Université Lille Nord de France
Pôle de Recherche et d'Enseignement Supérieur

Ecole doctorale régionale Sciences Pour l'Ingénieur
Lille Nord-de-France - 072

Sujets de thèse 2022

Titre : Intelligence Artificielle pour l'analyse prédictive des maladies à partir des données biologiques

Financement prévu : *Cocher au moins une des cases*

Contrat Doctoral (Ecole Centrale ☒ Université de Lille ☒ Président Univ Lille ☐)

Contrat Région ☐ avec co-financement: …… (à préciser)

ANR ☐ CIFRE ☐ DGA ☐ ADEME ☐ Co-tutelle ☐ : ……(à préciser)

Autre ☐ : …… (à préciser)

Directeur de thèse : Slim Hammadi
E-mail : slim.hammadi@centralelille.fr

Co-directeurs ou co-encadrants de thèse : Sarah Ben Othman , Marc Broucqsault
E-mail : sara.ben-othman@centralelille.fr , mbroucqsault@Altao.com

Laboratoire : CRIStAL – UMR 9189

Groupe Thématique : OPTIMA **Equipe : OSL**

Domaine de l'EDSPI : Automatique (AGITSI) ☒ **Informatique** ☐

Contexte et objectifs de la thèse

L'équipe OSL/CRISTAL travaille actuellement sur plusieurs projets dans le domaine de la santé en collaboration avec ALTAO, une entreprise spécialiste de l'introduction des innovations dans le milieu hospitalier. Cette entreprise encadre plusieurs projets visant à digitaliser les outils de la santé. Ce sujet de thèse consiste en la ***détection ultra précoce de maladies*** sur base des données de biologie sur plusieurs années grâce à des algorithmes d'Intelligence Artificielle. Ces données de biologie anonymes proviendront d'un grand laboratoire de biologie BIOPATH qui proposera également deux biologistes pour accompagner le doctorant sur ce sujet. L'objectif principal de ce sujet de thèse est de concevoir et développer un système d'aide à la détection précoce des maladies (SADUM) à partir d'examens biologiques de routine, et grâce à un entrepôt de données biologiques qui permet les apprentissages de algorithmes.
Concevoir et développer un SADUM, contenant des applications informatiques, pourra fournir aux biologistes et aux médecins au moment voulu les informations

décrivant précocement les risques de survenue de maladie ainsi que les connaissances appropriées, correctement filtrées.
Un énorme volume de données médicales est développé et conservé quotidiennement par les laboratoires d'analyse médicale. Avec le développement des plateformes bio-informatiques modernes, il est devenu essentiel pour les études biomédicales d'adopter une approche intégrative (combinée) afin d'utiliser pleinement ces données pour mieux comprendre l'origine et l'évolution de certaines maladies. Les valeurs mesurées font toutes partie du même pipeline d'informations, dont la sortie dépend des différentes entrées et de la régulation. Ces données peuvent être exploitées pour comprendre et analyser les caractéristiques et les complexités sous-jacentes des pathologies et leurs impacts sur le corps humain à l'aide d'algorithmes prédictifs basés sur l'apprentissage automatique. Ce dernier offre de nouvelles techniques pour intégrer et analyser les diverses données de santé, ce qui permet de déceler de nouveaux ensembles d'indicateurs. Ces ensembles d'indicateurs ont le potentiel de contribuer à la prédiction précise des maladies, à la stratification des patients et à la mise en place d'une médecine de précision.
Différentes méthodes d'apprentissage automatique intégratives peuvent être utilisées pour fournir une compréhension approfondie des indicateurs biologiques pendant le fonctionnement physiologique normal et en présence d'une maladie afin de fournir un aperçu et des recommandations aux professionnels interdisciplinaires. Certaines maladies sont étroitement liées à la génétique et à l'épigénétique, mais les mécanismes permettant de clarifier l'apparition et/ou la progression de la maladie n'ont parfois pas été entièrement maîtrisés. Ces dernières années et grâce au grand nombre d'études récentes, on sait que les modifications de l'équilibre des indicateurs biologiques peuvent être le signe d'une panoplie de maladies.
En pratique, la décision médicale est un processus complexe intégrant des données hétérogènes quant à leur origine (signes cliniques, biologiques, imagerie…) leur nature (objectives, subjectives interprétatives), leur degré de certitude et de péremption, et avant toute chose leur disponibilité conditionnelle (certains tests biologiques ne sont pas systématiquement réalisés, et cette réalisation n'est aucunement aléatoire). L'art médical consiste en une intégration de ces informations, volontiers contradictoires, incertaines ou manquantes. Cette intégration vise simultanément à porter un diagnostic ou prédire une pathologie et, au fil de l'eau, prescrire des examens complémentaires pour conforter ou écarter itérativement ce diagnostic. Cet art s'appuie de manière variable sur des connaissances et recommandations. Dans ce contexte d'incertitude permanente, les médecins et les biologistes peuvent se tromper tant dans l'interprétation des résultats des examens biologiques que dans l'orientation diagnostique.
Des travaux antérieurs ont atteint une certaine précision pour la prédiction de certaines maladies à partir des résultats des examens biologiques en utilisant les

différentes techniques d'apprentissage automatique, telles que : les réseaux de neurones, le Deep Learning, le classifieur SVM, etc. Cependant, les profils à risque pour certaines maladies et les résultats précis restent plutôt difficiles à obtenir.
Le but de ce sujet de thèse est de concevoir et développer un SADUM (Système d'Aide à la Détection Ultra-précoce des Maladies) basé sur une ontologie pour un diagnostic, un pronostic et une prédiction des maladies à partir des données biologiques des patients à l'aide d'une technique de clustering. Cette technique permettra d'identifier des modèles complexes et non linéaires d'expression et de relations au sein de l'ensemble des données afin d'extraire des connaissances intrinsèques sans aucune hypothèse biologique sur les données. Les ontologies médicales devraient contribuer à l'utilisation efficace des ressources d'informations médicales qui stockent une quantité considérable de données. Et pour cela l'ontologie ainsi conçue permettrait de modéliser et représenter les informations nécessaires au diagnostic et à la prédiction de certaines maladies.

Mots clés : Système d'aide à la décision, Intelligence artificielle, optimisation, données biologiques, ontologie.

Le programme et l'échéancier de travail du doctorant :

Ce sujet de thèse a pour but d'étudier, de concevoir et de développer un SADUM utilisant des approches alliant l'IA et la RO. Ce SADUM propose des modèles de clustering pour certaines maladies et aide à la décision thérapeutique en se basant sur les ontologies en utilisant des techniques de data mining et de raisonnement automatique.

Le programme et l'échéancier de travail pourraient dès lors porter sur :

- Etat de l'art sur les méthodes et les ontologies pour la recherche, la contextualisation et l'intégration de données médicales hétérogènes et distribuées, méthodes de fouille de données et d'extraction des connaissances. **(6 mois)**

- Etude et modélisation d'un SADUM basées sur les systèmes médicaux, les techniques d'apprentissage non supervisé (clustering) et les ontologies pour l'identification des classes de maladies. Établissement d'une formalisation de connaissances et le développement de l'ontologie de domaine. **(4 mois)**

- Conception et réalisation d'un SADUM. Ce système devra naturellement prendre en compte les spécificités du profil à risque ou bien de la maladie étudiée et exploiteront les technologies informatiques avancées (IA, optimisation,

apprentissage). Ce système d'aide doit s'adapter aux spécificités des patients concernés. **(18 mois)**

- Modèles et méthodes de simulation et résultats en tenant compte des ressources matérielles et humaines. Le but principal du SADUM est de fournir aux professionnels de soins et de santé (biologistes et médecins) la situation prédictive de l'état de santé d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées. **(8 mois)**

6 dernières publications dans le domaine de la santé

[1]	Fakhfakh K., Ben Othman S., Zgaya H., Jourdan L., Smith G., Renard J-M., Hammadi S. (2021) Fuzzy Ontology for Patient Emergency Department Triage. In: Paszynski M., Kranzlmüller D., Krzhizhanovskaya V.V., Dongarra J.J., Sloot P.M. (eds) Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science, vol 12744. Springer, Cham. https://doi.org/10.1007/978-3-030-77967-2_60
[2]	Fakhfakh K., Ben Othman S., Zgaya H., Jourdan L., Smith G., Renard J-M., Hammadi S. (2021) Ontology for Overcrowding Management in Emergency Department, MEDINFO 2021 (In proceedings).
[3]	Ajmi F., Zgaya H., Ben Othman S., Hammadi S., "Agent-Based dynamic optimization for the patient pathway workflow management", Journal : Simulation Modelling Practice and Theory, SIMPAT-D-18—695, 2019. HAL Id : hal-02439090
[4]	Ben Othman S., Ajmi F., Zgaya H., Hammadi S. (2018). "Cubic Chromosome representation for Patient'Sceduling in the Emergency Department", accepté pour publication dans RAIRO – Operation Research journal, EDP Sciences, 2018. https ://hal.archives-ouvertes.fr/hal-017
[5]	Ben Othman S., Zgaya H., Dotoli M., Hammadi S. (2017). "An Agent-Based Decision Support System for Resources' Scheduling in Emergency Sup Chains", Control Engineering Practice 2017, Volume 59, pp 27-43, IF : 2,602. doi http://dx.doi.org/10.1016/j.conengprac.2016.11.014
[6]	Ben Othman S., Hammadi S. (2017). "A Multi-criteria Optimization Approach to Health Care task Scheduling Under Resources Constrain, International journal of Computational Intelligence Systems (IJCIS). www.atlantis-press.com/php/download_paper.php?id=25865516