Projet 2

Détection automatique de métastases du cancer du sein dans des images histopathologiques

Rapport de Synthèse

Auteurs :

Yassin CHAAIRATE • Roa CHAR • Jamin CHOQUER • Alexis FAYAN

Janvier 2026

1. Introduction générale

1.1 Contexte du projet CAMELYON17

Le projet s'appuie sur le jeu de données CAMELYON17, issu d'un concours international dont la vocation est de stimuler l'innovation dans le domaine de la pathologie numérique. Ce challenge se concentre spécifiquement sur le développement de méthodes automatiques capables d'identifier des métastases dans les ganglions lymphatiques chez des patientes atteintes d'un cancer du sein.

La démarche associée à ce projet vise à concevoir une chaîne d'analyse d'images médicales réelles dans sa globalité, depuis la manipulation de la donnée brute jusqu'à la prédiction finale du diagnostic. Il ne s'agit pas uniquement de l'optimisation d'un algorithme isolé, mais de la construction d'un système robuste d'aide au diagnostic médical.

En mobilisant des notions avancées de traitement de l'information, ce projet place l'analyse de données réelles au centre d'un processus de décision médicale critique. L'objectif est de répondre aux exigences de la recherche clinique en proposant une solution automatisée capable d'assister les pathologistes dans une tâche complexe et chronophage.

1.2 Objectifs médicaux et enjeux cliniques

L'enjeu médical de ce projet est double et s'articule autour de la précision du diagnostic oncologique. Le premier objectif est d'identifier de manière automatisée la présence de cellules métastatiques au sein des ganglions lymphatiques. Cette détection est une étape préalable indispensable à la seconde mission du système : la prédiction du stade pN (Pathological Nodal stage) pour chaque patiente.

Le stade pN est une composante essentielle de la classification TNM, utilisée mondialement pour décrire l'extension du cancer. Il permet de quantifier l'atteinte ganglionnaire, allant d'une absence de métastase (pN0) à une dissémination plus importante (pN1, pN2, pN3). Cette information est déterminante pour l'équipe médicale car elle oriente directement le protocole thérapeutique, qu'il s'agisse de la nécessité d'une chimiothérapie, d'une radiothérapie ou d'un ajustement de la stratégie chirurgicale.

Les enjeux cliniques associés sont majeurs. En pathologie traditionnelle, l'examen manuel des lames est une tâche extrêmement chronophage et sujette à une certaine variabilité inter-observateur. Le risque principal réside dans les faux négatifs : l'omission d'une micro-métastase peut conduire à un sous-stadiage de la maladie, privant ainsi la patiente d'un traitement pourtant nécessaire. En proposant un système d'aide au diagnostic, l'objectif est d'accroître la fiabilité des résultats, de standardiser l'évaluation et de sécuriser la décision thérapeutique.

1.3 Vision du système

La philosophie de ce projet repose sur une approche intégrée. Il ne s'agit pas uniquement de développer un algorithme de prédiction, mais de concevoir un véritable système d'aide au diagnostic médical. Cette vision impose de traiter le problème sous un angle de bout en bout, où chaque brique technologique, de la gestion massive des données à la production du résultat clinique, est interconnectée et justifiée par sa finalité médicale.

Le système est envisagé comme un support à l'expertise humaine, visant à augmenter la capacité d'analyse du pathologiste face à des volumes de données gigapixéliques. En automatisant les tâches de détection les plus répétitives et en synthétisant l'information à l'échelle du patient, ce dispositif aspire à réduire la charge cognitive des praticiens tout en minimisant les risques d'erreur de diagnostic. Cette perspective systémique place la fiabilité, l'efficacité opérationnelle et l'intégration des contraintes réelles, qu'elles soient techniques, budgétaires ou cliniques, au cœur de la démarche de conception.

1.4 Organisation du rapport

Le présent rapport est structuré en plusieurs sections permettant de retracer la conception du système dans sa globalité. La deuxième partie est consacrée à la **compréhension et à la gestion des données**, où sont abordés les **enjeux cliniques**ainsi que les **contraintes techniques** liées aux **images de très haute résolution**. La troisième partie détaille l'**architecture cloud AWS**, véritable pilier du projet, qui a permis de lever les **verrous technologiques** liés au **stockage** et à la **puissance de calcul**.

La quatrième partie expose la **méthodologie de modélisation** par **réseaux de neurones convolutifs** et les **stratégies d'agrégation** nécessaires pour passer d'une analyse locale à un **diagnostic à l'échelle du patient**. Enfin, la cinquième partie traite des questions d'**interprétabilité**, d'**éthique** et de **robustesse** du système, avant de conclure sur les enseignements tirés et les **perspectives d'évolution** du projet.

---

2. Compréhension et Gestion des Données

2.1 Compréhension du contexte clinique

La transition d'une analyse locale à une conclusion clinique globale constitue le défi majeur de ce projet. En pathologie numérique, l'examen ne s'arrête pas à l'identification d'une cellule anormale sur une portion d'image. Le diagnostic repose sur la capacité à agréger des informations provenant de plusieurs ganglions pour déterminer le stade pN au niveau du patient.

Cette classification pN est le pivot du pronostic médical. Elle nécessite de distinguer avec précision les métastases, les micro-métastases et les cellules tumorales isolées. La compréhension de ce contexte clinique est indispensable pour calibrer le système, car une erreur d'interprétation à l'échelle d'un seul ganglion peut modifier radicalement le stade pN final et, par extension, le traitement administré à la patiente. Le système doit donc être conçu pour garantir une continuité parfaite entre l'analyse microscopique et la décision macroscopique.

2.2 Analyse des données d'imagerie médicale

Le projet repose sur l'exploitation des Whole Slide Images (WSI), qui constituent le standard actuel de la pathologie numérique. Contrairement aux images médicales traditionnelles, une WSI est une numérisation exhaustive d'une lame de tissu, capturée à un grossissement très élevé (généralement 20x ou 40x).

La principale caractéristique de ces données est leur volume exceptionnel : une seule image peut contenir plusieurs milliards de pixels et peser plusieurs gigaoctets. Cette dimension gigapixel interdit toute manipulation classique en mémoire vive et impose une structure de stockage spécifique. Les images sont ainsi organisées de manière pyramidale, contenant au sein d'un même fichier plusieurs niveaux de résolution. Cette architecture permet de naviguer de la vue d'ensemble (basse résolution) à l'observation fine des structures cellulaires (haute résolution), une propriété que nous exploitons pour optimiser l'accès aux données lors des phases de traitement.

2.3 Nature technique

La manipulation de ces images gigapixels repose sur l'exploitation de la coloration Hématoxyline et Éosine (H&E), qui permet de distinguer les structures biologiques par contraste colorimétrique. Sur le plan informatique, l'accès à ces données est réalisé via la bibliothèque OpenSlide, indispensable pour extraire des régions spécifiques sans saturer la mémoire vive du système.

L'aspect critique de cette phase réside dans la validation de la qualité des données extraites à travers une logique de filtrage rigoureuse implémentée dans la fonction `is_tissue_patch`. Pour garantir que seules les zones pertinentes sont conservées, le système applique des seuils stricts :

- **Détection du matériel biologique** : un calcul de ratio exclut les patchs dont la surface de tissu est inférieure à 40%, éliminant ainsi le fond de la lame.
- **Analyse de la structure locale** : une mesure de la variance du gradient (seuil à 5.0) assure la présence de détails morphologiques, tels que des membranes ou des noyaux, indispensables à l'apprentissage.
- **Validation colorimétrique** : le contrôle de la saturation et de la variance d'intensité (seuil à 25.0) permet d'écarter les zones floues, les artefacts de numérisation ou les régions vitreuses non informatives.

Parallèlement, la gestion des vérités terrain (ground truth) nécessite une précision géométrique élevée. Les annotations tumorales, fournies au format XML, sont converties en objets géométriques grâce à la bibliothèque Shapely. Cette approche permet de transformer des coordonnées textuelles en polygones manipulables, facilitant ainsi la labellisation automatique des zones extraites en fonction de leur intersection avec les zones métastatiques identifiées par les experts.

2.4 Justification du découpage

Le découpage des données en unités de traitement plus petites est une nécessité imposée par la nature même des images WSI. Étant donné qu'une seule lame peut contenir des milliards de pixels, il est techniquement impossible d'injecter l'image complète dans un réseau de neurones. Le système procède donc à une segmentation en patchs de taille fixe,

fixée ici à **224x224 pixels**, une dimension standard **optimisée** pour les architectures de Deep Learning.

Ce choix de découpage répond à trois objectifs fondamentaux :

- **Gestion de la mémoire** : l'extraction de patchs localisés permet un entraînement du modèle sur des échantillons de taille constante et gérable par les processeurs graphiques.
- **Précision de la détection** : le découpage permet d'isoler des micro-structures cellulaires, facilitant ainsi l'identification de micro-métastases qui seraient noyées dans une analyse globale de la lame.
- **Échantillonnage statistique** : en découpant la lame au niveau de résolution 2, le système génère un volume important d'exemples qui serviront de base à l'apprentissage.

Ce processus est piloté par le script `run_batch_pipeline.py`, qui organise le traitement par lots de patients. Pour chaque patient, le système parcourt les cinq ganglions associés et extrait systématiquement les zones de tissu valides. Cette approche granulaire est l'unique moyen de transformer une donnée biologique continue et massive en une série d'exemples discrets exploitables pour une classification binaire entre tissu sain et tissu métastatique.

AVANT / APRES SUR IMAGE

2.5 Analyse Exploratoire des Données (AED)

L'analyse des données extraites révèle une caractéristique fondamentale du jeu de données CAMELYON17 : un déséquilibre de classes extrêmement marqué. Dans le cadre d'un diagnostic oncologique, les zones saines, correspondant aux patchs normaux, représentent l'immense majorité de la surface tissulaire d'un ganglion, tandis que les foyers métastatiques sont souvent localisés et de taille réduite.

Ce déséquilibre se manifeste à deux niveaux. D'une part, au sein d'une lame dite positive, le ratio entre patchs tumoraux et patchs sains est souvent très faible, car la tumeur ne colonise qu'une fraction du ganglion. D'autre part, de nombreux patients ne présentent aucune métastase sur l'ensemble de leurs cinq ganglions, ce qui accentue la prédominance de la classe normale dans le volume total de données récoltées. Pour le modèle d'intelligence artificielle, cette distribution asymétrique constitue un défi majeur : sans stratégie de compensation, l'algorithme risquerait de développer un biais prédictif en faveur de la classe majoritaire, ignorant ainsi les signaux critiques des micro-métastases.

L'identification de ce déséquilibre lors de la phase d'extraction est une étape clé, car elle conditionne les choix méthodologiques futurs, notamment l'utilisation de fonctions de perte pondérées ou de techniques de rééchantillonnage lors de l'entraînement.

SCREEN DE L'AED

2.6 Enjeux et limites

La gestion de ces données comporte des enjeux cliniques et techniques complexes qui ont dicté la conception du pipeline de traitement. Le premier défi réside dans la variabilité inter-hospitalière. Le jeu de données CAMELYON17 provient de cinq centres hospitaliers différents, chacun utilisant ses propres scanners et protocoles de préparation de lames. Cette hétérogénéité introduit des variations dans la texture, la luminosité et les nuances de coloration des images, ce qui peut impacter la capacité de généralisation du système.

Le second enjeu est le coût médical élevé associé aux erreurs de diagnostic, particulièrement les faux négatifs. Dans ce contexte, l'omission d'une cellule cancéreuse est bien plus préjudiciable qu'une fausse alerte, car elle peut mener à un sous-stadiage de la maladie et priver la patiente d'un traitement vital. Enfin, une limite technique majeure est liée à la fragmentation de l'information. En découpant les lames en patchs de 224x224 pixels, le système perd le contexte architectural global du ganglion. La réussite du projet dépend donc de la capacité du modèle à extraire des caractéristiques cellulaires pertinentes malgré cette vision localisée, tout en sachant que le diagnostic final devra être reconstruit à partir de ces milliers de fragments isolés.

A ces défis scientifiques se sont ajoutées des contraintes matérielles majeures lors de la réalisation du projet. S'agissant d'une première expérience avec des images de cette dimension, l'appropriation du sujet médical a nécessité un temps d'étude important. Sur le plan logistique, le volume colossal du dataset a rapidement saturé nos capacités locales. L'absence de connexion fibre chez la personne hébergeant le projet a rendu le premier téléchargement impossible, créant un blocage immédiat. De plus, les premiers tests d'entraînement sur nos machines personnelles ont montré des temps d'exécution conséquents, rendant toute itération concrète inenvisageable. Ces limites opérationnelles ont rendu indispensable le passage vers une infrastructure Cloud et l'utilisation de machines virtuelles pour mener à bien le projet.

---

3. Architecture AWS et Infrastructure Cloud

3.1 Motivation du passage au cloud

Comme nous l'avons souligné, les **limites rencontrées** lors de la phase locale ont rendu le **passage au cloud indispensable**. L'**absence de fibre optique** chez l'hôte du projet empêchait tout simplement de récupérer les données, tandis que la **puissance de calcul** de nos ordinateurs personnels ne permettait pas d'entraîner le modèle dans des délais raisonnables. Pour résoudre ces blocages, nous avons choisi de **migrer notre environnement de travail sur Google Cloud Platform**.

Cette migration nous a permis de provisionner une **machine virtuelle configurée spécifiquement** pour nos besoins. L'un des avantages majeurs a été de pouvoir bénéficier d'un **débit réseau professionnel**. Grâce à cela, la machine virtuelle a pu se **connecter directement aux serveurs d'AWS** pour télécharger les images à une **vitesse extrêmement élevée**, ce qui était impossible depuis une connexion domestique. En plus de lever le verrou du téléchargement, Google Cloud nous a fourni la **puissance de calcul** nécessaire pour traiter les **centaines de milliers de patchs** d'images et réaliser l'**apprentissage profond** de manière fluide et efficace.

3.2 Architecture retenue

L'architecture de notre solution repose sur une **séparation claire** entre la **source des données** et l'**environnement d'exécution**. Bien que le jeu de données CAMELYON17 soit hébergé par **AWS** sur un **bucket S3**, nous avons fait le choix de centraliser l'intégralité de notre **infrastructure de calcul** sur **Google Cloud Platform**. Cette **approche hybride**nous a permis de tirer profit de l'accessibilité du **stockage d'Amazon** tout en utilisant la flexibilité des **instances de calcul de Google**.

Le socle de cette architecture est une **machine virtuelle Google Cloud** hautement performante. Cette instance a été dimensionnée pour gérer simultanément plusieurs **tâches lourdes** : le **transfert de données à haute vitesse**, l'**extraction de patchs par parallélisation** et l'**entraînement de nos modèles** de classification. Pour supporter ces opérations, nous avons rattaché à cette machine un **disque persistant** de **grande capacité**, capable de stocker les **centaines de milliers de fichiers PNG** générés durant le prétraitement, tout en garantissant une **latence d'écriture minimale**. Cette configuration nous a offert un **environnement stable et isolé**, indispensable pour mener à bien des calculs qui s'étalent sur **plusieurs heures**.

3.3 Pipeline de données

L'organisation du flux de données repose sur une **orchestration automatisée** entre le **bucket S3 d'Amazon** et notre **VM**. Pour **accélérer les transferts**, nous avons intégré l'**outil s5cmd**, qui permet de **saturer la bande passante** grâce au **téléchargement en parallèle**. Le pipeline est conçu pour fonctionner de **manière atomique** : pour chaque **lot de patients**, le script télécharge les images brutes, extrait les coordonnées des zones d'intérêt à partir des fichiers XML, puis **génère les patchs PNG**. Une fois l'extraction validée, les **fichiers sources** sont **immédiatement purgés**. Cette approche garantit que la machine virtuelle dispose **toujours de l'espace nécessaire** pour le lot suivant, créant ainsi un **flux de production continu** sans intervention manuelle.

IMAGE DE LA PIPELINE

3.4 Optimisation des ressources

Au-delà de la gestion du stockage, l'optimisation a porté sur la **réduction des temps de cycle** pour **limiter les coûts de location de la VM**. Nous avons mis en place une **parallélisation multi-niveaux** à l'aide de **ProcessPoolExecutor**, permettant de **traiter plusieurs ganglions en simultané**. Cette méthode a permis d'**exploiter la totalité des cœurs CPU**de l'instance Google Cloud, là où un script classique n'en aurait utilisé qu'un seul. En **réduisant drastiquement le temps**nécessaire pour générer nos **centaines de milliers de patchs**, nous avons pu libérer des ressources pour la phase la plus critique : l'**entraînement du modèle**. Cette rigueur dans l'**allocation des ressources** a transformé une contrainte budgétaire en un **exercice d'ingénierie logicielle efficace**.

4. Modèles d'IA et Méthodologie

4.1 Choix et conception des modèles

Pour la phase de modélisation, nous avons sélectionné une architecture de **réseau de neurones convolutifs** de type **ResNet-18**. Ce choix est motivé par la nécessité de trouver un **équilibre** entre la **performance de détection** et l'**efficacité de calcul** sur notre machine virtuelle. Le ResNet-18 est suffisamment profond pour capturer les **textures complexes** des tissus cellulaires tout en restant assez léger pour permettre des **itérations rapides** lors de l'entraînement.

Nous utilisons la bibliothèque **PyTorch** pour l'implémentation, en adaptant la structure du réseau pour une **classification binaire**. Les **patchs de 224x224 pixels** extraits lors de la phase précédente servent d'**entrée au modèle**. Un point important de notre méthodologie est l'utilisation du **transfert learning** : nous partons d'un **modèle pré-entraîné** sur la base de données **ImageNet**, ce qui permet au réseau de disposer déjà d'une capacité de **reconnaissance de formes élémentaires**. Nous **affinons ensuite ce modèle** sur nos données médicales pour qu'il se **spécialise** dans l'identification des **caractéristiques propres aux métastases**, comme la **densité nucléaire** ou les **anomalies de structure** des tissus.

4.2 Gestion du déséquilibre

L'analyse de notre jeu de données a révélé une **disparité importante** entre le nombre de **tissus sains** et le nombre de **tissus tumoraux**. Cette situation est **classique en imagerie médicale**, où les **zones d'intérêt pathologiques** occupent souvent une **surface réduite** par rapport aux tissus normaux. Si nous entraînions notre modèle sur ces **données brutes**, il risquerait de développer un **biais en faveur de la classe majoritaire**, obtenant ainsi une **précision élevée en apparence**tout en **échouant à détecter les métastases réelles**.

Pour corriger ce problème, nous avons instauré une **stratégie d'échantillonnage équilibré** au sein du chargement des données. Concrètement, nous utilisons un **système de poids** qui force le modèle à rencontrer **autant d'exemples positifs (tumoraux) que d'exemples négatifs (sains)** durant chaque phase d'apprentissage. En présentant ces **centaines de milliers de patchs** de **manière paritaire**, nous obligeons le réseau à se concentrer sur les **caractéristiques visuelles discriminantes** de la tumeur plutôt que sur la **fréquence statistique** des classes. Cette étape est cruciale pour garantir la **sensibilité du modèle**, le but final étant de **ne laisser passer aucune cellule cancéreuse**.

4.3 Entraînement et évaluation

L'entraînement de notre modèle ne repose pas sur une simple ingestion du dataset, mais sur une **stratégie de sélection extrêmement précise** visant à transformer les **100 patients initiaux** en un jeu d'apprentissage de haute qualité. Ce processus a suivi une logique d'entonnoir rigoureuse. Sur les 100 patients de départ répartis dans les 5 centres hospitaliers, nous avons d'abord effectué un **tri clinique**. Nous avons choisi d'**exclure 32 patients** présentant des **statuts intermédiaires**, comme les **pN0(i+)** ou les **pN1mi**, car leur faible charge tumorale et l'ambiguïté du signal histologique risquaient d'introduire un **bruit préjudiciable** à l'apprentissage du réseau. Nous nous sommes concentrés exclusivement sur les **stades pN1 et pN2** pour obtenir un **signal de vérité terrain incontestable**.

À cette contrainte clinique s'est ajoutée une limite technique majeure liée aux annotations. Sur les 44 patients pN1 ou pN2 restants, seuls 23 disposaient d'**annotations exploitables**. Les 21 patients positifs non annotés ont donc été écartés de l'apprentissage supervisé. Comme les 24 patients négatifs (pN0) ne nécessitent pas de détourage spécifique, notre capacité maximale théorique s'est fixée à 46 patients pour conserver un **scénario parfaitement équilibré**. Pour stabiliser l'apprentissage et éviter qu'un établissement ne domine statistiquement les autres, nous avons appliqué une **règle de plafonnement par centre** : le nombre de patients retenus est déterminé par le nombre de patients positifs annotés, auquel nous ajoutons un nombre équivalent de patients sains du même centre, avec un maximum de cinq par catégorie.

C'est lors de l'application de cette règle de sélection que le choix du centre à exclure s'est imposé. Le **Centre n°3**(Rijnstate Hospital) est apparu comme celui possédant le moins de patients exploitables après nos différents filtres. Nous avons donc décidé de le retirer de la phase d'entraînement pour en faire notre **jeu "Out-Of-Distribution" (OOD)**. Ce choix stratégique nous permet d'évaluer la **capacité de généralisation** de notre modèle sur un hôpital inconnu tout en impactant le moins possible le volume de données d'entraînement. Au final, notre jeu d'apprentissage est constitué de **34 patients (17 sains et 17 positifs)** répartis sur les **centres 0, 1, 2 et 4**.

Un aspect fondamental de notre méthodologie d'évaluation réside dans la séparation entre les ensembles d'entraînement et de validation. Nous avons opté pour un hold-out stratifié réalisé strictement au **niveau du patient**. Concrètement, le split est effectué sur une table récapitulative avant toute expansion des données en patchs. Cela garantit une répartition équilibrée des classes entre les deux ensembles et, surtout, assure qu'aucun patient ne soit présent simultanément dans l'entraînement et la validation. Cette stratégie est indispensable pour éviter toute **fuite de données (data leakage)** : le modèle ne peut pas "tricher" en reconnaissant la signature visuelle d'un patient qu'il aurait déjà vu. C'est une approche méthodologiquement équivalente à une **validation groupée par patient** (GroupK-Fold), mais adaptée à notre schéma de hold-out.

Sur le plan technique, cet entraînement est réalisé sur notre machine virtuelle **Google Cloud** via **PyTorch**. Pour piloter nos ressources et respecter les contraintes de budget, nous utilisons un échantillon initial de 200 lots pour calculer le temps moyen par batch et extrapoler la durée totale sur les **5 époques** prévues. L'optimisation des poids du réseau s'appuie sur la fonction de perte **CrossEntropyLoss** (entropie croisée) et l'**optimiseur Adam**. Ce dispositif nous permet de traiter les centaines de milliers de patchs avec une grande stabilité, en nous assurant que le modèle apprend réellement à discriminer les tissus malins avant l'étape de classification globale par patient.

4.4 Agrégation Patch vers Patient

Une fois que le modèle a prédit une probabilité pour chaque patch individuel, l'étape finale consiste à **consolider ces résultats** pour établir un **diagnostic à l'échelle du patient**. C'est une **phase délicate** car un patient peut avoir des **milliers de patchs sains** et seulement une **petite zone tumorale** qui détermine pourtant son stade clinique. Pour transformer ces **prédictions locales** en un **stade pN (pathological N-stage)**, nous utilisons une **méthode d'agrégation** basée sur des **règles logiques**.

Nous analysons la **répartition et la concentration** des patchs prédits comme "positifs" sur l'ensemble des ganglions du patient. Si le modèle détecte des **amas de patchs tumoraux** dépassant certains **seuils de confiance**, le ganglion est considéré comme **métastatique**. En combinant l'état de chaque ganglion d'un même patient, nous pouvons alors déduire son

stade global. Cette approche permet de **lisser les éventuelles erreurs isolées** du modèle (**faux positifs parsemés**) en se concentrant sur la **cohérence spatiale** des détections, indispensable pour fournir une **évaluation médicale exploitable**.

Interprétation clinique des stades pN :

Stade	Interprétation	Classe ML
pN0	aucun ganglion atteint	NÉGATIF
pN0(i+)	ITC seulement	⚠️ à discuter
pN1mi	micrométastases	POSITIF
pN1	métastases	POSITIF
pN2 / pN3	métastases étendues	POSITIF

---

5. Interprétation, Équité et Éthique : Vers une IA Responsable

Dans un contexte médical, la performance brute d'un modèle (comme l'AUC ou l'Accuracy) ne peut être considérée isolément. Il est indispensable de comprendre comment une prédiction est produite, sur quels signaux elle repose et quelles sont ses limites, afin d'éviter toute sur-interprétation des résultats. Cette exigence justifie l'intégration d'une démarche d'interprétabilité explicite, centrée non pas sur la technique pure, mais sur l'usage clinique réel.

5.1 Stratégie d'interprétabilité : Du Patch au Patient

La demande académique initiale suggérait l'utilisation d'outils de visualisation locale (type Grad-CAM). Cependant, nous avons fait le choix méthodologique délibéré de ne pas nous focaliser sur l'interprétabilité au niveau du patch individuel. Une prédiction isolée n'est qu'une variable intermédiaire dans notre pipeline : chercher à expliquer visuellement pourquoi le modèle s'active sur un patch spécifique n'apporte que peu de valeur clinique si cette information n'est pas contextualisée.

Notre approche de l'interprétabilité s'est donc déplacée vers **l'analyse des mécanismes d'agrégation**. L'objectif était de comprendre comment la somme des signaux locaux (les patchs) est synthétisée pour aboutir à une décision globale (le patient). Pour ce faire, nous avons mis en œuvre trois axes d'analyse :

1. **Contrôle des seuils de décision :** Nous avons testé différents seuils de confiance (`THR_PATCH` à 0.7, 0.8, 0.9) pour observer à quel point un signal local doit être robuste pour influencer le diagnostic final.

2. **Analyse de la proportion de patchs tumoraux :** Cette méthode intuitive mesure la proportion de patchs suspects par patient. Bien qu'offrant une bonne lisibilité clinique, elle s'est révélée très sensible au déséquilibre extrême des classes.
3. **Estimation de la surface tumorale :** Cette approche tente de reproduire le raisonnement morphologique du pathologiste (plus la surface atteinte est grande, plus le risque est élevé).

5.2 Analyse de Robustesse et Domain Shift

L'application de ces méthodes d'interprétabilité a mis en lumière le défi majeur de ce projet : le **Domain Shift**. Nos résultats montrent une divergence flagrante entre les performances d'entraînement (AUC proche de 0.99) et celles de validation (oscillant entre 0.53 et 0.76).

L'analyse spécifique sur le **Centre n°3** (notre jeu de données *Out-Of-Distribution*) confirme cette vulnérabilité. Les performances y sont proches de l'aléatoire, révélant que le modèle peine à généraliser ses acquis à de nouveaux hôpitaux. Ce phénomène s'explique par deux facteurs identifiés grâce à notre analyse des erreurs :

- **Biais technologique :** Le modèle semble avoir appris à reconnaître des spécificités techniques (protocoles de coloration, types de scanners) propres aux centres d'entraînement, plutôt que des signatures tumorales universelles.
- **Amplification des erreurs :** Dans un contexte de rareté absolue (seulement ~0.02% de patchs réellement tumoraux en validation), 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.

5.3 Limites Structurelles et IA Responsable

Le cœur de notre démarche d'IA responsable a consisté à ne pas masquer ces limites, mais à les documenter explicitement. L'analyse critique de nos résultats met en évidence plusieurs obstacles à un déploiement clinique :

- **Le paradoxe Sensibilité / Spécificité :** Nos méthodes d'agrégation favorisent une sensibilité élevée (proche de 100%), ce qui est crucial pour ne pas rater de cancer. Cependant, cela se fait au prix d'une spécificité faible. Le modèle génère une charge de travail trop importante en fausses alertes pour les pathologistes.
- **La limite de la cohorte patient :** Bien que nous ayons entraîné le modèle sur des centaines de milliers de patchs, ceux-ci ne proviennent que de **34 patients**. Cette redondance intra-patient crée un sur-apprentissage implicite : le réseau voit beaucoup d'images, mais une diversité biologique trop faible.
- **Fiabilité de l'agrégation :** Nous avons observé que les erreurs patch-level ne s'annulent pas, mais s'additionnent. Un faible taux d'erreur local peut suffire à biaiser totalement le diagnostic patient.

---

6. Conclusion et Perspectives

6.1 Bilan technique

Le bilan technique de ce projet souligne la complexité de traiter des données histopathologiques à grande échelle pour le diagnostic des métastases. Nous avons réussi à mettre en place un pipeline complet, allant de l'extraction de millions de patchs à la

classification finale. L'utilisation d'une architecture ResNet-18 a permis d'obtenir des performances remarquables au niveau local, avec une AUC d'entraînement proche de 0.99. Ces résultats démontrent que le modèle capte parfaitement les signaux morphologiques des cellules cancéreuses sur les données qu'il a déjà rencontrées lors de l'apprentissage.

Cependant, le passage à la validation a révélé la difficulté de la généralisation, avec une AUC oscillant entre 0.53 et 0.76 selon les configurations. Ce décrochage met en lumière l'importance de la variabilité inter-patients et inter-centres. Sur le plan de l'agrégation, l'utilisation du seuil THR_PATCH à 0.9 a permis d'isoler les signaux les plus robustes, même si la très faible proportion de patchs positifs en validation (environ 0.02 %) rend le système extrêmement sensible au moindre bruit visuel.

[Image d'un graphique montrant l'évolution de l'AUC entre le training et la validation]

L'infrastructure cloud a joué un rôle central dans cette réussite. Face à des volumes de données colossaux, l'utilisation d'une instance Google Cloud avec GPU a été le seul moyen de traiter les flux d'images en un temps raisonnable. Le pilotage rigoureux des ressources, notamment par l'extrapolation du temps de calcul via le paramètre sec_per_batch, a été la clé pour respecter nos contraintes budgétaires tout en garantissant la stabilité de l'entraînement sur les 5 époques prévues. Ce projet prouve que la maîtrise de l'outil informatique est indissociable de l'expertise en deep learning pour mener à bien un projet d'imagerie médicale numérique.

6.2 Enseignements

Ce projet nous a permis de comprendre que la rigueur méthodologique est plus cruciale que la performance brute en contexte médical. L'enseignement principal réside dans l'importance du partitionnement des données. Le choix d'un split stratifié au niveau patient, plutôt qu'au niveau patch, a été déterminant pour garantir l'absence de fuite de données et la crédibilité de nos évaluations. Sans cette précaution, nos résultats auraient été artificiellement gonflés, masquant les difficultés réelles de généralisation.

Nous avons également appris que le traitement de données médicales impose une gestion stricte de la variabilité. L'équilibre entre les centres hospitaliers et la gestion du déséquilibre des classes ne sont pas de simples étapes techniques, mais des conditions nécessaires pour obtenir un modèle éthique. La confrontation aux réalités du terrain, comme la rareté du signal tumoral (0.02 % de patchs positifs en validation), nous a enseigné l'importance de la précision dans la préparation des données avant même de lancer le premier entraînement.

6.3 Perspectives

L'analyse de nos résultats démontre qu'un modèle peut être techniquement performant au niveau patch tout en restant cliniquement risqué au niveau patient. Pour transformer ce prototype en un outil robuste, plusieurs axes d'amélioration peuvent être envisagés.

Évolution des modèles et de l'agrégation

La limite principale réside dans l'agrégation par seuillage manuel (THR_PATCH), qui est trop sensible au bruit. Pour dépasser cela, il serait nécessaire d'évoluer vers des approches

d'apprentissage multi-instance (MIL). Ces méthodes permettent au réseau de pondérer automatiquement les zones réellement informatives et d'apprendre directement au niveau du patient, s'affranchissant ainsi des seuils arbitraires. L'ajout de mécanismes d'attention permettrait également de visualiser quelles régions de la lame ont le plus influencé le diagnostic final, offrant une interprétabilité bien plus utile au clinicien que de simples scores statistiques.

Enrichissement des données et lutte contre le biais

Au-delà de l'architecture, la qualité des données reste le nerf de la guerre. Augmenter la taille de la cohorte patient est indispensable pour réduire le sur-apprentissage constaté. Nous pourrions aussi mettre en œuvre des techniques de normalisation de couleur (stain normalization) pour harmoniser les différences de coloration entre les hôpitaux. Cela permettrait de réduire le Domain Shift sans avoir besoin de milliers de nouveaux patients. De plus, l'intégration de techniques d'augmentation de données plus agressives (rotations, variations de contraste, flou) aurait pu aider le modèle à se focaliser sur les formes cellulaires plutôt que sur les caractéristiques techniques du scanner.

Optimisation du pipeline et calibration

Sur le plan opérationnel, l'implémentation de modèles de fond de lame (backbone) plus récents ou pré-entraînés spécifiquement sur des données pathologiques (comme les modèles SSL en histopathologie) pourrait offrir une meilleure extraction de caractéristiques que le ResNet-18 classique. Enfin, une collaboration étroite avec des pathologistes pour calibrer les scores de confiance constituerait l'étape ultime. Un modèle d'IA ne doit pas seulement donner une réponse, il doit être capable d'exprimer son incertitude, surtout face à des tissus sains mais atypiques qui provoquent actuellement nos faux positifs.

6.4 Conclusion Finale

Ce projet avait pour objectif de développer et d’évaluer une approche de détection de métastases à partir de lames histopathologiques numériques, en s’appuyant sur un modèle de classification patch-level et différentes stratégies d’agrégation patient-level. Il s’inscrit dans un contexte réaliste de données massives, fortement déséquilibrées et soumises à des exigences cliniques élevées.

Sur le plan méthodologique, une chaîne complète a été mise en place, couvrant le prétraitement des données, l’entraînement du modèle, l’inférence patch-level et l’agrégation patient-level, avec une séparation rigoureuse entre les ensembles d’entraînement, de validation et de test, ainsi qu’une distinction entre centres vus et centres jamais vus. Le modèle parvient à détecter des motifs tumoraux localisés, ce qui indique qu’il apprend des caractéristiques histopathologiques pertinentes au niveau patch.

Néanmoins, les performances obtenues doivent être analysées à la lumière d’une contrainte majeure: **la taille et la nature du jeu de données**. Avec un total d’environ 40 patients, le dataset n’était pas favorable à l’obtention de résultats optimaux au niveau patient. Or, la phase d’entraînement est l’élément déterminant de tout modèle d’apprentissage automatique: elle conditionne directement la qualité de la validation, puis celle du test. Un

volume de données limité et une diversité clinique restreinte réduisent la capacité du modèle à généraliser, en particulier dans un contexte inter-centres.

Les stratégies d'agrégation patient-level explorées, bien qu'interprétables et cohérentes d'un point de vue conceptuel, ont montré des limites importantes, notamment une forte sensibilité accompagnée d'une spécificité faible et d'un nombre élevé de faux positifs. Ces résultats illustrent la difficulté de passer d'une détection locale efficace à une décision clinique fiable à l'échelle du patient, surtout lorsque les données d'entraînement sont peu nombreuses.

Dans ce contexte, l'utilisation d'outils d'interprétabilité a été essentielle. Ils ont permis d'analyser le comportement du modèle, de comprendre les mécanismes sous-jacents aux prédictions et d'identifier les sources de biais et d'erreurs, en particulier liées à la distribution des patches et aux seuils utilisés. Cette démarche s'inscrit pleinement dans une approche d'IA responsable, en évitant toute sur-interprétation des performances et en mettant en évidence les limites du modèle.

En conclusion, ce travail montre que la performance d'un modèle de détection de métastases ne dépend pas uniquement de l'architecture choisie, mais avant tout de la **qualité, de la quantité et de la diversité des données d'entraînement**, ainsi que de la capacité à interpréter et à critiquer les décisions produites. Les résultats obtenus constituent une base expérimentale solide, tout en soulignant la nécessité de jeux de données plus larges et d'approches plus avancées, telles que l'apprentissage multi-instance hiérarchique, les modèles attentionnels ou la calibration clinique, pour envisager une transférabilité réelle vers la pratique médicale.

— Fin du Rapport —