Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

Sparse ProtoPatient: Interactive Multi-Prototype Explanations for Clinical Diagnosis Prediction

Conor Fallon[1] , **Bogdan Kostić**[1] , **Betty van Aken**[1,2] , **Jens-Michalis Papaioannou**[1] , **Alexei Figueroa**[1] and **Keno Bressem**[3] , **Alexander Löser**[1]
[1]Berliner Hochschule für Technik (BHT), Berlin, Germany
[2]Grammarly
[3]Klinikum rechts der Isar, Department of Radiology, Technical University of Munich (TUM), Germany
conor.fallon@bht-berlin.de

Abstract

We present the Sparse ProtoPatient Demo, a publicly available interactive system for interpretable ICD-10 diagnosis prediction from clinical admission notes. The system is designed for clinicians in training, researchers, and educators exploring prototype-based diagnostic reasoning.

The demo links predictions to learned prototypical patient representations and token-level evidence, allowing users to input custom text or select preset cases, inspect predicted ICD-10 codes, visualize label-wise saliency, retrieve supporting prototype notes, and compare alternative prototype cohorts. The demo provides a reproducible platform for interactive inspection of prototype-based clinical reasoning, enabling complementary opinion exploration, model auditing, and teaching of interpretable diagnosis prediction. The deployed model is trained on the publicly released CodiEsp corpus (1000 clinical notes, 955 ICD-10 labels) using a sparse multi-prototype architecture with five prototypes per label. We use the official machine-translated English CodiEsp-MT release to support English-language interaction. It achieves a macro-AUROC of 0.92 on a held-out test set and supports real-time interaction (300ms per query). The system is fully containerized for public research and educational use.

1 Introduction

Clinical outcome prediction is a central task in clinical NLP [van Aken *et al.*, 2021]. We focus on diagnosis prediction, where a model predicts discharge ICD codes from admission-time clinical text. Modern neural classifiers provide limited insight into why a diagnosis was predicted, while real-world code distributions are highly skewed, with many clinically relevant conditions appearing only sparsely in training data. In high-stakes settings, systems must perform well and expose the evidence supporting their decisions.

Prototype-based models address this tension by representing each diagnosis through learned example-like embeddings rather than opaque weight vectors [van Aken *et al.*, 2022]. During training, prototypical patient representations are learned from archived clinical data. At inference time, a new patient note serves as a query, activating the most similar prototypes and retrieving representative archived patient cases. Label-wise attention highlights clinically relevant concepts in the input text. Sparse ProtoPatient (S-Proto) [Figueroa *et al.*, 2024] extends this approach by learning multiple prototypes per label, capturing heterogeneous disease subtypes and presentation patterns.

We release a live demo interface for ICD-10 prediction from clinical admission notes that combines multi-prototype reasoning, token-level saliency visualization, and prototype cohort switching, revealing (i) which tokens influenced each prediction, (ii) which prototypical patients were retrieved, and (iii) how alternative prototype cohorts support the diagnosis. The service performs stateless inference and does not store user input.

Prior clinical coding demos [Hajialigol *et al.*, 2023; Kim *et al.*, 2022] and attention visualization tools have typically exposed either token-level attention or exemplar retrieval in isolation. Unlike these systems, the Sparse ProtoPatient Demo unifies both within a single interactive interface centered on multi-prototype cohort exploration, enabling comparison of multiple supporting patient groups for the same diagnosis and inspection of heterogeneous diagnostic patterns rather than a single explanation pathway.

The Sparse ProtoPatient Demo supports interactive analysis of diagnostic predictions by clinicians, researchers, and NLP practitioners through the following capabilities:

- **Interactive multi-prototype exploration:** Clinicians and medical trainees can explore multiple learned prototypes per diagnosis and compare heterogeneous disease sub-cohorts.
- **Integrated token-level evidence visualization:** Users see which terms influenced each prediction, supporting verification of clinically meaningful cues and detection of spurious associations.
- **Transparent long-tail reasoning:** For infrequent diagnoses, the system retrieves similar archived patient cases with expert-coded labels and highlights the clinical indicators supporting the prediction.

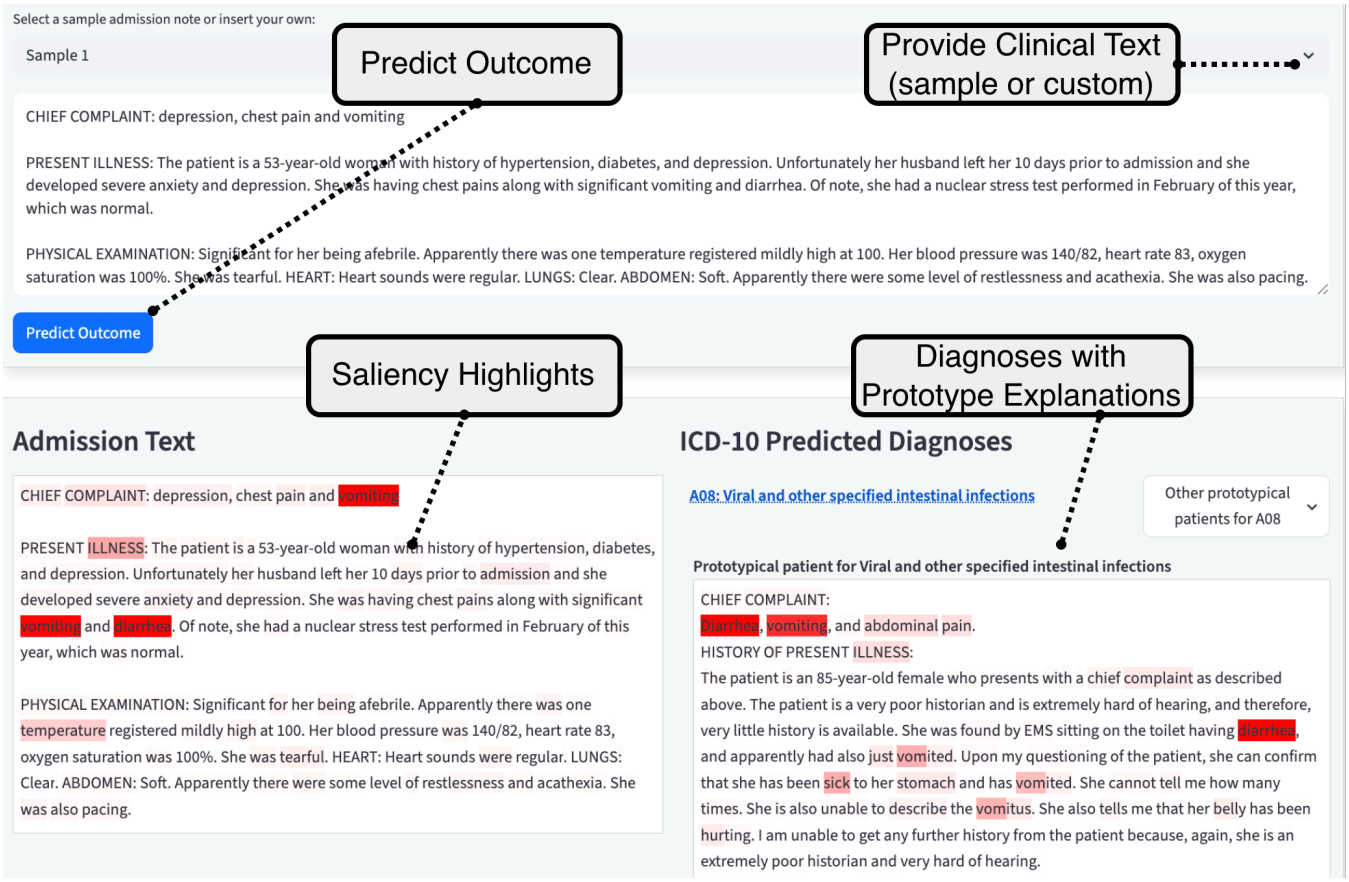

Figure 1: Overview of the Sparse ProtoPatient demo interface. Users enter or select clinical text, trigger a prediction, and interactively explore token-level saliency and prototype-based explanations. Clicking on any predicted diagnosis reveals the most similar prototypical patients and the evidence linking them to the input note.

2 System Overview

Task, Dataset, and Model. The task is multi-label ICD-10 diagnosis prediction from free-text clinical admission summaries. The deployed model is trained on the publicly available CodiEsp corpus, consisting of 1000 de-identified clinical notes annotated with 955 ICD-10 codes. We use an 80/10/10 train–validation–test split.

While prior S-Proto results were reported on larger private corpora such as MIMIC-IV, the public demo uses the official English CodiEsp-MT release (machine-translated from the original Spanish notes) to enable licensing-compliant public deployment and full reproducibility, while preserving the explanation mechanism demonstrated in the interface.

The backend serves a frozen PubMedBERT [Gu *et al.*, 2020] encoder with Sparse ProtoPatient (S-Proto) prototypes, packaged in a single Docker deployment with a React frontend and Flask API. The trained model checkpoint is publicly available on Hugging Face[1] to support reproducibility and further experimentation.

Model Configuration. The deployed S-Proto model [Figueroa *et al.*, 2024] uses a frozen PubMedBERT encoder and five prototypes per label, following prior ablation experiments in the original S-Proto study [Figueroa *et al.*, 2024], where five prototypes provided a stable trade-off between long-tail F1 gains and parameter efficiency. Each label has an independent validation-tuned decision threshold to maximize per-label F1.

Explanations Exposed in the Interface. As shown in Figure 1, users enter clinical text, receive ranked ICD-10 predictions, and inspect prototype-based explanations in a dedicated panel. Selecting a diagnosis reveals the highest-activated prototype, with optional switching between alternative cohorts. Token-level attention weights are displayed on both the input and retrieved notes, aggregated to word level for readability.

3 Demonstration Workflow

Stage 1: Input. A medical student or junior doctor enters findings as free-text notes or selects predefined cases from the CodiEsp dataset. Inference is performed in real time, and no user-entered text is stored.

Stage 2: Outcome Prediction. The system returns a ranked list of predicted ICD-10 codes. Each diagnosis is evaluated using its own validation-tuned decision threshold,

[1]https://huggingface.co/DATEXIS/sproto

enabling calibrated multi-label predictions rather than a fixed global threshold.

Stage 3: Prototype Retrieval and Evidence Highlighting. For each predicted diagnosis, the system retrieves the most strongly activated prototypical patient. Label-wise attention weights are projected onto the input note, highlighting clinically salient tokens that drive the prediction. Attention-based highlights are also displayed within the prototype note, revealing shared clinical aspects (e.g., symptoms, comorbidities, medications) that link the current case to prior patients.

Stage 4: Zoom into Selected Prototypical Patients. Users can open a detailed prototype view to inspect the retrieved patient note in full. Side-by-side comparison supports qualitative analysis of similarities and differences.

Stage 5: Cohort Switching. When multiple prototypes exist for a diagnosis, users can toggle between alternative representative patients. This reveals distinct sub-cohorts corresponding to different disease presentations. Alternative prototypes are surfaced only when the associated diagnosis passes its validation-tuned threshold, ensuring faithfulness to the multi-label decision process.

3.1 Representative Use Cases

Clinician. A clinician inputs a clinical note and inspects predicted ICD-10 codes with highlighted evidence. By comparing salient tokens and retrieved prototypes, they can assess whether the prediction reflects a clinically coherent presentation, including rare diagnoses, supporting rapid qualitative auditing.

Researcher. A researcher uses the demo to examine how prototype-based models organize clinical patterns. By toggling alternative prototypes for the same diagnosis, they can observe how heterogeneous disease presentations are partitioned into distinct learned cohorts. Saliency overlays support targeted error analysis and inspection of attention patterns.

Long-tail inspection. For rare diagnoses, users can examine how predictions are grounded in specific prototype cohorts. Distinct prototypes often correspond to different clinical subtypes within the same ICD-10 code, making sparse predictions directly inspectable.

4 Evaluation and Validation

Predictive Performance. On the held-out CodiEsp test set, the deployed S-Proto-5P model achieves **91.95** macro-AUROC and **36.90** F1 on long-tail diagnoses. The model outperforms a strong PubMedBERT encoder baseline, with particularly pronounced improvements on rare labels. Performance is sufficient to support stable interactive inspection.

4.1 Clinical Validation and Faithfulness

A clinician co-author qualitatively reviewed all preset demo cases and found that highlighted spans frequently correspond to clinically meaningful cues, including explicit diagnoses, symptoms, medications, and comorbidities. Occasional mismatches occurred when the model overweighted explicit mentions, reflecting expected limitations of a compact public checkpoint.

As a sanity check, removing highly attended tokens typically reduced prediction confidence, suggesting that highlighted evidence reflects model-internal decision signals rather than arbitrary artifacts [Figueroa *et al.*, 2024]. The public CodiEsp-based deployment preserves the same attention mechanism; large-scale robustness experiments are beyond the scope of this system demonstration.

4.2 Latency and Reproducibility

The system performs inference in approximately 300ms per query, enabling smooth interactive exploration. The full pipeline is containerized with Docker and can be rebuilt end-to-end with a single command. This enables reproducible experimentation and portable deployment on modest hardware.

5 Limitations and Responsible Use

Attention weights reflect model-internal relevance signals and do not constitute causal or human-like explanations [Bibal *et al.*, 2022]; they are provided as transparency tools rather than clinical evidence. The deployed system uses a sparse multi-prototype model trained on the publicly available CodiEsp dataset (1000 clinical notes, 955 ICD-10 labels), limiting coverage of rare diagnoses and generalization compared to larger clinical corpora. The interface currently supports single-note inference only and does not include longitudinal trajectories or multi-note aggregation. Freezing the encoder improves stability and reproducibility but constrains domain adaptation. Allowing fine-tuning may improve predictive performance at the cost of comparability and deployment simplicity. The prototype count is fixed at five per label; while empirically effective, this may not fully capture intra-class heterogeneity. The system is intended strictly for research and educational use. The demo illustrates model behavior and explanation mechanisms, not clinical correctness.

6 Availability and Conclusion

The live system, source code, and fully reproducible Docker deployment are publicly available.[2]

Sparse ProtoPatient makes multi-prototype clinical reasoning inspectable in real time. Instead of returning a single opaque prediction, it exposes multiple supporting patient cohorts and the token-level evidence linking them to the input case. This reframes heterogeneous ICD-10 prediction from a black-box output into an interactive reasoning process.

By unifying saliency visualization, prototype retrieval, and cohort switching within a single interface, the demo provides a practical environment for auditing, analysis, and education in clinical NLP. As a fully reproducible, public platform, it demonstrates how interpretable prototype-based reasoning can be operationalized in safety-critical healthcare AI.

Ethical Statement

All data used in this demonstration are publicly available and fully de-identified. The system does not store or log user-submitted input. The demo is intended strictly for research

[2]Live Demo · Walkthrough Video · Source Code + Docker.

and educational purposes and must not be used for clinical decision-making. Saliency visualizations reflect model-internal attention signals and do not constitute causal medical reasoning.

Acknowledgments

Our work is funded by the German Federal Ministry of Education and Research (BMBF) under the grant agreements 01—S23013C (More-with-Less), and 01—S23015A (AI4SCM). This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 528483508 - FIP 12, as well as the European Union under the grant project 101079894 (COMFORT - Improving Urologic Cancer Care with Artificial Intelligence Solutions). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. Also funded with the project SOOFI: Large Reasoning Models, Project-ID 13IPC040D.

References

[Bibal *et al.*, 2022] Adrien Bibal, Romain Cardon, David Alfter, Judith Wilke, Johannes Schaeffer, and Robert West. Is Attention Explanation? An Introduction to the Debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.

[Figueroa *et al.*, 2024] Alexei Figueroa, Jens-Michalis Papaioannou, Conor Fallon, et al. Boosting long-tail data classification with sparse prototypical networks. In Albert Bifet, Jesse Davis, Tomas Krilavičius, Meelis Kull, Eirini Ntoutsi, and Indrė Žliobaitė, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*. Springer Nature Switzerland, 2024.

[Gu *et al.*, 2020] Yu Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23, 2020.

[Hajialigol *et al.*, 2023] Daniel Hajialigol, Derek Kaknes, Tanner Barbour, et al. Drgcoder: Explainable clinical coding for the early prediction of diagnostic-related groups. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2023.

[Kim *et al.*, 2022] Juyong Kim, Abheesht Sharma, Suhas Shanbhogue, Jeremy Weiss, and Pradeep Ravikumar. Anemic: A framework for benchmarking icd coding models. In Wanxiang Che and Ekaterina Shutova, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2022.

[van Aken *et al.*, 2021] Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

[van Aken *et al.*, 2022] Betty van Aken, Jens-Michalis Papaioannou, Marcel Naik, et al. This patient looks like that patient: Prototypical networks for interpretable diagnosis prediction from clinical text. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.