Intitulé de l'offre de stage

CUP classifieur

Stagiaire en	Informatique (Deep Learning)
Affectation	CHU Toulouse
Durée	5 à 6 mois
Rémunération	environ 670 € net mensuel
Date de la publication	15/10/2025
Date d'embauche prévue	mars 2026 suivant disponibilités
Lieu	CHU, 2 rue Charles Viguerie, 31300 Toulouse, France

Le Centre Hospitalier Universitaire de Toulouse

Le Centre Hospitalier Universitaire (CHU) de Toulouse est constitué de plusieurs sites (les principaux étant les sites de Rangueil, Larrey et Purpan, ainsi que celui de l'oncopole en commun avec l'institut Claudius Régaud); il comprend 4 000 médecins et 12 000 personnels hospitaliers. La mission de recherche et d'innovation du CHU fait partie intégrante à la fois de son activité quotidienne et de sa stratégie pour l'avenir. Elle est menée en collaboration étroite avec les facultés et les organismes de recherche que sont notamment l'Institut national de la santé et de la recherche médicale (Inserm) et le Centre national de la recherche scientifique (CNRS).

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, le traitement des données et l'intelligence artificielle sont appelés à prendre une place croissante dans le monde de la recherche médicale. Dans ce cadre, le CHU de Toulouse a créé le Centre de Données pour la Santé et la Recherche (CDSR) dont le but est le recueil, l'analyse et le traitement de la donnée médicale, en support aux activités de recherche médicales du CHU. Elle collabore activement avec l'IRIT afin d'appliquer les techniques les plus récentes de la science des données et de l'intelligence artificielle dans le domaine de la santé.

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.

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. Les tumeurs ORL rares sont parfois difficiles à classer et pourraient bénéficier d'une classification par méthylome. Dans le cadre d'un projet européen (TRANSCAN SPELCASTER Pr D. Capper) puis d'un projet national PRTK fédérateur (REFCORomics Dr F-R Ferrand), nous allons réaliser 400 méthylomes correspondant à des tumeurs ORL rares. Environ 1000 cas sont déjà disponibles en open source sur des bases de données publiques. A partir de cette base organisée en dataframe , nous voulons évaluer l'efficience de différents algorithmes (Random forest, LOGREG, SVM) pour prédire le diagnostic histologique.Il faudra aussi :

- tester d'autres algorithmes : XGBoost, réseaux de neurones en privilégiant des méthodes permettant une évaluation/interprétation des résultats (Shapley Value...)
- optimiser le fonctionnement (i) en améliorant la base de données : équilibrer les classes, (ii) paralléliser l'utilisation des CPU avec Spark (iii) enrichir la dataframe avec les anomalies chromosomiques.
- valider les résultats en validation croisée et comparer les résultats obtenus avec la clusterisation en t-SNE, u-MAP ou ACP.
- identifier les CpG les plus importants en utilisant l'ACP, le clustering hiérarchique et la sélection de features (Boruta...) pour répondre à des questions biologiques et également permettre une utilisation de cet outil avec des technologies plus légères et plus rapides.

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: cup@stages-medecine-numerique.fr
- Date limite de candidature : 01/01/2026
- Encadrants: Professeur Emmanuelle Uro-Coste (CHU de Toulouse/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.