UNIVERSITI TEKNOLOGI MARA

A HYBRID NATURAL LANGUAGE PROCESSING METHOD FOR INTERPRETABLE ICD CLASSIFICATION FROM ELECTRONIC MEDICAL RECORDS CLINICAL NOTES

NURUL ANIS BALQIS BINTI IQBAL BASHEER

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Computer Science)

Faculty of Computer and Mathematical Sciences

December 2025

ABSTRACT

Accurate interpretation of Electronic Medical Records (EMRs), especially clinical notes, is crucial for effective healthcare communication and achieving accurate patient outcomes. The main challenges in hybrid Natural Language Processing (NLP) methods include integrating various techniques while maintaining contextual understanding, resolving ambiguous abbreviations, and reducing misinterpretations of clinical narratives. The dataset in this study consisted of cardiovascular-related clinical notes containing medical abbreviations, diagnoses, and discharge summaries. Before analysis, the data underwent preprocessing steps such as text normalization, abbreviation extraction, and punctuation cleaning to ensure consistency and readiness for the model. This study addresses abbreviation ambiguity, diagnosis prediction, and International Classification of Diseases (ICD) classification using a hybrid NLP approach. The objectives are to extract and expand abbreviations, develop a hybrid framework for diagnosis prediction and ICD mapping, and evaluate its performance. The methodology integrates the Text-to-Text Transfer Transformer (T5) model with enhanced inference combining cosine similarity and beam search for abbreviation expansion, MedBioClinicalBERT, an integration of BioClinicalBERT and MedBERT for diagnosis prediction, and Semantic Role Labeling (SRL) for explainability. The enhanced elicitive inference achieved 95.38% BLEU and 97.96% ROUGE-L scores on abbreviation expansion. For diagnosis prediction, the hybrid input framework with MedBioClinicalBERT attained 90.00% accuracy with precision, recall, and F1 scores of 0.9530, 0.9470, and 0.9000, respectively, outperforming BioClinicalBERT and MedBERT individually. Standardization to ICD-10 codes was refined using fuzzy matching to improve mapping accuracy. The overall performance for the hybrid NLP method is 94.89% of precision, 94% of recall, and 95% of F1 score. Although limitations persist due to the multimodal data nature of clinical notes and the cardiovascular-specific dataset, the proposed method demonstrates substantial improvements. Overall, this study highlights the effectiveness of combining hybrid NLP methods with advanced abbreviation expansion to enhance EMR interpretation and ICD-10 classification, paving the way for broader applications in medical text analysis.

ACKNOWLEDGEMENT

Firstly, I wish to thank God for giving me the opportunity to embark on my master's and for completing this long and challenging journey successfully. My gratitude and thanks go to my supervisors, Dr. Sharifalillah Nordin and Madam Nurzeatul Hamimah Abdul Hamid, and my domain experts, Prof. Dr. Sazzli Shahlan Kasim.

My appreciation goes to the Hospital Al-Sultan Abdullah (HASA), which assisted with data collection and evaluation. Special thanks to my colleagues and friends for helping me with this project.

Finally, this thesis is dedicated to the loving memory of my very father for his vision and determination to educate me. This piece of victory is dedicated to him. Alhamdulilah.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Research Background

Electronic Medical Records (EMRs) are the digital versions of traditional paper charts used in doctors' offices, clinics, and hospitals. They store patients' medical information, including clinical notes, diagnoses, and treatment histories, which are recorded and used by healthcare professionals to support diagnosis, treatment, and ongoing care (Adamson et al., 2023; Jia et al., 2024; Merchant et al., 2024). EMRs contain multimodal data like structured and unstructured data, which require different methods of processing and analysis (Anshari, 2019; Jia et al., 2024; Gu et al., 2025). Structured data refers to data that has a predefined format, such as codes, dates, and numbers (Tayefi et al., 2021). Unstructured data refers to data that has no fixed format, such as free-text notes, images, and audio, which are usually called clinical notes (Benicio et al., 2021; Swaminathan et al., 2023). Unstructured data accounts for about 80% of the data in EMRs, but it is often underutilized or ignored due to its complexity and variability (Adamson et al., 2023; Kong 2019; Merchant et al., 2024; Gu et al., 2025). Unstructured clinical notes often contain numerous clinical abbreviations, which are highly context-dependent and can vary between specialties, making automatic interpretation and ICD classification difficult (Rajkomar et al., 2022; Zaretsky et al., 2024). Misinterpreting these abbreviations can lead to incorrect diagnosis coding and hinder automated mapping to ICD classifications. Therefore, accurately expanding and interpreting abbreviations in EMRs is a crucial step toward improving the reliability of diagnosis prediction and ICD code assignment (Merchant et al., 2024; Lin et al., 2025).

Furthermore, knowing the International Classification of Diseases (ICD) code for a specific disease allows clinicians to quickly access relevant information about the condition (Yan et al., 2022). This can aid in making timely and informed decisions about patient care, treatment plans, and referrals (Dai et al., 2024). This is due to the large volume of unstructured data in EMRs, which creates difficulties for critical tasks such as automated diagnosis generation and ICD classification, which are crucial for patient care, billing, and healthcare management (Merchant et al., 2024). The International Classification of Diseases (ICD) is the globally recognized system developed by the