

Ruban rose

“Grâce au dépistage anticipé et aux traitements modernes, le taux de survie à 5 ans pour un cancer du sein localisé est d'environ 99 %.”

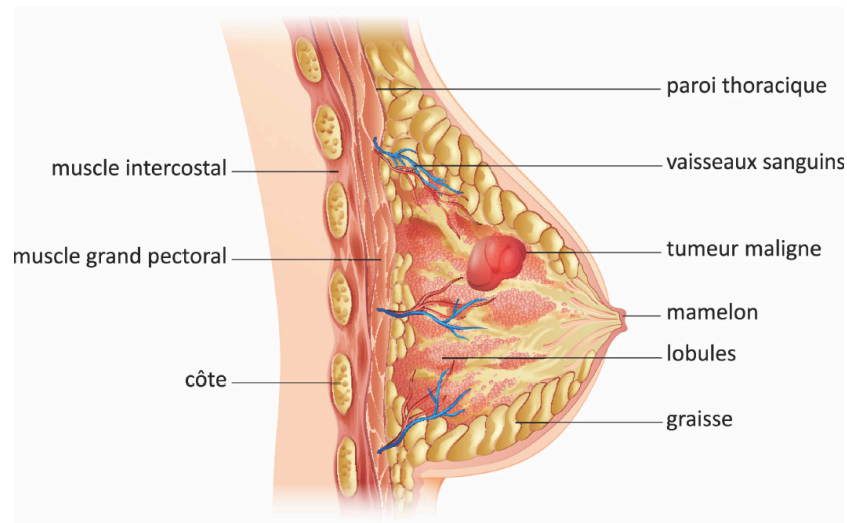


Dans le cadre du prochain Octobre Rose, un **centre de recherche** spécialisé dans l'innovation en imagerie médicale et en intelligence artificielle organise un événement réunissant experts, chercheurs et passionnés. L'objectif est de développer **des solutions concrètes pour renforcer la lutte contre le cancer du sein**. Vous décidez de participer et de proposer une solution permettant d'améliorer la précision du diagnostic et facilitant une prise en charge rapide et adaptée des patientes

Grâce à un outil de détection basé sur **la computer vision**, les images médicales peuvent être analysées avec une précision accrue, permettant d'**identifier les premiers signes du cancer du sein et d'augmenter les chances de rémission**.



Invasive ductal carcinoma



Les carcinomes canauxaux invasifs (Invasive ductal carcinoma) **sont dans près de 80% des cas les plus communs de cancer du sein.** Ils se trouvent être particulièrement dangereux car des formes malignes et des métastases peuvent se développer.

Une biopsie est souvent réalisée pour prélever des échantillons de tissus par la technique dite du « ganglion sentinelle » où un produit est injecté afin de révéler la présence de tumeur. Un médecin spécialisé doit ensuite décider si la patiente est atteinte d'IDC, d'un autre type de cancer du sein, ou si elle est en bonne santé.

Ce travail, effectué manuellement, est long et dépend fortement de l'expertise du spécialiste et de la qualité de son matériel.

L'apprentissage automatique pourrait être d'une grande aide pour **détecter automatiquement une tumeur** dans les tissus cellulaires et accélérer le processus de diagnostic.



Afin de réaliser cette tâche, vous modélisez les données issues du jeu de données **Breast Histopathology Images**, qui contient des images de biopsies mammaires annotées pour la détection du cancer.

Ce jeu de données est composé d'**images de tissus prélevés et classés en catégories représentant des cellules cancéreuses et non cancéreuses**. Il permet d'entraîner des modèles de classification pour l'identification automatique de lésions malignes.

1. Téléchargez le jeu de données **Breast Histopathology Images** depuis [ce lien](#).
2. Vous importez le dataset et vous affichez un **échantillon des patches sains et des patches cancéreux**.
3. Réalisez une exploration de vos données. Vous devrez en **tirer des observations** et les noter pour la suite des traitements.
4. Vous cherchez **des méthodes de pré-traitements** d'images. Effectuez un prétraitement des données adapté. Vous pouvez :
 - **Redimensionner** les images à une taille uniforme (ex. 50x50),
 - Appliquer des techniques d'**augmentation de données** (rotation, retournement, ajustements de contraste, etc),
 - **Normaliser** les valeurs des images,
 - Autre forme de traitement.
5. Vous devrez entraîner **au moins deux des modèles** suivants :



- **CNN,**
 - **ResNet,**
 - **Faster R-CNN,**
 - **Vision Transformer (ViT).**
6. Utilisez **Optuna** afin d'**optimiser automatiquement** les **hyper-paramètres** des modèles entraînés (Learning Rate, Batch Size, nombre d'Epochs, Weight decay Dropout rate, nombre de neurones-couches, Optimizer).
 7. Évaluez chaque modèle selon les **métriques de classification supervisée** que vous connaissez bien. Cherchez puis utilisez une métrique adaptée à la problématique étudiée.
 8. Toujours à l'aide d'Optuna, **vous identifiez le modèle le plus performant pour la classification** (et éventuellement la détection) des images de biopsies mammaires annotées pour la détection du cancer.
 9. Vous **concluez sur les résultats obtenus** et le meilleur pipeline pour le développement de votre outil.

Compétences visées

→ Apprentissage profond



Rendu

L'évaluation de ce projet se fera sur les aspects suivants :

1. Une présentation explicative de votre travail sous forme de diapositives.
2. Un repository github public nommé **ruban-rose**, contenant les éléments suivants :
 - a. Un **notebook Python propre et commenté** (introduction, titres des sections, interprétation des visuels, justification des résultats, conclusion, etc) contenant le procédé de développement de votre outil, du nettoyage à la modélisation des données, en passant par l'**analyse exploratoire**. **Vous pouvez avoir jusqu'à 3 notebooks pour développer votre outil.**
 - b. Un fichier **README.md** présentant le contexte du projet, les données et leur analyse, les algorithmes utilisés et une conclusion sur votre travail. Pensez à inclure la veille réalisée.

Base de connaissances

- [Multiclass classification of breast cancer histopathology images using multilevel features of deep convolutional neural network](#)
- [Deep Residual Learning for Image Recognition](#)
- [Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks](#)
- [Vision Transformer \(ViT\) - Hugging Face](#)