

Une stratégie de données efficace pour la détection des anévrismes cérébraux avec l'apprentissage profond

Youssef Assis, Liang Liao, Fabien Pierre, René Anxionnat, Erwan Kerrien

▶ **To cite this version:**

Youssef Assis, Liang Liao, Fabien Pierre, René Anxionnat, Erwan Kerrien. Une stratégie de données efficace pour la détection des anévrismes cérébraux avec l'apprentissage profond. ORASIS 2021 - 18èmes journées francophones des jeunes chercheurs en vision par ordinateur, Centre National de la Recherche Scientifique [CNRS], Sep 2021, Saint Ferréol, France. hal-03339672v2

HAL Id: hal-03339672
https://hal.science/hal-03339672v2

Submitted on 5 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Une stratégie de données efficace pour la détection des anévrismes cérébraux avec l'apprentissage profond

An efficient data strategy for the detection of brain aneurysms with deep learning

Youssef Assis[1] Liang Liao[2] Fabien Pierre[1] René Anxionnat[2,3] Erwan Kerrien[1]

[1] Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
[2] Département de neuroradiologie diagnostique et thérapeutique, Université de Lorraine, Nancy, France
[3] IADI, INSERM U1254, Université de Lorraine, Nancy, France

Résumé

La détection des anévrismes intracrâniens à partir d'images d'angiographie par résonance magnétique (MRA) est un problème dont l'importance clinique croît rapidement, mais qui est par ailleurs extrêmement difficile à automatiser. Cependant, ces 3 dernières années ont connu une augmentation du recours à des réseaux de neurones convolutifs et suscité un corpus de méthodes qui ont surmonté cette impasse technologique et avec des performances de détection convaincantes. Le problème majeur à résoudre est le très déséquilibre de classes présenté dans les données. Les travaux précédents ont concentré leurs études sur l'architecture du réseau et la fonction de coût. De manière complémentaire, cet article se concentre sur les données. Une annotation approximative et rapide est proposée : chaque anévrisme est approché par une sphère définie par deux points. Par ailleurs, une approche par patchs est adoptée afin d'augmenter le nombre d'échantillons. Enfin, les échantillons sont générés par une combinaison de sélection de données (les patchs négatifs sont centrés pour moitié sur les vaisseaux sanguins et pour moitié sur le parenchyme) et de synthèse de données (les patchs contenant un anévrisme sont dupliqués et déformés par une transformation spline 3D). Cette stratégie est appliquée pour entraîner un réseau standard 3D U-net, avec l'entropie-croisée binaire comme fonction de coût, sur un ensemble de données de 111 patients. De très bonnes performances ont été évaluée par validation croisée à 5 blocs (sensibilité 0.82, nombre de faux positifs 0.61, selon les critères du challenge ADAM [1]). L'étude fournit également une comparaison avec la fonction de coût Focal, et le coefficient Kappa de Cohen s'avère être une meilleure métrique que Dice pour ce problème de détection très déséquilibré.

Mots Clef

Détection des anévrismes cérébraux, échantillonnage de données, CNN.

1. http ://adam.isi.uu.nl/

Abstract

The detection of intracranial aneurysms from Magnetic Resonance Angiography images is a problem of rapidly growing clinical importance, but also extremely challenging to automate. However, in the last 3 years, the raise of deep convolutional neural networks has instigated a streak of methods that have convincingly removed the technological deadlock and show promising performance. The major issue to address is the very severe class imbalance. Previous authors have focused their efforts on the network architecture and loss function. This paper tackles the data. A rough but fast annotation is considered : each aneurysm is approximated by a sphere defined by two points. Second, a small patch approach is taken so as to increase the number of samples. Third, samples are generated by a combination of data selection (negative patches are centered half on blood vessels and half on parenchyma) and data synthesis (patches containing an aneurysm are duplicated and deformed by a 3D spline transform). This strategy is applied to train a 3D U-net model, with a binary cross entropy loss, on a data set of 111 patients. A 5-fold cross-validation evaluation provides state of the art results (sensitivity 0.82, false positive count 0.61, as per ADAM challenge criteria). The study also reports a comparison with the focal loss, and Cohen's Kappa coefficient is shown to be a better metric than Dice for this highly unbalanced detection problem.

Keywords

Brain aneurysm detection, data sampling, CNN.

1 Introduction

Les anévrismes intracrâniens sont des dilatations locales des vaisseaux sanguins cérébraux. Leur rupture représente 85% des hémorragies sub-arachnoïdiennes, et est associée à des taux élevés de mortalité et de morbidité [1]. La généralisation des examens radiologiques dans le processus de diagnostic a révélé un nombre insoupçonné d'anévrismes non rompus. Leur détection est donc un problème dont l'importance clinique ne cesse pas de croître. Cependant, l'exploration des examens 3D d'angiographie par tomodensitométrie (CTA) ou par imagerie par résonance magnétique (IRM), de volume toujours plus grand et dans un contexte clinique de plus en plus contraint par le temps, conduit fatalement à des erreurs. L'innocuité de l'IRM temps de vol en 3D (TOF – *Time-of-Flight*) rend cette imagerie particulièrement adaptée au dépistage, même si la détection de petits anévrismes (< 5 mm) peut être difficile [2]. Par conséquent, une méthode automatisée fiable serait un atout précieux pour aider les radiologues dans leur routine clinique.

Le premier système de détection assistée par ordinateur signalé dans la littérature [3] était basée sur les opérations traditionnelles de traitement d'images. Un filtre sélectif 3D d'amélioration multi-échelle suivi par des techniques de multi-seuillage en niveaux de gris pour détecter les caractéristiques liées aux anévrismes dans les images. Récemment, les réseaux de neurones convolutifs (CNN – *Convolutional Neural Networks*) ont prouvé leur l'efficacité dans de nombreuses tâches visuelles, y compris l'analyse d'images médicales.

La détection des anévrismes cérébraux est une tâche très difficile car les anévrismes sont un signal rare (quelques dizaines à des centaines de voxels positifs parmi des millions dans les données IRM), et leur nombre est indéfini a priori (1 à quelques unités par patient). Par conséquent, ce n'est que très récemment que des approches d'apprentissage profond ont été étudiées dans ce contexte. Un modèle ResNet-18 a été entraîné pour détecter les anévrismes dans les images 2D source (axiales) parmi une collection de classes d'anomalies [4]. Mais rapidement, des images de projection d'intensité maximale (MIP – *Maximum Intensity Projection*) ont été utilisées pour mieux capturer les informations vasculaires en 3D. Une seule image a été définie comme entrée d'un réseau U-net dans [5], tandis que [6] a utilisé une collection d'images MIP d'un *patch* (sous-volume) comme entrée d'un classificateur binaire à 5 couches dans une approche 2.5D. Mais les méthodes les plus récentes sont entièrement en 3D. Les performances du modèle DeepMedic multi-échelles à double chemin ont été jugées prometteuses [7] mais en complément à un expert [8]. L'année dernière, le challenge ADAM a permis une comparaison objective de diverses autres approches 3D. Les 3 principales méthodes pour la tâche de détection étaient basées sur 3D U-net [9] pour échapper au problème du nombre indéfini d'anévrismes grâce à la génération d'une carte de chaleur. La rareté, et donc le déséquilibre de classes élevé, a été abordée soit par la fonction de coût et/ou par l'architecture du réseau utilisé. La somme du coefficient de similarité Dice et de la fonction de coût d'entropie binaire croisée (BCE) a été combinée avec la somme des fonctions de coût Dice et TopK pour concevoir une nouvelle fonction de coût dans [10]. Une approche différente a été adoptée dans [11]. Quatre modèles basés sur la variante New-Net de 3D U-net ont été entraînés et la segmentation finale a été décidée par un vote majoritaire de ces modèles. La méthode leader [12] s'est concentrée plus sur l'architecture du modèle avec un modèle Retina U-net qui agrège un réseau d'encodeurs et un réseau pyramidal pour guider la détection de haute résolution avec des informations sémantiques fortes à basse résolution. L'émergence de ces informations sémantiques a été facilitée en associant l'image angiographique TOF avec l'imagerie structurelle comme entrée du réseau. L'impact réel de toutes ces variantes a été remis en question par l'émergence récente d'un nouveau leader avec un modèle U-net 3D entraîné avec une fonction de coût combinant Dice et BCE, et une prédiction basée sur un ensemble de 5 modèles [13]. Nous pensons qu'une différence majeure réside dans la génération des données d'entrée. Si les 3 premières méthodes utilisaient de gros patchs ($\{192,224,256\} \times 256 \times 56$ voxels), cette dernière utilisait des patchs isotropes de taille 128^3 avec un processus d'augmentation de données riche. Une autre limitation que nous constatons est le petit nombre de données sur les patients dans les bases de données actuelles et la corvée de l'annotation de données, voxel par voxel.

L'étude actuelle se concentre sur la stratégie de données pour générer des échantillons d'entrée qui sont mieux conçus pour faire face au problème de déséquilibre de classes et à la rareté des données médicales. Nous avons utilisé un réseau standard 3D U-net dans toutes nos expériences, mais avec de petits patchs isotropes (48^3). L'annotation des anévrismes est approximative mais rapide. Cette approximation est suffisamment précise pour la tâche de détection. Une combinaison de sélection guidée et de synthèse des patchs échantillons est également proposée, et les fonctions de coût BCE et Focal sont comparées. Finalement, nous préconisons l'utilisation du coefficient Kappa de Cohen comme meilleure métrique que Dice dans les situations de déséquilibre de classes.

2 Matériels et méthodes

2.1 Collection et annotation de données

Un total de 111 examens TOF-IRM (56 femmes, 55 hommes) ont été collectés dans notre établissement médical entre avril 2015 et janvier 2020, dont chacune présente au moins un anévrisme. Tous les anévrismes sont de type sacculaire. Les critères d'exclusion étaient les anévrismes pré-traités et les gros anévrismes (> 20 mm), ces derniers étant extrêmement rares et évidents à détecter par les experts. Les images ont été acquises sur une machine 3T (GE Healthcare) avec les paramètres suivants : TR = 28 ms, TE = 3.4 ms, épaisseur de coupe = 0.8 mm,

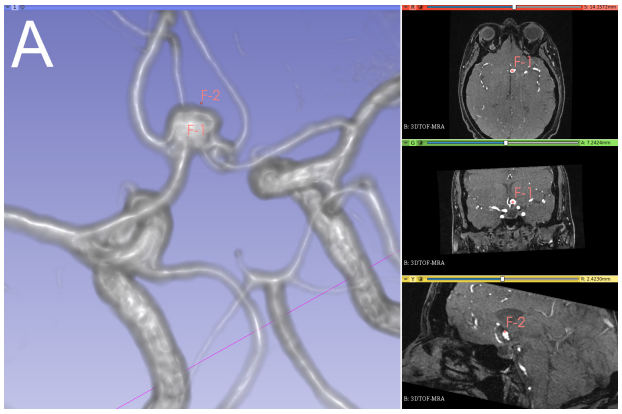


FIGURE 1 – Annotation d'anévrisme comme sphère approximative avec 2 points dans Slicer

FOV = 24, angle de bascule = 17°, 4 blocs (54 coupes / bloc), temps d'acquisition = 6min 28s, ce qui donne des volumes de $512 \times 512 \times 254$ voxels avec une taille de $0.47 \times 0.47 \times 0.4\text{mm}^3$ voxel. Chaque donnée DICOM a été anonymisée puis convertie au format Nifty dans le site clinique avant le traitement. Chaque examen contenait de un (81/111) à cinq anévrismes (1 cas) pour un total de 155 anévrismes d'un diamètre moyen de 3.86 mm $\pm$ 2.39 mm (min : 1.23 mm, max : 19.63 mm). Il s'agissait pour la plupart de petits anévrismes puisque 60 étaient inférieurs à 3mm et 66 entre 3 et 5mm, ce qui constitue une base de données variée et de grande difficulté.

Les travaux précédents s'appuient sur des bases de données où les anévrismes ont été annotés voxel par voxel. Cette annotation est à la fois fastidieuse et liée à une grande variabilité inter-observateurs. Comme nous ne visons que la détection des anévrismes, nous avons utilisé un processus d'annotation moins précis mais beaucoup plus facile et rapide : chaque anévrisme a été approchée par une sphère, définie en plaçant deux points, l'un au centre du collet et l'autre au niveau du dôme. Cette annotation a été réalisée par un radiologue avec 10 ans d'expérience. Le logiciel *3D Slicer* a été utilisé [14] pour placer les points dans une vue 3D en rendu volumique, avec un contrôle visuel dans les plans de coupe canoniques (voir Fig. 1).

Le crâne et ce qui l'entoure a été supprimé de manière entièrement automatique dans chaque volume de la base. Nous avons constaté que la méthode d'effacement de Brett [15] avait tendance à supprimer trop d'informations vasculaires, en particulier autour du siphon carotidien. En conséquence, nous avons conçu une méthode approximative mais simple : l'image de magnitude du gradient a été seuillée à son 80^e percentile et un masque de sélection a été généré en conservant les segments entre le premier et le dernier voxels non nuls de chaque ligne du volume. Ensuite, ce masque a été érodé 20 fois à l'aide d'une boule 3D d'un rayon de 2 voxels.

2.2 Implémention du modèle

Notre code est écrit en Python (3.8.5) en utilisant le framework TensorFlow (2.4.0), avec Keras (2.4.3). Il est basé sur l'implémentation open-source de 3D U-net par D.G. Ellis [16] avec 4 couches, chacune avec 2 blocs de convolution (transposés dans le chemin d'expansion). Il prend des patchs multimodaux en entrée, mais accepte également une entrée de modalité unique. Nos premières investigations avec l'échantillonnage de patchs régulier n'ont pas montré de convergence du modèle. En conséquence, le générateur de patchs décrit dans la section suivante a été intégré en entrée du modèle. Les hyperparamètres suivants ont été utilisés : 100 époques, taux d'apprentissage constant $= 10^{-4}$, fonction de coût BCE, optimiseur Adam, taille de batch = 10, avec une normalisation par batch. Chaque volume d'entrée a été normalisé entre 0 et 1. L'entraînement et la prédiction ont été effectuées sur une carte graphique NVIDIA RTX 1080Ti (11 Go). Le modèle est décrit par environ 19 millions de paramètres et occupe 7.8 Go en mémoire GPU.
Un volume complet est prédit par reconstruction par patch : le volume initial est rééchantillonné à une taille de voxel isotrope de 0.4 mm ; des patchs de taille $48 \times 48 \times 48$ sont extraits de manière à couvrir tout le volume ; une prédiction est calculée pour chaque patch ; puis les patchs sont recollés et le volume résultant est rééchantillonné à la résolution d'origine. Afin d'éviter les effets de bord dus aux opérations de convolutions sur les petits patchs, un chevauchement de 8 voxels a été considéré entre les patchs voisins et seule la partie centrale $32 \times 32 \times 32$ des patchs a été juxtaposée pour couvrir le volume final.

2.3 Génération et augmentation de données

Le pouvoir discriminant d'un classificateur dépend de sa capacité à modéliser statistiquement à la fois le fond (échantillons négatifs) et l'objet (échantillons positifs). Dans les approches par grands patchs, un anévrisme est présent dans la plupart des patchs, ce qui nécessite des patients en bonne santé dans la base de données. Dans notre approche par petits patchs, les patchs négatifs (sans anévrisme) sont très courants en dehors de l'environnement de l'anévrisme, mais plusieurs instances doivent être extraites de chaque donnée patient pour construire des statistiques fiables sur le fond. Dans l'autre côté, un seul patch positif existe pour chaque anévrisme. L'échantillonnage de données adapté est une stratégie efficace pour gérer ce problème [17]. Notre première stratégie d'échantillonnage consiste démultiplier les patchs positifs en dupliquant plusieurs fois chaque patch positif, centré sur chaque anévrisme. Une variété de formes est synthétisées en appliquant une distorsion aléatoire à chaque copie : un treillis $3 \times 3 \times 3$ de points est positionné pour englober le patch, et chaque point de contrôle, à l'exception du point central, est déplacé aléatoirement de 4 mm dans les 3 directions de l'espace. Un champ de distorsion dense 3D est calculé par interpolation spline cubique (fonction

`scipy.ndimages.map_coordinates`).

Mais un autre déséquilibre de classes doit être traité. En effet, les informations vasculaires ne représentent que 3 à 5 % du fond. Afin de guider le modèle pour discriminer les vaisseaux sains de la pathologie, notre deuxième stratégie d'échantillonnage consiste à prélever la moitié des échantillons négatifs centrés sur un vaisseau sanguin, et l'autre moitié à l'extérieur (parenchyme). Les 100 voxels les plus brillants ont été sélectionnés comme centres de patch, contraints avec une distance minimale de 20 mm entre centres de patch pour éviter tout recouvrement. Cela capture la partie la plus brillante de 3 à 5% de chaque volume IRM. 100 autres centres ont été sélectionnés au hasard dans les valeurs de voxel entre le 20^e et le 80^e percentile, avec la même contrainte de distance inter-centres.

En conséquence, nous avons utilisé 200 patchs négatifs, et nous avons expérimenté avec 5 et 50 doublons de patchs positifs. Nous avons utilisé des patchs de taille $48 \times 48 \times 48$ avec une taille de voxel isotrope de 0.4 mm, la plus proche de la résolution nominale, de sorte que les patchs étaient des cubes de côté 19 mm. Pendant l'apprentissage, la même augmentation de données à été appliquée à tous les patchs (positifs et négatifs) : rotations (0 à 180°) et translations (10 mm) aléatoires dans les 3 directions de l'espace. En raison de la taille limitée du patch, chaque instance de patch a été ré-extraite du volume d'origine correspondant au moment de l'entraînement. L'apprentissage requiert environ 1h par époque.

2.4 Mesures et évaluation des performances

L'évolution de l'apprentissage a été surveillée en utilisant le coefficient de Dice. Cependant, comme les anévrismes sont rares et de petite taille dans le volume, cette métrique est très sensible même aux petites erreurs de détection. Nous avons également calculé le coefficient Kappa de Cohen (κ) [18], qui est plus robuste au déséquilibre de classes. Il compare la probabilité d'accord observée (P_o) à la probabilité d'accord aléatoire (P_e) : $\kappa = (P_o - P_e)/(1 - P_e)$. Il peut être négatif, avec un maximum de 1 et vaut 0 lorsque le classificateur binaire choisit systématiquement une classe. Ces métriques sont calculées sur les voxels de la collection de patchs extraits d'une base de validation (seuil de détection=0.5).

Les performances du modèle ont été évaluées sur un ensemble de test, en utilisant les scores de la sensibilité moyenne et le nombre de faux positifs par cas (FP/cas) tels que définis pour la tâche 1 dans le challenge ADAM [19], et que nous rappelons brièvement. Les composantes connexes (CC) sont extraites à la fois dans la vérité terrain et les volumes prédits. Un vrai positif (TP) est une CC dans la vérité terrain qui contient le centre de gravité d'une CC prédite. Un faux négatif (FN) est une CC dans la vérité terrain sans CC prédite. Un faux positif (FP) est une CC prédite dont le centre n'est contenu dans aucune CC de la vérité terrain.

Cependant, les métriques ci-dessus ne calculent pas le nombre de vrais négatifs (TN) et par suite empêchent le calcul de la spécificité. C'est pour cette raison que nous avons également calculé des statistiques relatives aux patchs (pas de duplication pour les patchs positifs) : pour cette deuxième évaluation, un patch est considéré comme positif s'il contient un voxel positif, sinon il est négatif. Cela permet de calculer une matrice de confusion complète.

3 Expériences et résultats

3.1 Étude initiale

Un premier ensemble d'expériences vise à évaluer la pertinence des différentes parties du modèle, en procédant par ablation. On note *Model0*, la stratégie d'échantillonnage de patchs proposée, telle que décrite dans les sections 2.2 et 2.3, avec 50 copies de patchs positifs. 4 variantes ont été testées.

- *Model1* : le coût BCE est très sensible au déséquilibre de classes [20]. Afin de voir l'efficacité de notre stratégie de données pour agir contre le déséquilibre de classes, nous avons entraîné le même modèle avec la stratégie *Model0*, mais avec la fonction de coût Focal [21], qui a été conçue pour concentrer l'apprentissage sur la classe minoritaire.
- *Model2* : comme *Model0* avec seulement 5 copies pour chaque patch positif.
- *Model3* : utilise 50 copies mais sans aucune distorsion aléatoire appliquée.
- *Model4* : comme *Model0* mais ne considère que 100 patchs négatifs par patient (50 sur les vaisseaux, 50 à l'extérieur).

L'ensemble de données a été divisé en 3 ensembles utilisés pour : l'entraînement (78 cas, 70%), la validation (22 cas, 20%) et le test (11 cas, 10%). Tous les modèles ont été entraînés sur l'ensemble d'entraînement et surveillés avec l'ensemble de validation. Le tableau 1 rapporte les résultats de cette étude. Les coefficients Dice et κ sont calculés sur l'ensemble de validation à la fin de chaque époque. Les autres métriques de performance (voir Sec. 2.4) sont calculées sur l'ensemble de données de test.

Le meilleur résultat est obtenu avec la stratégie que nous proposons (*Model0*). Aucune réelle amélioration n'a pu être observée avec la fonction de coût Focal (*Model1*), et la sensibilité est même meilleure avec BCE, ce qui démontre l'efficacité de notre stratégie d'échantillonnage. Notez que le Dice a resté très bas car les CC prédites étaient très petites. Cependant, κ a mieux rendu compte les performances relativement bonnes de ce modèle. *Model2* n'a pas convergé : le déséquilibre de classes est en effet un problème. *Model3* a fourni de bons résultats mais avec trop de FP, en raison du manque de diversité dans les formes d'anévrisme montrées au modèle pendant l'entraînement (pas de synthèse de nouvelles formes par distorsion). L'excès de voxels positifs conduit à des TP plus grands, ce qui explique les meilleurs scores Dice et κ. Enfin, l'équilibre de classes est amélioré dans *Model4*, mais il n'atteint pas la

Modèle	*Validation set* Dice	κ	*Métriques ADAM* Sensibilité	FPs/case
Model0	0.339	0.665	0.970	0.454
Model1	0.089	0.527	0.803	0.190
Model2	0.038	-1.21e-8	0	0
Model3	0.434	0.772	0.879	1.545
Model4	0.245	0.589	0.833	1.0

Modèle	*Métriques par patch* TP	FP	FN	TN
Model0	14	5	2	2194
Model1	12	2	4	2197
Model2	0	0	16	2199
Model3	14	11	2	2188
Model4	13	8	3	2191

TABLE 1 – Comparaison des variantes du modèle : *Model0* est notre modèle proposé. Dice et κ (coefficient Kappa de Cohen) ont été évalués sur l'ensemble de validation à l'issue de l'entraînement (100 époques, 22 patients). Les métriques d'ADAM et par patchs ont été mesurées sur l'ensemble de test (11 patients).

performance de *Model0* car la taille de l'échantillon pour les patchs négatifs est trop petite pour modéliser de manière fiable les statistiques du fond.

3.2 Validation croisée

La performance globale de notre modèle proposé (*Model0*) a été évaluée en utilisant une validation en 5 blocs. L'ensemble de données a été divisé en 4 blocs de 22 cas et un avec 23 cas. 5 modèles ont été entraînés, chaque fois avec 4 blocs pour l'entraînement (88 ou 89 cas) et en laissant un seul bloc pour le test. Des prédictions ont été générées pour chaque patient dans chaque ensemble de tests, fournissant une prédiction pour chaque patient. Les diamètres moyens des anévrismes dans les 5 blocs étaient : 3.82 mm, 3.74 mm, 3.96 mm, 3.84 mm et 3.93 mm.

La figure 2 rend compte des scores de sensibilité et de FP/cas, selon le challenge ADAM, calculés sur les 111 patients de notre ensemble de données, et comparés aux scores rapportés pour les 4 premières méthodes du challenge. Même si ces derniers ont été évalués sur une base de données différente, les caractéristiques de notre base sont similaire et permettent une comparaison. Notre modèle atteint une excellente sensibilité de 0.82 avec 0.61 FP/cas. Nous indiquons par ailleurs la courbe FROC (*Free-response Receiver Operating Characteristic*) calculée pour cette expérience en faisant varier le seuil de détection. Par comparaison avec la méthode abc, notre méthode atteint une sensibilité de 0.80 à un taux de 0.40 FP/cas, et, par rapport à la méthode mibaumgartner, une sensibilité de 0.70 à un taux de 0.13 FP/cas. La surface sous la courbe (AUC) est de 85.24 %. Notez que en choisissant le seuil de détection optimal qui correspond au point le plus proche du coin supérieur à gauche de la courbe FROC, notre modèle

Méthode	Sensibilité	FP/ cas
abc [13]	0.68	0.40
mibaumgartner [12]	0.67	0.13
joker [11]	0.63	0.16
junma [10]	0.61	0.18
Notre modèle	0.82	0.61

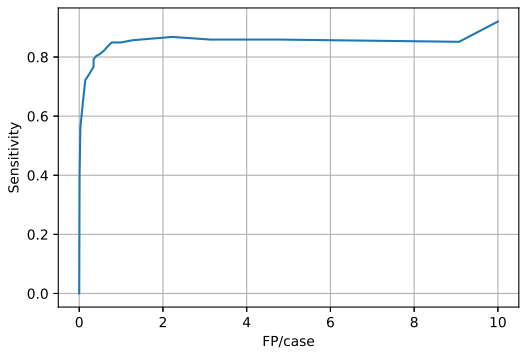


FIGURE 2 – (Haut) Comparaison avec 4 méthodes principales du challenge ADAM (par ordre décroissant). À noter que la base de données est différente, bien que similaire, de la nôtre. (Bas) Courbe FROC pour notre modèle : AUC = 85,24 %.

atteint une sensibilité de 0,72 associée à un taux de 0,14 FP/cas.

4 Discussion

Dans cette étude et afin de déterminer l'impact de notre stratégie de données, nous avons volontairement utilisé un réseau 3D U-net standard avec la fonction de coût BCE et un processus d'optimisation simple (par exemple, taux d'apprentissage et nombre d'époques fixés). L'accent a été mis sur les données afin d'évaluer l'impact de divers aspects de leur préparation sur l'apprentissage.

Tout d'abord, une annotation rugueuse mais rapide a été utilisée, ce qui permet d'annoter rapidement un grand nombre de volumes IRM. En outre, une approche par petits patchs a été choisie. Les petits patchs sans intersection peuvent être supposés indépendants, ce qui permet une exploitation efficace même d'un petit ensemble d'images IRM originales (111). Plus de 2000 échantillons de patches ont généralement été utilisés dans le processus d'apprentissage (voir le tableau 1). En outre, le modèle résultant a une faible empreinte mémoire.

Deuxièmement, nous avons proposé un processus d'échantillonnage de données adapté en deux étapes. D'un côté, un échantillonnage guidé : les patchs négatifs (sans anévrisme) sont extraits par moitié centrés sur les vaisseaux sanguins et l'autre moitié ailleurs. Nous avons montré que 200 patchs étaient plus capables que 100 de capturer les statistiques du fond (*Model0* vs *Model4*). De l'autre côté, la synthèse des données : les patchs positifs sont démultipliés, ce qui permet de faire face au déséquilibre de la

classe majoritaire (*Model0* vs *Model2*), et diverses formes nouvelles sont synthétisées en appliquant des distorsions aléatoires non rigides. Ainsi, les statistiques caractérisant les anévrismes sont décrites avec plus de précision, ce qui s'exprime par une réduction des faux positifs (*Model0* vs *Model3*).

Le modèle proposé a une sensibilité de 0.82, avec un nombre de FP/cas de 0.61. L'analyse de la courbe FROC montre qu'elle est compétitive par rapport aux meilleurs méthodes actuelles du le challenge ADAM. Notre méthode devra cependant être adaptée aux conditions du challenge ADAM pour une comparaison définitive. Mais il apparaît également que les FP sont encore trop nombreux. Les tests avec la fonction de coût Focal (*Model1*) ont généré des CC plus petites, ce qui a réduit le score FP mais au détriment de la sensibilité. Outre les cas faciles à éliminer pour un radiologue, les FP les plus difficiles sont situés là où une petite artère, proche de la limite de résolution, se branche sur une grande artère. Ceux-ci sont confondus avec de petits anévrismes (voir Fig. 3, en haut). En effet les performances de notre modèle sont moins bonnes sur les petits anévrismes. Sur les 155 anévrismes, 34 n'ont pas été détectés (FN). Mais 18 de ces FN avaient un diamètre inférieur à 2 mm, et 10 autres étaient inférieurs à 3 mm. La sensibilité de notre modèle est de 0.53 pour les anévrismes inférieurs à 2 mm, mais atteint 0.89 pour les plus gros anévrismes. Cette difficulté de détection des petits anévrismes en imagerie par résonance magnétique est bien documentée [2]. Notez que lors de l'examen visuel des résultats par un radiologue de 30 ans d'expérience, 8 FP se sont avérés être de vrais anévrismes, qui avaient été omis lors de la phase d'annotation initiale (voir Fig. 3, en bas).

Dans nos expériences, nous avons observé une augmentation soudaine du score κ qui était corrélée à une convergence satisfaisante de la phase d'apprentissage (voir Fig. 4). Nous l'avons interprétée comme une meilleure sensibilité de la métrique κ par rapport à Dice, même à de petites intersections entre la prédiction et la vérité terrain.

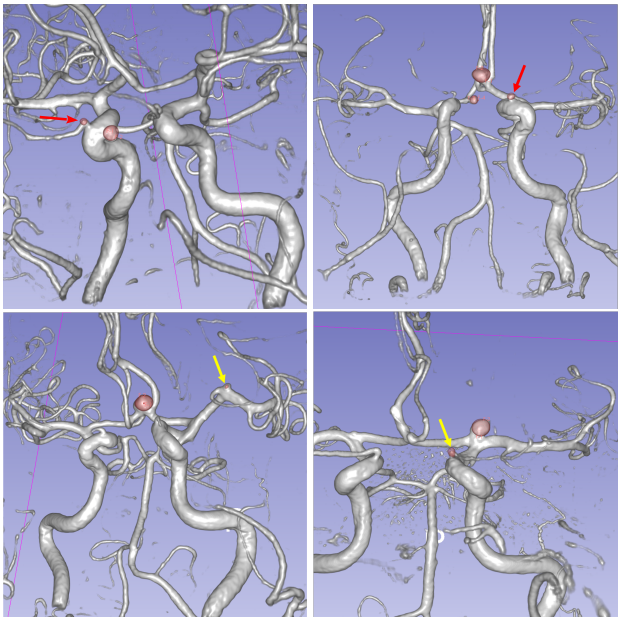


FIGURE 3 – (Haut) La ramification de petites artères peut être confondue avec un anévrisme (flèches rouges, les CC prédites sont en rouge, les points d'annotation sont présents). (Bas) Exemples d'anévrismes omis lors de la phase d'annotation initiale (flèches en jaune).

5 Conclusion

Dans cet article, nous avons présenté une stratégie d'échantillonnage de données efficace pour détecter les anévrismes intracrâniens à partir d'images IRM, qui est capable d'atteindre une sensibilité de 0.82 pour 0.61 FP/cas. Cette stratégie peut facilement être combinée avec des architectures plus sophistiquées et des fonctions de coût qui ont démontré leur efficacité, notamment à travers le challenge ADAM. Nous espérons réduire le nombre de FP pour concevoir un classificateur plus spécifique. Un travail en progression étudie notamment le score κ comme une fonction de coût, pour tirer parti de sa capacité à évaluer la qualité d'un classificateur malgré le déséquilibre de classes.

6 Remerciements

Nous remercions la région Grand Est et le centre hospitalier régional et universitaire (CHRU) de Nancy pour le financement de ce travail. Les expériences présentées dans cet article ont été réalisées en utilisant la plateforme expérimentale Grid'5000, issue de l'Action de Développement Technologique (ADT) Aladdin pour l'INRIA, avec le support du CNRS, de RENATER, de plusieurs Universités et autres contributeurs (https ://www.grid5000.fr).

Références

[1] Z. Shi, B. Hu, U.J. Schoepf, et al. Artificial intelligence in the management of intracranial aneurysms : current status and future perspectives. *American Journal of Neuroradiology*, 41(3) :373–379, 2020.

[2] M. Jang, J.H. Kim, J.W. Park, et al. Features of "false positive" unruptured intracranial aneurysms on screening magnetic resonance angiography. *PloS one*, 15(9) :e0238597, 2020.

[3] H. Arimura, Q. Li, Y. Korogi, et al. Computerized detection of intracranial aneurysms for three-dimensional MR angiography : Feature extraction of small protrusions based on a shape-based difference image technique. *Medical physics*, 33(2) :394–401, 2006.

[4] D. Ueda, A. Yamamoto, M. Nishimori, et al. Deep learning for MR angiography : automated detection of cerebral aneurysms. *Radiology*, 290(1) :187–194, 2018.

[5] J.N. Stember, P. Chang, D.M. Stember, et al. Convolutional neural networks for the detection and measurement of cerebral aneurysms on magnetic resonance

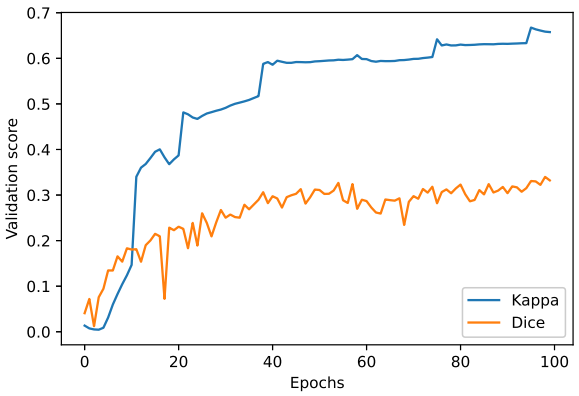


FIGURE 4 – L'augmentation typique du score κ est un bon prédicteur de la convergence finale (vers l'époque 10)

angiography. *Journal of digital imaging*, 32(5) :808–815, 2019.

[6] T. Nakao, S. Hanaoka, Y. Nomura, et al. Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. *Journal of Magnetic Resonance Imaging*, 47(4) :948–953, 2018.

[7] T Sichtermann, A Faron, R Sijben, et al. Deep learning–based detection of intracranial aneurysms in 3D TOF-MRA. *American Journal of Neuroradiology*, 40(1) :25–32, 2019.

[8] A. Faron, T. Sichtermann, N. Teichert, et al. Performance of a deep-learning neural network to detect intracranial aneurysms from 3D TOF-MRA compared to human readers. *Clinical neuroradiology*, 30(3) :591–598, 2020.

[9] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, et al. 3D U-net : learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention (MICCAI)*, pages 424–432, 2016.

[10] J. Ma and X. An. Loss ensembles for intracranial aneurysm segmentation : An embarrassingly simple method. In *Automatic Detection And SegMentation Challenge (ADAM)*, 2020.

[11] Y. Yang, Y. Lin, Y. Li, et al. Automatic aneurysm segmentation via 3D U-net ensemble. In *Automatic Detection And SegMentation Challenge (ADAM)*, 2020.

[12] M. Baumgartner, P.F. Jaeger, F. Isensee, et al. Retina U-net for aneurysm detection in MR images. In *Automatic Detection And SegMentation Challenge (ADAM)*, 2020.

[13] H. Yu, Y. Fan, and H. Shi. Team abc. In *Automatic Detection And SegMentation Challenge (ADAM)*, 2020.

[14] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, et al. 3D Slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9) :1323–1341, 2012. PMID : 22770690.

[15] M. Brett, A.P. Leff, C. Rorden, and J. Ashburner. Spatial normalization of brain images with focal lesions using cost function masking. *Neuroimage*, 14(2) :486–500, 2001.

[16] D.G. Ellis. 3D U-net convolution neural network with Keras, 2017.

[17] J.M. Johnson and T.M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1) :1–54, 2019.

[18] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1) :37–46, 1960.

[19] A.A. Taha and A. Hanbury. Metrics for evaluating 3D medical image segmentation : analysis, selection, and tool. *BMC medical imaging*, 15(1) :1–28, 2015.

[20] S. A. Taghanaki, K. Abhishek, J.P. Cohen, et al. Deep semantic segmentation of natural and medical images : a review. *Artificial Intelligence Review*, pages 1–42, 2020.

[21] T.-Y. Lin, P. Goyal, R. Girshick, et al. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2980–2988, 2017.