Évaluation et adaptation de réseaux VGG16 pour la détection de carcinome sur des lames d'histologie: de la résection à la biopsie

Qinghe Zeng, Nicolas Loménie, Christophe Klein, Julien Calderaro

Évaluation et adaptation de réseaux VGG16 pour la détection de carcinome sur des lames d'histologie : de la résection à la biopsie
Evaluation and adaptation of VGG16 networks for the detection of carcinoma on histology slides : from resection to biopsy

Qinghe Zeng[1,2] Christophe Klein[2] Julien Calderaro[3,4] Nicolas Loménie[1]

[1]Laboratoire d'informatique Paris Descartes (LIPADE), Université de Paris, Paris, France
[2]Centre d'Imagerie, Histologie et Cytométrie (CHIC), Centre de recherche des Cordeliers, Paris, France
[3]INSERM U955, Team "Pathophysiology and Therapy of Chronic Viral Hepatitis and Related Cancers", Créteil,
[4]APHP, Department of Pathology, Hôpital Henri Mondor, Université Paris-Est, Créteil, France

Résumé

La classification assistée par ordinateur des images histologiques est un sujet de plus en plus populaire pour les applications d'apprentissage profond. Les images histologiques sont multi-gigapixels, hiérarchiques et de dimensions non uniformes, alors que les images de biopsie sont de même nature mais contiennent une quantité très limitée de tissus, ce qui pose de nouveaux défis à la pathologie computationnelle. L'analyse de la biopsie est une tâche clinique difficile, mais elle joue parfois un rôle décisif dans le diagnostic et la sélection de la gestion du traitement. Dans cette étude, nous avons effectué une optimisation de l'architecture et des paramètres pour VGG16, et avons montré qu'ils avaient la capacité d'identifier les tumeurs sur les images histologiques de résection. Nous avons également démontré leur généralisabilité aux images histologiques des biopsies de carcinome hépatocelluclaire et proposé en outre une stratégie de chevauchement pour améliorer les résultats.

Mots Clef

Pathologie Computationnelle, Apprentissage Profond, Généralisation de Domaine.

Abstract

The computer-assisted classification of histological images is an increasingly popular subject for deep learning applications. Histological images are multi-gigapixels, hierarchical, and of non-uniform dimensions, whereas biopsy images are of the same nature but have very limited amount of tissue thus bringing new challenges to computational pathology. Biopsy analysis is a difficult clinical task, but sometimes it plays a decisive role in diagnosis and selection of treatment management. In this study, we performed architetcure and parameter optimization for VGG16, and showed that they had the ability to identify tumors on resection histological images. We also demonstrated their generalizability to histological images of Hepatocelluclar Carcinoma biopsies and further proposed an overlapping strategy to improve results.

Keywords

Computational Pathology, Deep Learning, Domain Generalization.

1 Introduction

Dans cet article, nous présentons une expérience complète d'adaptation des techniques de deep learning à un domaine émergent appelé histopathologue computationnelle et en particulier le traitement de biopsies qui sont scannées à très haute résolution sous forme d'une énorme image de plusieurs Go de données pixels pour un patient. L'examen au microscope de coupes de tissus colorées à l'hématoxyline et à l'éosine (H&E) par des pathologistes permet de diagnostiquer la grande majorité des maladies néoplasiques. L'avènement des techniques de numérisation de lames entières permet une analyse assistée par ordinateur qui peut transformer la procédure de qualitative à quantitative, et les lames histologiques numérisées, appelées Whole Slide Images (WSIs), sont plus faciles à conserver, à partager et à annoter. De plus, des études ont confirmé la haute concordance inter et intra-observateur dans la performance diagnostique de la WSI numérisée par rapport aux lames de verre au microscope optique [1, 2], quelle que soit l'origine du tissu : résection ou biopsie (voir FIGURE 1).
Nous travaillons sur le carcinome hépatollulaire (CHC) qui est le sixième cancer le plus fréquent dans le monde [3]. La résection chirurgicale (ablation de la plus grande partie de la tumeur) est l'une des options curatives de première ligne promettant d'allonger la durée de la rechute avec un taux de transformation maligne réduit. Bien que la résection chirurgicale minimise la probabilité d'un faux diagnostic (par exemple les erreurs d'échantillonnage) en fournissant une quantité adéquate de tissu pour l'analyse histopathologique, elle expose intrinsèquement le patient à des risques de morbidité et de mortalité plus élevés. Il est également

possible d'utiliser des techniques d'imagerie non invasives telles que l'échographie, la tomographie par ordinateur et l'imagerie par résonance magnétique. Cependant, avec ces techniques, les caractéristiques du CHC précoce peuvent être similaires à celles des nodules hyperplasiques non malins [5, 6, 7] et le diagnostic peut donc être ambigu. Par conséquent, les biopsies à l'aiguille sont généralement recommandées pour les nodules de plus de 1,0 cm [8] afin de limiter les résections non essentielles. Une confirmation par biopsie est également nécessaire pour évaluer la réponse au traitement dans les études de phase III des nouvelles thérapies ciblées [9]. Contrairement à la résection qui vise à enlever la tumeur, la biopsie extrait chirurgicalement une quantité limitée de tissu tumoral et, généralement, les échantillons de tissu de résection et de biopsie sont cliniquement disponibles. Pourtant, la question de la meilleure option de gestion est controversée.

Avec une longue histoire dans le domaine de l'oncologie, la biopsie de tumeurs devient maintenant une composante importante des techniques d'imagerie médicale et s'avère être un domaine d'application prometteur pour l'apprentissage profond [10, 11]. Récemment, de plus en plus d'études ont appliqué les réseaux convolutifs profonds à la classification et au grading des biopsies, mais principalement pour le cancer de la prostate [12, 13, 14, 15, 16]. Dans cet article, nous proposons une optimisation de l'architecture de VGG16 [17] pour la détection de tissus tumoraux sur notre cas spécifique de WSI de résection de CHC. L'objectif est d'évaluer l'adaptabilité et la généralisation des modèles appris sur les WSIs de résection pour la classification des WSIs de biopsie de CHC sur une très petite quantité de tissu. En plus de cette tâche de classification, nous avons également testé des modèles de segmentation classiques par patch sur des WSIs de biopsie, car ils peuvent aider à la compréhension du grading final par le pathologiste ce qui est promordial dans le domaine de la santé clinique.

2 Matériaux et méthodes

2.1 Données image

Des échantillons de tissus, des images et des annotations cliniques anonymes ont été récupérés dans les archives de pathologie de l'hôpital Henri-Mondor (Créteil, France). Les échantillons de tissus comprenaient 728 lames de verre provenant de 365 tumeurs de résection chirurgicale de CHC et 107 lames de verre provenant de biopsies de CHC. Tous les échantillons de tissus ont été colorés en H&E FFPE et numérisés par un scanner Hamamatsu NanoZoomer au format *ndpi*. Le plus haut grossissement des WSIs de résection est soit 20X (0,46 mpp) soit 40X (0,23 mpp), tandis que tous les WSIs de biopsie sont à 40X. Des ensembles de données ont été extraits à des grossissements réduits, qui seront décrits dans la sous-section suivante.

Toutes les lames de résection ont été grossièrement annotées par un pathologiste expert à l'aide de régions d'intérêt (ROI) polygonales associées à des étiquettes tumorales et non tumorales (FIGURE 1.a). Sept lames de biopsie de tumeur (21 biopsies) ont été finement annotées par le pathologiste avec des ROIs tumorales (Figure 1.b). On rappelle qu'une seule lame peut faire jusque 10Go de données et environ 100 000 par 70 000 pixels sur les scanners modernes.

2.2 Pré-traitement de l'image

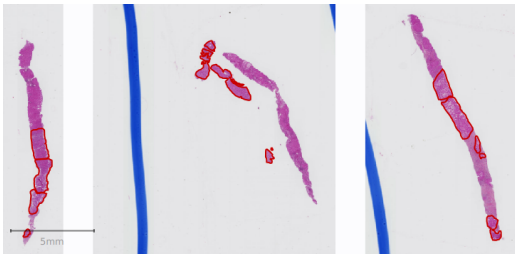
Pour la résection, 583 (80%) WSIs ont été utilisés pour les entraînements et pour les validations tandis que les 145 restants (20%) ont été utilisés comme données de test indépendantes. Le fractionnement a été effectué au niveau de la tumeur afin de s'assurer que les WSIs appartenant à la même tumeur se trouvent dans la même fraction. Les 21 biopsies annotées ont été retenues comme ensemble de validation externe pour tester la généralisation des modèles.

Pour l'ensemble de données rassemblés en un ensemble de patchs (ou imagettes), nous avons extrait la même taille physique de 117,76 μm^2 au grossissement 5X (1,84 mpp) et 10X (0,92 mpp). Pour l'ensemble de données de résection à 5X, dans Visiopharm [18] nous avons recadré les WSIs en patchs RVB au format TIFF de 512x512 pixels (en raison de la limitation de la taille d'exportation du logiciel) avec les masques correspondants. Les patchs ont été à nouveau divisés en images de 64x64 pixels et les patchs d'arrière-plan ont été éliminés. Une patch a été défini comme arrière-plan s'il y avait plus de 50 % de pixels dont la valeur d'intensité moyenne des RVB canaux était supérieure à 200 (critères de sélection des patchs de contour). Pour un ensemble de données de résection à 10X, nous avons recadré les WSIs en images de 1024x1024 pixels puis les avons découpés en image de 128x128 pixels.

Pour les ensembles d'entraînement et de validation, les patchs dont plus de 75% des pixels sont annotés comme étant des tumeurs dans leurs masques ont été classés dans la classe tumeur. Les patchs dont plus de 75% des pixels sont annotés comme non-tumoraux ont été classés dans la classe non-tumeur et les autres patchs ont été mis au rebut. De plus, nous avons utilisé un filtre laplacien et un seuil de somme de contours détectés à 20 comme détecteur de frontières pour éliminer les zones floues. Nous avons échantillonné de manière aléatoire 200 000 patchs avec deux classes équilibrées (tumeur vs non-tumeur), à partir des 583 WSI de résection, puis nous les avons divisés en ensembles complets d'entraînement (80%) et de validation (20%). Pour ce qui est de l'ensemble de test, un patch a reçu l'étiquette de tumeur s'il contenait plus de 50% de pixels de tumeur, sinon il était considéré comme une classe de non-tumeur. Enfin, pour les données de résection, nous avons échantillonné de manière aléatoire 200 000 patchs avec deux classes équilibrées (tumeur vs non-tumeur) pour l'ensemble d'entraînement et l'ensemble de test. Un ensemble de test complet de 200 000 patchs à classe équilibrée sans floues a été échantillonné à partir des 145 WSI de résection de test. Les ensembles d'entraînement, de validation et de test ont été préparés comme décrit pour la tâche de classification au grossissement 5X et 10X.

Pour la tâche de segmentation en analyse comparative,

a. resection WSI



b. biopsy WSI

FIGURE 1 – Exemples de régions tissulaires. Rouge : tumeur ; vert : non-tumeur. a. Une WSI de résection comportant 1 ou 2 sections de tissu. Une résection est définie comme l'ablation chirurgicale d'une partie d'un organe, en conservant les parties saines et en rétablissant, s'il y a lieu, leur continuité. b. Une WSI de biopsie contentant 1 ou plusieurs tissus avec des marqueurs bleus pour distinguer les tumeurs. On observe que la quantité de tissu est alors drastiquement réduite pour la classification, l'apprentissage ou le diagnostic.

nous avons échantillonné de manière aléatoire 10 000 patchs tumoraux et 5 000 patchs non tumoraux à partir de l'ensemble d'entraînement de résection 5X, ainsi que les masques correspondants lissés par un filtre médian. Les patchs mélangés avec leurs étiquettes (pour classification) ou leurs masques (pour segmentation) ont été introduits dans un modèle de réseau neuronal convolutif (CNN) pour entraîner et évaluer. Pour l'évaluation de la résection et des biopsies au niveau des lames, nous avons découpé les WSIs annotées en patches 64x64 à 5X en utilisant le paquet OpenSlide [19] en Python.

2.3 Architectures apprises

Nous avons testé le modèle VGG16 couramment utilisé en pathologie numérique et ses versions modifiées. Tous les modèles sont pré-entraînés sur ImageNet fourni par Keras [20]. VGG16 est composé de 5 blocs de convolution. Les 2 premiers blocs et les 3 derniers blocs sont respectivement composés de 2 et 3 couches de convolution répétées de 3x3 avec padding, puis d'une couche de pooling maximale de 2x2 pour le sous-échantillonnage.
Comme modèle de base (baseline) (FIGURE 2), après la couche Flatten, les trois couches de tête (top layers) du modèle VGG16 ont été remplacées par des couches denses à 4096 nœuds et 512 nœuds, chacune étant suivie d'une couche Batch-Normalization avec activation ReLu et d'une couche Dropout de 0,5. Enfin, une couche dense à 2 nœuds avec activation Softmax a été ajoutée pour la classification. Le modèle a été entraîné en minimisant la perte d'entropie croisée catégorielle (categorical cross-entropy fonctions de perte) à l'aide de l'optimiseur SGD avec une taille de batch de 128. Le taux d'apprentissage initial de $1e^{-3}$ a été réduit d'un facteur dix à chaque fois qu'il n'y avait pas d'amélioration supérieure à $2e^{-4}$ en précision de validation, sur 8 époques après 3 époques de refroidissement. L'option d'arrêt précoce (early stopping) a été activée pour arrêter l'entraînement lorsque aucune amélioration supérieure à $5e^{-4}$ en précision de validation n'était observée sur 15 époques. Après l'entraînement, les poids correspondant à la meilleure fonction de perte en validation ont été retournés.

Nous avons également instancié plusieurs versions de VGG16 pour étudier l'impact des 4 facteurs : backbone, tête, taille du batch et optimiseur.

Architecture de backbone convolutif [21]. Il s'agit de la partie principale du réseau VGG16, de la couche d'entrée à la dernière couche de convolution incluse. Il est utilisé pour extraire des caractéristiques sur une image entière. Nous avons conçu 8 versions : **VGG16.** Backbone par défaut fourni par Keras et **VGG16Base** comme implémentation du backbone VGG16 à partir duquel nous avons dérivé 5 nouvelles architectures : **VGG16DrV1.** Une couche Dropout de 0,1, 0,15, 0,2 et 0,25 a été ajoutée après chaque couche de convolution dans les blocs 1, 2, 3 et 4/5, respectivement. **VGG16DrV2.** Une couche Dropout de 0,1, 0,15 et 0,2 a été ajoutée après chaque couche de convolution dans les blocs 3, 4 et 5, respectivement. **VGG16DrV3.** Une couche Dropout de 0,1 a été ajoutée après chaque couche de pooling maximale dans les blocs 3 et 4. **VGG16BnV1.** Une couche Batch-Normalization a été ajoutée après chaque couche de pooling maximale dans les 4 premiers blocs. **VGG16BnV2.** Une couche Batch-Normalization a été ajoutée après chaque couche de convolution. Enfin, **VGG16DrBnV1,** basé sur VGG16DrV3, pour lequel une couche de Batch-Normalization a été ajoutée après chaque couche de convolution dans les 4 derniers blocs.

Architecture de tête [21]. Il s'agit des couches de tête du réseau VGG16, à l'exception du backbone. Elle est utilisée pour générer des probabilités pour chaque catégorie avec les caractéristiques extraites. Nous avons testé 3 types de têtes : **FC512.** Il s'agit de la tête du modèle de base décrit ci-dessus. **FC512L2.** Basé sur FC512, une régularisation L2 de $1e^{-5}$ a été activée pour le noyau des deux premières couches denses, ce qui devrait réduire le sur-apprentissage. **GAP.** Nous avons utilisé une couche de pooling moyen global (global average popling, GAP) [23] pour remplacer les couches denses et les autres couches, suivie uniquement d'une couche de classification avec activation Softmax. La

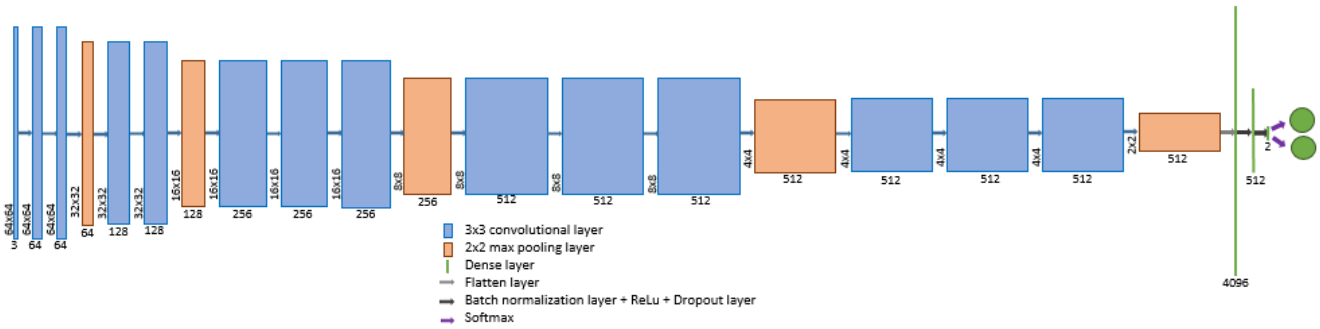


FIGURE 2 – Architecture du modèle de base

couche GAP calcule la valeur moyenne de chaque carte de caractéristiques produite par la couche précédente et sans paramètres entraînables, ce qui réduit considérablement la complexité du modèle.

La taille du batch. Nous avons testé 4 valeurs qui sont 32, 64, 128 et 256.

Optimiseurs. 2 optimiseurs ont été appliqués, à savoir SGD et Adam.

3 Expériences et résultats

3.1 Optimisation de l'architecture

Nous avons testé les instances VGG16 avec les combinaisons des 4 facteurs. Pour accélérer les expériences, les instances ont été entraînées sur un sous-ensemble de 50 000 patchs de résection à classe équilibrée sélectionnés au hasard avec 20% comme validation et testés sur l'ensemble de test complet. Pour tester la stabilité des performances, nous avons répété le processus d'entraîner 3 fois pour chaque instance. Les modèles ont été évalués par l'aire sous la courbe ROC (AUC). La meilleure combinaison a été observée comme étant le backbone VGG16DrV2, la tête Gap, la taille de batch de 32 et l'optimiseur SGD avec une AUC de test supérieure à 0,86 (FIGURE 3).

Backbones. Avec arrêt précoce, les backbones VGG16 et VGG16Base ont montré une tendance au sur-apprentissage en augmentant à la fois la précision et la fonction de perte en validation à la fin de leur entraînement. (FIGURE S1.1) Les probabilités en sortie étaient extrêmement élevées ou très faibles, ce qui a permis d'améliorer la précision dans une certaine mesure, alors que de nombreuses distances par rapport aux vérités terrain deviennent plus grandes et conduisent donc à des fonctions de pertes plus élevées. Les backbones avec des couches de Batch-Normalization (VGG16BnV1, VGG16BnV2 et VGG16DrBnV1) ont été observés comme ayant un problème de sur-apprentissage encore plus sévère (FIGURE S1.1) et ont eu de mauvaises performances en particulier VGG16BnV2 (FIGURE 3). **L'ajout de couches de dropout dans le backbone** a permis de surmonter ce problème et VGG16DrV2 a obtenu les meilleures performances, suivi de VGG16DrV3 (FIGURE 3). **Tête.** FC512 et FC512L2 ont été observés avec des performances médiocres tandis que GAP a obtenu les meilleures et les plus mauvaises performances(FIGURE 3). Les courbes de validation des modèles avec FC512L2 étaient les plus fluctuantes. (FIGURE S1.2) **Tailles des batchs.** Les modèles les plus performants ([0.86, 0.87]) ont été entraînés avec 32. Au contraire, les modèles les moins performants ont utilisé 128. **Optimiseurs.** Les meilleures performances ([0.86, 0.87]) ont été obtenues avec SGD. Adam a considérablement accéléré la convergence mais a un peu nui aux performances. L'optimiseur Adam a également causé un problème de sur-apprentissage, la fonction de perte en validation augmentant rapidement après le passage du meilleur point. (FIGURE S1.3).

3.2 Taille de l'ensemble d'entraînement

Nous avons entraîné le modèle de base avec différentes tailles d'ensemble d'entraînement, de 10 à 200 000, avec 2 classes équilibrées et 20% de split comme ensemble de validation. Après avoir répété le processus d'entraînement trois fois, les modèles ont été testés sur l'ensemble de test complet (FIGURE 4). Les performances du modèle ont été déstabilisées avec moins de 2 000 échantillons d'entraînement. Elle a grimpé à plus de 0,8 AUC à 5 000 échantillons et a continué à augmenter régulièrement. Elle s'est transformée en une tendance plafond à partir de 100 000 échantillons et a finalement atteint 0,904 avec 200 000 échantillons.

3.3 Résultats quantitatifs

Le modèle de base a été capable d'identifier les tissus à partir d'arrière-plan et de segmenter le contour tumoral, au moins à basse résolution. (FIGURE S2) Sur chaque lame de test, nous avons évalué le modèle de base en utilisant 8 métriques différentes (TABLE 1) pour les performances moyennes) avec un seuil optimal par lame calculé à partir des courbes ROC. L'AUC moyenne a atteint 0,934.Nous avons également calculé l'Intersection sur l'Union (Intersection over Union, IoU) au niveau du pixel à 5X, qui a atteint 0,762, même si les prédictions sont faites sur des patchs à gros grain de 64x64 pixels et que les annotations sont des ROI grossières sur la haute résolution. Même si nous avons utilisé le seuil optimal pour chaque lame, les

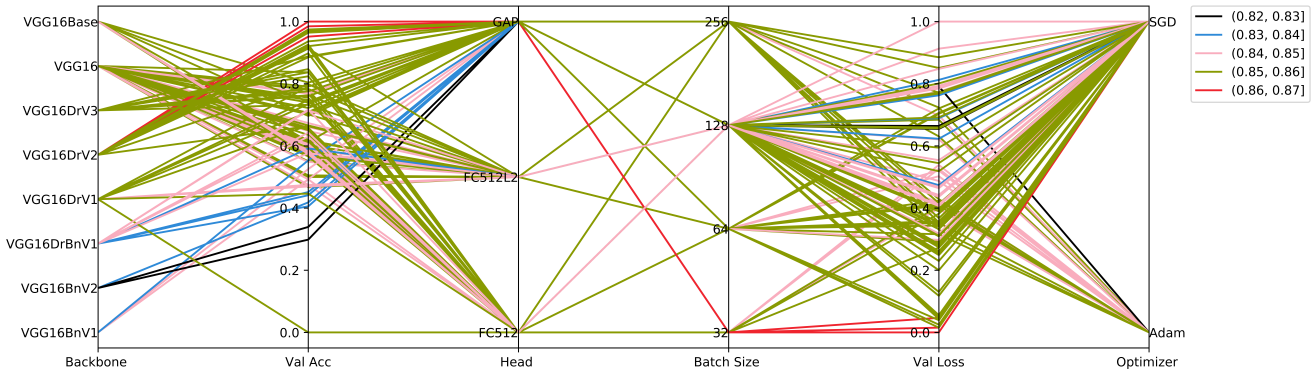


FIGURE 3 – Performance parallèle des instances avec la combinaison des 4 facteurs

métriques évaluées varient considérablement et de nombreuses valeurs aberrantes ont été détectées, impliquant l'existence de lames anormales. Nous avons également validé le modèle de base (TABLE 1) et la FIGURE 5 sur les lames d'entraînement pour vérifier la qualité de l'ensemble d'entraînement.

3.4 Généralisation aux biopsies

Au grossissement de 5X, le modèle de base a détecté les zones tumorales avec quelques faux positifs, notamment sur le bord du tissu (FIGURE 6.b). Cela est probablement dû à un problème de fixation, comme c'est le cas pour les lames de résection. Pour améliorer la précision de la prédiction, nous avons testé le modèle de base avec un overlap de 25% en horizontal et vertical, ce qui signifie que le pas d'échantillonnage était de 16 lors du découpage des patchs. Le contour tumoral prédit était plus précis, même s'il y a très peu de cellules dans certaines tumeurs (FIGURE 6.c). Le problème des faux positifs sur les contours a été résolu. Cependant, il y avait encore quelques faux positifs à faible probabilité qui peuvent être éliminés en fixant un seuil approprié pour la catégorie de tumeur. Nous avons également testé les pas de 8, 16, 32 et 48 et il n'y avait pas de différence significative dans la performance du modèle, sauf pour la résolution de la carte de probabilité.
Nous avons testé le modèle de base entraîné sur l'ensemble de données 10X avec chevauchement de 25% et les problèmes de faux positifs ont été considérablement améliorés par rapport au modèle 5X. Non seulement les faux positifs des petits tissus (par exemple, le tissu situé en bas à gauche dans la FIGURE 6.d), mais la robustesse à la couleur et à la texture de l'arrière-plan, ainsi qu'à la marque du marqueur, a été considérablement améliorée (FIGURE 6.d).

3.5 Analyse comparative et discussion

Nous avons testé les modèles U-Net [24] et Fully Convolutional Network (FCN) [25] entraînés sur les patchs de résection à 5X. Par rapport au modèle de base, les deux modèles de segmentation présentaient moins de vrais positifs et plus de faux positifs (FIGURE 6.e&f). Cela est prévisible car nos annotations ne sont pas au niveau du pixel et ne correspondent donc pas aux exigences d'une tâche de segmentation. De plus, les patchs d'entrée de petite taille et de faible résolution perdent une grande quantité d'informations lors du traitement dans les structures de réseaux profonds.
Notre approche a utilisé un modèle d'apprentissage profond modéré, facile à entraîner et à déployer (notre station de travail est équipée d'un GPU NVIDIA GTX 1080). Le sur-apprentissage est une stratégie simple qui peut être appliquée à n'importe quel modèle de classification. Sa nature non intensive en calcul lui permet d'être utilisée cliniquement, soit pour éliminer les données de mauvaise qualité, soit comme un filtrage préliminaire pour un nouveau balayage. Elle présente de bonnes performances de prédiction et peut être utilisée comme pré-entraînement pour des méthodes coûteuses ou des tâches complexes afin de se concentrer sur les régions suspectes, afin de réduire la charge de travail de calcul et la charge mémoire. Comme cliniquement les pathologistes n'ont pas besoin d'une analyse au niveau du pixel, le modèle peut également être utilisé pour une double inspection. Nous devons souligner ici que l'entraînement a été effectué sur l'ensemble de données de résection et la validation sur un ensemble de données complètement différent, à savoir les WSIs de biopsie actuellement en cours de consolidation, avec plus de lames annotées à venir.

4 Conclusion

Nous avons effectué divers tests sur l'optimisation de l'architecture et des paramètres de VGG16, et nous avons fourni les meilleurs paramètres qui sont VGG16DrV2 bqckbone et Gap head avec une taille de lot de 32 et un optimiseur SGD. Notre approche peut exploiter l'annotation polygonale grossière des ROIs (ce qui est souvent le cas si un médecin doit en annoter des centaines) et ne nécessite pas d'annotations au niveau des pixels. Notre approche est un pipeline exploratoire qui utilise un réseau d'apprentissage profond modéré et un ensemble de données simple, tumeur vs. non-tumeur, dans le cas des patients du CHC. Afin de l'appliquer directement sur les WSIs de biopsie,

Split	Sensitivity	Specificity	Precision	Fall-Out	AUC	précision	F1-Score	IoU
Test	0.858	0.913	0.869	0.087	**0.934**	0.881	0.854	0.762
Train	0.887	0.9213	0.886	0.079	**0.956**	0.900	0.878	0.797

TABLE 1 – Performance moyenne des WSIs du modèle de base avec les seuils optimaux par WSI sur les WSI de résection de test et d'entraînement. Le seuil optimal a été calculé par la courbe AUC de tous les patchs extraits d'un WSI.

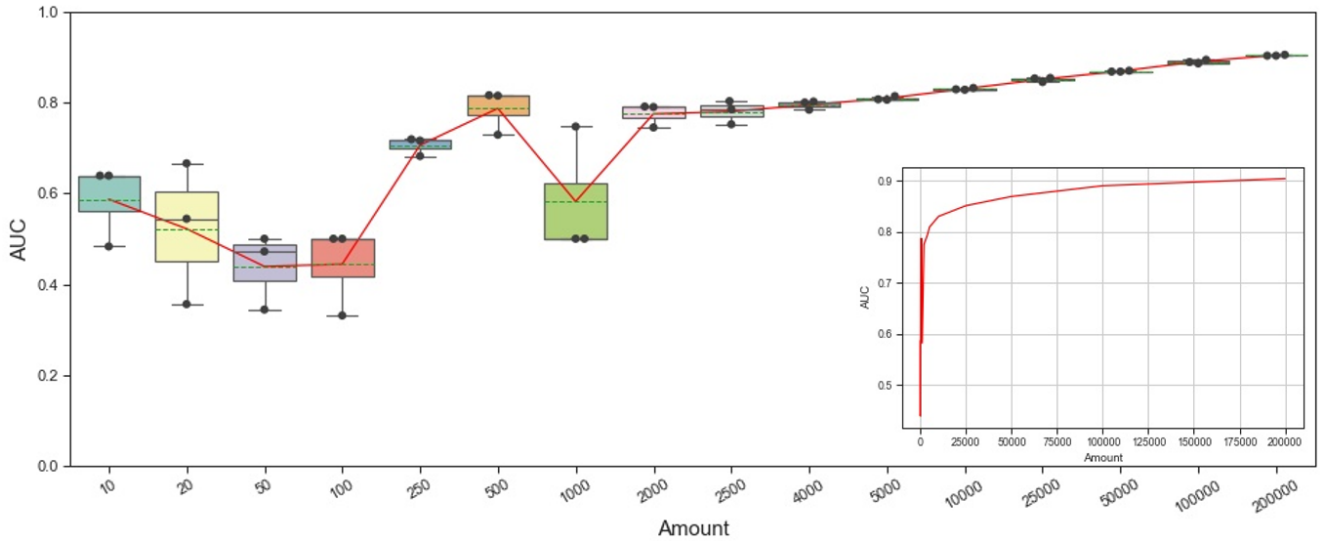


FIGURE 4 – AUC sur l'ensemble de test complet, avec différentes tailles d'ensembles de données (entraînement 80% et validation 20%). Les valeurs de l'axe x correspondent à la taille de l'ensemble de données testé. La courbe AUC avec un espacement uniforme entre les points de l'axe x est représentée dans le coin inférieur droit.

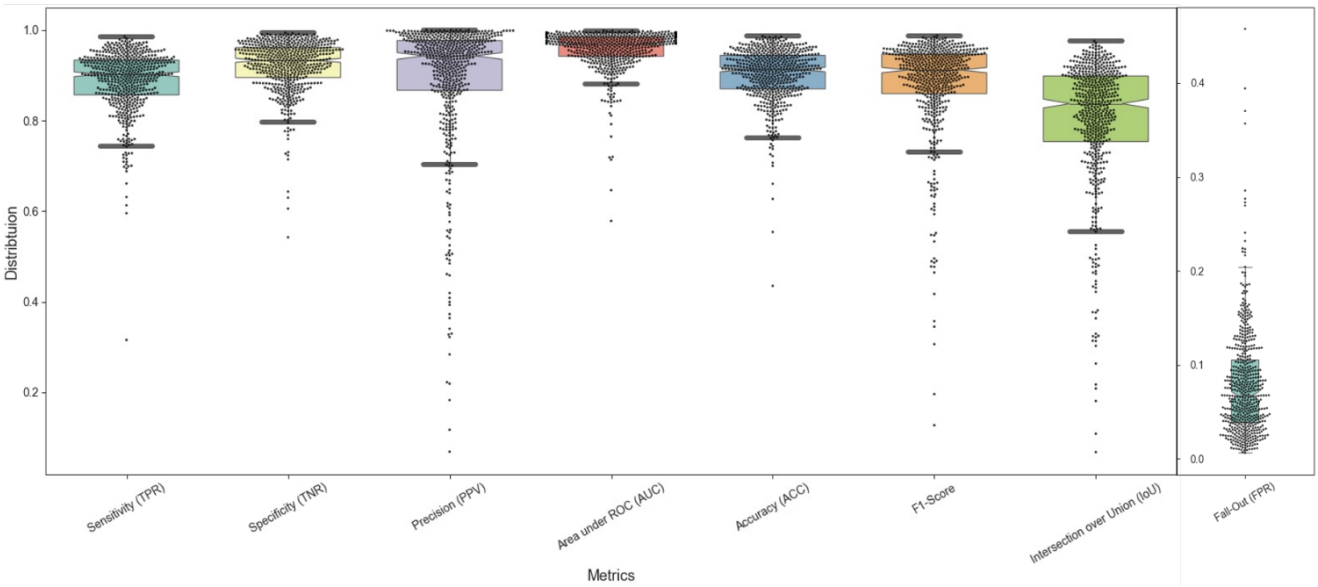
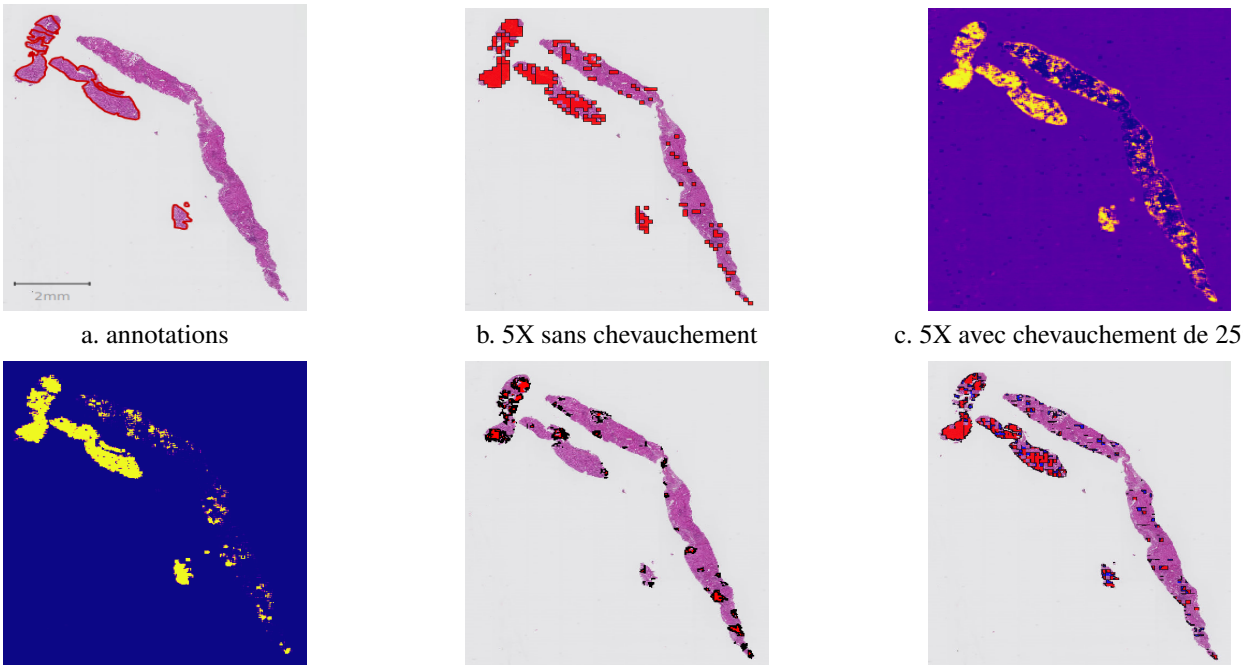


FIGURE 5 – Distribution des valeurs métriques pour les WSIs d'entraînement en utilisant le modèle de base avec les seuils optimaux par WSI.

nous avons proposé un pipeline pour généraliser les modèles basés par patchs entraînes sur les WSI de résection en utilisant une stratégie de chevauchement pour améliorer la précision de prédiction limitée à la petite quantité de tissus cancéreux. Dans le cadre de travaux ultérieurs, nous visons à discriminer les tumeurs malignes et bénignes et les sous-



a. annotations b. 5X sans chevauchement c. 5X avec chevauchement de 25%

d. 10X avec chevauchement de 25% e. U-Net f. FCN

FIGURE 6 – Vérité terrain et résultats d'un exemple de biopsie

types sur les biopsies, voire à prédire les classes immunitaires génétiques à partir de ces biopsies WSI. Nous appliquerons ce pipeline à l'apprentissage par transfert pour divers cancers et types de tissus.

Disponibilité du code

Tous les codes sources sont disponibles sur `https://github.com/qinghezeng/DetectHCC`.

Remerciements

Les auteurs tiennent à exprimer leurs remerciements à l'hôpital Henri-Mondor pour leurs supports de données. Q.Z est financé par le China Scholarship Council (CSC) du ministère de l'Éducation de la République populaire de Chine.

Références

[1] Araújo, A.L.D., Arboleda, L.P.A., Palmier, N.R., Fonsêca, J.M., Pauli Paglioni, M. de, Gomes-Silva, W., et al. : The performance of digital microscopy for primary diagnosis in human pathology : a systematic review. Virchows Archiv **474**(3), 269–287 (2019)

[2] Mukhopadhyay, S., Feldman, M.D., Abels, E., Ashfaq, R., Beltaifa, S., Cacciabeve, N.G., et al : Whole slide imaging versus microscopy for primary diagnosis in surgical pathology : a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). The American journal of surgical pathology **42**(1), 39 (2018)

[3] Berretta M., Cavaliere C., Alessandrini L., Stanzione B., Facchini G., Balestreri L., Perin T., Canzonieri V. : Serum and tissue markers in hepatocellular carcinoma and cholangiocarcinoma : clinical and prognostic implications. Oncotarget **8**(8), 14192 (2017)

[4] Lindman, B. R., Clavel, M. A., Mathieu, P., Iung, B., Lancellotti, P., Otto, C. M., Pibarot, P. : Calcific aortic stenosis. Nature reviews Disease primers **2**(1), 1–28 (2016)

[5] CADRANEL, J. F., Buffet, C., Cauquil, P., Ink, O., Pariente, D. : Nodules pseudo-tumoraux du foie chez le cirrhotique : étude de 7 cas. Gastroentérologie clinique et biologique **12**(11), 833–840 (1988)

[6] Miller, W. J., Baron, R. L., Dodd 3rd, G. D., Federle, M. P. : Malignancies in patients with cirrhosis : CT sensitivity and specificity in 200 consecutive transplant patients. Radiology **193**(3), 645–650 (1994)

[7] Torzilli, G., Minagawa, M., Takayama, T., Inoue, K., Hui, A. M., Kubota, K., et al : Accurate preoperative evaluation of liver mass lesions without fine-needle biopsy. Hepatology **30**(4), 889–893 (1999)

[8] Omata, M., Cheng, A. L., Kokudo, N., Kudo, M., Lee, J. M., Jia, J., et al : Asia–Pacific clinical practice guidelines on the management of hepatocellular carcinoma : a 2017 update. Hepatology international **11**(4), 317–370 (2017)

[9] Duffy, A. G., Ulahannan, S. V., Makorova-Rusher, O., Rahma, O., Wedemeyer, H., Pratt, D., et al : Tremelimumab in combination with ablation in patients with

advanced hepatocellular carcinoma. Journal of hepatology **66**(3), 545–551 (2017)

[10] Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., et al : Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports, **6**(1), 1-11 (2016)

[11] Janowczyk, Andrew, Anant Madabhushi : Deep learning for digital pathology image analysis : A comprehensive tutorial with selected use cases. Journal of pathology informatics **7** (2016).

[12] Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D. M., et al : Artificial intelligence for diagnosis and grading of prostate cancer in biopsies : a population-based, diagnostic study. The Lancet Oncology **21**(2), 222–232 (2020)

[13] Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., et al. : Automated deep-learning system for Gleason grading of prostate cancer using biopsies : a diagnostic study. The Lancet Oncology **21**(2), 233–241 (2020)

[14] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., et al. : Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)

[15] Lucas, M., Jansen, I., Savci-Heijink, C. D., Meijer, S. L., de Boer, O. J., van Leeuwen, T. G., et al. : Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. Virchows Archiv **475**(1), 77–83 (2019)

[16] Raciti, P., Sue, J., Ceballos, R., Godrich, R., Kunz, J. D., Kapur, S., et al. : Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. Modern Pathology **33**(10), 2058–2066 (2020)

[17] Simonyan, K., Zisserman, A. : Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv :1409.1556 (2014)

[18] Visiopharm Homepage, `https://visiopharm.com/`. Last accessed 4 Mar 2021

[19] Goode, A., Gilbert, B., Harkes, J., Jukic, D., Satyanarayanan, M. : OpenSlide : A vendor-neutral software foundation for digital pathology. Journal of pathology informatics **4** (2013)

[20] Visiopharm Homepage, `https://keras.io/`. Last accessed 4 Mar 2021

[21] K. He, G. Gkioxari, P. Dollár and R. Girshick : Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 386–397 (2020)

[22] Cortes, C., Mohri, M., Rostamizadeh, A. : L2 regularization for learning kernels. arXiv preprint arXiv :1205.2653 (2012)

[23] Lin, M., Chen, Q., Yan, S. : Network in network. arXiv preprint arXiv :1312.4400 (2013)

[24] Ronneberger, O., Fischer, P., Brox, T. : U-net : Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer, Cham (2015). https ://doi.org/10.1007/978-3-319-24574-4_28

[25] Long, J., Shelhamer, E., Darrell, T. : Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440 (2015). https ://doi.org/CVPR.2015.7298965.