A Multi-Component System for Multi-Label ICD-10 Classification of Greek Cardiology Discharge Summaries

BioASQ ElCardioCC at CLEF 2026

Nikolaos Strafiotis[1], Panteleimon Stanimeros[1], Vasiliki Katsara[1], Georgios Chalkias[1], Glykeria Tsavlidou[1], Stelios Magalios[1], Effrosyni Nalmpanti[1] and Panteleimon Stamatakis[1]

[1]*MSc in Artificial Intelligence & Data Analytics*
Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece

Abstract

We present a multi-component system for the ElCardioCC 2026 shared task on automatic ICD-10 coding of Greek cardiology discharge summaries. Our approach builds and trains six complementary components from scratch: a fine-tuned Greek BERT model, two XLM-RoBERTa variants (base and large), a hybrid information retrieval module, a dictionary-based rule system, and a named entity recognition with entity linking pipeline. Individual models are trained with task-specific loss functions (Asymmetric Loss, ZLPR) and per-label threshold tuning on the validation set. Rather than selecting a single model, we fuse all six outputs through a declarative metaheuristic ensemble that searches over eleven base strategies and three composition operators via random restarts, hill-climbing, and Variable Neighbourhood Search (VNS). Our best submission achieves a micro-F1 of **0.8667** on the official test set, ranking first among all participating teams and improving by approximately 1.8 percentage points over the best standalone model.

Keywords

automatic clinical coding, multi-label classification, metaheuristic ensemble, Greek BERT, cardiology discharge summaries, low-resource NLP

1. Introduction

Clinical coding assigns standardised diagnostic codes to free-text medical records, a mandatory step in hospital administration, epidemiological surveillance, and insurance reimbursement. Manual coding is slow, costly, and prone to inter-annotator disagreement, motivating automatic approaches [5]. The ElCardioCC 2026 shared task, part of the fourteenth BioASQ challenge at CLEF 2026 [9, 10], targets this problem for Greek cardiology discharge summaries, requiring systems to predict all applicable ICD-10 codes per document, a multi-label classification problem over 115 codes.

Three challenges define the setting. First, Greek is a morphologically rich, low-resource language for clinical NLP, with discharge summaries mixing Greek terminology and transliterated abbreviations (PCI, TAVI, CABG). Second, some summaries exceed the 512-token BERT limit. Third, the 115 ICD-10 codes span frequencies from a single instance to over 1,000.

No single architecture dominates across all labels: neural models excel on frequent labels, dictionary rules on procedural codes, retrieval on rare coverage, and NER+EL on mention-level disambiguation. We therefore build and train six complementary components, each targeting different failure modes, and fuse their outputs through a declarative metaheuristic ensemble that searches over 11 base strategies and 3 composition operators on the validation set.

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany
aid26012@uom.edu.gr (N. Strafiotis); aid26006@uom.edu.gr (P. Stanimeros); aid26009@uom.edu.gr (V. Katsara); aid26004@uom.edu.gr (G. Chalkias); aid26005@uom.edu.gr (G. Tsavlidou); aid26002@uom.edu.gr (S. Magalios); aid26014@uom.edu.gr (E. Nalmpanti); aid26013@uom.edu.gr (P. Stamatakis)

2. Background

Automatic ICD coding from clinical text has been addressed with CNN/RNN architectures [14], BERT-based models [4, 7], and hybrid retrieval-classification pipelines [5]. For Greek clinical NLP, Greek BERT [6] provides a domain-relevant starting point, while XLM-RoBERTa [1] handles Greek–Latin code-switching in clinical abbreviations. Label imbalance is the central challenge in multi-label coding: Asymmetric Loss (ASL) [13] addresses this by suppressing easy negatives more aggressively than positives. Ensemble methods [3] and VNS-based search [8] over non-differentiable objectives complement neural approaches for rare-label coverage.

3. Task and Dataset

3.1. Task Definition

ElCardioCC 2026 is a multi-label ICD-10 coding shared task for Greek cardiology discharge summaries, organised as part of BioASQ at CLEF 2026 [9, 10], with 115 target codes.

3.1.1. Evaluation metric.

The primary metric is a group-level F1 score that reflects clinical equivalence among related codes. Each gold annotation is a *group*, a set of interchangeable ICD-10 codes. Scoring is defined as follows:

- TP: a gold group is matched if at least one of its member codes is predicted.
- FN: gold groups with no predicted member.
- FP: predicted codes that do not belong to any gold group.

Multiple predictions within the same group do not increase TP. Precision, recall, and F1 are then:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad F1 = \frac{2PR}{P + R} \tag{1}$$

This differs from standard multi-label micro-F1: predicting any one equivalent code fully satisfies the corresponding group, whereas predicting codes outside any gold group is penalised as false positives. **Example.** Given gold groups $\{\{\text{R55}\}, \{\text{I10}, \text{I11}\}, \{\text{I44}\}, \{\text{Z95}\}\}$:

- Predicting {R55, I10, I11, I44, Z95}: TP=4, FP=0, F1=1.0; both I10 and I11 satisfy the same group without penalty, since neither falls outside a gold group.
- Predicting {R55, I10, I44, Z95, I20}: TP=4, FP=1, F1=0.889; I20 does not belong to any gold group and is counted as a false positive.

3.2. Dataset

The dataset consists of 2,500 anonymised discharge summaries. We split these into training (2,000), validation (250), and test (250) sets; a separate blind test set is held by the organisers. Documents are moderately long: mean 280 tokens (≈2,000 characters), maximum 1,218 tokens. Only 2.4% of documents exceed the 512-token BERT limit.

The 115 codes are highly skewed. The five most frequent (Z95, I25, I21, I48, E11) account for ∼41% of all annotations; 30 codes appear fewer than 10 times and 4 codes appear only once (Table 1). This long-tailed distribution requires explicit strategies for rare-label coverage.

Table 1
Label frequency distribution across 115 ICD-10 codes.

Frequency band	# codes	Approx. % of instances
≥ 500	5	41%
100–499	22	36%
10–99	58	21%
< 10	30	<2%

3.3. Preprocessing and Data Augmentation

Text cleaning is shared across all splits: whitespace collapse and removal of non-essential special characters (preserving hyphens, commas, slashes, parentheses). Greek BERT additionally receives lowercasing and accent stripping; XLM-RoBERTa receives whitespace-only normalisation to preserve abbreviation signals ("PCI", "TAVI"). ICD-10 annotations are encoded as 115-dimensional binary vectors.

Stratified splitting uses a two-stage Multilabel Stratified Shuffle Split [15] to produce an 80/10/10 partition, ensuring every validation and test label also appears in training. Of the 2,500 samples, 1,453 include mention-level annotations enabling NER+EL training.

Synonym augmentation (XLM-RoBERTa base only) draws directly on the dictionary baseline's term-to-code table (Section 4.4): for each gold code on a document with at least two known surface forms, one occurrence of a matched term is replaced, with probability $p = 0.35$, by another randomly chosen term mapped to the same code. Frequency-adaptive multipliers generate additional augmented copies per document ($\times 4$ for codes with 30–99 training instances, $\times 3$ for 100–199, $\times 2$ for 200–299; codes with fewer than 30 or at least 300 instances are not augmented). For example, for code I21 (myocardial infarction), a discharge summary sentence translating to "*Diagnosis: acute myocardial infarction (STEMI) of the inferior wall*" (original in Greek) may have the matched term "STEMI" replaced by an alternative I21 surface form such as "NSTEMI" drawn from the same term list, yielding "*Diagnosis: acute myocardial infarction (NSTEMI) of the inferior wall*"; because substitution is a case-insensitive regex replacement rather than a grammar-aware rewrite, it can occasionally produce minor casing or spacing artefacts around the substituted span, which we treat as an acceptable trade-off given the resulting recall gains on underrepresented codes.

4. Methodology

The system pipeline is: (i) train six independent components; (ii) tune per-model decision thresholds on the validation set; (iii) search for the best ensemble fusion strategy. Figure 1 provides an overview.

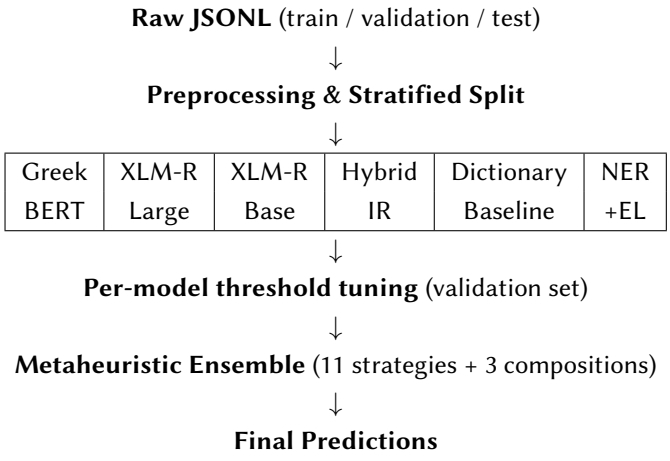

Figure 1: High-level system pipeline.

4.1. Greek BERT Multi-Label Classifier

4.1.1. Architecture.

We fine-tune `nlpaueb/bert-base-greek-uncased-v1` (Greek BERT) [6] with a two-layer MLP head: mean+CLS concatenation pooling (1536-dim) → Dropout(0.2) → Linear($1536 \to 768$, GELU) → Dropout(0.2) → Linear($768 \to 115$). Concatenating mean-pooled hidden states and the `[CLS]` vector outperformed plain CLS extraction by ∼0.8 percentage points in our ablation, as diagnostic evidence is distributed across the full document. Max sequence length: 256 tokens.

4.1.2. Training.

Table 2 lists the final hyperparameters. ASL [13] replaced binary cross-entropy (BCE) after Phase 2, raising validation micro-F1 from ≈0.74 to 0.788 by suppressing easy negatives. ASL applies asymmetric focusing to positive and negative terms:

$$\mathcal{L}_{\text{ASL}} = \sum_{y} \begin{cases} (1-p)^{\gamma^+} \log p & y = 1 \\ (p_m)^{\gamma^-} \log(1-p_m) & y = 0 \end{cases} \tag{2}$$

where $p_m = \max(p - m, 0)$ shifts and clips easy negatives, $\gamma^+ = 1$, $\gamma^- = 4$, $m = 0.05$. **LLRD** (factor 0.90) updates lower encoder layers at LR $\times\ 0.90^l$, preserving pre-trained Greek representations while adapting upper layers to the clinical domain. The model evolved across four phases: BCE baseline (0.74) → ASL (0.788) → LLRD + MLP (0.811 at t=0.5) → final P4 tuning (0.7984 at t=0.6, 0.8062 after per-label threshold tuning). The drop from 0.811 to 0.7984 reflects the stricter active threshold (t=0.6 vs t=0.5); per-label tuning then recovers most, though not all, of this gap (0.8062), approaching but not exceeding the 0.811 milestone, which was obtained under a less conservative Phase 3 configuration.

Table 2
Greek BERT final hyperparameters.

Hyperparameter	Value
Max length / batch	256 tok / 32 (16 + grad. accum. 2)
Learning rate / LLRD	3×10^{-5} / 0.90 per layer
Warmup / weight decay	6% / 0.01
Loss	ASL ($\gamma^- = 4$, $\gamma^+ = 1$, clip $= 0.05$)
Epochs / patience	30 / 7 (BF16)
Pooling / head	mean+CLS concat / MLP 1536→768→115
Dropout	0.2
Active eval threshold	0.6

4.1.3. Threshold tuning.

A two-pass sweep (step 0.01) selects $t_l^* = \arg\max_{t \in [0.45, 0.75]} F1_l(t; \mathcal{D}_{\text{val}})$ per label independently. Pass 2 accepts per-label values only for labels with ≥15 positive validation examples and ΔF1 > 0.0015. This adds ≈0.8 percentage points over the active threshold (t=0.6) and ≈3.7 percentage points over a fixed t=0.5 baseline, bringing the tuned validation micro-F1 to 0.8062.

4.2. XLM-RoBERTa Classifiers

XLM-RoBERTa's 250K multilingual vocabulary handles Greek–Latin code-switching better than Greek BERT's 35K vocabulary.

4.2.1. Long-document strategy.

A sliding window (stride 256) handles documents exceeding 512 tokens. At inference, chunk logits are fused as $\hat{y} = 0.5\,\bar{y}_{\text{mean}} + 0.5\,y_{\max}$, balancing average evidence with the most confident chunk signal.

4.2.2. Large variant.

ZLPR loss, effective batch 32 (batch 2, grad. accum. $\times 16$), LR 2×10^{-5}, weight decay 0.05, dropout 0.15, 40 epochs (patience 5). Validation F1 ceiling 0.7538; slower convergence and memory constraints (560M parameters) led us to focus development on Greek BERT.

4.2.3. Base variant.

Implements MedicalModelWithDescriptions: $\mathbf{h} \in \mathbb{R}^{768}$ is the `[CLS]` hidden state of the encoder's last layer, which passes through 5 parallel dropout layers ($p = 0.1i$, $i \in \{1 \dots 5\}$) whose classifier logits are averaged (multi-sample dropout) to obtain $W\mathbf{h}$. A semantic anchoring term adds cosine similarity between the document embedding and frozen ICD-10 description embeddings, scaled by a learnable scalar α:

$$\hat{y} = W\mathbf{h} + \alpha \cdot \tfrac{\mathbf{h}}{\|\mathbf{h}\|}\mathbf{D}^{\top} \tag{3}$$

where rows of $\mathbf{D}$ are unit-normalised `[CLS]` embeddings of the Greek ICD-10 descriptions, encoded once with a frozen copy of the same pretrained encoder, so the term is a true cosine similarity. $\hat{y}$ are logits; `BCEWithLogitsLoss` applies the sigmoid internally during training, and it is applied explicitly at inference. α converges from 0.099 to $\approx$0.12, confirming a consistent but small semantic correction. Training: AdamW LR 2×10^{-5}, early stopping (patience 4), per-class positive weights (ceiling 20). Post-processing: dynamic threshold tuning + ICD-10 hierarchy enforcement + co-occurrence rules. Validation micro-F1: **0.8101**. Despite using a smaller backbone, the Base model outperforms the Large variant (0.8101 vs. 0.7538). The primary factors are the semantic anchoring mechanism and task-specific post-processing absent in Large; GPU memory constraints (560M parameters) also forced micro-batch size 2 with gradient accumulation, limiting effective per-step gradient quality.

4.3. Information Retrieval Module

Three retrieval channels (BM25 [12], TF-IDF, dense MiniLM embeddings) are fused by Reciprocal Rank Fusion [2] (RRF$(d) = \sum_i 1/(k + r_i(d))$, $k = 60$). Codes with RRF score $\geq 0.22 \times$ top score, capped at 12 per document, are predicted. The cap is critical: raising it to 25 collapses F1 from 0.662 to 0.140 (precision drops from 0.57 to 0.10 while recall rises modestly) because additional hits that pass the 0.22$\times$ score threshold are predominantly false positives.

4.4. Dictionary Baseline

4.4.1. Dictionary construction.

The term dictionary is built automatically from the gold mention-level annotations supplied with the training data (character-span, code-labelled clinical mentions, available for 1,453 of the 2,500 documents), rather than sourced from an external terminology or hand-typed from scratch. For every observed (mention, code) pair we count occurrences of each normalised mention string and retain it as a term if its normalised length is at least 4 characters and it does not appear in a manually curated blacklist; a term is assigned to every code accounting for at least 20% of its occurrences, so an ambiguous surface form can map to more than one code (22 of 1,680 terms do, e.g. a phrase shared between I21 and I24). This yields a CSV of 1,680 term-to-code entries covering 83 of the 115 target codes.

4.4.2. Normalisation and matching.

Both dictionary terms and document text are normalised identically: lowercasing, Unicode NFD decomposition with removal of combining marks (stripping Greek tonos and Latin accents), removal of non-alphanumeric characters, and whitespace collapse; no stemming, lemmatisation, or stopword removal is applied. Matching is exact substring matching, implemented with an Aho–Corasick automaton built once over the full term set so all 1,680 terms are located in a single linear pass rather than via repeated per-term scans; this was an efficiency refactor partway through the project and left matching semantics, and validation micro-F1, unchanged.

4.4.3. Rule layers.

A manually curated 27-term blacklist removes short or highly ambiguous fragments identified by per-term precision analysis on the training set (e.g. a two-letter abbreviation that fired on all 2,500 documents with near-zero precision). Nineteen hand-authored clinical-pattern rules fire on procedure, device, and syndrome keywords not reliably captured by literal mention matching, e.g. PCI/PPCI/PTCA $\rightarrow$ Y84, TAVI $\rightarrow$ {Y84, Z95, I35}, Takotsubo $\rightarrow$ I43. Three co-occurrence rules mined from training-set code statistics enforce clinical consistency: I10 $\leftrightarrow$ I11 (100% co-occurrence in training), I22 $\Rightarrow$ I21 (100%), and I25 $\Rightarrow$ Z95 (75%, applied as a recall heuristic since most I25 patients in this corpus have vascular implants); the last introduces false positives that are corrected downstream by merge_and. A code is predicted whenever any dictionary term or rule fires; there is no confidence score or ranking threshold, since this component is a deterministic lexical baseline rather than a scored classifier. An optional word-boundary matching mode, section/negation filters, and a description-level fuzzy fallback are implemented in the codebase but disabled by default and were not used for the reported results.

Validation micro-F1 **0.7176** (precision 0.65, recall 0.80), covering 99.76% of documents.

4.5. Named Entity Recognition and Entity Linking

Greek BERT is fine-tuned as a BIO sequence tagger with a single entity type (MED) and a Partial CRF [16]. The coarse type maximises recall; the Partial CRF enforces label consistency via Viterbi decoding and enables learning from documents with only document-level annotations. Training: 6 epochs, batch 8, LR 2×10^{-5}, early stopping patience 2.

NER spans are augmented by Aho–Corasick dictionary matching; containment-based deduplication retains the longest span. Each mention m is linked to ICD-10 code c by:

$$\text{score}(m, c) = 0.6 \cdot P_{\text{prior}}(c \mid m) + 0.4 \cdot \cos(\mathbf{e}_{\text{ctx}}, \mathbf{e}_c) \tag{4}$$

where P_{prior} is the (mention, code) co-occurrence frequency from training and $\mathbf{e}_{\text{ctx}}$, $\mathbf{e}_c$ are MiniLM embeddings of a $\pm$200-char context window and the Greek ICD-10 description respectively. Full document-level dictionary boosting is applied after mention linking. Best validation micro-F1: **0.7223** at epoch 5.

4.6. Metaheuristic Ensemble

Rather than fixing a single fusion rule [3], we define a declarative strategy space in YAML and evaluate all strategies on the validation set, selecting the best-scoring configuration without code changes.

4.6.1. Base strategies.

Eleven base strategies cover the main fusion paradigms. The primary strategy is the weighted ensemble, a convex combination of all component probability vectors:

$$\hat{y} = \sum_{i=1}^{6} w_i \, \hat{y}^{(i)}, \quad w \in \mathcal{W} = \left\{ w \in \mathbb{R}^6 : \textstyle\sum_i w_i = 1,\ w_i \geq 0 \right\} \tag{5}$$

Additional strategies address complementary failure modes: majority vote pools predictions across all weight-search restarts (both classic and VNS); gated secondary falls back to IR/NER for documents where the weighted ensemble is uncertain; top-K pruning caps predictions per document; frequency-bucketed thresholding applies separate thresholds per label-frequency band; per-label champion routes each label to its highest-validation-F1 component; per-label champion + vote adds a label if the champion fires or a minimum number of other models agree; per-patient score routing selects the best-scoring model per document; and label correction mines co-occurrence rules $(a \to b)$ from validation predictions, predicting b whenever a is predicted above a mined threshold.

4.6.2. Composition operators.

Three label-set algebra operators defined in `strategy_compositions.yaml` compose pairs of base strategy outputs $\{\hat{L}^{(s)}\}_{s\in\mathcal{S}}$, where $\hat{L}^{(s)} = \{l : \hat{y}_l^{(s)} \geq t_l\}$:

$$\hat{L}^{\text{or}} = \bigcup_{s\in\mathcal{S}} \hat{L}^{(s)} \qquad \text{(maximises recall)} \tag{6}$$

$$\hat{L}^{\text{and}} = \bigcap_{s\in\mathcal{S}} \hat{L}^{(s)} \qquad \text{(maximises precision)} \tag{7}$$

$$\hat{L}^{k\text{-of-}n} = \left\{ l : \sum_{s\in\mathcal{S}} \mathbb{1}[l \in \hat{L}^{(s)}] \geq k \right\} \qquad \text{(balances precision/recall)} \tag{8}$$

The best submission applies merge_and over the weighted ensemble and the label correction module, letting the correction step remove false positives from the ensemble output.

4.6.3. Weight optimisation.

Weights are optimised directly on the non-differentiable validation micro-F1 objective:

$$w^* = \arg\max_{w\,\in\,\mathcal{W}} \text{F1}_\mu\Big(\hat{y}(w),\, y^{(\text{val})}\Big) \tag{9}$$

Table 3 reports the weights found by the classic optimiser (validation micro-F1 0.8488, matching the VNS optimum to four decimal places). `ner_el` receives the lowest weight under both optimisers (0.10–0.16), consistent with its role as a secondary recall source rather than a primary driver of the weighted fusion; Greek BERT and XLM-RoBERTa Base receive the two highest weights, matching their standalone ranking.

Table 3
Weighted-ensemble weights w_i found by the classic optimiser (validation set).

Component	Weight w_i
mlc_greek_bert	0.6635
xlm_r_base	0.6365
dictionary_baseline	0.5151
information_retrieval	0.4247
xlm_r_large	0.1197
ner_el	0.1000

Two optimisers run independently, each for 2 restarts of 10,000 evaluations: (1) *classic*: random search over $\mathcal{W}$ followed by pairwise hill-climbing; (2) *VNS* [8]: shaking (radius-growing perturbation, $k_{\max}$=5) followed by a short hill-climb, with k reset on improvement and a random restart when all neighbourhoods are exhausted. The optimiser achieving higher validation micro-F1 provides the final weights; restart predictions from both are pooled for majority-vote strategies.

4.6.4. Submitted strategies.

The five submissions span the precision–recall trade-off: standalone Greek BERT (80/10 split and 100% data) establish the single-model ceiling; three ensemble submissions apply merge_and (precision-oriented), merge_k2 (balanced), and a safer-threshold merge_k2 variant. The merge_and composition achieves the best F1 by combining the weighted ensemble with the label correction module's precision boost.

5. Experimental Results

All micro-F1 values reported in this section use the group-level metric defined in Section 3: a gold group is satisfied by any one of its member codes, and over-predicting within a group counts as false positives.

5.1. Individual Component Performance

Figure 2 shows standalone group-level micro-F1 on our internal test split for each component. Greek BERT is the strongest single model (0.8196 on the internal test split), ahead of XLM-RoBERTa Base (0.8050). Dictionary and NER+EL perform comparably; the IR module has the lowest F1 but the highest recall (0.83), making it a key coverage contributor in the ensemble.

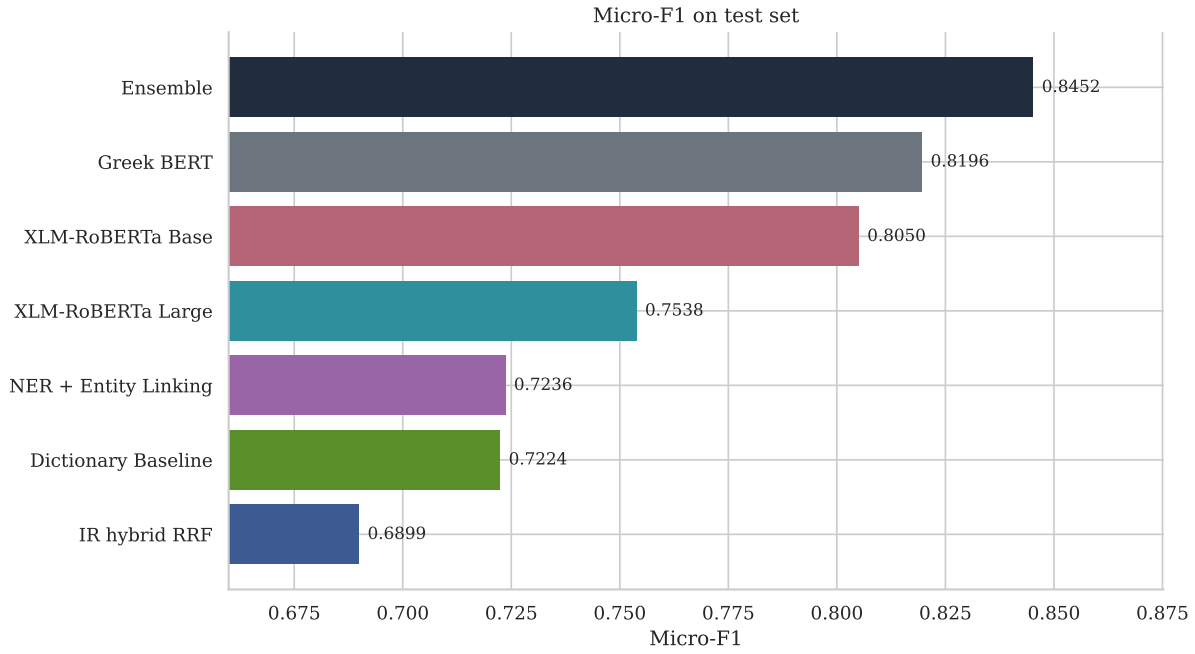

Figure 2: Standalone group-level micro-F1 per component on our internal 250-document test split. XLM-RoBERTa Large reports validation F1 (no separate test submission). Official test set scores appear in Table 6.

5.2. Ablation Studies

5.2.1. Threshold tuning.

Table 4 shows the effect of progressively finer thresholding. Per-label tuning adds ≈3.7 percentage points over a fixed t=0.5 baseline and ≈0.7 percentage points over the global optimum.

Table 4
Threshold tuning ablation for Greek BERT.

Strategy	Val. micro-F1
Fixed $t = 0.5$	0.7693
Fixed $t = 0.6$ (active)	0.7984
Global sweep $[0.45, 0.75]$	0.7989
Per-label sweep (2-pass)	0.8062

5.2.2. Impact of LLRD.

Layer-wise learning rate decay (factor 0.90) preserves pre-trained lower-layer representations while adapting upper layers to the clinical domain, most notably benefiting rare labels.

5.2.3. Loss function comparison.

ASL outperformed both binary cross-entropy and focal loss through better recall on rare labels. BCE gave good precision but lower recall; focal loss improved recall over BCE but less than ASL due to its symmetric treatment of positive and negative examples.

5.2.4. Ensemble composition.

Table 5 compares the composition operators. The intersection (merge_and) achieves the best F1 by combining the weighted ensemble with the label correction module's precision boost.

Table 5
Ensemble composition operator comparison (validation set).

Strategy	Recall	Precision	F1 trend
Union (merge_or)	highest	lowest	moderate
Weighted ensemble	mid	mid	strong
k-of-n (merge_k2)	mid	mid+	strong
Intersection (merge_and)	mid−	highest	best

5.3. Official Test Set Results

Table 6 reports all five submissions on the official test set. The best submission uses the merge_and strategy (weighted ensemble intersected with the label correction module).

Table 6
Official ElCardioCC 2026 test set results.

Submission	Recall	Precision	F1
ensemble (merge_and: weighted + correction)	0.8510	0.8830	**0.8667**
ensemble (merge_k2: weighted + per-label + corr.)	0.8687	0.8539	0.8613
ensemble (safer thr, merge_k2)	0.8767	0.8431	0.8596
mlc_greek_bert_100 (100% data)	0.8194	0.8811	0.8491
mlc_greek_bert (80/10/10 split)	0.8310	0.8675	0.8489

The ensemble outperforms the best standalone model by +1.8 percentage points (0.8667 vs. 0.8491), confirming the complementarity of the six components. The full-data Greek BERT achieves virtually the same F1 as the split version (0.8491 vs. 0.8489), suggesting the validation set contributes little additional signal at this scale.

The precision-oriented merge_and yields the highest overall F1 (0.8667), while merge_k2 variants achieve higher recall at the cost of precision, consistent with the 2-of-3 voting design. The safer-threshold merge_k2 reaches the highest recall (0.8767) but the lowest precision (0.8431).

5.3.1. Statistical significance.

To confirm that the ensemble's gain over the best standalone model is not an artefact of the small validation set, we compare the weighted ensemble (Table 3) against the best single component (Greek BERT) on the 250-document validation split using two paired tests. A document-level paired bootstrap (10,000 resamples with replacement) gives a mean micro-F1 difference of $+0.0323$ (95% CI $[+0.0212, +0.0437]$), with zero of the 10,000 resamples favouring the standalone model ($p < 10^{-4}$). A McNemar test over gold-group-level outcomes (each gold group counted as a matched pair: covered by the ensemble only, by Greek BERT only, by both, or by neither) gives 88 groups recovered only by the ensemble against 21 recovered only by Greek BERT, a highly significant asymmetry ($p \approx 2.6 \times 10^{-10}$). Both tests indicate the ensemble's improvement over the strongest individual component is statistically robust rather than due to chance variation across documents.

5.4. Computational Complexity

Table 7 reports backbone parameter counts and observed single-run wall-clock training time for each component. The neural components dominate training cost: XLM-RoBERTa Large (560M parameters) takes the longest to converge despite plateauing at a lower validation F1 (Section 4.2), while Greek BERT (113M parameters) reaches its best checkpoint in under 15 minutes over 30 epochs. The dictionary baseline and information retrieval module involve no gradient-based training: the former builds its term automaton and applies rule layers in a few seconds, and the latter's cost is dominated by one-off dense MiniLM encoding of the document collection, completing in under a minute. At inference time, all components process the full validation set (250 documents) in well under a minute per component on commodity hardware (dictionary and IR: single-digit seconds; neural components: one forward pass per document/chunk), and the metaheuristic ensemble search itself (10,000 evaluations $\times$ 2 restarts $\times$ 2 optimisers over precomputed score matrices) completes in under 5 minutes on CPU alone, since it operates on cached prediction scores rather than raw text.

Table 7
Backbone parameter counts and single-run training wall-clock time per component.

Component	Backbone params	Training time
XLM-RoBERTa Large	560M	2h 14m 52s
XLM-RoBERTa Base	278M	1h 35m 43s
Greek BERT (mlc)	113M	14m 50s
NER+EL (Greek BERT)	113M	3m 05s
Information Retrieval	–	≈30s (no training; one-off dense encoding)
Dictionary Baseline	–	≈5s (no training; rule construction only)

Training times were logged from independent single runs (Weights & Biases run duration for the four neural components; direct wall-clock measurement for the non-neural components) and are not strictly comparable across hardware, but the order-of-magnitude gap between the neural components and the rule-based/retrieval components is stable regardless of environment.

5.5. Error Analysis

Rare labels. The 30 codes with <10 training instances are the primary error source. For 10 of these, per-label thresholding is not applicable. The dictionary recovers some (e.g. E85 via the amyloidosis rule), but codes with no lexical surface form (e.g. M35, R57) remain unrecovered.

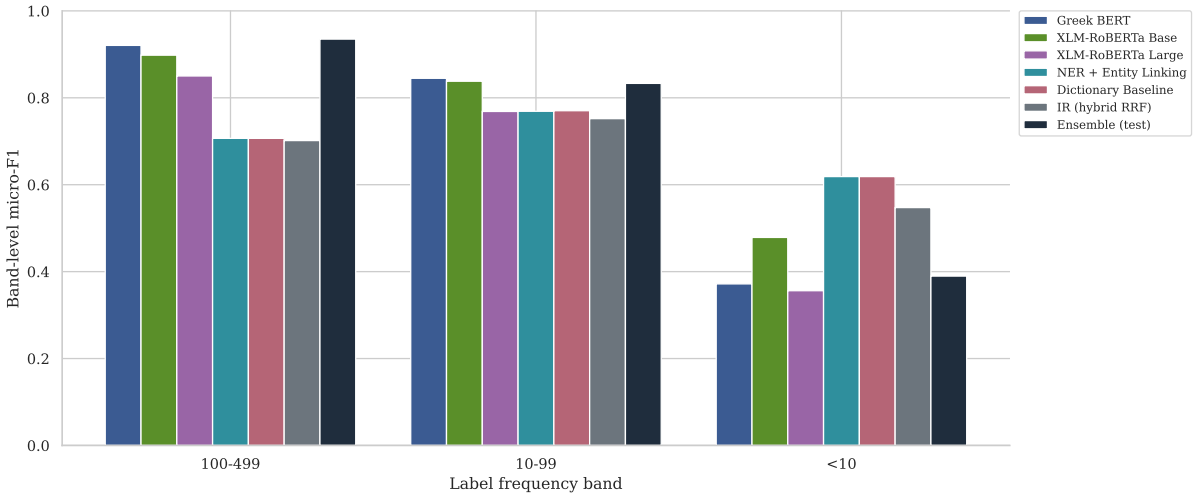

Figure 3: Macro-F1 per component across label frequency bands (three bands shown; no code reaches $\geq$500 instances in the 250-document validation set). All models degrade sharply on rare labels ($<$10 instances); the ensemble provides no recovery for this band.

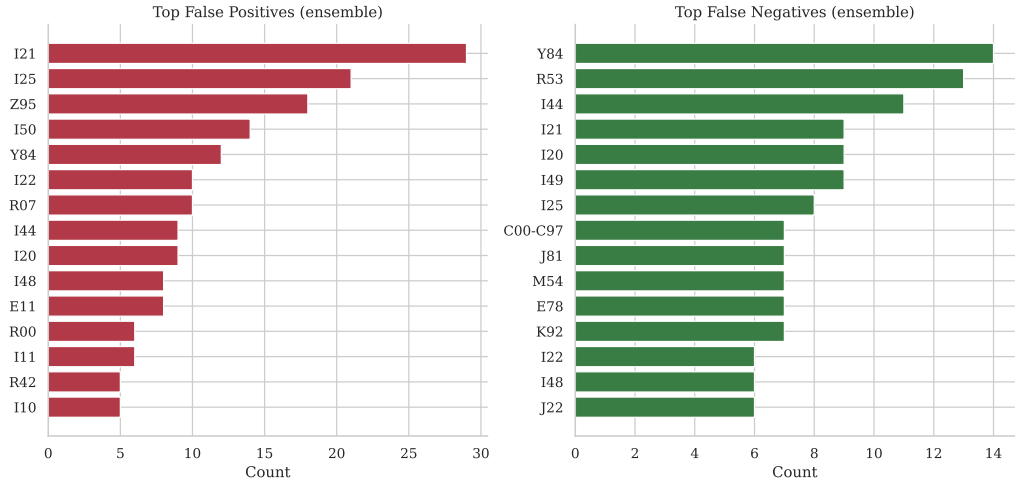

Figure 4: Top false-positive and false-negative codes for the best ensemble submission.

Coverage gaps. 31 label types lie outside the dictionary's coverage, including metabolic (E00), haematological (D47), and musculoskeletal codes, which require inference from laboratory values or implicit multi-sentence evidence.

Label confusion. I21/I22/I25 (acute MI, subsequent MI, chronic ischaemic disease) frequently co-occur and require temporal reasoning to assign correctly, beyond current bag-of-token models. Z95 (vascular implants) is the main false-positive source due to ubiquitous stent and pacemaker mentions, motivating the merge_and correction step.

Calibration. Greek BERT under-predicts rare labels: probabilities for rare codes often fall below 0.5. Per-label thresholding helps but becomes unreliable for labels with fewer than 15 positive validation examples.

6. Conclusion

We presented a multi-component system for ElCardioCC 2026, achieving micro-F1 **0.8667** on the official test set and ranking first among all participating teams. The central finding is that paradigm diversity

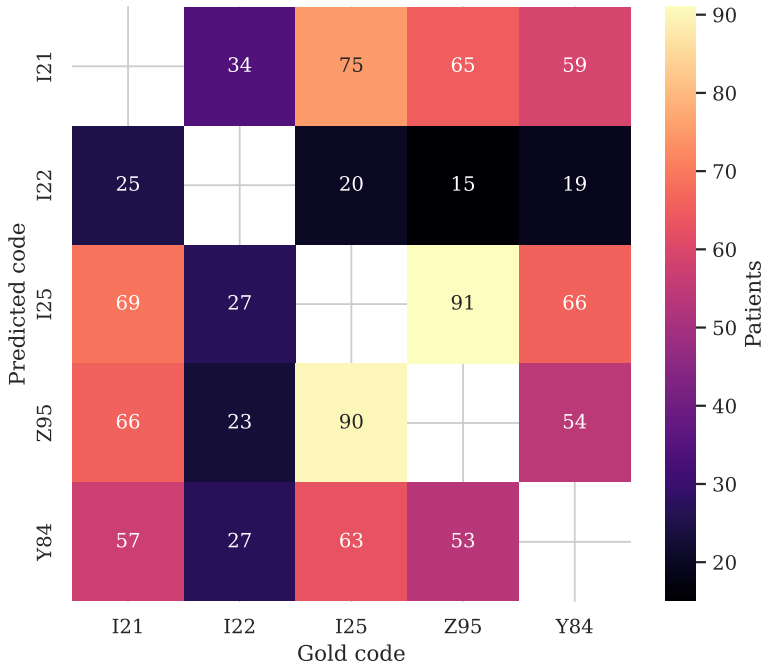

Figure 5: Co-prediction counts for the MI/procedure cluster (I21, I22, I25, Z95, Y84).

(neural classifiers, dictionary rules, retrieval, and NER+EL) matters more than tuning any single model, and that a metaheuristic search over fusion strategies can reliably exploit that diversity without manual engineering.

Several directions are worth exploring beyond the immediate fixes suggested by our error analysis. First, large language models (LLMs) have not yet been evaluated as ensemble components; few-shot prompting or retrieval-augmented generation could provide complementary coverage for rare and context-dependent codes. A related direction is to leverage the ICD-10 hierarchy explicitly, e.g. via a graph convolutional network over label co-occurrence and parent-child structure [11], which could improve zero- and few-shot coverage for the long tail of rare codes. Second, the metaheuristic ensemble framework is task-agnostic: applying it to other clinical coding benchmarks (e.g. English, multilingual, or full ICD-10 hierarchies) would reveal whether the diversity benefit generalises or is specific to low-resource Greek. Third, the current pipeline treats each document independently; a session-level model that considers a patient's prior discharge summaries could resolve temporal ambiguities such as I21/I22/I25 that are intractable with bag-of-token representations. Finally, deploying such a system in a real hospital workflow raises open questions around active learning, human-in-the-loop correction, and concept drift as coding guidelines evolve over time.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Claude, and Gemini in order to: (1) identify and correct grammatical errors, typos, and other writing mistakes; and (2) rephrase sentences and paragraphs to improve clarity, conciseness, and style. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

This work was carried out within the "MSc in Artificial Intelligence and Data Analytics" of the Department of Applied Informatics at the University of Macedonia, in the context of the Natural Language Processing course taught by Prof. Ioannis Refanidis.

One member of our team contributed the XLM-RoBERTa Base component (Section 4.2.3) to this system. The same author also co-authored a separate ElCardioCC 2026 submission, in which the corresponding component was fine-tuned independently using a different training methodology; the remaining co-authors of this paper did not have access to the specific methodological details of that separate effort.

References

[1] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020* (pp. 8440–8451).

[2] Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of SIGIR 2009* (pp. 758–759).

[3] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems* (pp. 1–15). Springer.

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186).

[5] Dong, H., Falis, M., Whiteley, W., Alex, B., Matterson, J., Ji, S., Chen, J., & Wu, H. (2022). Automated clinical coding: What, why, and where we are? *npj Digital Medicine*, 5(1), 159.

[6] Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). GREEK-BERT: The Greeks visiting Sesame Street. In *Proceedings of the 11th Hellenic Conference on Artificial Intelligence* (pp. 110–117).

[7] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.

[8] Mladenović, N., & Hansen, P. (1997). Variable neighborhood search. *Computers & Operations Research*, 24(11), 1097–1100.

[9] Nentidis, A., Katsimpras, G., Krithara, A., Krallinger, M., Rodríguez-Ortega, M., Rodriguez-López, E., Loukachevitch, N., Rozhkov, I., Tutubalina, E., Dimitriadis, D., Patsiou, V., Tsoumakas, G., Giannakoulas, G., Bekiaridou, A., Samaras, A., Di Nunzio, G. M., Ferro, N., Marchesin, S., Martinelli, M., Silvello, G., & Paliouras, G. (2026). Overview of BioASQ 2026: The fourteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Lecture Notes in Computer Science. Springer.

[10] Dimitriadis, D., Patsiou, V., Stoikopoulou, E., Toumpas, A., Bekiaridou, A., Samaras, A., Giannakoulas, G., & Tsoumakas, G. (2026). Overview of ElCardioCC task on clinical coding in cardiology at BioASQ 2026. In *CLEF 2026 Working Notes*.

[11] Rios, A., & Kavuluru, R. (2018). Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3132–3142).

[12] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.

[13] Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., & Zelnik-Manor, L. (2021). Asymmetric loss for multi-label classification. In *Proceedings of ICCV 2021* (pp. 82–91).

[14] Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., & Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. In *Proceedings of NAACL-HLT 2018* (pp. 1101–1111). ACL.

[15] Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In *Proceedings of ECML PKDD 2011* (pp. 145–158).

[16] Tsuboi, Y., Kashima, H., Mori, S., Oda, H., & Matsumoto, Y. (2008). Training conditional random fields using incomplete annotations. In *Proceedings of COLING 2008* (pp. 897–904).