

أَلْجُمْهُورِيَّةُ الْجَزَائِرِيَّةُ الدِّيمُقْرَاطِيَّةُ الشَّعْبِيَّةُ

République Algérienne Démocratique et Populaire

وَزَارَةُ التَّعْلِيمِ الْعَالِي وَالْبَحْثِ الْعِلْمِيِّ

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

الْمَدْرَسَةُ الْعُلْيَا لِلْإِعْلَامِ الْإِلَهِيِّ - 08 مَآيُو 1945 - بَسِيدِي بَلْعَبَّاس

Ecole Supérieure en Informatique

-08 Mai 1945- Sidi Bel Abbès



MÉMOIRE DE MASTER

En Vue de l'obtention du diplôme de **Master**

Filière : **Informatique**

Spécialité : **Intelligence Artificielle et Science des Données (IASD)**

Thème

Design of a Hybrid Clinical Pre-Analysis Module for Rare Disease Diagnostic Support

Présenté par :

- Mr **Ourred** Islem Charaf Eddine
- Mr **Ouddane** Youcef Fahed

Soutenu le : **07/07/2026**

Devant le jury composé de :

- | | |
|----------------------------------|--------------|
| • Dr. Nemdili Amel Kahina | Présidente |
| • Dr. Belmekki Ghizlene | Examinatrice |
| • Dr. Mezrar Samiha | Encadreur |
| • Pr. Keskes Nabil | Co-Encadreur |
| • Dr. Kornyshova Elena | Co-Encadreur |
| • Dr. Pons Olivier | Co-Encadreur |

Année Universitaire : 2025 / 2026

ملخص

تصيب الأمراض النادرة أكثر من 400 مليون شخص حول العالم وتتميز برحلة تشخيصية طويلة يبلغ متوسطها 4.7 سنوات. البيانات السريرية في السجلات الطبية الإلكترونية غير منظمة في غالبيتها، مما يجعل من الصعب على المتخصصين في الرعاية الصحية استخراج وتنظيم المعلومات ذات الصلة اللازمة للتشخيص.

تقترح هذه المذكرة تصميم وحدة تحليل أولي سريري هجينة تجمع بين الفهم السياقي لنماذج اللغة الطبية الحيوية المدربة مسبقاً ودقة الأنظمة القائمة على القواعد والأنطولوجيات الطبية المعيارية. ينفذ النظام خط أنابيب كاملاً يعالج النص السريري الخام عبر خمس مراحل: المعالجة المسبقة مع كشف اللغة وإزالة التعريف، والتعرف ثنائي اللغة على الكيانات المسماة باستخدام مجموعة من النماذج المتخصصة، والتحقق السريري مع كشف النفي، والتطبيع متعدد المستويات وفقاً لأنطولوجيات HPO و ORDO و UMLS، وتوليد المخرجات المهيكلة وفقاً لمعيار Phenopackets الدولي -- ISO 4454.

تم تقييم النظام على 112 حالة سريرية عبر خمس مجموعات بيانات، بما في ذلك 17 تقريراً طبياً حقيقياً من مستشفى الأطفال بكنستال بوهران. على هذه التقارير السريرية الفرنسية، حقق النظام درجة F1 بنسبة 60.0% لاستخراج الأنماط الظاهرية، ودقة 88.2% لتحديد الأمراض، ومعدل 100% لتوليد حزم الأنماط الظاهرية.

الكلمات المفتاحية: معالجة اللغة الطبيعية، الأمراض النادرة، التعرف على الكيانات المسماة، التحليل الأولي السريري، الأنظمة الهجينة، HPO، ORDO، UMLS، Phenopackets، SapBERT، ClinicalBERT.

Résumé

Les maladies rares touchent plus de 400 millions de personnes dans le monde et se caractérisent par un parcours diagnostique prolongé dont la durée moyenne est de 4,7 ans. Les données cliniques contenues dans les dossiers médicaux électroniques sont majoritairement non structurées, ce qui rend difficile pour les professionnels de santé l'extraction et l'organisation efficaces des informations pertinentes nécessaires au diagnostic.

Ce mémoire propose la conception d'un **module de pré-analyse clinique hybride** combinant la compréhension contextuelle de modèles de langage biomédicaux pré-entraînés avec la précision de systèmes à base de règles et d'ontologies médicales standardisées. Le système implémente un pipeline complet qui traite le texte clinique brut à travers cinq étapes : le prétraitement avec détection de langue et dé-identification, la reconnaissance d'entités nommées bilingue utilisant un ensemble de modèles spécialisés, la validation clinique avec détection de négation, la normalisation multi-niveaux selon les ontologies HPO, ORDO et UMLS, et la génération de sortie structurée suivant le standard Phenopackets (ISO 4454).

Le système a été évalué sur 112 cas cliniques répartis sur cinq jeux de données, incluant 17 rapports hospitaliers réels de l'Hôpital pédiatrique de Canastel d'Oran. Sur ces rapports cliniques français, le système atteint un score F1 de 60,0% pour l'extraction de phénotypes, une précision de 88,2% pour l'identification des maladies, et un taux de 100% pour la génération de Phenopackets.

Mots-clés : Traitement Automatique du Langage Naturel, Maladies Rares, Reconnaissance d'Entités Nommées, HPO, ORDO, UMLS, Phenopackets, ClinicalBERT, SapBERT, Pré-analyse Clinique, Systèmes Hybrides.

Abstract

Rare diseases affect over 400 million people worldwide and are characterized by a prolonged diagnostic journey that averages 4.7 years. Clinical data in Electronic Medical Records is predominantly unstructured, making it difficult for healthcare professionals to efficiently extract and organize the relevant information needed for diagnosis.

This thesis proposes the design of a **hybrid clinical pre-analysis module** that combines the contextual understanding of pre-trained biomedical language models with the precision of rule-based systems and standardized medical ontologies. The system implements a complete pipeline that processes raw clinical text through five stages: preprocessing with language detection and de-identification, bilingual Named Entity Recognition using an ensemble of specialized models, clinical validation with negation detection, multi-level normalization against HPO, ORDO, and UMLS ontologies, and structured output generation following the Phenopackets standard (ISO 4454).

The system was evaluated on 112 clinical cases across five datasets, including 17 real hospital reports from the Hôpital pédiatrique de Canastel (Children’s Hospital) of Oran. On these French clinical reports, the system achieved an F1 score of 60.0% for phenotype extraction, 88.2% accuracy for disease identification, and 100% successful Phenopacket generation.

Keywords: Natural Language Processing, Rare Diseases, Named Entity Recognition, HPO, ORDO, UMLS, Phenopackets, ClinicalBERT, SapBERT, Clinical Pre-analysis, Hybrid Systems.