Université Claude Bernard Lyon 1

DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **11 octobre 2024**

Nom de famille et prénom de l'auteur : **Monsieur Jean-Baptiste GUIMBAUD**

Titre de la thèse : Amélioration de Scores de Risque Environmental par Machine Learning Informé et AI Explicable

Résumé

Dès la conception, des facteurs environnementaux tels que la qualité de l'air ou les habitudes alimentaires peuvent significativement influencer le risque de développer diverses maladies chroniques. Dans la littérature épidémiologique, des indicateurs connus sous le nom de Scores de Risque Environnemental (Environmental Risk Score, ERS) sont utilisés non seulement pour identifier les individus à risque, mais aussi pour étudier les relations entre les facteurs environnementaux et la santé. Une limite de la plupart des ERSs est qu'ils sont exprimés sous forme de combinaisons linéaires d'un nombre limité de facteurs. Cette thèse de doctorat vise à développer des indicateurs ERSs capables d'investiguer des relations non linéaires et des interactions à travers un large éventail d'expositions tout en découvrant des facteurs actionnables pour guider des mesures et interventions préventives, tant chez les adultes que chez les enfants. Pour atteindre cet objectif, nous exploitons les capacités prédictives des méthodes d'apprentissage automatique non paramétriques, combinées avec des outils récents d'IA explicable et des connaissances existantes du domaine. Dans la première partie de cette thèse, nous calculons des scores de risque environnemental basés sur l'apprentissage automatique pour la santé mentale, cardiométabolique et respiratoire de l'enfant. En plus d'identifier des relations non linéaires et des interactions entre expositions, nous avons identifié de nouveaux prédicteurs de maladies chez les enfants. Les scores peuvent expliquer une proportion significative de la variance des données et leurs performances sont stables à travers différentes cohortes. Dans la deuxième partie, nous proposons SEANN, une nouvelle approche intégrant des connaissances expertes sous forme d'Effet Agrégées (Pooled Effect Size, PES) dans l'entraînement de réseaux neuronaux profonds pour le calcul de scores de risque environnemental informés (Informed ERS). SEANN vise à calculer des ERSs plus robustes, généralisables à une population plus large, et capables de capturer des relations d'exposition plus proches de celles connues dans la littérature.

Nous illustrons expérimentalement les avantages de cette approche en utilisant des données synthétiques. Par rapport à un réseau neuronal agnostique, nous obtenons une meilleure généralisation des prédictions dans des contextes de données bruitées et une fiabilité améliorée des interprétations obtenues en utilisant des méthodes d'Intelligence Artificielle Explicable (Explainable AI - XAI). Dans la dernière partie de cette thèse, nous proposons une application concrète de SEANN en utilisant les données d'une cohorte espagnole composée d'adultes. Comparé à un score de risque environnemental basé sur un réseau neuronal agnostique, le score obtenu avec SEANN capture des relations mieux alignées avec les associations de la littérature sans détériorer les performances prédictives. De plus, les expositions ayant une couverture littéraire limitée diffèrent significativement de celles obtenues avec la méthode agnostique de référence en bénéficiant de directions d'associations plus plausibles. En conclusion, nos scores de risque démontrent un indubitable potentiel pour la découverte informée de relation environnement-santé non linéaires peu connues, tirant parti des connaissances existantes sur les relations bien connues. Au-delà de leur utilité dans la recherche épidémiologique, nos indicateurs de risque sont capables de capturer, de manière holistique, des relations de risque au niveau individuel et d'informer les praticiens sur des facteurs de risque actionnables identifiés. Alors que dans l'ère post-génétique, la prévention en médecine personnalisée se concentrera de plus en plus sur les facteurs non héréditaires et actionnables, nous pensons que ces approches seront déterminantes pour façonner les futurs paradigmes de la santé.

Mots-clés : Exposome,IA explicable,Scores de risques environnementaux,Machine Learning Informé,Réseaux de Neurones Profonds,