Offre de stage de fin d'étude (Bac+5), mention Informatique/Intelligence Artificielle

> LIEU DU STAGE

INSERM (Institut National de la Santé et de la Recherche Médicale)

CépiDc (Centre d'épidémiologie sur les causes médicales de décès)

80, rue du Général Leclerc, 94270 Le Kremlin-Bicêtre

> INTITULE DU STAGE

Reconnaissance d'écriture médicale à l'aide de méthodes d'intelligence artificielle

> DOMAINE(S) COUVERT(S) PAR LE STAGE

Statistique, Mathématiques, Informatique, Intelligence artificielle, Calcul haute performance

Contexte

Les données sur les causes médicales de décès sont considérées comme les données de santé de référence au niveau national et international. Utilisée par la recherche ainsi que par les multiplies acteurs de la santé publique, l'évaluation de la qualité de ces données est un enjeu primordial. Le CépiDc est l'unité de service Inserm chargée de produire annuellement la Base des Causes Médicales de Décès (BCMD), de diffuser et d'assurer un support technico-scientifique relatif à son exploitation.

La production de ces données contient une première étape de saisie et numérisation des certificats de décès, une seconde étape de codage du texte libre rédigé sur ces certificats, et une troisième étape décisionnelle de sélection de la cause initiale de décès. La seconde et la troisième étape sont aujourd'hui traitées à l'aide d'outils de traitement automatique des langues et d'apprentissage profond (deep learning).

Le présent stage porte sur la première étape de saisie numérique des certificats rédigés par les médecins. Des traitements de l'image ont déjà été développés pour reconnaitre le type de certificat et les principaux items. La reconnaissance de l'écriture médicale à l'aide de méthode d'apprentissage profond sera à développer au cours de ce stage.

Objectifs

Dans un premier temps, le stagiaire aura pour objectifs de s'approprier et d'adapter un jeu de données à l'exploitation par différents types de méthodes d'apprentissage telles que les réseaux de neurones. Dans un second temps, il sera chargé de mettre au point plusieurs méthodes de type apprentissage profond dédiées à la reconnaissance optique de caractères ainsi que d'évaluer leurs performances.

Méthodes

Source : A raison d'environ 500 000 certificats par an sur les 10 dernières années, une base d'apprentissage d'environ 5 millions d'images annotées permet de développer des outils experts avec une précision très élevées.

Résultats attendus :

- Un jeu de données ayant une granularité assez fine pour permettre l'adaptation d'algorithmes d'apprentissage variés.
- L'implémentation d'algorithmes d'apprentissage machine, dont l'apprentissage profond (réseaux à convolutions, modèles seq2seq à attention ou CTC)

Des critères d'évaluation de performance et fiabilité de ces derniers.

Le cas échéant, degré prévisible de confidentialité du rapport de stage

☐ **extrême** ☐ **moyen** ■ **faible**

Connaissances et aptitudes recherchées chez le stagiaire :

Connaissances des outils suivants :

- *Principes de l'apprentissage statistique et applications*
- *Méthodes de traitement d'image*

Aptitudes :

- *Logiciels : Python, openCV, Tensorflow*
- *Aisance en programmation*
- *Manipulation de bases de données volumineuses*
- *Traitement sur données médicales confidentielles*
- *Anglais lu et écrit courant*

> ENVIRONNEMENT DE LA MISSION

Intitulé, activité, compétences statistiques de l'unité d'accueil et du maître de stage :

INSERM-CépiDc (Centre d'épidémiologie sur les causes médicales de décès). Les missions du CépiDc sont :

- la production de la base nationale des causes médicales de décès,
- la diffusion de cette base pour des objectifs de recherche et de santé publique,
- la production d'analyses statistiques et de recherche sur cette base de données.

Cette dernière mission a donné lieu à l'application des méthodologies statistiques adaptées pour de nombreuses publications dans des revues scientifiques internationales.

Le stage sera co-encadré par :

- Walid Ghosn, statisticien-épidémiologiste, diplômé en Ingénierie mathématique et docteur en épidémiologie-santé publique,
- Karim Bounebache, spécialiste en apprentissage statistique, docteur en probabilité.

Il bénéficiera de l'expertise de Louis Falissard, expert en apprentissage profond actuellement en doctorat au CépiDc.

Ressources mises à la disposition du stagiaire :

Le stagiaire disposera d'un bureau, d'un ordinateur puissant, et du logiciel Python. Il aura accès à un serveur de calcul distribué.

La gratification du stage est d'environ 500€ / mois

Durée du stage : 6 mois

> PERSONNE(S) A CONTACTER

M GHOSN Walid,

INSERM-CépiDc

walid.ghosn@inserm.fr

01 49 59 53 37