Offre de stage « Interprétation des modèles d'apprentissage profond d'aide au diagnostic appliqués aux données d'électrocardiogrammes »

Mots-clés : apprentissage profond, explicabilité, intégration de connaissances, électrocardiogramme, diagnostic médical

Contexte : L'apprentissage profond (*deep learning*) est une avancée majeure de l'intelligence artificielle (IA) de ces dernières années. Cette approche de l'apprentissage automatique consiste à apprendre à un réseau de neurones de grande taille à réaliser une tâche de prédiction à l'aide d'un ensemble de données d'apprentissage. L'apprentissage profond s'est rapidement imposé comme un standard dans plusieurs domaines en pulvérisant les records des précédentes méthodes de l'état de l'art. Ses domaines de prédilection sont principalement l'analyse d'images et le traitement du langage naturel. Un des futurs enjeux majeurs de cette approche est son application à la santé.
Au sein du projet ANR DeepECG4U, nous nous intéressons à l'application des outils d'IA sur de maladies cardiovasculaires comme le syndrome du QT long pouvant aller jusqu'à la provocation d'une arythmie mortelle, la Torsade de Pointes (TdP). L'objectif est ainsi de diagnostiquer suffisamment tôt le patient pour réduire ce type de risque.

Sujet : Un outil de prédiction à base d'apprentissage profond a déjà été développé [1]. Cependant, l'un des verrous majeurs empêchant son déploiement est son manque d'interprétation. En effet, les modèles d'apprentissage profond ainsi que d'autres méthodes d'apprentissage automatique comme les machines à vecteurs de support, sont considérés comme des « boîtes noires », dans lesquelles les données des patients sont injectées en entrée puis une prédiction est calculée en sortie sans aucune explication. Or, l'Union Européenne a récemment adopté un texte imposant aux utilisateurs d'algorithmes d'apprentissage automatique d'être capables d'expliquer les décisions d'un modèle prédictif [2]. Il y a donc un réel besoin de rendre les modèles d'apprentissage profond plus explicables notamment dans le domaine médical pour deux raisons principales. Premièrement, il est important de s'assurer que le modèle base ses prédictions sur une représentation fiable des patients et ne se concentre pas sur des artefacts non pertinents présents dans les données d'apprentissage. Sans explications des prédictions, les médecins ne peuvent pas faire confiance au modèle quelques soit ses performances. Deuxièmement, un modèle performant pour la prédiction d'une certaine maladie, peut avoir identifié une signature dans les données qui pourrait être une piste de recherche pour les médecins.
Pour répondre à ces enjeux, des méthodes d'interprétation ont été développées pour interpréter à posteriori des modèles d'apprentissage profond déjà appris. Elles consistent généralement à identifier les parties de l'entrée du modèle ayant influencé le plus sa sortie [3]. Dans le cadre du projet, nous avons développé une première approche inspirée des travaux de Been Kim et al. [4], qui essaye d'identifier d'une part si le modèle parvient à capturer des concepts basés sur des annotations des données que les experts ont l'habitude d'utiliser, et d'autre part, leurs possibles liens d'implication avec les sorties du modèle.
Une autre approche consisterait à forcer le modèle à apprendre ces concepts [5] et ainsi faire du tout-en-un sans avoir à utiliser une méthode à posteriori. L'objectif de ce stage est donc d'explorer cette piste.

Le déroulement sera le suivant :

- Lecture et synthèse de l'état de l'art scientifique du domaine
- Sélection et adaptation des méthodes les plus pertinentes pour l'application
- Programmation de la solution proposée en Python / PyTorch de préférence
- Test et évaluation sur jeux de données publiques et privés
- Comparaison avec l'approche a posteriori
- Rédaction d'un article scientifique si les résultats sont concluants

Profil recherché :

- Master 2 ou dernière année d'école d'ingénieur en informatique ou mathématiques appliquées.
- De solides compétences en apprentissage profond sur une diversité d'architectures et de problématiques appliquées aux données (signaux, images, texte, etc.)
- Connaissances des librairies de *machine learning* (Scikit-learn) et de *deep learning* (PyTorch)
- Si possible un goût et une expérience sur les problématiques méthodologiques liées au domaine de la santé
- Motivé·e par la recherche scientifique
- Travail en équipe, communication et créativité

Cadre :

- Encadrement : Victoria Bourgeais (MCF, LaBRI) et Blaise Hanczar (PU, IBISC, Paris-Saclay)
- Lieu du stage : Laboratoire LaBRI à Bordeaux (33). 1 à 2 déplacements professionnels seront éventuellement à prévoir en région parisienne pour rencontrer et faire l'état d'avancement avec le reste des membres du projet ANR
- Durée : 5-6 mois (flexible, idéalement commençant en février-mars 2024)
- Gratification : environ 600€/mois

Modalités de candidature : envoyer son CV, une lettre de motivation et ses deux derniers bulletins de notes à victoria.bourgeais@u-bordeaux.fr

---

Références :

[1] Prifti, E., Fall, A., Davogustto, G., Pulini, A., Denjoy, I., Funck-Brentano, C., Khan, Y., Durand-Salmon, A., Badilini, F., Wells, Q. S., Leenhardt, A., Zucker, J.-D., Roden, D. M., Extramiana, F., & Salem, J.-E. (2021). Deep learning analysis of electrocardiogram for risk prediction of drug-induced arrhythmias and diagnosis of long QT syndrome. *European Heart Journal*, *42*(38), 3948-3961. https://doi.org/10.1093/eurheartj/ehab588

[2] Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". AI Magazine, 38(3), 50-57. https://doi.org/10.1609/aimag.v38i3.2741

[3] Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2019). Gradient-Based Attribution Methods. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Éds.), *Explainable AI: Interpreting,*

Explaining and Visualizing Deep Learning (p. 169-191). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_9

[4] Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., & Sayres, R. (2018). Interpretability Beyond Feature Attribution : Quantitative Testing with Concept Activation Vectors (TCAV). In J. G. Dy & A. Krause (Éds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (Vol. 80, p. 2673-2682). PMLR. http://proceedings.mlr.press/v80/kim18d.html

[5] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept Bottleneck Models. *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, *119*, 5338-5348. http://proceedings.mlr.press/v119/koh20a.html