Assisting Clinicians with AI: Hierarchical Multi-Label Approaches to International Classification of Diseases Coding

Riccardo Gibello[1] and Enrico Gianluca Caiani *[1,2]

[1]Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milan, Italy
[2]IRCCS Istituto Auxologico Italiano, San Luca Hospital, Milan, 20149, Italy

October 3, 2025

1 Context

The International Classification of Diseases (ICD) is a standard taxonomy widely used for diagnostic coding in healthcare systems. Manual assignment of ICD codes from clinical narratives (e.g. discharge summaries, pathology reports) is time-consuming and expensive. Automated ICD coding aims to assist professionals by predicting the appropriate set of ICD codes for a given clinical text. Recent reviews highlight that this is a hot research area, with challenges including large label sets, severe class imbalance, and the hierarchical dependencies among codes [1]. Moreover, many modern proposals try to better exploit taxonomic relationships between codes (for example via graph neural networks, label embedding methods, or hierarchical regularization) to improve prediction accuracy and consistency [2].

2 Objective

This thesis aims to evaluate, compare, and possibly extend state-of-the-art methods for automatic ICD coding under a *multi-label hierarchical classification* framework. Specific goals include:

- Systematically survey recent methods and publicly available codebases for ICD coding, including attention-based, graph-based, and hierarchical-aware architectures.
- Select a few strong baseline implementations from `GitHub` or related repositories, adapt them to your clinical dataset, and benchmark their performance in terms of micro / macro F1, hierarchical consistency, and calibration.
- (Optional) Propose a slight architectural modification (e.g. incorporate hierarchical contrastive learning or label-aware attention) and evaluate whether it gives improvements over baselines.

*Advisor

3 Rationale

Automatic ICD coding is naturally a *multi-label task*, and the ICD system is organized hierarchically, so a plain "flat" multi-label classifier may ignore important parent–child relations. Hierarchical multi-label classification methods can exploit structure to improve performance [6,10]. Some recent works reframe hierarchical text classification as sequence-to-sequence tasks [7,8] to better capture label dependencies. Other architectures (e.g., hierarchical multi-label classification networks [9]) explicitly model hierarchical relations in their layers. In the medical domain, regularization and hierarchical constraints (e.g. parent–child consistency) have been integrated into neural ICD coding models to reduce spurious predictions [4,5]. By combining insights from hierarchical classification theory and domain-specific ICD coding literature, this thesis can both assess practical performance and explore possible extensions.

4 Approach

The proposed research pipeline consists of the following steps:

- **Literature and code survey**: Conduct a systematic review of recent approaches to multi-label hierarchical text classification, with a particular focus on automatic ICD coding. Identify publicly available implementations and benchmark repositories that can serve as baselines.
- **Dataset preparation**: Select and preprocess one or more clinical text corpora annotated with ICD codes (e.g., MIMIC-III, pathology reports). This step includes data cleaning, tokenization, and splitting into training, validation, and test sets.
- **Baseline experimentation**: Train a range of representative models (e.g., flat multi-label classifiers, hierarchy-aware methods, and graph-based models) on the chosen dataset(s). Evaluate their performance using hierarchy-aware metrics [3] in addition to standard multi-label measures.
- **Error and performance analysis**: Investigate model behavior by analyzing error patterns, challenges with rare codes, prediction calibration, and consistency across hierarchical levels.
- **Architectural extension (optional)**: Design and test a lightweight modification to an existing baseline (e.g., hierarchical contrastive loss, label-aware attention, or sequence-to-tree decoding). Compare its performance against unmodified baselines to assess its added value.

5 Requirements

The candidate is expected to bring, or be willing to acquire, the following competencies:

- A background in computer science, biomedical engineering, or a related discipline.
- Good programming skills in Python. Familiarity with NLP libraries (e.g. `HuggingFace`, `PyTorch`) is a plus.
- Basic understanding of machine learning, deep learning, and natural language processing. *Bonus*: Completion or attendance of courses in Natural Language Processing or Artificial Neural Networks and Deep Learning.

- Ability to read and understand others' code, adapt it for experiments, and analyze results quantitatively.
- Write clean code, document the work, and present findings in a reproducible fashion.

If you think you might be a good fit for this project or would like more information, please contact us at riccardo.gibello@polimi.it. The listed requirements are meant to provide a clear picture of the project and are not strict prerequisites. If you are motivated and eager to learn, we encourage you to reach out regardless of prior experience.

References

[1] Seyyedeh Fatemeh Mousavi Baigi, Masoumeh Sarbaz, Ali Darroudi, Fatemeh Dahmardeh Kemmak, Reyhane Norouzi Aval, and Khalil Kimiafar. Artificial intelligence-based automated international classification of diseases coding: A systematic review. *Journal of Medical Signals & Sensors*, 15(8):22, 2025.

[2] Gonçalo Gomes, Isabel Coutinho, and Bruno Martins. Accurate and well-calibrated icd code assignment through attention over diverse label embeddings. *arXiv preprint arXiv:2402.03172*, 2024.

[3] Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015.

[4] Anthony Rios, Eric Durbin, Isaac Hands, and Ramakanth Kavuluru. Assigning ICD-O-3 Codes to Pathology Reports using Neural Multi-Task Training with Hierarchical Regularization. volume 2021, pages 1–10, 2 2021.

[5] Anthony Rios and Ramakanth Kavuluru. Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142, Brussels, Belgium, 10 2018. Association for Computational Linguistics.

[6] Carlos N Silla Jr and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22(1):31–72, 2011.

[7] Fatos Torba, Christophe Gravier, Charlotte Laclau, Abderrhammen Kammoun, and Julien Subercaze. A study on hierarchical text classification as a seq2seq task. In *European Conference on Information Retrieval*, pages 287–296. Springer, 2024.

[8] Fatos Torba, Christophe Gravier, Charlotte Laclau, Abderrhammen Kammoun, and Julien Subercaze. Decoding the hierarchy: A hybrid approach to hierarchical multi-label text classification. In *European Conference on Information Retrieval*, pages 405–420. Springer, 2025.

[9] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International conference on machine learning*, pages 5075–5084. PMLR, 2018.

[10] Alessandro Zangari, Matteo Marcuzzo, Matteo Rizzo, Lorenzo Giudice, Andrea Albarelli, and Andrea Gasparetto. Hierarchical text classification and its foundations: A review of current research. *Electronics*, 13(7):1199, 2024.