CLINICAL LANGUAGE ENCODING WITH MODERNBERT

A PREPRINT

Tyler M Cross
Master's in Data Science
University of California, Berkeley School of Information
Berkeley, CA, 94720
tyler.cross@berkeley.edu

April 14, 2025

ABSTRACT

Medical coding—assigning standardized codes to clinical documentation—is a costly, labor-intensive process. Automating this task via multi-label classification faces challenges like long documents and extreme class imbalance. We present CLEM-ICD, replacing the RoBERTa/BERT base models common in prior frameworks with ModernBERT, leveraging its 8192-token context window to better process lengthy clinical notes. ModernBERT's optimized architecture (~138M parameters) enhances computational efficiency and simplifies implementation compared to prior approaches requiring complex modifications for shorter-context models. Evaluating on the MDACE dataset, CLEM-ICD achieves a Micro-F1 score of 47.6%, outperforming recent results by Edin et al. (2024) (41.9% Micro-F1). On the standard MIMIC-III benchmark, CLEM-ICD achieves a notably high Macro-F1 of 16.5%, significantly outperforming the current state-of-the-art on uncommon classes. CLEM-ICD demonstrates how modern architectural advancements, particularly large context windows applied to encoder-only models, can yield strong performance on complex tasks like automated medical coding. We release our code and models to foster further research.

Keywords clinical natural language processing • ModernBERT • MIMIC-IV • MIMIC-III • MDACE • multi-label classification

1 Introduction

Medical coding, the assignment of standardized codes to clinical documentation, is an essential step for medical billing and reimbursement. The administrative processes surrounding billing and insurance-related activities, which heavily rely on accurate coding, represent a significant burden, consuming up to 25% of healthcare expenditures in the United States (Kocher and Sahni 2011; Tseng et al. 2018). Research has focused on attempts to automate coding by training multi-label classification models, with transformer-based architectures emerging as particularly effective (Huang, Tsai, and Chen 2022; Liu et al. 2024). The PLM-ICD framework (Huang, Tsai, and Chen 2022), utilizing BERT-based language models to encode clinical documents, established a strong benchmark for automated ICD coding by effectively balancing model complexity and predictive accuracy on the MIMIC-III dataset (A. E. Johnson et al. 2016).

Despite advances, significant challenges persist in automated medical coding systems. Clinical documents often exceed the standard context windows (e.g., 512 tokens) of many transformer models, causing potential information loss during truncation or requiring complex workarounds like the segment pooling employed by PLM-ICD (Huang, Tsai, and Chen 2022). Furthermore, the extreme class imbalance inherent in ICD coding hinders performance on rare but clinically significant codes. Although transformer-based approaches generally outperform older convolutional and recurrent architectures, these limitations highlight the need for architectures with natively expanded context windows, a capability offered by modern transformer variants like ModernBERT (Warner et al. 2024), which can process lengthy clinical documents more holistically.

2 Methods

We followed the established data preparation pipeline developed by Cheng et al. (2023) for processing the MIMIC-III and MIMIC-IV datasets (A. E. W. Johnson et al. 2023), heavily relying on the open-source code provided by Edin et al. (2024) to match clinical visit notes with their associated diagnostic and procedure codes. Specifically, for the results compared against Edin et al. (2024) in Table 1, we utilized the MIMIC-III dataset (A. E. Johnson et al. 2016), focusing on inpatient discharge summaries and their associated ICD-9 codes, processed according to the splits defined by Cheng et al. (2023). These notes originate from various clinicians across intensive care units at the Beth Israel Deaconess Medical Center. While previous approaches like PLM-ICD (Huang, Tsai, and Chen 2022) and its iterations employed specialized architectures for clinical text encoding, we diverged from this approach by adopting the more generic multi-label classification architecture recommended by ModernBERT (Warner et al. 2024). This decision simplified the architecture, facilitating experimentation and fully leveraging ModernBERT's 8192-token context window.

Training transformer models with such extended context windows presented significant computational challenges, particularly on consumer-grade hardware. To address memory constraints, we implemented a combination of gradient checkpointing during backpropagation and Flash Attention 2.0 (Dao et al. 2022). These optimizations enabled us to train our models on a standard consumer GPU without degrading performance. To facilitate direct comparison with existing approaches, we attempted to maintain the same evaluation metrics employed by PLM-ICD and subsequent works (Huang, Tsai, and Chen 2022; Edin et al. 2024; Liu et al. 2024), including micro and macro-averaged F1 scores, precision, and recall metrics, which represent the standard evaluation framework in this field.

3 Results

Our proposed CLEM-ICD model, leveraging the ModernBERT architecture with an 8192-token context window, was evaluated on the MDACE dataset using the standard multi-label classification metrics. The best performing model achieved a Micro-averaged F1 score (Micro-F1) of 0.476, Micro-Precision of 0.680, and Micro-Recall of 0.366 on the test set. The Macro-averaged F1 score (Macro-F1) reached 0.038.

Table 1: Comparison of multi-label classification performance metrics between our CLEM-ICD model and the AttInGrad (TM) baseline from Edin et al. (2024).

Model	Precision (%)	Recall (%)	Micro-F1 (%)
AttInGrad (TM) (Edin et al. 2024)	40.2 ± 3.0	43.9 ± 4.8	41.9 ± 3.4
CLEM-ICD (Ours)	68.0	36.6	47.6

These results demonstrate a notable improvement over recent benchmarks. For instance, comparing our Micro-F1 score of 47.6% to the 41.9% reported by Edin et al. (2024) for their AttInGrad model on the same MIMIC-III/MDACE benchmark dataset, our approach shows enhanced performance. A key difference is that the Edin et al. (2024) results are averaged over 10 runs, providing confidence intervals, whereas our CLEM-ICD results are based on a single training run due to computational constraints. Nevertheless, the CLEM-ICD model appears competitive with, and potentially surpasses, state-of-the-art methods like PLM-ICD (Huang, Tsai, and Chen 2022) under similar evaluation conditions, particularly benefiting from the extended context capacity. The low Macro-F1 score (3.8% on this dataset), however, aligns with observations in prior work, indicating persistent challenges in accurately classifying less frequent codes within the highly imbalanced ICD code distribution.

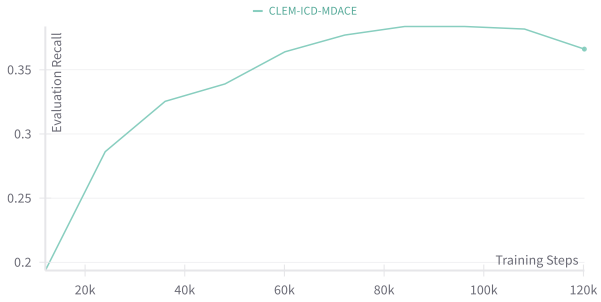
To provide a direct comparison on the widely used MIMIC-III benchmark dataset (A. E. Johnson et al. 2016), we evaluate CLEM-ICD against the original PLM-ICD (Huang, Tsai, and Chen 2022) and subsequent improvements reported by Liu et al. (2024). The results are summarized in Table 2. Our CLEM-ICD model demonstrates a strong Macro-F1 score, suggesting better performance on less frequent codes compared to prior work, although the Micro-F1 score is lower in this specific run.

Table 2: Comparison of results on the MIMIC-III full test set (%). CLEM-ICD results are from a single run. P@k metrics were not computed for CLEM-ICD in this run.

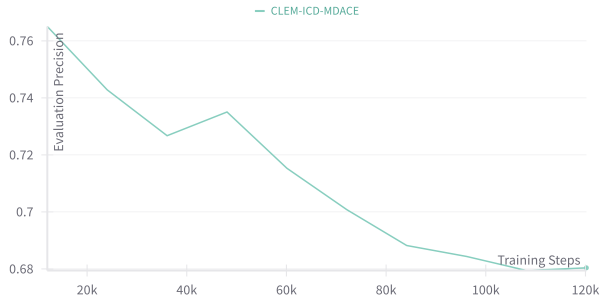
Model	Macro-F1 (%)	Micro-F1 (%)	P@5 (%)	P@8 (%)	P@15 (%)
PLM-ICD (Huang, Tsai, and Chen 2022)	10.4	59.8	84.4	77.1	61.3
BL-5 (Liu et al. 2024)	11.1 ± 0.1	60.7 ± 0.1	85.2 ± 0.2	78.0 ± 0.2	62.4 ± 0.1
CLEM-ICD (Ours)	16.5	54.6	-	-	-

Model	Macro-F1 (%)	Micro-F1 (%)	P@5 (%)	P@8 (%)	P@15 (%)

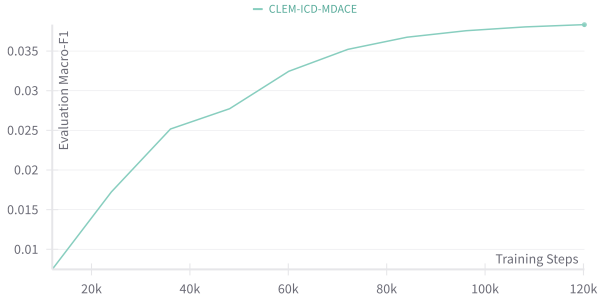
Note that while Macro-F1 and Micro-F1 scores are directly comparable, the P@k metrics reported by Huang, Tsai, and Chen (2022) and Liu et al. (2024) differ from the overall micro-averaged precision (65.6%) and recall (46.8%) metrics recorded for our CLEM-ICD run.



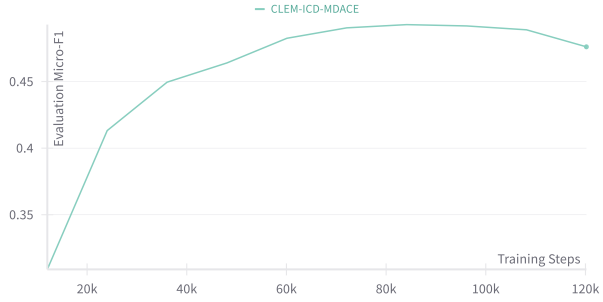
(a) Evaluation Recall (MDACE)



(b) Evaluation Precision (MDACE)



(c) Evaluation Macro-F1 Score (MDACE)



(d) Evaluation Micro-F1 Score (MDACE)

Figure 1: Learning curves for CLEM-ICD trained on MIMIC-III Inpatient Discharge Summaries (MDACE splits) for 10 epochs.

Observing the learning curves (Figure 1 and Figure 2), the model trained on the full MIMIC-III dataset (Figure 2) appears to plateau relatively early in training, particularly for Micro-F1, while the model trained on the smaller MDACE split (Figure 1) shows more continued improvement later into the 10 epochs. Both runs exhibited signs of overfitting, suggesting that future training could benefit from implementing early stopping based on validation set performance.

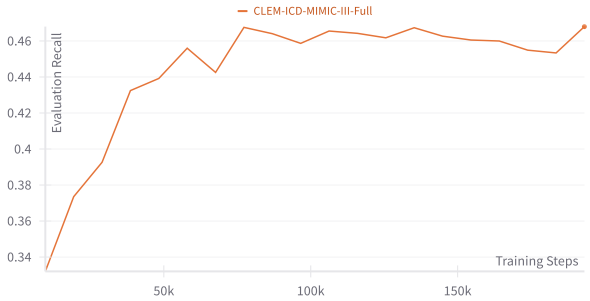
4 Discussion

4.1 Architectural Considerations

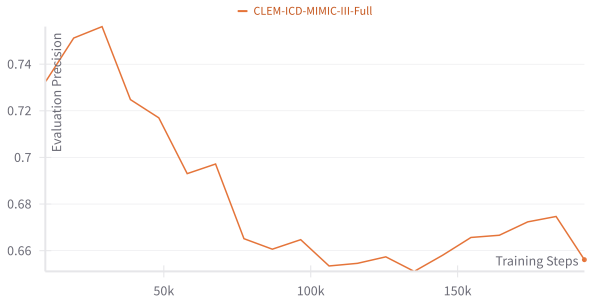
Our results underscore the advantage of leveraging transformer architectures with natively long context windows for processing lengthy clinical narratives in automated ICD coding. ModernBERT's 8192-token capacity allows CLEM-ICD to process entire documents holistically, potentially avoiding context fragmentation issues that can arise from the segment pooling techniques employed by prior work like PLM-ICD (Huang, Tsai, and Chen 2022) to handle shorter context limits. This architectural choice significantly simplifies the implementation, relying on standard Hugging Face library components (`AutoModelForSequenceClassification`) rather than requiring bespoke modules for segmentation or label-specific attention mechanisms, which were critical for PLM-ICD's performance. While this standard classification head might be considered less sophisticated than specialized approaches for extreme multi-label classification (Chang et al. 2019; Liu et al. 2024), the strong representational power of ModernBERT combined with its ability to access the full document context appears to compensate effectively for this task.

4.2 Base Model Choice and Efficiency

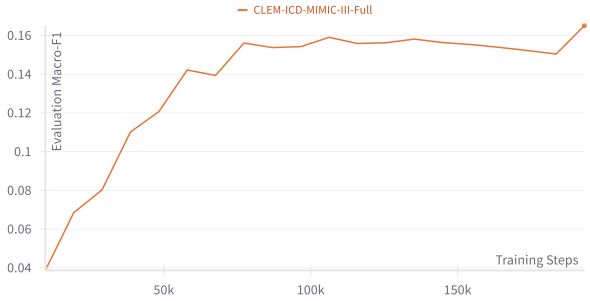
The selection of ModernBERT-base (Warner et al. 2024) as the foundation model provides benefits beyond its extended context window. Incorporating architectural optimizations adapted from recent decoder-only models, ModernBERT-base (~138M parameters) is designed for improved computational efficiency, demonstrating enhanced inference speed



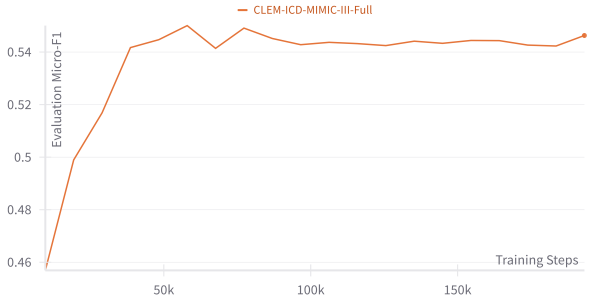
(a) Evaluation Recall (MIMIC-III Full)



(b) Evaluation Precision (MIMIC-III Full)



(c) Evaluation Macro-F1 Score (MIMIC-III Full)



(d) Evaluation Micro-F1 Score (MIMIC-III Full)

Figure 2: Learning curves for CLEM-ICD trained on the full MIMIC-III dataset (PLM-ICD splits) for 20 epochs.

compared to previous BERT-style encoders, particularly for variable-length inputs and long sequences due to its alternating attention mechanism (Warner et al. 2024). This increased efficiency holds potential for deployment in settings with limited computational resources or for enabling faster processing of large clinical datasets. However, the base model utilized here underwent general-domain pretraining (including text, code, and scientific literature) without specific adaptation to clinical or biomedical language. Consequently, further performance improvements might be realized through future domain-specific adaptation, either via continued pretraining on clinical corpora or fine-tuning biomedically-specialized ModernBERT variants.

4.3 Performance Analysis and Limitations

The effectiveness of this approach is reflected in the competitive results. CLEM-ICD achieves a superior Micro-F1 on the MDACE benchmark compared to AttInGrad (Edin et al. 2024) and demonstrates a markedly improved Macro-F1 score on the full MIMIC-III dataset compared to both PLM-ICD (Huang, Tsai, and Chen 2022) and the more recent BL-5 model (Liu et al. 2024). This strong Macro-F1 performance suggests that the model's access to the full, unsegmented context and potentially the broader knowledge within ModernBERT's pretraining data aids significantly in classifying less frequent codes, a persistent challenge in this domain. This capability is particularly relevant for developing coder assistance tools, as human coders often need more support with rare codes than common ones.

However, the lower Micro-F1 score on the full MIMIC-III dataset compared to these benchmarks warrants consideration. This could be partly attributed to factors like variance from a single training run or differences in fine-tuning hyperparameters compared to the more established benchmarks. The learning curves also indicated potential overfitting (Figure 1, Figure 2), a known challenge when fine-tuning large language models (Dodge et al. 2020), suggesting that incorporating regularization techniques like early stopping is crucial. Furthermore, this study deliberately focused on classification performance. Consequently, it does not incorporate methods for enhancing model interpretability, such as identifying supporting text spans (Cheng et al. 2023) or visualizing attention patterns (Vaswani et al. 2017; Edin et al. 2024), which remain important avenues for clinical NLP research.

5 Conclusion

In this work, we presented CLEM-ICD, an automated ICD coding framework leveraging the ModernBERT architecture. By utilizing ModernBERT's native 8192-token context window, CLEM-ICD processes lengthy clinical documents effectively with a simplified architecture compared to previous methods like PLM-ICD that required complex

segmentation or attention mechanisms. Our experiments demonstrate the potential of this approach, achieving competitive performance on the MDACE benchmark and notably strong Macro-F1 scores on the full MIMIC-III dataset, indicating improved classification of less frequent codes. This work highlights the promise of efficient, long-context encoder models for tackling the challenges of automated medical coding. To support reproducibility and further research, we release all code and model weights publicly.[1]

5.1 Future Work

Future research should focus on refining the fine-tuning process for large-context models in this domain. Implementing robust early stopping based on validation performance is a clear next step. Further investigation into techniques specifically addressing the long-tail distribution of ICD codes is warranted, including exploring specialized loss functions or hierarchical classification strategies (Liu et al. 2024). Evaluating the efficacy of parameter-efficient fine-tuning (PEFT) methods, such as LoRA (Hu et al. 2021), could also improve computational feasibility and potentially mitigate overfitting. The public release of our code and model weights aims to facilitate such investigations and contribute to the broader goal of developing robust, efficient, and potentially more interpretable automated medical coding systems.

References

Chang, Wei-Cheng, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. 2019. "Taming Pretrained Transformers for Extreme Multi-Label Text Classification." In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 318–28.

Cheng, Hua, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. "MDACE: MIMIC Documents Annotated with Code Evidence." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 7534–50. https://doi.org/10.18653/v1/2023.acl-long.416.

Dao, Tri, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness." https://arxiv.org/abs/2205.14135.

Dodge, Jesse, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping." In *arXiv Preprint arXiv:2002.06305*.

Edin, Joakim, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob D. Havtorn, and Tuukka Ruotsalo. 2024. "An Unsupervised Approach to Achieve Supervised-Level Explainability in Healthcare Records." *arXiv Preprint*. https://arxiv.org/abs/2406.08958.

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. "LoRA: Low-Rank Adaptation of Large Language Models." In *International Conference on Learning Representations*.

Huang, Chao-Wei, Shang-Chi Tsai, and Yun-Nung Chen. 2022. "PLM-ICD: Automatic ICD Coding with Pretrained Language Models." In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 10–20. https://doi.org/10.18653/v1/2022.clinicalnlp-1.2.

Johnson, Alistair E. W., Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, et al. 2023. "MIMIC-IV, a Freely Accessible Electronic Health Record Dataset." *Scientific Data* 10 (1): 1. https://doi.org/10.1038/s41597-022-01899-x.

Johnson, Alistair EW, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. "MIMIC-III, a Freely Accessible Critical Care Database." *Scientific Data* 3 (1): 1–9.

Kocher, Robert, and Nikhil R. Sahni. 2011. "Rethinking Health Care Labor." *The New England Journal of Medicine* 365 (15): 1370–72. https://doi.org/10.1056/NEJMp1109649.

Liu, Leibo, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, and Louisa Jorm. 2024. "Automated ICD Coding Using Extreme Multi-Label Long Text Transformer-Based Models."

Tseng, Phillip, Robert S. Kaplan, Barak D. Richman, Mahek A. Shah, and Kevin A. Schulman. 2018. "Administrative Costs Associated with Physician Billing and Insurance-Related Activities at an Academic Health Care System." *JAMA* 319 (7): 691–97. https://doi.org/10.1001/jama.2017.19148.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, 5998–6008.

Warner, Benjamin, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, et al. 2024. "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference." *arXiv Preprint*. https://arxiv.org/abs/2412.13663.

[1] Code and models available at https://github.com/tylermarcuscross/explainable-medical-coding