Intégration de données médicales hétérogènes

PMIND 2026

Encadrants

Hubert Naacke (hubert.naacke@lip6.fr) et Garance Lucas (garance.lucas@lip6.fr)

Contexte. Le cancer colorectal est enjeu de santé publique [2]. L'examen diagnostique de référence du côlon est la vidéocoloscopie [1] qui est réalisée en s'appuyant sur la méta-classification CONECCT9 (Colorectal Endoscopic Classification to Choose the Treatment) afin d'analyser les marqueurs dans toutes leurs propriétés connues et de caractériser les polypes de manière plus fine et plus précise [3].

Le projet s'inscrit dans un travail doctoral dont l'objectif est de concevoir une méthode hybride de classification des polypes colorectaux par apprentissage multitâche utilisant la classification CONECCT9.

Défis. Les données d'imagerie médicale servant de base à ce travail sont hétérogènes : il y a des fortes différences en termes de format d'image et leur modalité, de type, de catégorie, d'annotation, de description textuelle, de nombres d'images, etc. La préparation des données est actuellement traitée de manière *ad hoc* afin de produire des jeux d'entraînement et de validation spécifiques dont les caractéristiques dépendent du modèle de classification considéré. Cette approche *ad hoc* ne convient plus lorsqu'un grand nombre de modèles sont considérés car chaque modèle peut nécessiter un jeu de données différent afin d'apprendre à classer les images selon différents aspects. De plus, la conception de modèles de classification généralisables nécessite de disposer de jeux d'entraînement suffisamment riches en termes de classes représentées. Or certaines classes ne sont présentes que dans certaines bases d'images ; l'absence de description unifiée des images empêche de constituer des jeux d'entraînement enrichis qui réuniraient des images provenant de diverses bases.

L'objectif de ce projet est de concevoir une solution pour intégrer les divers jeux de données et faciliter leur accès et leur gestion à long terme.

Le travail à effectuer consistera à :

- Décrire de manière plus unifiée les bases d'images existantes et leurs propriétés. Les descriptions devant être flexibles pour tenir compte de l'évolution possible des propriétés.
- Sélectionner des images provenant de plusieurs bases, à l'aide d'un langage de requêtes.
- Définir un pipeline d'opérations qui produit des jeux d'entraînement et de validation, ceci de manière reproductible pour garantir la pérennité des jeux produits.

Références

[1] Institut National du Cancer. La situation du cancer en france en 2010. Technical report, www.e-cancer.fr, 2010.

[2] Institut National du CAncer. Panorama des cancers en france 2025. Technical report, 2025.

[3] Clementine Brule et al. The colorectal neoplasia endoscopic classification to choose the treatment classification for identification of large laterally spreading lesions lacking submucosal carcinomas : A prospective study of 663 lesions. *United European gastroenterology journal*, 10(1) :80–92, 2022.