

Projet 2 - Détection automatique de métastases du cancer du sein dans des images histopathologiques

Contexte

Le projet utilise les données du <https://camelyon17.grand-challenge.org/> un concours international visant à développer des méthodes automatiques pour détecter les métastases dans les ganglions lymphatiques de patientes atteintes d'un cancer du sein.

L'objectif est d'aider les étudiants à concevoir une chaîne complète d'analyse d'images médicales réelles, depuis la donnée brute jusqu'à la prédiction finale du diagnostic.

L'objectif principal est de :

1. Détecter automatiquement la présence et l'étendue des métastases dans ces lames histologiques.
2. Prédire pour chaque patiente un stade pN, correspondant à la classification clinique indiquant le nombre de ganglions lymphatiques affectés.

Vous ne construisez pas "juste un modèle", vous concevez un **système d'aide au diagnostic médical**. Pensez aux notions apprises en module 4.

Jeu de données

1. Nature des données

Le jeu de données CAMELYON17 est constitué de lames histologiques numérisées (*Whole Slide Images – WSI*) de ganglions lymphatiques provenant de patientes atteintes d'un cancer du sein.

- Images issues de microscopes numériques
- Coloration H&E (Hématoxyline & Éosine)
- Très haute résolution (plusieurs gigapixels par lame)

Les WSI ne peuvent pas être traitées directement par un CNN classique.

2. Découpage en patches

Pour rendre les données exploitables par des réseaux convolutifs :

- Les WSI sont découpées en **patchs d'images**
- Taille standard : **224 × 224 pixels**
- Format : **.png** ou **.tif**
- Images **RGB**

Chaque patch correspond à une **petite région locale du tissu**.

3. Labels au niveau patch

Chaque patch est annoté avec un label binaire ou multiclass :

- **0** → tissu normal
- **1** → tissu tumoral
- L'annotation est issue d'un masque de vérité terrain fourni par des experts pathologistes.

4. Labels au niveau patient

En plus des labels patch :

- Chaque patiente est associée à un stade pN
- Le pN-stage correspond à la classification clinique indiquant :
 - le nombre de ganglions lymphatiques atteints
 - l'étendue de la maladie

Exemples (simplifiés) :

- pN0 : aucun ganglion atteint
- pN1 : quelques ganglions positifs
- pN2 / pN3 : atteinte plus étendue

Le stade pN n'est pas directement prédit par le CNN, mais dérivé par agrégation des prédictions patch.

5. Particularités

- Les données proviennent de 5 hôpitaux différents, permettant d'étudier la robustesse et le transfert de domaine (*domain shift*) des modèles.
- Le jeu de données est fortement déséquilibré :
 - beaucoup plus de patchs normaux que tumoraux
 - certaines patientes sans métastases
 - Contraintes spécifiques aux données médicales : Coût élevé des faux négatifs (métastases non détectées)

Les choix de modèle doivent être médicalement justifiés, pas uniquement optimisés sur une métrique.

Travail demandé aux étudiants

1. Compréhension du problème médical

Comprendre le contexte clinique :

- détection de métastases dans les ganglions lymphatiques
- prédiction du stade pN au niveau patient

Identifier les enjeux médicaux :

- coût des faux négatifs
- variabilité inter-hôpitaux
- limites d'un modèle automatique

2. Exploration et préparation des données

- normalisation des images
- gestion du déséquilibre des classes

3. Modélisation – Classification au niveau patch

- Concevoir et entraîner un CNN
- Évaluer les performances

4. Agrégation patch → patient

- Proposer une stratégie d'agrégation :
 - proportion de patchs tumoraux
 - surface totale prédictive comme tumorale
 - règles heuristiques ou modèle simple
- Justifier le lien entre :
 - prédictions locales (patch)
 - décision clinique globale (stade pN)

5. Évaluation clinique et robustesse

- Évaluer les performances :
 - au niveau patch
 - au niveau patient (pN)
- Étudier l'impact du domain shift :
 - performances par hôpital
 - généralisation du modèle
- Discuter les limites :
 - biais
 - sur-apprentissage
 - transférabilité clinique

6. Interprétabilité et IA responsable

- Utiliser un outil d'interprétabilité

Objectifs pédagogiques

Ce projet permet aux étudiants de :

- Manipuler un jeu de données médical réel et complexe.
- Construire et entraîner un modèle de classification par réseau de neurones convolutifs (CNN).
- Mettre en place une stratégie d'agrégation pour passer du niveau « patch » (petite image) au niveau « patient »
- Évaluer les performances avec des métriques adaptées
- Discuter des limites, du biais de domaine (différences entre hôpitaux) et de la généralisabilité des modèles IA en santé.

Livrables attendus

- Code et pipeline reproduit
- Rapport écrit (max 15 pages) : Méthodologie, résultats, interprétation, discussion.

- Présentation orale : 15 minutes maximum !
 - **Notation finale : 20% (follow-up) + 80% (rendu et soutenance)**
-

Utilisation de l'intelligence artificielle

L'utilisation d'outils d'intelligence artificielle est autorisée. Toutefois :

1. Chaque membre doit contribuer activement au projet.
2. Chaque étudiant du groupe doit maîtriser l'intégralité des aspects techniques, scientifiques et méthodologiques du projet (code, données, interprétations, résultats, visualisations, etc.).
3. Lors de la soutenance orale, tout membre doit être capable d'expliquer clairement le travail réalisé, les choix effectués, et de reproduire le raisonnement menant aux résultats.

Délais et rendus

Tout rendu (rapport, code, présentation, dashboard, etc.) doit être transmis avant la date limite indiquée. Aucun rendu en retard ne sera accepté.

Tout rendu remis après la deadline recevra la note de 0, sans exception.

Attentes pour les livrables

Le travail doit être original, documenté et reproductible.

Les sources de données, de code et d'outils externes doivent être clairement citées.

Toute forme de plagiat, d'utilisation non transparente d'IA, ou de reproduction non justifiée de code externe entraînera une note de 0.

Accent sur la compréhension et l'interprétation

L'évaluation met l'accent sur :

- La qualité de l'analyse, la rigueur méthodologique et la clarté des explications ;
- L'interprétation des résultats, leur cohérence et leur mise en perspective ;
- La justification des choix techniques et scientifiques.

Les valeurs numériques finales (scores, métriques, performances) ne constituent pas le critère principal d'évaluation.

La compréhension, la pertinence des approches et la capacité à expliquer les résultats priment sur l'obtention d'un "bon score".