

Monsieur Hadrien SALEM

Spécialité "Informatique, Automatique"

Soutiendra publiquement ses travaux de thèse intitulés

intelligence artificielle pour le dépistage précoce de maladies à partir de données biologiques

dirigés par Monsieur Slim HAMMADI

Soutenance prévue le **lundi 15 décembre 2025** à 10h00

Lieu : Avenue Paul Langevin, Cité Scientifique 59650 Villeneuve-d'Ascq, École Centrale de Lille, Amphi Cuccaroni

Salle : Amphi Cuccaroni

Composition du jury proposé

M. Slim HAMMADI	Centrale Lille Institut	Directeur de thèse
M. Nicolas LOMÉNIE	Université Paris Cité	Rapporteur
M. Jean-Paul BOUFFLET	Université de technologie de Compiègne	Rapporteur
Mme Sarah BEN OTHMAN	Université de Lille	Co-encadrante de thèse
M. Emmanuel CHAZARD	Université de Lille	Examineur
M. Salah ZIDI	Université de Gabes	Examineur
M. Marc BROUCQSAULT	Groupe MLAB	Invité

Mots-clés : Système d'aide à la décision, Intelligence Artificielle, Réseaux de neurones, Données biologiques, Apprentissage automatique,

Résumé :

La détection précoce des maladies dans des populations asymptomatiques représente un enjeu majeur et un levier déterminant pour l'amélioration de la santé publique mondiale. Bien que les systèmes de santé modernes produisent d'immenses volumes de données cliniques et biologiques, leur exploitation reste encore largement limitée pour le dépistage systématique. Les méthodes d'apprentissage automatique (ML) offrent aujourd'hui une opportunité sans précédent : elles permettent d'extraire des signaux faibles à partir de données massives et hétérogènes, ouvrant la voie à une identification plus rapide et plus précise des patients à risque, bien au-delà des approches traditionnelles. Ces avancées méthodologiques favorisent non seulement des suivis cliniques plus ciblés et des interventions précoces, mais également une optimisation des coûts de prise en charge, conciliant impact médical et efficacité économique. Cette thèse explore de manière approfondie le potentiel du ML pour la prédiction précoce des maladies à partir de données recueillies en routine, et en particulier de données biologiques. Dans une première étude, nous avons démontré qu'il est possible de prédire des valeurs anormales de Prostate-Specific Antigen (PSA) sans le mesurer directement, à partir de biomarqueurs courants. Cette approche novatrice établit la faisabilité de la détection de pathologies via des marqueurs indirects, ouvrant des perspectives pour un dépistage non invasif et à moindre coût. L'intégration d'une analyse longitudinale a ensuite permis de tirer parti des variations temporelles des biomarqueurs, malgré les difficultés liées à leur échantillonnage irrégulier. Pour surmonter ces limites, nous avons conçu un réseau neuronal hybride associant couches convolutionnelles et involutionnelles, capable de capter à la fois les dépendances locales et globales dans des séries temporelles complexes. Sur le cas de la maladie rénale chronique (CKD), nos modèles ont atteint des performances remarquables (AUROC = 90%, F1-score = 77%, rappel = 84%) jusqu'à 15 mois avant la progression vers les stades avancés. De façon surprenante, un simple perceptron multicouche (MLP) s'est révélé plus performant que des architectures plus sophistiquées, soulignant l'importance d'une conception pragmatique des modèles, où la simplicité peut se traduire par une meilleure généralisation. Dans un second temps, nous avons développé MIPHA (Modular data Integration for Predictive Healthcare Analysis), un framework innovant et flexible pour la conception, l'évaluation et le partage de modèles ML en santé. Grâce à ses composants modulaires, MIPHA facilite le prototypage rapide, la fusion multimodale et l'apprentissage par transfert. À travers ce cadre, nous avons mené de multiples expérimentations démontrant notamment que l'intégration de caractéristiques tirées d'ECG améliore la prédiction précoce du diabète de type 2. Cette approche a également révélé la pertinence de l'architecture multi-résolutions exploitant plusieurs Feature Extractors, ainsi que l'intérêt, mais aussi les limites, de l'apprentissage par transfert : efficace entre tranches d'âge dans des sous-populations restreintes, mais moins probant entre maladies ou bases de données distinctes. Enfin, l'ensemble de ces travaux s'appuie sur une base de données exceptionnelle, issue d'un laboratoire français, comprenant près de deux milliards d'analyses biologiques anonymisées collectées sur plus de douze ans. Ce dataset riche confère une grande robustesse aux résultats obtenus et illustre le potentiel immense de l'exploitation des données de routine pour transformer la pratique clinique. Cette thèse met en lumière le rôle clé de l'apprentissage automatique dans la médecine prédictive. Elle souligne que l'intégration du ML dans les systèmes de santé pourrait mener à des dépistages préventifs plus performants et plus accessibles.