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—the assignment of standardized ICD-10 codes to clinical documentation—remains a labor-intensive process requiring expert manual review of extensive narratives against 150,000+ potential classifications. We present CLEM-ICD, an extension of the PLM-ICD framework that replaces RoBERTa with ModernBERT to leverage its 8192-token context window for improved automated medical coding. Our architectural enhancement preserves the multi-label classification paradigm while significantly expanding the model's capacity to process lengthy clinical text, capturing distant dependencies crucial for accurate code assignment. Experiments on the MIMIC-IV dataset demonstrate that CLEM-ICD achieves superior performance compared to previous approaches, as measured by precision, recall, and F1-score. This work addresses a critical bottleneck in healthcare administration through advanced NLP techniques, offering a scalable solution for reducing the cognitive burden of medical coding while maintaining diagnostic accuracy.

Keywords: clinical natural language processing, medical coding automation, transformer models, multi-label classification, long-context language models, ModernBERT, MIMIC-IV, ICD-10, healthcare informatics, biomedical text classification

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

*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

150,000 possible diagnostic and procedural codes, creating significant administrative overhead (Tseng et al., 2018). Non-doctor workers outnumber doctors in healthcare roughly 16 to 1 (Kocher and Sahni, 2011), with medical coding representing a significant portion of this burden.

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 preprint URL: https://arxiv.org/abs/2406.08958, arXiv:2406.08958.

Kocher, R., Sahni, N.R., 2011. Rethinking health care labor. The New England Journal of Medicine 365, 1370–1372. URL: https://doi.org/10.1056/NEJMp1109649, doi:10.1056/NEJMp1109649.

Tseng, P., Kaplan, R.S., Richman, B.D., Shah, M.A., Schulman, K.A., 2018. Administrative costs associated with physician billing and insurance-related activities at an academic health care system. JAMA 319, 691–697. URL: https://doi.org/10.1001/jama.2017.19148, doi:10.1001/jama.2017.19148.

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. arXiv preprint URL: https://arxiv.org/abs/2412.13663, arXiv:2412.13663.