Intitulé de l'offre de stage

Création de jeux de données synthétiques pour la classification des tumeurs cérébrales

Stagiaire en	Informatique (Deep Learning)
Affectation	CHU Toulouse
Durée	5 à 6 mois
Rémunération	environ 670 € net mensuel
Date de la publication	15/11/2026
Date d'embauche prévue	mars 2027 suivant disponibilités
Lieu	IUCT Oncopole, 1 avenue Irène Joliot-Curie, Toulouse

L'IUCT Oncopole

L'Oncopole regroupe sur un même site plus de 2000 professionnels engagés dans la lutte contre le cancer. Ils sont répartis entre l'Institut Universitaire du Cancer de Toulouse - Oncopole (IUCT-O)) et le Centre de Recherche en cancérologie de Toulouse (CRCT). L'IUCT-Oncopole réunit l'expertise et l'excellence de l'Institut Claudius Regaud (Centre de Lutte Con tre le Cancer) et de plusieurs services d'oncologie du CHU de Toulouse. Ses missions sont le soin, la recherche et l'enseignement.

L'Institut de Recherche en Informatique de Toulouse

L'Institut de Recherche en Informatique de Toulouse (IRIT), une des plus imposantes Unité Mixte de Recherche au niveau national, est l'un des piliers de la recherche en Occitanie avec ses 700 membres, permanents et non-permanents. De par son caractère multi-tutelle (CNRS, Universités toulousaines), son impact scientifique et ses interactions avec les autres domaines, le laboratoire constitue une des forces structurantes du paysage de l'informatique et de ses applications dans le monde du numérique, tant au niveau régional que national.

L'équipe d'accueil

L'informatique et l'intelligence artificielle sont appelés à prendre une place croissante dans le monde de la recherche médicale, et en particulier dans le monde de la recherche contre le cancer. L'équipe commune IRIT/CRCT, co-localisée sur le site de l'Oncopole de Toulouse, a pour but de faire travailler ensemble et sur le même site chercheurs en informatique, chercheurs dans le domaine du cancer et médecins. Elle peut s'appuyer sur le plateau technique du CRCT (analyses biologiques, séquençages génomique,...) et sur les moyens de calcul de la région Occitanie (CALMIP).

Le stagiaire sera accueilli au sein de l'équipe d'Emmanuelle Uro-Costes, professeur de médecine au CHU de Toulouse, chercheuse au Centre de Recherche en Cancérologie de Toulouse (CRCT/INSERM), et sera co-encadré par des chercheurs de l'IRIT dans le cadre de l'équipe commune IRIT-CRCT.

Objet du stage

Les tumeurs humaines sont très variées. Par exemple, derrière le terme grand public « cancer du sein » se cachent plus d'une vingtaine d'entités tumorales différentes, fréquentes ou rares. L'identification de chaque entité est capitale, car elle a des implications pronostiques et thérapeutiques. L'identification des types tumoraux repose sur l'analyse au microscope de la morphologie des cellules anormales, effectuée par des médecins spécialistes, les pathologistes. Ces dernières années, l'OMS a intégré des mutations de l'ADN tumoral au diagnostic microscopique, rendant la classification plus fiable et plus précise (exemple de diagnostic intégré : astrocytome IDH muté). Plus récemment encore, l'OMS a intégré une technique de classification des tumeurs cérébrales (brain classifier), développée par l'Université de Heidelberg, se basant sur le profil de méthylation de l'ADN tumoral et utilisant le Random Forest.

Le principe du méthylome est le suivant : sur les milliards de bases (A,T, C, G) de notre ADN, 28 millions de Cytosine (C), toujours à côté d'une Guanine (G) (on parle de doublets CpG) peuvent-être méthylées ou non. Cette méthylation régule la transcription des gènes. En particulier, si plusieurs CpG au niveau d'un promoteur (séquence d'ADN qui peut initier l'expression d'un gène) portent des cytosines méthylées, un blocage de la transcription du gène est généralement observé. Les cellules tumorales détournent ce système pour inactiver des anti-oncogènes et activer des oncogènes. A partir d'un set de gènes choisis pour leur implication dans le cancer, des puces d'hybridation pour étudier la méthylation des cellules tumorales ont été développées. Le pattern de méthylation de 450 000 CpG suffit amplement à identifier un type tumoral. Voici le lien vers une conférence de vulgarisation que nous avions faite sur ce sujet et qui est disponible en replay sur inscription.

Cette démarche de classification par le méthylome est en train de se développer pour d'autres tumeurs (tissus mous, leucémies, tumeurs cutanées etc...). Des fichiers informatiques (.idat) correspondant aux données brutes du pattern de méthylation sont disponibles en open source pour de nombreux cancers. À partir de ces bases de données on peut reconstruire des classifiers grâce à l'IA. La seule base de données publique pour les tumeurs cérébrales correspond à environ 3000 tumeurs et aux 81 classes de tumeurs cérébrales, qui étaient connues en 2018. Or de nombreuses nouvelles entités tumorales ont été décrites depuis et nous souhaiterons éduquer une IA avec une base de données actualisée comportant 120 entités. Pour construire cette base de données nous pourrions récolter les profil de méthylation de l'ADN de plusieurs centaines de patients réalisé dans différents CHU de France. Hors cette mise en commun des données au niveau de nationale rencontre des obstacles

juridiques jusqu'ici infranchissable. Une des solutions pourrait être de créer des profil de méthylation synthétique représentatif d'une classe tumorale. Ces profils synthétiques ne correspondant pas réellement à des tumeurs d'une personne physique, ils pourraient ainsi être échangés librement entre les CHU. Le travail de ce stage sera donc d'évaluer les techniques de génération de données synthétiques à partir des profils existants et de production de profils représentatifs pour obtenir une base de données de profils synthétiques de méthylation de l'ADN. Il faudra ensuite développer des classifiers avec ces données synthétiques et comparer l'efficacité de ces classifiers par rapport à celles obtenues avec des profils réels de méthylation de l'ADN. À l'issue du stage, les différents CHU de France pourraient donc recevoir ce logiciel transformateur de leur données de méthylation réelles pour les transformer en batchs de données synthétiques exportables.

Le but sera donc :

> de tester différents algorithmes de génération de données synthétiques (Generative Adversarial Networks, Variable Auto-Encoders, perturbation de voisinage, ...)
> d'évaluer l'efficience de différents algorithmes (Random forest, LOGREG, SVM, XGBoost,...) sur les données synthétiques et les données réelles.
> d'étudier les données synthétiques avec la clusterisation en t-SNE, u-MAP ou ACP.

Formation

École d'ingénieur, de préférence avec spécialisation en informatique ou mathématiques (en année de césure ou stage long). Master 2 informatique ou mathématiques appliquées.

Compétences attendues

Des connaissances en science des données, apprentissage et réseaux de neurones seront appréciées.

Modalité de candidature

> CV à envoyer à: cups3@stages-medecine-numerique.fr
> Date limite de candidature : 01/01/2027
> Encadrants: Professeur Emmanuelle Uro-Coste (CHU de Toulouse/CRCT), Yvan Nicaise (CRCT) Sandrine Mouysset (IRIT), Daniel Ruiz (IRIT)

Une première sélection sera effectuée sur la base des CVs reçus. Les candidats seront immédiatement informés du résultat, et ceux qui seront retenus à l'issue de la première sélection seront invités à un ou deux entretiens individuels en téléconférence avant sélection définitive.

L'ensemble des offres de stage est disponible sur http://www.stages-medecine-numerique.fr.