réside dans son manque de «sens commun», c'est-à-dire que cet algorithme peut faire une classification erronée qu'un humain ne ferait jamais. Par exemple, les prévisions de grande confiance ont été attribuées à des images non reconnaissables (62). Cette idée a été montrée d'une manière différente: l'ajout d'un petit «adversarial patch» (un petit cercle d'images non reconnaissables) (63) a modifié les prévisions. Des résultats similaires ont été obtenus en ajoutant des variations non perceptibles dans l'image (64).

Dans nos résultats, nous pouvons relier ce concept au fait que de meilleurs résultats ont été observés avec une répartition aléatoire. Les modèles ont capturé les informations relatives aux hôpitaux et les ont utilisées pour prédire l'âge.

Architectures avancées de CNN : perspectives

L'auto Machine learning (Auto ML) est une évolution prometteuse de l'apprentissage automatique. À l'heure actuelle, les data scientists consacrent beaucoup de temps au tuning (réglage) d'hyperparamètres des modèles, ou à la recherche d'une architecture numérique adaptée à un problème donné à résoudre. Une automatisation de ces aspects pourrait accélérer le processus de ML. Par exemple, un réseau de neurones pourrait être utilisé ... pour découvrir des architectures de réseau de neurones (pour une revue récente, voir (65).

Les Capsule networks (66) ont montré une performance *state-of-the-art* sur une tâche "*benchmark*" de classification par vision par ordinateur (classification de nombre écrits à la main MNIST). Contrairement aux CNN, son architecture prend en compte la corrélation spatiale dans l'image (par exemple, dans le cadre d'une tâche de détection de visage, le nez est classé dans la catégorie nez du fait de sa proximité avec les yeux, etc.). Un des avantages de cette méthode serait de pouvoir avoir des performances avec des data sets moins conséquents en taille que nécessitent encore l'utilisation des CNN classiques, ce qui aurait un intérêt en médecine, où les data sets sont souvent relativement petits.

Une autre méthode, le deep reinforcement learning inspirée du système de récompense dopaminergique du cerveau, a récemment eu un fort écho médiatique du fait de la défaite du champion du monde d'alphago face à un algorithme entraîné par l'équipe d'ingénieur de Deepmind (53). Cela pourrait être une technique intéressante, mais la principale limite réside (encore) dans le nombre élevé d'exemples requis pour un entraînement efficace du modèle. Nous pouvons citer un article axé sur la détection de nodules pulmonaires (67), comme preuve de concept, mais sans résultats exploitables pour le moment (Accuracy sur l'ensemble de test de 64,4% (sensibilité de 58,9%, spécificité de 55,3%, VPP de 54,2%, et VPN 60,0%)).

CONCLUSION :

Pour conclure, les CNN sont une puissante innovation en vision par ordinateur et ses applications semblent prometteuses en imagerie médicale.
Nous avons pu par ce travail identifié les forces de cette méthodologie, en dépassant les performances de l'état de l'art sur la tâche de prédiction de l'âge cérébral, et en allant plus loin dans la compréhension de cette performance : (i) le gain réel des CNN par rapport aux algorithmes moins complexes de machine learning, (ii) l'importance d'identifier des biais inhérents à une étude multicentrique, et le moyen d'y pallier par une validation croisée adaptée, (iii) l'utilisation de méthodes d'interprétabilité pour extraire les régions d'importances de l'image étudiée dans les prédictions et ainsi faire le lien avec la biologie (en l'occurrence ici l'importance de la substance grise, des régions périventriculaires et des ventricules pour prédire l'âge du cerveau) et enfin (iv) l'étude de la généralisation du modèle par la validation de nos résultats sur une cohorte indépendante, montrant son applicabilité dans la maladie d'Alzheimer.

Les méthodes de machine learning avancées constituent donc un outil puissant, dans le cadre d'une approche rigoureuse et reproductible. Un travail de collaboration entre experts en machine learning et médecins semble être une méthode efficace et souhaitable pour construire des biomarqueurs robustes, puissants et explicables.

BIBLIOGRAPHIE :

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May;521(7553):436–44.

2. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. 2014 Sep 1 [cited 2018 May 23]; Available from: https://arxiv.org/abs/1409.0575

3. Monticciolo DL, Newell MS, Moy L, Niell B, Monsees B, Sickles EA. Breast Cancer Screening in Women at Higher-Than-Average Risk: Recommendations From the ACR. J Am Coll Radiol. 2018 Mar;15(3):408–14.

4. BAKER M. A Nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help. :3.

5. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011 Apr;64(4):401–6.

6. Bruno MA, Walker EA, Abujudeh HH. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. Radiogr Rev Publ Radiol Soc N Am Inc. 2015 Oct;35(6):1668–76.

7. Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and System Factors Contributing to Diagnostic Errors in Radiology. Am J Roentgenol. 2013 Aug 23;201(3):611–7.

8. Reiner BI, Krupinski E. The Insidious Problem of Fatigue in Medical Imaging Practice. J Digit Imaging. 2012 Feb;25(1):3–6.

9. Muenzel D, Engels H-P, Bruegel M, Kehl V, Rummeny EJ, Metz S. Intra- and inter-observer variability in measurement of target lesions: implication on response evaluation according to RECIST 1.1. Radiol Oncol. 2012 Jan 2;46(1):8–18.

10. Suzuki C, Torkzad MR, Jacobsson H, Aström G, Sundin A, Hatschek T, et al. Interobserver and intraobserver variability in the response evaluation of cancer therapy according to RECIST and WHO-criteria. Acta Oncol Stockh Swed. 2010 May;49(4):509–14.

11. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017 Dec;14(12):749–62.

12. Sacconi B, Anzidei M, Leonardi A, Boni F, Saba L, Scipione R, et al. Analysis of CT features and quantitative texture analysis in patients with lung adenocarcinoma: a correlation with EGFR mutations and survival rates. Clin Radiol. 2017 Jun;72(6):443–50.

13. Zhou H, Vallières M, Bai HX, Su C, Tang H, Oldridge D, et al. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. Neuro-Oncol. 2017 01;19(6):862–70.

14. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016 May 23;6:26286.

15. Courtiol P, Tramel EW, Sanselme M, Wainrib G. Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. ArXiv180202212 Cs Stat [Internet]. 2018 Feb 1 [cited 2018 May 23]; Available from: http://arxiv.org/abs/1802.02212

16. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb;542(7639):115–8.

17. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 Dec 13;316(22):2402–10.

18. Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Comput Biol Med. 2018 May 17;98:126–46.

19. Zhu B, Z. Liu J, R. Rosen B, S. Rosen M. Image reconstruction by domain transform manifold learning. Nature. 2017 Apr 28;555.

20. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver Fibrosis: Deep Convolutional Neural Network for Staging by Using Gadoxetic Acid–enhanced Hepatobiliary Phase MR Images. Radiology. 2017 Dec 14;287(1):146–55.

21. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer [Internet]. 2018 May 17 [cited 2018 May 23]; Available from: http://www.nature.com/articles/s41568-018-0016-5

22. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, et al. Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. Can Assoc Radiol J. 2018 May 1;69(2):120–35.

23. Artificial intelligence and medical imaging 2018: French Radiology Community white paper. Diagn Interv Imaging. 2018 Nov 1;99(11):727–42.

24. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ArXiv160204938 Cs Stat [Internet]. 2016 Feb 16 [cited 2018 May 22]; Available from: http://arxiv.org/abs/1602.04938

25. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. Computer Vision – ECCV 2014 [Internet]. Cham: Springer International Publishing; 2014 [cited 2018 May 29]. p. 818–33. Available from: http://link.springer.com/10.1007/978-3-319-10590-1_53

26. Holzinger A, Plass M, Holzinger K, Crisan GC, Pintea C-M, Palade V. A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop. ArXiv170801104 Cs Stat [Internet]. 2017 Aug 3 [cited 2018 Aug 21]; Available from: http://arxiv.org/abs/1708.01104

27. Hutton C, Draganski B, Ashburner J, Weiskopf N. A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. NeuroImage. 2009 Nov 1;48(2):371–80.

28. Good CD, Johnsrude IS, Ashburner J, Henson RNA, Friston KJ, Frackowiak RSJ. A Voxel-Based Morphometric Study of Ageing in 465 Normal Adult Human Brains. NeuroImage. 2001 Jul 1;14(1):21–36.

29. Grueter BE, Schulz UG. Age-related cerebral white matter disease (leukoaraiosis): a review. Postgrad Med J. 2012 Feb;88(1036):79–87.

30. Lin J, Wang D, Lan L, Fan Y. Multiple Factors Involved in the Pathogenesis of White Matter Lesions. BioMed Res Int. 2017;2017:9372050.

31. Anthony M, Lin F. A Systematic Review for Functional Neuroimaging Studies of Cognitive Reserve Across the Cognitive Aging Spectrum. Arch Clin Neuropsychol Off J Natl Acad Neuropsychol. 2017 Dec 13;

32. Gaillard F. Medial temporal lobe atrophy score | Radiology Reference Article | Radiopaedia.org [Internet]. Radiopaedia. [cited 2018 May 16]. Available from: https://radiopaedia.org/articles/medial-temporal-lobe-atrophy-score

33. Segev Y. Morphometric study of the midsagittal MR imaging plane in cases of hydrocephalus and atrophy and in normal brains. - PubMed - NCBI [Internet]. 2001 [cited 2019 May 8]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/11673160

34. Fazekas F, Chawluk J, Alavi A, Hurtig H, Zimmerman R. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. Am J Roentgenol. 1987 Aug 1;149(2):351–6.

35. Haller S, Vernooij MW, Kuijer JPA, Larsson E-M, Jäger HR, Barkhof F. Cerebral Microbleeds: Imaging and Clinical Significance. Radiology. 2018 Apr 1;287(1):11–28.

36. Cole JH, Poudel RPK, Tsagkrasoulis D, Caan MWA, Steves C, Spector TD, et al. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage. 2017 Dec 1;163:115–24.

37. Koutsouleris N, Davatzikos C, Borgwardt S, Gaser C, Bottlender R, Frodl T, et al. Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophr Bull. 2014 Sep;40(5):1140–53.

38. Raffel J, Cole J, Record C, Sridharan S, Sharp D, Nicholas R. Brain Age: A novel approach to quantify the impact of multiple sclerosis on the brain (P1.371). Neurology [Internet]. 2017 Apr 18;88(16 Supplement). Available from: http://n.neurology.org/content/88/16_Supplement/P1.371.abstract

39. Cole JH, Leech R, Sharp DJ, Alzheimer's Disease Neuroimaging Initiative. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. Ann Neurol. 2015 Apr;77(4):571–81.

40. Gaser C, Franke K, Klöppel S, Koutsouleris N, Sauer H, Alzheimer's Disease Neuroimaging Initiative. BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease. PloS One. 2013;8(6):e67346.

41. Franke K, Gaser C. Longitudinal Changes in Individual BrainAGE in Healthy Aging, Mild Cognitive Impairment, and Alzheimer's Disease 1Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. GeroPsych. 2012 Jan 1;25(4):235–45.

42. Cole JH, Ritchie SJ, Bastin ME, Valdés Hernández MC, Muñoz Maniega S, Royle N, et al. Brain age predicts mortality. Mol Psychiatry [Internet]. 2017 Apr 25 [cited 2018 Jan 9]; Available from: http://www.nature.com/doifinder/10.1038/mp.2017.62

43. Yang C, Rangarajan A, Ranka S. Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification. ArXiv180302544 Cs Stat [Internet]. 2018 Mar 7 [cited 2018 May 31]; Available from: http://arxiv.org/abs/1803.02544

44. Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, Peters TM. 3D statistical neuroanatomical models from 305 MRI volumes. In: 1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference. 1993. p. 1813–7 vol.3.

45. Ruigrok ANV, Salimi-Khorshidi G, Lai M-C, Baron-Cohen S, Lombardo MV, Tait RJ, et al. A meta-analysis of sex differences in human brain structure. Neurosci Biobehav Rev. 2014 Feb 1;39:34–50.

46. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: Improved N3 Bias Correction. IEEE Trans Med Imaging. 2010 Jun;29(6):1310–20.

47. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging. 2001 Jan;20(1):45–57.

48. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. ArXiv170609516 Cs [Internet]. 2017 Jun 28 [cited 2018 May 30]; Available from: http://arxiv.org/abs/1706.09516

49. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci. 1997 Aug 1;55(1):119–39.

50. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017 May 24;60(6):84–90.

51. Graves A, Mohamed A, Hinton G. Speech Recognition with Deep Recurrent Neural Networks. ArXiv13035778 Cs [Internet]. 2013 Mar 22 [cited 2018 May 30]; Available from: http://arxiv.org/abs/1303.5778

52. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv14090473 Cs Stat [Internet]. 2014 Sep 1 [cited 2018 May 30]; Available from: http://arxiv.org/abs/1409.0473

53. Silver D, Huang A, Maddison C, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016 Jan 27;529:484–9.

54. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. ArXiv151203385 Cs [Internet]. 2015 Dec 10 [cited 2017 Nov 27]; Available from: http://arxiv.org/abs/1512.03385

55. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ArXiv150203167 Cs [Internet]. 2015 Feb 10 [cited 2018 May 30]; Available from: http://arxiv.org/abs/1502.03167

56. Hamer M, Batty GD. Association of body mass index and waist-to-hip ratio with brain structure: UK Biobank study. Neurology. 2019 Feb 5;92(6):e594–600.

57. Fama R, Sullivan EV. Thalamic structures and associated cognitive functions: Relations with age and aging. Neurosci Biobehav Rev. 2015 Jul;54:29–37.

58. Carter, et al., "Activation Atlas", Distill, 2019. [Internet]. [cited 2019 May 1]. Available from: https://distill.pub/2019/activation-atlas/

59. Kaufmann T, van der Meer D, Doan NT, Schwarz E, Lund MJ, Agartz I, et al. Genetics of brain age suggest an overlap with common brain disorders. 2018 Apr 17 [cited 2018 Aug 27]; Available from: http://biorxiv.org/lookup/doi/10.1101/303164

60. Becker BG, Klein T, Wachinger C. Gaussian Process Uncertainty in Age Estimation as a Measure of Brain Abnormality. ArXiv180401296 Cs Q-Bio [Internet]. 2018 Apr 4 [cited 2018 Apr 6]; Available from: http://arxiv.org/abs/1804.01296

61. Bouts MJRJ, Möller C, Hafkemeijer A, van Swieten JC, Dopper E, van der Flier WM, et al. Single Subject Classification of Alzheimer's Disease and Behavioral Variant Frontotemporal Dementia Using Anatomical, Diffusion Tensor, and Resting-State Functional Magnetic Resonance Imaging. J Alzheimers Dis JAD. 2018;62(4):1827–39.

62. Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. ArXiv14121897 Cs [Internet]. 2014 Dec 5 [cited 2018 May 31]; Available from: http://arxiv.org/abs/1412.1897

63. Brown TB, Mané D, Roy A, Abadi M, Gilmer J. Adversarial Patch. ArXiv171209665 Cs [Internet]. 2017 Dec 27 [cited 2018 May 31]; Available from: http://arxiv.org/abs/1712.09665

64. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. ArXiv13126199 Cs [Internet]. 2013 Dec 20 [cited 2018 May 31]; Available from: http://arxiv.org/abs/1312.6199

65. Elsken T, Metzen JH, Hutter F. Neural Architecture Search: A Survey. ArXiv180805377 Cs Stat [Internet]. 2018 Aug 16 [cited 2018 Aug 30]; Available from: http://arxiv.org/abs/1808.05377

66. Sabour S, Frosst N, Hinton GE. Dynamic Routing Between Capsules. ArXiv171009829 Cs [Internet]. 2017 Oct 26; Available from: http://arxiv.org/abs/1710.09829

67. Ali I, Hart GR, Gunabushanam G, Liang Y, Muhammad W, Nartowt B, et al. Lung Nodule Detection via Deep Reinforcement Learning. Front Oncol. 2018;8:108.