Fiche de poste : stagiaire informatique niveau M2 en intelligence artificielle/Machine learning/Analyse de données

Sujet du stage : "Utilisation de techniques d'apprentissage automatique dans le contexte de données décrites dans un espace à grande dimension et avec données manquantes, pour le diagnostic/pronostic de la sclérose latérale amyotrophique"

Employeur : ATOS (Orléans)
Grade : Stagiaire ingénieur de 5ème année, ou M2
Rémunération : sera déterminée par ATOS
Environnement hiérarchique : Le stagiaire sera sous la responsabilité d'ATOS et associé au LIFAT/RFAI pour la partie scientifique (http://www.rfai.lifat.univ-tours.fr)
Interlocuteurs : Laurent Garriga (ATOS), Nicolas RAGOT (LIFAT-Tours), Julien Olivier (LIFAT Blois)
Poste à pourvoir : stage à temps plein d'une durée de 4 à 6 mois.

Contexte du stage :
Ce stage s'inscrit dans le cadre du projet OPTIMEDIAS qui vise à développer des outils d'IA pour structurer et exploiter des données de santé massives dans un but d'aide au diagnostic, pronostic et à la décision thérapeutique, notamment pour les Troubles du Neuro-Développement (TND), les Troubles du Spectre de l'Autisme (TSA), la Sclérose Latérale Amyotrophique (SLA) et la Pneumopathie Acquise sous Ventilation Mécanique (PAVM). Ce projet, financé par la région Centre, fait intervenir le CHRU de Tours, l'INSERM, Le LIFAT (Laboratoire d'Informatique Fondamentale et Appliquée de Tours), le LIFO (Laboratoire d'informatique Fondamental d'Orléans) et ATOS. Ce stage, financé par ATOS et codirigé par le LIFAT, donnera la possibilité de candidater pour une poursuite en thèse CIFRE en octobre 2023, toujours dans le cadre du projet OPTIMEDIAS. Le stage se concentrera sur la SLA.

Contexte médical :
La sclérose latérale amyotrophique (SLA) est une maladie neurodégénérative rare et grave qui affecte les neurones moteurs et entraîne la paralysie progressive des muscles volontaires. Le délai diagnostique est estimé à 9 mois et le pronostic est effroyable, avec un décès après en moyenne 36 mois d'évolution imprévisible. La grande diversité de présentations cliniques (les symptômes peuvent être variés et progresser de manière différente d'une personne à l'autre) et l'absence de marqueurs diagnostiques (le diagnostic repose souvent sur l'analyse de plusieurs éléments, tels que les antécédents médicaux de la personne, les résultats de tests neurologiques et de laboratoire, et l'observation des symptômes) rendent le diagnostic de la SLA difficile à établir et tardif, retardent la prise en charge symptomatique. Il existe donc un enjeu important autour de la définition de modèles d'IA capables d'aider à établir le diagnostic de la SLA. Les méthodes de l'état de l'art reposent sur différentes variables, notamment cliniques, telles que l'âge de début de la maladie, la rapidité de progression des symptômes, la présence de troubles respiratoires ou de dysphagie (difficulté à avaler), et la présence de réflexes anormaux. Ces variables peuvent être numériques ou catégorielles. Néanmoins, ces méthodes d'IA restent limitées quant à leurs performances (Grollemund et al. 2019). La nécessité d'utiliser d'autres types de données et des outils d'IA plus avancés est donc une étape cruciale.

Objectifs :
La littérature fait état de quelques biomarqueurs isolés prometteurs mais peu consensuels (Lanznaster et al., 2020) et l'accès aux approches omiques (métabolomique, lipidomique) permettent un enrichissement de la caractérisation biologique des patients. Ainsi, après une étude pilote modeste (Blasco et al., 2010) montrant la spécificité métabolique des patients SLA, et après évolution des méthodes d'acquisition, de traitement de données et d'analyses statistiques, nous avons pu montrer des performances de prédiction diagnostique (Blasco et al., 2014), de caractéristiques phénotypiques et d'évolution de la maladie (Blasco et al., 2017, Blasco et al., 2016) très prometteuses.

L'objectif de ce stage sera donc d'étudier dans quelle mesure un spectre plus large de variables (notamment les données biologiques et omiques collectées par le CHRU peut permettre d'améliorer le diagnostic SLA. Ces données ont déjà été partiellement analysées et traitées lors d'un stage précédent. Les autres difficultés auxquelles nous sommes confrontés sont notamment :
- le nombre de patients (taille des cohortes) limité (1600 patients environ dont 1045 atteints de SLA) par rapport à la taille de l'espace de description potentiel, notamment dans le cadre d'apprentissage automatique et fortiori avec des méthodes type *deep learning*
- le grand nombre de valeurs manquantes ou acquises à des instants différents.
L'objectif sera donc de reprendre ces données et de voir quelles méthodes de l'état de l'art peuvent permettre de rajouter des variables tout en prenant en compte les contraintes ci-dessus (faible nombre de patients et données manquantes (Smieja et al. 2018, Tlamelo et al. 2021). Par extension, on s'intéressera aux méthodes d'apprentissage machine plus récentes comme les mécanismes d'attention par exemple dans les transformers (TIPIRNENI et al. 2022) pour voir l'aptitude de ces modèles à traiter des espaces de description larges.

Méthodologie :
La méthodologie proposée reposera sur les étapes suivantes :

1. Réalisation d'une revue de la littérature approfondie sur les méthodes existantes de d'apprentissage automatique dans le contexte de données à valeurs manquantes, en particulier dans le cadre de l'usage de modèles de *deep learning* avec mécanismes d'attention et transformers.
2. Sélection des méthodes les plus prometteuses.
3. Reprise des données en suivant le processus de préparation et nettoyage préconisé et en intégrant la prise en compte des nouvelles variables.
4. Préparation des données pour l'apprentissage automatique.
5. Implémentation (si nécessaire) et application des méthodes sélectionnées sur les données.
6. Evaluation de l'efficacité des méthodes en termes de précision du diagnostic en fonction des variables utilisées.
7. Mise en forme des livrables facilitant la reprise : code, documentation, rapport de stage détaillant les résultats obtenus et les conclusions tirées de l'étude.

Qualifications souhaitées : le candidat doit être en M2 ou en 5ème année d'un diplôme d'ingénieur dans le domaine de l'informatique avec une expérience solide analyse de

données/statistique et des connaissances en machine learning. Une bonne expérience en deep learning et des outils associés est un véritable plus.

Compétences requises :
· Qualités relationnelles, ouverture et curiosité afin de dialoguer et comprendre les chercheurs en médecine
· Sens de l'initiative et force de proposition
· Sens de l'organisation, autonomie
· Capacité à faire du *reporting*

Candidatures : Lettre de motivation, CV, notes et recommandations par courrier électronique à nicolas.ragot@univ-tours.fr et julien.olivier2@insa-cvl.fr avant le 27/01/2023. Une deuxième phase de sélection aura lieu avec ATOS.

Références :
Grollemund V, Pradat P-F, Querin G,Delbot F, Le Chat G, Pradat-Peyre J-Fand Bede P (2019) Machine Learningin Amyotrophic Lateral Sclerosis:Achievements, Pitfalls, and FutureDirections. Front. Neurosci. 13:135.

Patin F, Corcia P, Vourc'h P, Nadal-Desbarats L, Baranek T, Goossens JF, Marouillat S, Dessein AF, Descat A, Madji Hounoum B, Bruno C, Leman S, Andres CR, Blasco H.Omics to Explore Amyotrophic Lateral Sclerosis Evolution: the Central Role of Arginine and Proline

D. Lanznaster, G. Dingeo, R.A. Samey, P. Emond and H. Blasco. *Metabolomics: A Tool to Understand the Impact of Genetic Mutations in Amyotrophic Lateral Sclerosis*. Metabolites 2022, 12, 864. https://doi.org/10.3390/metabo12090864

Blasco H, Corcia P, Moreau C, Veau S, Fournier C, et al. (2010) *H-NMR-Based Metabolomic Profiling of CSF in Early Amyotrophic Lateral Sclerosis*. PLoSONE 5(10): e13223. doi:10.1371/journal.pone.0013223

Blasco H, Błaszczyński J, Billaut JC, Nadal-Desbarats L, Pradat PF, Devos D, Moreau C, Andres CR, Emond P, Corcia P, Słowiński R. Comparative analysis of targeted metabolomics: dominance-based rough set approach versus orthogonal partial least square-discriminant analysis. J Biomed Inform. 2015 Feb;53:291-9. doi: 10.1016/j.jbi.2014.12.001. Epub 2014 Dec 11. PMID: 25499899.

Patin F, Corcia P, Vourc'h P, Nadal-Desbarats L, Baranek T, Goossens JF, Marouillat S, Dessein AF, Descat A, Madji Hounoum B, Bruno C, Leman S, Andres CR, Blasco H. Omics to Explore Amyotrophic Lateral Sclerosis Evolution: the Central Role of Arginine and Proline Metabolism. Mol Neurobiol. 2017 Sep;54(7):5361-5374. doi: 10.1007/s12035-016-0078-x. Epub 2016 Sep 2. PMID: 27590138.

H. Blasco, P. Vourc'h, P. F. Pradat, P. H. Gordon, C. R. Andres & P. Corcia (2016) Further development of biomarkers in amyotrophic lateral sclerosis, Expert Review of Molecular Diagnostics, 16:8, 853-868, DOI: 10.1080/14737159.2016.1199277

Blasco, H., Patin, F., Madji Hounoum, B., Gordon, P.H., Vourc'h, P., Andres, C.R. and Corcia, P. (2016), Metabolomics in amyotrophic lateral sclerosis: how far can it take us?. Eur J Neurol, 23: 447-454. https://doi.org/10.1111/ene.12956

Marek Smieja, Łukasz Struski, Jacek Tabor, Bartosz Zielinski, and Przemysław Spurek. Processing´ of missing data by neural networks. In Advances in Neural Information Processing Systems, pp. 2719–2729, 2018.

Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago & Oteng Tabona. A survey on missing data in machine learning. Journal of Big Data volume 8, Article number: 140 (2021)

S. TIPIRNENI, C. K REDDY, *Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series*, ACM Trans Knowl Discov Data Vol I. No. 1., 2022, Arxiv :2107.14293v2.