CLEM-ICD - Clinical Language Encoding with ModernBERT

Using ModernBERT to Improve Automated ICD-10 Classification

Tyler Cross[a,1,*]

[a] *University of California, Berkeley, School of Information, 102 South Hall, Berkeley, 94720*

Abstract

Medical coding is a critical yet complex process that requires translating clinical documentation into standardized ICD-10 codes. Currently, healthcare providers rely on dedicated medical coders who manually review extensive clinical notes to assign appropriate codes from 150,000+ possible diagnostic and procedural codes. This project builds upon the work of Edin et al. (2024)] with two key innovations: (1) replacing RoBERTa with ModernBERT (Warner et al., 2024) as the encoder to potentially improve performance through its enhanced context window, and (2) adding a LLM-powered reasoning layer that provides natural language explanations for code predictions, bridging the gap between model predictions and human validation. Using the MIMIC-IV dataset, we implement and evaluate this approach using standard metrics (precision, recall, F1) and explanation quality.

Keywords: ModernBERT, natural language processing, MIMIC-IV, ICD-10

Medical coding is a critical yet complex process in healthcare administration. The translation of clinical documentation into standardized ICD-10 codes is essential for healthcare reimbursement, clinical research, and public health statistics. Currently, healthcare providers rely on dedicated medical coders who manually review extensive clinical notes to assign appropriate codes from over 150,000 possible diagnostic and procedural codes, creating significant administrative overhead.

For every $1 of revenue collected by a hospital, approximately $0.25 is spent on administrative tasks necessary to collect it. Non-doctor workers outnum-

[*] Corresponding author
Email address: `tyler.cross@berkeley.com` (Tyler Cross)

[1] Final project submission for the UC Berkeley Master of Information and Data Science program, DATASCI 266: Natural Language Processing with Deep Learning

ber doctors in healthcare roughly 16 to 1, with medical coding representing a significant portion of this burden.

This project aims to improve automated medical coding by leveraging ModernBERT's enhanced context window capabilities, potentially reducing the administrative burden while maintaining or improving coding accuracy.

1. Background

Edin et al. (2024) recently demonstrated success in medical coding using a RoBERTa-based approach. Their work showed promising results but left room for improvement in both performance and contextual understanding. Other notable contributions in this field include [additional related work to be filled in].

ModernBERT (Warner et al., 2024) represents an advancement over previous BERT-like models, offering an enhanced context window of 8192 tokens compared to the 512 tokens in traditional models. This increased context capacity is particularly valuable for medical documents, which tend to be lengthy and contain important information distributed throughout the text.

1.1. Using CSL

2. Methods

This project utilizes the MIMIC-IV dataset, a large, freely available database comprising de-identified health data associated with hospital stays. The dataset includes detailed clinical notes, diagnostic codes, procedural information, and other health-related data from real hospital encounters.

2.1. System Architecture

Our approach replaces the RoBERTa model used in previous work (Edin et al., 2024) with ModernBERT (Warner et al., 2024), leveraging its enhanced context window to better process lengthy clinical notes. We hypothesize that this will lead to improved code prediction by enabling the model to capture more comprehensive context from medical documents.

2.2. Implementation Details

[To be completed with specific implementation details]

2.3. Evaluation Approach

We evaluate our system using standard performance metrics: precision, recall, F1-score, and accuracy compared to gold-standard human coding, following the evaluation protocol established in previous work.

3. Results and Discussion

[This section will be completed after obtaining experimental results]

References

Edin, J., Maistro, M., Maaløe, L., Borgholt, L., Havtorn, J.D., Ruotsalo, T., 2024. An Unsupervised Approach to Achieve Supervised-Level Explainability in Healthcare Records. arXiv:2406.08958.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., Poli, I., 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. URL: https://arxiv.org/abs/2412.13663, arXiv:2412.13663.