Enhancing Clinical Coding through Interactive Machine Learning

Yidong Gan[1,2], Maciej Rybinski[1,2], Ben Hachey[1], Jonathan K. Kummerfeld[1],
[1]The University of Sydney
[2]CSIRO Data61
yidong.gan@sydney.edu.au

Abstract

Clinical coding involves the classification of medical diagnoses and procedures using alphanumeric codes. Manual coding is labour-intensive and error-prone, motivating research towards automating the process as a multi-label classification problem. However, current automatic approaches fall short of human-level accuracy and do not align with real clinical practice. This paper proposes a semi-automatic approach, framing the task as a human-in-the-loop multi-class classification problem. Given an input medical note, our model predicts the associated codes one by one and receives feedback from a human domain expert (e.g., a professional coder) after each prediction. This enables users to participate in building the model, resulting in a more accurate and trustworthy model for real clinical applications.

1 Problem Definition

Clinical coding is essential for hospital reimbursement and the study of disease prevalence. It is a demanding task that requires expert knowledge and experience, where coders need to review relevant medical documents and assign codes sequentially. However, most current automated coding research overlooks this practice and instead makes a one-shot multi-label prediction (Edin et al., 2023).

Existing one-shot multi-label classifiers are not yet practically useful due to the following reasons: **(i) Low Accuracy**: Our preliminary results show that the state-of-the-art (SOTA) multi-label classifier, PLM-ICD (Huang et al., 2022), yields an error rate almost twice as high as that of an average human coder. This suggests that current automated solutions are too immature for reducing manual workload. **(ii) Misalignment with Real Clinical Practice**: Human coders must adhere to specific coding guidelines, which vary by country and include rules regarding code usage, such as which codes should or should not be used together (CMS and NCHS, 2024). This highlights the importance of modelling code dependencies, an aspect often overlooked in previous research.

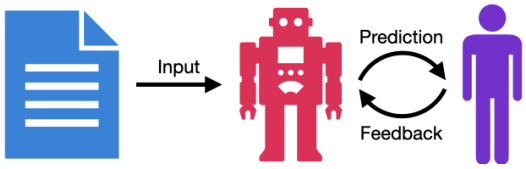

Figure 1: Overview of the interactive learning process.

2 Proposed Solution

Interactive machine learning (IML) enhances model performance through real-time user feedback and has been applied to tasks such as image segmentation (Amershi et al., 2014) and semantic parsing (Yan et al., 2023). We propose to apply the technique and re-frame clinical coding as an interactive multi-class classification problem. This allows clinical coders to refine a code classifier with their expertise. We believe this approach could lead to a more practical solution sooner than relying solely on automated coding. Figure 1 presents an overview of the interactive learning process, in which the model functions as a code recommendation system, receiving feedback after each prediction (code recommendation). This design offers four key benefits:

- **Human-in-the-loop Error Correction**: Unlike non-interactive training methods, the proposed design allows the model to correct errors in real time based on feedback from clinical coders, ultimately leading to higher model accuracy in the long run.
- **Improved Trust**: The model's decision-making process becomes more transparent, making clinical coders more likely to trust the model's recommendations as they contribute to its learning process (Amershi et al., 2014).

- **Handling Code Dependencies**: ICD codes are interdependent (CMS and NCHS, 2024), a crucial property that has been neglected in most previous research. Per-prediction feedback allows the model to gradually learn to capture these code dependencies over time.

- **Adaptability to Coding Preferences**: Coding practices vary between hospitals and may change due to guideline updates (Cheng et al., 2023). As the model receives feedback on a per-prediction basis, it can adapt to fine-grained preference changes without the need for complete retraining.

2.1 Evaluation

We plan to use the preprocessed MIMIC datasets published by Edin et al. (2023) and evaluate the model's effectiveness from three perspectives.

Information-retrieval-based We will measure the model's precision at full recall (P@FR) and the number of iterations at full recall (NI@FR). Both metric scores will be averaged over the samples. The baselines are the SOTA multi-label approaches, where their i-th highest scored code is considered equivalent to their i-th prediction.

Learning-curve-based We will measure the model's convergence rate, i.e., how quickly it improves through feedback.

User-based We will develop an UI application and engage real clinical coders to conduct a user study. This study measures user satisfaction (e.g., ease of use and response time), as well as how much the application improves their overall coding accuracy and completion time.

3 Progress

We are exploring different feedback incorporation approaches, such as adding a vector to represent the feedback as part of the model's input. Specifically, our model receives both the text (medical note) and the feedback vector to make iterative predictions. In contrast, the multi-label baseline model, PLM-ICD, uses only the text input and makes a one-shot prediction. Both models are optimised using binary cross-entropy loss. Table 1 presents the preliminary results. After training on 20% of the training data for two epochs, our model demonstrates approximately twice the performance of PLM-ICD on information-retrieval metrics, indicating that our iterative feedback approach positively impacts accuracy. The fully-trained PLM-ICD represents the upper bound of current SOTA multi-label approaches. Next, we will fully train our model and compare it to this upper bound for further analysis.

	Train (%)	**P@FR (%)**	**NI@FR**
PLM-ICD (2 epoch)	20	0.45	3413
Ours (2 epoch)		1.18	1792
PLM-ICD (fully-trained)	100	17.61	325

Table 1: Evaluation on the MIMIC-III Clean Dataset. The 'Train' value indicates the percentage of training data used. Fully-trained means the model has been trained until it no longer improves on the validation set.

References

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine*, 35(4):105–120.

Hua Cheng, 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 (Volume 1: Long Papers)*, pages 7534–7550, Toronto, Canada. Association for Computational Linguistics.

CMS and NCHS. 2024. Icd-10-cm official guidelines for coding and reporting. FY 2024 ICD-10-CM Coding Guidelines.

Joakim Edin, Alexander Junge, Jakob D Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated medical coding on mimic-iii and mimic-iv: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2572–2582.

Chao-Wei Huang, 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*, pages 10–20, Seattle, WA. Association for Computational Linguistics.

Hao Yan, Saurabh Srivastava, Yintao Tai, Sida I. Wang, Wen-tau Yih, and Ziyu Yao. 2023. Learning to simulate natural language feedback for interactive semantic parsing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3149–3170, Toronto, Canada. Association for Computational Linguistics.