PathologistAI

Détection automatique de métastases

Cancer du sein - Images histopathologiques

Projet CAMELYON17

Yassin CHAAIRATE • Roa CHAR • Jamin CHOQUER • Alexis FAYAN

Janvier 2026

Contexte médical et enjeux cliniques

Challenge CAMELYON17

Détection automatique de métastases dans les ganglions lymphatiques de patientes atteintes d'un cancer du sein

Enjeux cliniques majeurs

- Coût élevé des faux négatifs
- Variabilité inter-observateur
- Tâche chronophage pour les pathologistes
- Impact direct sur le protocole thérapeutique

Objectif 1

Identifier automatiquement la présence et l'étendue des métastases dans les lames histologiques

Objectif 2

Évaluer les performances au niveau patch et au niveau patient pN (classification TNM indiquant le nombre de ganglions affectés)

Données : Whole Slide Images (WSI)

Caractéristiques des WSI

- Images gigapixels (plusieurs Go par lame)
- Coloration H&E (Hématoxyline & Éosine)
- Structure pyramidale multi-résolution
- 5 hôpitaux différents → variabilité

Découpage en patchs

- Taille standard : 224 × 224 pixels
- Filtrage intelligent du tissu
- Labels binaires : normal (0) vs tumoral (1)
- Annotations XML converties en polygones

Défi majeur : Déséquilibre extrême des classes

~0.8% de patchs tumoraux en validation • Beaucoup plus de tissu sain que de zones métastatiques • Risque de biais du modèle vers la classe majoritaire

Pipeline de prétraitement et filtrage

Fonction is_tissue_patch : validation rigoureuse de la qualité

1. Détection du tissu

Ratio tissu > 40%

Élimine le fond de la lame et les zones sans matériel biologique pertinent

2. Structure locale

Variance gradient > 5.0

Assure la présence de détails morphologiques (membranes, noyaux) essentiels à l'apprentissage

3. Validation colorimétrique

Variance intensité > 25.0

Écarte les zones floues, les artefacts de numérisation et les régions non informatives

Architecture cloud : résolution des contraintes

Problématiques initiales

- Volume colossal : ~1 To de données WSI
- Absence de puissance de calcul (téléchargement impossible)
- Capacités locales insuffisantes pour l'entraînement

Solution : Google Cloud Platform

- VM 1 (Préprocessing) : 16 CPU haute performance, 64 Go RAM, SSD 250 Go
- VM 2 (Deep Learning) : 16 CPU, 1 GPU NVIDIA L4 pour entraînement
- Traitement par batchs : 5-6 patients × 120 Go à la fois
- Parallélisation : 3 patients // × 5 workers par patient

Maîtrise des coûts : 40€ pour 4 jours d'utilisation continue

Optimisation par séparation CPU/GPU • Activation ciblée des ressources • Calcul du temps par batch (sec_per_batch) pour extrapolation

Workflow reproductible

Environnements uv • Fichiers de mapping (tools/) • Pipeline unifié train/test • Traçabilité complète

Stratégie de sélection des données d'entraînement

Logique d'entonnoir rigoureux : de 100 patients à 34 patients d'entraînement

100 patients initiaux (5 centres)

68 patients (exclusion pN0(i+), pN1mi)

47 patients annotés exploitables

34 patients finaux (4 centres)

Critères de sélection

- Exclusion stades ambigus : pN0(i+), pN1mi (faible signal)
- Conservation pN1 et pN2 uniquement (signal clair)
- Limite technique : seuls 23 patients positifs annotés
- Équilibrage par centre : max 5 sains + 5 positifs
- Centre 3 exclu → jeu Out-Of-Distribution (OOD)

Résultat : 34 patients (17 sains + 17 positifs)
• Centres 0, 1, 2, 4

Modélisation : CNN et gestion du déséquilibre

Architecture : ResNet-18

- Transfer Learning depuis ImageNet
- Adaptation pour classification binaire
- Input : patchs 224×224 RGB
- Entraînement sur GPU L4 (PyTorch-CUDA)

Stratégie d'entraînement

- 5 époques avec optimiseur AdamW
- Fonction de perte : CrossEntropyLoss
- Split stratifié au niveau patient (StratifiedShuffleSplit)
- OverSampling + Data augmentation

Gestion du déséquilibre extrême des classes

Solution : Échantillonnage équilibré avec pondération

Utilisation d'un système de poids pour forcer le modèle à rencontrer autant d'exemples positifs (tumoraux) que négatifs (sains) durant chaque phase d'apprentissage. Cette stratégie oblige le réseau à se concentrer sur les caractéristiques visuelles discriminantes de la tumeur plutôt que sur la fréquence statistique des classes.

Agrégation patch → patient et prédiction pN

Du signal local (patch) à la décision clinique globale (stade pN)

Méthodes d'agrégation explorées

Top-K avec seuil élevé

La prédiction au niveau patient repose sur une agrégation **top-k (k = 10)** des patchs les plus confiants, avec un **seuil strict à 0,988**, afin de ne conserver que les signaux locaux les plus robustes, dans une logique exploratoire et non clinique.

Proportion tumorale

Mesure intuitive : % de patchs suspects par patient. Lisibilité clinique mais très sensible au déséquilibre

Surface tumorale

Approche morphologique reproduisant le raisonnement du pathologiste basé sur l'étendue de l'atteinte

Résultats et performances

Performances au niveau patch

Entraînement : ROC-AUC ≈ 0.99

Le modèle capture parfaitement les signaux morphologiques des cellules cancéreuses sur les données vues pendant l'apprentissage

Validation : ROC-AUC = 0.83

Décrochage important révélant les difficultés de généralisation inter-patients et inter-centres

Performances au niveau patient

Points positifs

- Sensibilité ≈ 100% (aucune métastase manquée)
- THR_PATCH = 0.9 isole les signaux robustes

Limites majeures

- Spécificité faible → nombreux faux positifs
- 0.8% patchs tumoraux → sensibilité au bruit

Domain Shift : défi majeur de généralisation

Performances sur le Centre 3 (Out-Of-Distribution)

Performances proches de l'aléatoire sur le centre jamais vu

Le modèle ne généralise pas ses acquis aux nouveaux hôpitaux, révélant une vulnérabilité critique au changement de domaine

Facteurs explicatifs identifiés

1. Biais technologique

Le modèle a appris à reconnaître des spécificités techniques propres aux centres d'entraînement (protocoles de coloration H&E, types de scanners, artefacts spécifiques) plutôt que des signatures tumorales universelles et biologiquement pertinentes

2. Amplification des erreurs

Dans un contexte de rareté extrême (~0.8% de patchs tumoraux), le moindre biais de domaine génère des faux positifs locaux. Ces erreurs s'accumulent lors de l'agrégation, transformant des patients sains en faux positifs cliniques

Interprétabilité et IA responsable

Choix méthodologique : interprétabilité au niveau agrégation, pas patch

Notre approche

- Focus sur les mécanismes d'agrégation patch → patient
- Test de différents seuils de décision (THR_PATCH)
- Analyse de la proportion de patchs tumoraux
- Estimation de la surface tumorale (mimant le pathologiste)

Justification

Une prédiction patch isolée n'a que peu de valeur clinique si elle n'est pas contextualisée. Notre démarche d'interprétabilité s'est déplacée vers l'analyse de comment la somme des signaux locaux est synthétisée pour aboutir à une décision globale exploitable médicalement.

Enseignements de l'analyse d'interprétabilité

- Révélation du Domain Shift : divergence flagrante entraînement (ROC-AUC 0.99) vs validation (ROC-AUC 0.83)
- Identification du paradoxe Sensibilité/Spécificité : 100% sensibilité mais spécificité faible → trop de fausses alertes
- Mise en évidence que les erreurs patch-level s'accumulent au lieu de s'annuler lors de l'agrégation

Limites structurelles et obstacles au déploiement

Paradoxe Sensibilité / Spécificité

Sensibilité ≈ 100% (crucial) mais spécificité faible → charge de travail excessive en fausses alertes pour les pathologistes

Limite de la cohorte patient

Centaines de milliers de patchs mais seulement 34 patients → diversité biologique trop faible → sur-apprentissage implicite

Fiabilité de l'agrégation

Les erreurs patch-level s'additionnent au lieu de s'annuler. Un faible taux d'erreur local suffit à biaiser totalement le diagnostic patient

Domain Shift critique

Performances quasi-aléatoires sur Centre 3 (OOD) → modèle non déployable en l'état sur de nouveaux hôpitaux

Notre démarche d'IA responsable : documenter explicitement ces limites plutôt que les masquer

Perspectives d'amélioration

Évolutions architecturales

- Apprentissage Multi-Instance (MIL) : pondération automatique des zones informatives, apprentissage direct au niveau patient
- Mécanismes d'attention : visualisation des régions influençant le diagnostic (interprétabilité clinique)
- Calibration clinique : c'est la méthode standard en histopathologie pour produire un score interpretable cliniquement et pas juste mathématiquement

Enrichissement des données

- Augmentation de la cohorte patient : réduire le sur-apprentissage, améliorer la diversité biologique
- Normalisation de couleur (stain normalization) : harmoniser les différences entre hôpitaux, réduire le Domain Shift
- Data augmentation avancée : rotations, variations de contraste, flou → focalisation sur les formes cellulaires

Collaboration avec pathologistes pour calibration des scores et expression de l'incertitude du modèle

Enseignements clés du projet

1 **Rigueur méthodologique > Performance brute**

Le split stratifié au niveau patient (pas patch) est crucial pour éviter la fuite de données et garantir la crédibilité des évaluations

2 **Gestion stricte de la variabilité**

L'équilibre inter-centres et la gestion du déséquilibre des classes ne sont pas techniques mais éthiques et nécessaires

3 **Infrastructure = fondation du succès**

Sans le cloud (GCP, GPU L4, traitement par batchs), ce projet aurait été impossible. La maîtrise technique est indissociable de l'expertise en deep learning

4 **Performance ≠ Déployabilité clinique**

Un modèle peut être techniquement performant au niveau patch tout en restant cliniquement risqué au niveau patient (Domain Shift, faux positifs)

La qualité des données d'entraînement conditionne directement la validation et le test

Conclusion

Réalisations

- Pipeline complet end-to-end sur données médicales réelles et massives
- Infrastructure cloud optimisée (40€, 4 jours) pour traitement de 1 To de données
- Détection efficace au niveau patch (ROC-AUC 0.99 en entraînement)
- Analyse rigoureuse des limites via approche d'IA responsable

La performance ne dépend pas uniquement de l'architecture, mais avant tout de la qualité, quantité et diversité des données d'entraînement

Base expérimentale solide nécessitant jeux de données plus larges et approches avancées (MIL, attention, calibration clinique) pour transférabilité vers la pratique médicale

Merci pour votre attention

Questions ?