

# Classification multi-label de données médicales par LSTM temporel et clustering flou

Abdelhamid Gaddari<sup>\*,\*\*</sup>, Haytham Elghazel<sup>\*</sup>, Rakia Jaziri<sup>\*\*\*</sup>, Mohand Hacid-Saïd<sup>\*</sup>  
Pierre-Henri Comble<sup>\*\*</sup>

<sup>\*</sup> Université Lyon 1, LIRIS, UMR CNRS 5205, F-69622

prenom.nom@liris.cnrs.fr

<sup>\*\*</sup> CEGEDIM Insurance Solutions, France

prenom.nom@cegedim.com

<sup>\*\*\*</sup> Université Paris 8, Laboratoire Paragraphe, EA 349

rjaziri@univ-paris8.fr

**Résumé.** La prévention de la santé est un aspect très important de la recherche en informatique médicale grâce à la prédiction d'événements médicaux. Dans ce travail, nous proposons une approche d'apprentissage profond pour effectuer des prédictions multi-label sur les actes de soins médicaux. L'approche proposée utilise un réseau LSTM temporel (*time-aware long short-term memory*) et l'étend avec des informations supplémentaires à partir d'un clustering flou du même portefeuille. Le premier mécanisme (*time-aware*) est utilisé pour gérer l'irrégularité temporelle entre les éléments d'une trajectoire médicale, tandis que le second mécanisme (*fuzzy clustering*) aide à modéliser l'hétérogénéité entre les patients et les traitements. En utilisant un large portefeuille d'actes médicaux remboursés (plus de 16 millions d'actes médicaux consommés) par une assurance médicale en France, nous montrons que notre approche surpasse les méthodes traditionnelles et d'apprentissage profond dans la classification médicale multi-label. Notre travail a pour objectif de soutenir la prévention médicale et plus largement améliorer la qualité des services de santé et des assurances.

## 1 Introduction

La recherche en informatique médicale a connu un développement rapide au cours de ces dernières années, et plus particulièrement l'utilisation de modèles d'apprentissage automatique pour différentes tâches médicales (Hong et al., 2018), telles que l'évolution de maladies ou la prédiction de diagnostics. Dans ce papier, nous sommes confrontés à un sujet commun de la recherche biomédicale : la prédiction multi-label des actes de soins médicaux. Nous utilisons les données d'une assurance médicale privée en France, représentant tous les actes de soins médicaux consommés par les assurés. Cependant, nous devons faire face au défi de l'irrégularité temporelle qui réside dans les données de santé, puisque les intervalles entre les différents actes de soins varient de quelques jours à plusieurs mois et dépendent de la pathologie médicale concernée. Un autre défi motivant notre travail est l'hétérogénéité que présentent nos données, puisque le portefeuille est composé d'assurés de tous les âges et avec des situations médicales