Baymax at CLEF 2026: Leveraging XLM-RoBERTa and Semantic Label Embeddings for Greek Clinical Coding

Baymax at CLEF 2026

Georgios Chalkias[1], Panagiotis Amanatiadis[1], Glykeria Tsavlidou[1] and Sotirios Malamas[2]

[1]Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece
[2]Department of Computer Science, Aristotle University, Thessaloniki, Greece

Abstract

This paper describes the system developed by team Baymax for the ELCardioCC task at CLEF 2026, which addresses automated multi-label ICD-10 classification of Greek cardiology clinical records. Our architecture, Baymax (implemented as MedicalModelWithDescriptions), employs an XLM-RoBERTa Large backbone augmented with a semantic label anchoring mechanism that projects document-level representations into a label-description manifold derived from Greek ICD-10 terminology. To mitigate the severe class imbalance characteristic of power-law distributed clinical coding frequencies, we introduce a knowledge-informed data expansion strategy combining frequency-aware augmentation multipliers with stochastic synonym replacement drawn from a curated medical lexicon. Optimization is conducted via Layer-wise Learning Rate Decay (LLRD) with Stochastic Weight Averaging (SWA), enabling convergence to flatter minima in the high-dimensional multi-label parameter space. Post-processing incorporates per-label dynamic threshold tuning and ontological hierarchy enforcement. Our final system achieved a mean Micro-F_1 score of 0.8455 across stratified 5-fold cross-validation on 2,500 clinical records spanning 115 ICD-10 codes.

Keywords

clinical coding, ICD-10, multi-label classification, XLM-RoBERTa, semantic label embeddings, Greek NLP

1. Introduction

Automated clinical coding—the task of assigning standardized diagnosis codes from the International Classification of Diseases (ICD-10) to free-text medical records—is a critical component of healthcare information systems. When applied to Greek-language cardiology records, this task presents compounded challenges: the morphological richness of the Greek language, pervasive use of medical abbreviations and acronyms in hospital settings, and the heavily skewed frequency distribution of ICD-10 codes, which follows a power-law pattern where a small number of prevalent diagnoses dominate the training signal.

The ELCardioCC shared task at CLEF 2026 is organized within the broader BioASQ lab [1], and formalizes this problem as a multi-label classification task over cardiology-related clinical narratives; full details on task design, data collection, and annotation protocols are provided in the official task overview [2]. Each document may be annotated with one or more ICD-10 codes drawn from a fixed label set $\mathcal{L}$ of 115 codes. Let $\mathcal{D} = \{(x_i, Y_i)\}_{i=1}^{N}$ denote the dataset, where x_i is a clinical text and $Y_i \subseteq \mathcal{L}$ is its associated label set. The objective is to learn a function $f : \mathcal{X} \to 2^{\mathcal{L}}$ that maximizes the Micro-F_1 score over the prediction space.

In this paper, we present Baymax[1], a system that addresses these challenges through three principal contributions: (1) a transformer-based architecture with a semantic label anchoring module that aligns document representations with ICD-10 description embeddings via a learnable cosine similarity mechanism; (2) a frequency-aware, knowledge-informed data augmentation strategy that leverages a curated clinical dictionary for stochastic synonym replacement; and (3) a multi-faceted optimization

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
george.chalkias.duth@gmail.com (G. Chalkias); panagiotis.amanatiadis.ai@gmail.com (P. Amanatiadis); aid26005@uom.edu.gr (G. Tsavlidou); sotirismalamas@gmail.com (S. Malamas)

[1]The system name is inspired by Baymax, the personal healthcare companion robot from the movie *Big Hero 6* (2014), reflecting our system's design goal of providing careful, systematic, and supportive assistance to clinical coding personnel.

pipeline combining Layer-wise Learning Rate Decay, Stochastic Weight Averaging, and dynamic post-processing.

We further note that our submission shares foundational baseline components with another participating team from the University of Macedonia (UOM), following an informal exchange of preliminary infrastructure within our research lab. Beyond this starting point, the two efforts diverged significantly. The collaborating team pursued a large-scale ensembling strategy across multiple pre-trained encoders; interestingly, their winning ensemble incorporated the XLM-RoBERTa Base model, sharing the exact same architectural framework as the single XLM-RoBERTa Large model that secured 2nd place. Our contribution focused exclusively on this single-model pipeline, refining the architecture, hyper-parameter tuning, and post-processing strategy described in this paper. Despite utilizing a standalone model rather than a complex ensemble, our submission achieved higher recall than the top-ranked UOM ensemble and a highly competitive Micro-F_1 score (trailing by a marginal 0.0028 points). This demonstrates that our proposed architectural choices offer a highly effective and computationally lighter alternative to large-scale ensembling.

2. System Description

2.1. Problem Formulation

Given a clinical document x_i, the model must predict a subset $\hat{Y}_i \subseteq \mathcal{L}$ from the label space $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$, where $|\mathcal{L}| = 115$. We decompose this into $|\mathcal{L}|$ independent binary classification decisions, each governed by label-specific sigmoid activations and optimized through a weighted binary cross-entropy objective.

2.2. Model Architecture: Baymax (MedicalModelWithDescriptions)

The core architecture extends the standard transformer classification paradigm through two mechanisms: multi-sample dropout regularization and semantic label anchoring.

2.2.1. Encoder and Multi-Sample Dropout

We employ XLM-RoBERTa Large [3] as the backbone encoder ϕ_θ, which maps tokenized input sequences to contextualized representations. For a document x_i tokenized into $\mathbf{t}_i = (t_1, t_2, \ldots, t_T)$ with $T \leq 512$, we extract the [CLS] token representation:

$$\mathbf{h}_i = \phi_\theta(\mathbf{t}_i)_{[\text{CLS}]} \in \mathbb{R}^H, \quad H = 1024 \tag{1}$$

where H denotes the hidden dimensionality of the encoder.

For the minority of clinical records exceeding this 512-subword budget, we apply truncation from the *left*, discarding tokens from the beginning of the document while retaining the final 512 tokens. This design choice reflects the structural convention of Greek cardiology reports, in which the diagnostic impression, final assessment, and discharge summary—the segments most informative for ICD-10 code assignment—are typically located toward the end of the narrative, whereas introductory sections (e.g., patient history and admission circumstances) are comparatively less predictive of the target labels.

To stabilize gradient estimation under the high variance inherent in multi-label classification with 115 targets, we apply multi-sample dropout [4]. We define $K = 5$ dropout operators $\{\text{Drop}_k\}_{k=1}^{5}$ with rates $p_k = 0.1k$, yielding:

$$\hat{\mathbf{z}}_i = \frac{1}{K} \sum_{k=1}^{K} \mathbf{W}_c \, \text{Drop}_k(\mathbf{h}_i) + \mathbf{b}_c \tag{2}$$

where $\mathbf{W}_c \in \mathbb{R}^{|\mathcal{L}| \times H}$ and $\mathbf{b}_c \in \mathbb{R}^{|\mathcal{L}|}$ are the parameters of the classification head.

2.2.2. Semantic Label Anchoring

A central innovation of our approach is the formalization of label semantics as anchor points in the encoder's representation space [5]. Let $\mathbf{D} = \{d_1, d_2, \ldots, d_{|\mathcal{L}|}\}$ denote the official Greek-language ICD-10 descriptions for each code in $\mathcal{L}$, sourced from a curated lookup table. During initialization, we pre-compute normalized description embeddings:

$$\mathbf{e}_j = \frac{\phi_\theta(d_j)_{[\text{CLS}]}}{\|\phi_\theta(d_j)_{[\text{CLS}]}\|_2} \in \mathbb{R}^H, \quad \forall\, l_j \in \mathcal{L} \tag{3}$$

forming a label-description embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathcal{L}| \times H}$.

Importantly, $\mathbf{E}$ is computed once prior to fine-tuning and subsequently held *fixed* throughout training—implemented as a non-trainable buffer rather than a learnable parameter—so that the label-description manifold acts as a stable semantic reference frame against which document representations are aligned. The document encoder ϕ_θ and the scalar α, by contrast, continue to receive gradient updates, allowing the model to adapt how strongly document representations are drawn toward this fixed manifold without letting the manifold itself drift during optimization.

The semantic anchoring score is defined as the cosine similarity between the normalized document representation and the description manifold, modulated by a learnable scalar:

$$\mathbf{s}_i = \alpha \cdot \bar{\mathbf{h}}_i \, \mathbf{E}^\top \tag{4}$$

$$\bar{\mathbf{h}}_i = \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|_2} \tag{5}$$

where $\alpha \in \mathbb{R}$ is a learnable parameter initialized to 0.1, governing the magnitude of the semantic alignment signal relative to the discriminative logits.

The final output logits combine both components:

$$\hat{\mathbf{y}}_i = \hat{\mathbf{z}}_i + \mathbf{s}_i \tag{6}$$

This formulation establishes a dual-pathway architecture: the linear head $\hat{\mathbf{z}}_i$ operates as a purely discriminative classifier, while $\mathbf{s}_i$ provides a semantically grounded inductive bias. The learnable α allows the model to adaptively calibrate the contribution of each pathway during training.

2.3. Preprocessing: Knowledge-Informed Data Expansion

Clinical coding datasets exhibit a pronounced power-law distribution over label frequencies, where a small number of prevalent codes (e.g., `I10`, `I25.5`) account for the majority of annotations, while many codes occur in fewer than 30 training instances. Standard oversampling methods (e.g., SMOTE) are poorly suited for textual data. We therefore propose a *knowledge-informed expansion* strategy that combines frequency-adaptive replication [6] with domain-specific lexical perturbation.

2.3.1. Frequency-Aware Augmentation Multiplier

Let $\text{freq}(l)$ denote the occurrence count of label l in the training partition. We define a piecewise multiplier function:

$$m(l) = \begin{cases} 0 & \text{if } \text{freq}(l) < 30 \text{ (too rare for reliable augmentation)} \\ 4 & \text{if } 30 \leq \text{freq}(l) < 100 \\ 3 & \text{if } 100 \leq \text{freq}(l) < 200 \\ 2 & \text{if } 200 \leq \text{freq}(l) < 300 \\ 0 & \text{if } \text{freq}(l) \geq 300 \text{ (sufficiently represented)} \end{cases} \tag{7}$$

The decision to set $m(l) = 0$ for labels with $\text{freq}(l) < 30$—rather than assigning them the highest multiplier, as might seem intuitive for the rarest classes—is deliberate. Codes in this regime are typically

represented by only a handful of clinical narratives, often sharing highly idiosyncratic phrasing; aggressively replicating and lexically perturbing such a small pool of examples risks overfitting to narrow surface patterns and introducing synonym-induced noise that fails to generalize, rather than genuinely enriching the model's exposure to the underlying diagnostic concept. We therefore restrict synonym-based augmentation to the medium-frequency range (30–300 occurrences), where the curated dictionary offers sufficient lexical diversity to produce non-degenerate augmented variants, prioritizing the stabilization of these medium-frequency classes. The extremely rare codes are instead handled primarily through the positive-class weighting scheme described in Section 3.3.

For a multi-label record (x_i, Y_i), the effective multiplier is $M_i = \max_{l \in Y_i} m(l)$, ensuring that records containing at least one under-represented code receive augmentation proportional to the rarity of their most infrequent label. This strategy yielded, for example, an expansion from 2,000 original records to 7,312 training samples in Fold 1.

2.3.2. Stochastic Synonym Replacement as Regularization

Each augmented copy undergoes a stochastic lexical perturbation process parameterized by a curated clinical dictionary $\mathcal{S} : \mathcal{L} \to 2^{\mathcal{V}}$, which maps ICD-10 codes to sets of clinically equivalent terms, including synonyms, abbreviations, and morphological variants specific to Greek hospital documentation practice.

For each code $l \in Y_i$ with associated terms $\mathcal{S}(l)$, if a term $w \in \mathcal{S}(l)$ with $|w| \geq 4$ characters is found in the text, it is replaced with probability $p = 0.35$ by a uniformly sampled alternative $w' \sim \text{Uniform}(\mathcal{S}(l) \setminus \{w\})$. This stochastic replacement serves a dual purpose: it acts as a data-level regularizer that prevents the model from memorizing surface-level lexical patterns, while simultaneously teaching the encoder that semantically equivalent clinical expressions should map to similar regions of the representation space. Crucially, the restriction to curated medical synonyms ensures that the clinical meaning of each record is preserved under perturbation, unlike generic augmentation methods such as random word deletion or back-translation [7].

3. Training Strategy

3.1. Layer-wise Learning Rate Decay

Fine-tuning large pre-trained transformers on relatively small clinical datasets ($N = 2{,}500$) risks catastrophic forgetting of pre-trained representations in the lower encoder layers, which capture general linguistic features. We employ Layer-wise Learning Rate Decay (LLRD) [8] to address this. For XLM-RoBERTa Large with $L = 24$ transformer layers, the learning rate assigned to layer i is:

$$\eta_i = \eta_{\text{base}} \cdot \gamma^{L-i-1}, \quad i \in \{0, 1, \ldots, L-1\} \tag{8}$$

where $\eta_{\text{base}} = 7 \times 10^{-6}$ and $\gamma = 0.9$. The embedding layer receives $\eta_{\text{base}} \cdot \gamma^{L}$, while the classification head receives $5 \times \eta_{\text{base}}$. This exponential decay schedule ensures that lower layers encoding universal linguistic knowledge are updated conservatively, while task-specific upper layers and the classification head are adapted aggressively.

3.2. Stochastic Weight Averaging

Optimization in high-dimensional multi-label spaces is characterized by a complex loss landscape with many local minima of comparable F_1 performance but varying generalization properties. We employ Stochastic Weight Averaging (SWA) [9] to navigate toward flatter regions of the loss surface, which are empirically associated with improved generalization.

Starting at epoch $T_{\text{SWA}} = 30$ (out of 35 total epochs), we maintain a running average of model parameters:

$$\bar{\theta}_n = \frac{n \cdot \bar{\theta}_{n-1} + \theta_n}{n+1} \tag{9}$$

where θ_n denotes the parameters at epoch $T_{\text{SWA}} + n$ and $\bar{\theta}_0 = \theta_{T_{\text{SWA}}}$. This averaging over 6 checkpoints per fold was conducted in conjunction with Cosine Annealing with Warm Restarts ($T_0 = 10$, $T_{\text{mult}} = 1$), which cyclically explores the loss landscape before SWA aggregation consolidates the traversed solutions.

3.3. Loss Function and Class Imbalance Handling

The training objective is binary cross-entropy with logits, augmented with label-specific positive weights to address class imbalance:

$$\mathcal{L} = -\frac{1}{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} \Big[w_j^+ \cdot y_j \log(\sigma(\hat{y}_j)) + (1 - y_j) \log(1 - \sigma(\hat{y}_j)) \Big] \tag{10}$$

where $w_j^+ = \min\Big(\frac{N - n_j}{n_j}, 20.0\Big)$, with n_j denoting the number of positive instances for label j in the training partition. The cap at 20.0 prevents numerical instability for extremely rare codes. Additional training details include gradient accumulation over 4 micro-batches (effective batch size of 16), mixed-precision training via `torch.amp`, and gradient norm clipping at 1.0.

4. Post-Processing

4.1. Per-Label Dynamic Threshold Tuning

Rather than applying a global decision threshold $\tau = 0.5$ across all labels, we optimize label-specific thresholds $\{\tau_j\}_{j=1}^{|\mathcal{L}|}$ on the validation fold by grid search over the interval $[0.05, 0.95]$ with 50 uniformly spaced candidates, selecting $\tau_j^* = \arg\max_\tau F_{1,j}(\tau)$ for each label independently.

4.2. ICD-10 Hierarchy Enforcement

We enforce parent-child relationships defined by the ICD-10 ontological structure [10]. For known hierarchical pairs (e.g., `I11` $\rightarrow$ `I10`, `I22` $\rightarrow$ `I21`), the parent code's predicted probability is set to the maximum of its own value and that of any active child:

$$\hat{p}_{\text{parent}} \leftarrow \max\big(\hat{p}_{\text{parent}},\, \hat{p}_{\text{child}}\big) \tag{11}$$

This ontological consistency enforcement is applied after probability averaging across the 5-fold ensemble but before thresholding, ensuring logical coherence in the final predictions.

5. Results

The system was evaluated under stratified 5-fold cross-validation on the ELCardioCC training set comprising 2,500 Greek cardiology records annotated with 115 ICD-10 codes.

Table 1
Per-fold performance of the Baymax system. "Best Micro-F_1" is the peak validation F_1 at a fixed threshold of 0.5. "Tuned F_1" incorporates per-label threshold optimization and ICD-10 hierarchy enforcement.

Fold	Best Micro-F_1	Tuned F_1 (+Hierarchy)
Fold 1	0.8339	0.8484
Fold 2	0.8304	0.8442
Fold 3	0.8339	0.8495
Fold 4	0.8275	0.8447
Fold 5	0.8253	0.8405
Mean	**0.8302**	**0.8455**

The results demonstrate consistent performance across folds (standard deviation $\sigma = 0.0035$ on the tuned metric), indicating robust generalization. The post-processing pipeline contributed an average improvement of $\Delta F_1 = +0.0153$ over raw predictions, with per-label threshold tuning accounting for the majority of this gain. SWA evaluation yielded competitive results (e.g., SWA $F_1 = 0.8310$ vs. best checkpoint $F_1 = 0.8304$ in Fold 2), confirming its effectiveness in finding well-generalized weight configurations. The training dynamics showed convergence within 35 epochs per fold, with the cosine annealing schedule providing three full learning rate cycles ($T_0 = 10$) that enabled progressive refinement of the model parameters.

5.1. Learned Semantic Anchoring Weight

Across the five folds, the learnable scalar α, initialized at 0.1, converged to a mean final value of $\alpha \approx 0.1028$. This stability indicates that the semantic anchoring pathway maintained a meaningful, supplementary contribution throughout the training process. Rather than being suppressed by the primary discriminative head, the fixed description embeddings consistently provided a beneficial baseline signal for label representation, effectively acting as a regularization mechanism that prevented the network from overfitting solely on the classification layer's weights.

5.2. Official Test Set Results

Beyond the 5-fold cross-validation results reported above, our final system was evaluated on the official, held-out ELCardioCC test set. Predictions were produced by averaging probabilities across the five fold models, followed by ICD-10 hierarchy enforcement and per-label threshold tuning, as described in Section 4. Under the submission identifier `Baymax_submission_v1` (team `Georgios_1`), our system achieved a Precision of 0.861783, a Recall of 0.865845, and a Micro-F_1 of 0.863809 on the official test set. These results are consistent with our cross-validation estimates, confirming that the proposed architecture and post-processing pipeline generalize well to unseen data.

6. Conclusion

We presented the Baymax system for Greek clinical coding at the ELCardioCC task of CLEF 2026. Our approach combines semantic label anchoring with frequency-aware knowledge-informed augmentation and a carefully designed optimization pipeline. The system achieved a mean Micro-F_1 of 0.8455, demonstrating the effectiveness of integrating label semantics as an inductive bias for multi-label clinical text classification in a morphologically rich, low-resource language setting.

Acknowledgments

This work was conducted as part of the CLEF 2026 shared task evaluation campaign. Computational resources were provided by the University of Macedonia. We thank the ELCardioCC task organizers for providing the annotated dataset and evaluation framework. Furthermore, we express our profound gratitude to Professor Ioannis Refanidis, Director of the MSc in Artificial Intelligence and Data Analytics at the University of Macedonia. His dedicated teaching in Natural Language Processing, introduction to transformer-based architectures, and valuable guidance in entering the ELCardioCC competition were instrumental to the completion of this work.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini and Claude to support specific phases of the research and writing process, in accordance with the CEUR-WS policy on AI usage. Specifically:

- **Grammar and spelling check, Paraphrase and reword:** The tools were employed to assist with grammar correction, syntactic polishing, and ensuring the semantic clarity of the final manuscript.
- **Code generation and debugging:** The tools were utilized as supplemental programming aids to partially generate, optimize, and debug scripts within the experimental coding pipeline.

Following the application of these tools, the authors thoroughly reviewed, verified, and edited all generated text and source code to ensure scientific accuracy. The authors take full responsibility for the integrity and ultimate content of this publication.

References

[1] A. Nentidis, A. Krithara, G. Paliouras, et al., Overview of bioasq 2026: The fourteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.

[2] ELCardioCC Task Organizers, Overview of the ELCardioCC task on automated ICD-10 coding of greek cardiology clinical records at CLEF 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR-WS.org, 2026.

[3] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[4] H. Inoue, Multi-sample dropout for accelerated training and better generalization, arXiv preprint arXiv:1905.09788 (2019).

[5] Z. Yuan, C. Tan, S. Wang, F. Huang, S. Huang, Code synonyms do matter: Multiple synonyms matching network for automatic icd coding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2022, pp. 82–88.

[6] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6382–6388.

[7] Y. Papanikolaou, A. Pierleoni, DARE: Data augmented relation extraction with GPT-2, arXiv preprint arXiv:2004.13845 (2020).

[8] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: International Conference on Learning Representations, 2020.

[9] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, A. G. Wilson, Averaging weights leads to wider optima and better generalization, in: Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence, 2018, pp. 876–885.

[10] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: Proceedings of the 2018 Conference of the North Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, pp. 1101–1111.