LSI_UNED at ELCardioCC 2026: Leveraging Clinical Entity Extraction for ICD-10 Coding in Greek Cardiology Reports

Alicia Ramirez-Arrabe[1,*], Andres Duque[1,2] and Juan Martinez-Romo[1,2]

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Madrid, 28040, Spain

[2]Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS), Madrid, 28029, Spain

Abstract

This paper describes the participation of the LSI_UNED team in the ELCardioCC 2026 shared task, organized within the BioASQ Workshop at CLEF. The task focuses on automatic ICD-10 clinical coding from Greek discharge letters, a challenging scenario due to the complexity of medical terminology and the limited availability of annotated resources in low-resource languages. We addressed the task using a two-phase pipeline architecture. First, a Named Entity Recognition (NER) model was employed to identify text spans corresponding to potential ICD-10 mentions. Then, a text classification model assigned an ICD-10 code to each detected entity. Several multilingual and Greek-specific Transformer-based models were fine-tuned and evaluated. Results show that the proposed approach achieves competitive performance, obtaining a best F1-score of 0.83 on the test set and ranking third out of the seven participating teams.

Keywords

Automated Clinical Coding, ICD-10 codes, Biomedical Natural Language Processing, Named Entity Recognition, Greek Electronic Health Records

1. Introduction

This work presents the participation of our team, LSI_UNED, in the shared task ELCardioCC 2026 [1], developed under the 14th BioASQ Workshop [2] in the Conference and Labs of the Evaluation Forum (CLEF). This task focuses on automated clinical coding in a low-resource language setting, namely Greek. Clinical coding is a key NLP task in the biomedical domain and involves assigning standardized medical codes to information contained in clinical reports, improving healthcare information management for professionals, researchers, and institutions [3]. Since manual coding is complex, time-consuming, and error-prone [4], automated solutions based on NLP techniques are increasingly necessary. These systems identify medical terminology in clinical texts and map it to standardized coding systems such as Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [5], Current Procedural Terminology (CPT) [6], and the International Classification of Diseases (ICD) [7]. The ELCardioCC task specifically addresses ICD-10, which was introduced in the 1990s and is the most widely used system for coding diagnoses (ICD-10-CM) and medical procedures (ICD-10-PCS).

The system developed by our team follows a two-phase pipeline architecture. First, a Named Entity Recognition (NER) model detects text spans associated with potential ICD-10 mentions in the clinical documents. Then, a text classification model assigns the corresponding ICD-10 code to each detected entity. This formulation transforms the original extreme multi-label classification problem into a simpler single-label classification task at the mention level. To evaluate the proposed approach, several Transformer-based models were fine-tuned and compared under the same experimental setup.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of related work, including relevant tasks and corpora, as well as the best-performing systems from the previous edition of ELCardioCC. Section 3 describes the dataset and the proposed approach, detailing the models

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

aramirez@lsi.uned.es (A. Ramirez-Arrabe); aduque@lsi.uned.es (A. Duque); juaner@lsi.uned.es (J. Martinez-Romo)

0009-0001-8550-6275 (A. Ramirez-Arrabe); 0000-0002-0619-8615 (A. Duque); 0000-0002-6905-7051 (J. Martinez-Romo)

and tools employed. Section 4 presents the experimental results and performance analysis. Finally, Section 5 concludes the paper and discusses directions for future work.

2. Related work

In recent years, several ICD-10 coding corpora and shared tasks have been released across multiple languages, contributing to the development of automated clinical coding systems. In English, the MIMIC-III corpus [8] and its extended version MIMIC-IV [9] provide large-scale collections of discharge summaries and radiology reports from more than 350,000 patients, annotated with ICD-10 diagnosis and procedure codes. Multilingual efforts include the CLEF eHealth evaluation lab shared tasks conducted between 2016 and 2018 on coding death certificates in French, English, Hungarian, and Italian [10, 11, 12], which focused on mapping diagnostic statements to ICD-10 codes. For Spanish, relevant resources include the CARES corpus [13], consisting of radiology reports annotated with ICD-10 codes, as well as the CodiEsp shared task [14], which introduced explainable ICD coding for diagnoses and procedures, and Cantemist [15], focused on neoplasm coding in clinical narratives. In German, CLEF eHealth 2019 released a corpus of over 8,000 non-technical summaries of planned animal experiments annotated with ICD-10 codes [16], complemented by the BRONCO corpus [17] of discharge summaries with diagnosis labels. Finally, for Greek, the ELCardioCC corpus [18], published for the first edition of the shared-task and released in CLEF BioASQ 2025, provides cardiology discharge letters annotated for explainable ICD-10 coding in a low-resource setting.

The first edition of ELCardioCC [18] was divided into three subtasks: Named Entity Recognition (NER), Entity Linking (EL), and Multi-Label Clinical Coding with Explainability (MLC-X). NER involved identifying five types of clinical mentions with span boundaries, EL required mapping these mentions to ICD-10 codes, and MLC-X aimed at predicting all relevant codes per document with explainability through supporting text spans. In the NER subtask of the previous edition, teams droidlyx [19] and enigma [20] used fine-tuned BERT-style architectures, while bhuang [21] applied multilingual LLMs with zero-shot prompting and filtering. Performance was led by droidlyx, which achieved the highest F1-score of 0.7328, followed closely by enigma with 0.7167 and bhuang with 0.5761. For the EL subtask, droidlyx combined translation, SapBERT embeddings [22], and MLP classification; bhuang used a BM25 retrieval stage followed by a MedCPT cross-encoder reranker [23] along with ensemble prompting strategies; and enigma applied dictionary matching and BGE-M3 bi-encoder retrieval [24] trained with ranking loss. Again, droidlyx obtained the best result with an F1-score of 0.6778, followed by enigma with 0.6693 and bhuang with 0.5336. Finally, in the MLC-X subtask, droidlyx and bhuang derived ICD-10 codes from their EL pipelines, while enigma used a direct document-level multi-label classification approach. In this setting, droidlyx achieved an F1-score of 0.8472 and bhuang reached 0.7543, while enigma did not participate. This year's edition of ELCardioCC is not divided into separate subtasks and does not include an explainability component. Instead, participating teams are required to directly predict the ICD-10 codes for each document.

3. Methodology

3.1. Dataset

The ELCardioCC corpus contains 3,000 discharge letters from the cardiology department of a Greek hospital. Each document is written in Greek and annotated with ICD-10 codes. There are different variants of the ICD coding system, each focused on different medical concepts, such as diagnoses (ICD-10-CM), procedures (ICD-10-PCS) or neoplasms (ICD-O). In this case, ELCardioCC focuses on diagnosis coding, one of the most widely used variants. The ICD-10 codes in this variant follow a hierarchical structure, ranging from a minimum of three to a maximum of seven characters. Nonetheless, the task has been simplified and only the first three characters need to be identified.

The corpus is split into a training set of 2,500 documents and a test set containing the remaining

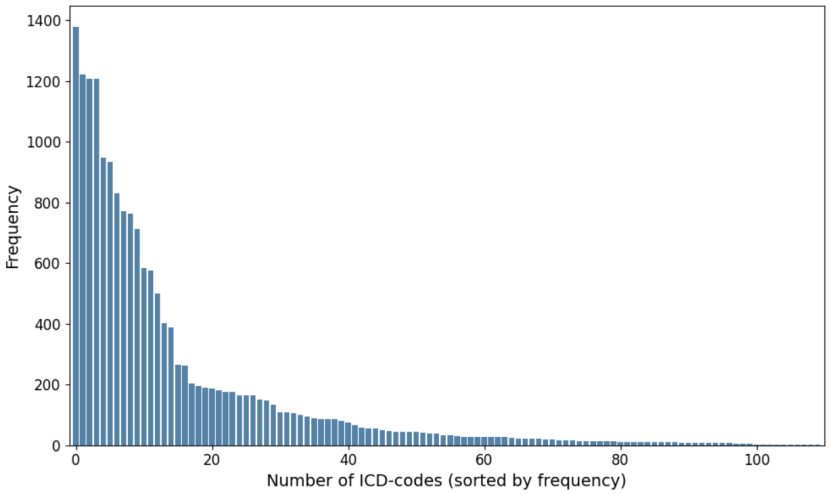

Figure 1: Frequency distribution of ICD-10 codes in the training set, sorted in descending order of occurrence.

500 texts. It extends the dataset published for the ELCardioCC 2025 shared task [18], which originally consisted of 1,500 documents. Table 1 presents the main statistics of the training set. All 2,500 documents are annotated at the document level with the ICD-10 codes they contain, while a subset of 1,493 texts is also annotated at the mention level; that is, these texts include the spans that justify the presence of each ICD-10 code. The last row of the table indicates that the number of unique codes included in the document-level annotations (110) exceeds by 25 the number of codes annotated at the mention level (85). These additional 25 codes account for 564 of the 17,555 occurrences in the training set, corresponding to 3.21%.

Type of annotations	Document-level	Mention-level
Total documents	2500	1493
Total codes	17555	11979
Mean codes per document	7.02	8.02
Unique codes	110	85

Table 1
Statistics of the ELCardioCC 2026 training set. All 2,500 documents include document-level ICD-10 annotations, while a subset of 1,493 documents additionally contains mention-level annotations identifying the text spans supporting each code.

Figure 1 shows the frequency distribution of ICD-10 codes in the corpus, sorted in descending order of occurrence. As can be observed, the distribution is highly unbalanced, with a small number of codes appearing very frequently and a large number of codes appearing only a few times. This phenomenon is common in ICD-10 clinical coding corpora, as the number of possible ICD-10 codes is extremely large (more than 98,000 [14]). For this reason, automatic clinical coding is typically considered an eXtreme Multi-Label Text Classification (XMTC) task [25].

Furthermore, the frequency distribution of the 25 ICD-10 codes that are not considered by the system is presented in Figure 2. As previously mentioned, these codes account for 3.21% of the total ICD-10 code occurrences and cannot be included because their mention-level annotations are unavailable. The most frequent is code E00, corresponding to the diagnosis *Congenital iodine-deficiency syndrome*, which appears in 165 documents. In contrast, 18 of these codes occur in fewer than 20 documents.

3.2. System description

The system designed and implemented by our team is structured into two phases, as shown in Figure 3. First, a Named Entity Recognition (NER) model identifies spans of text that may potentially describe an ICD-10 code. Then, a text classification model assigns a code to each detected entity. To do so, we used

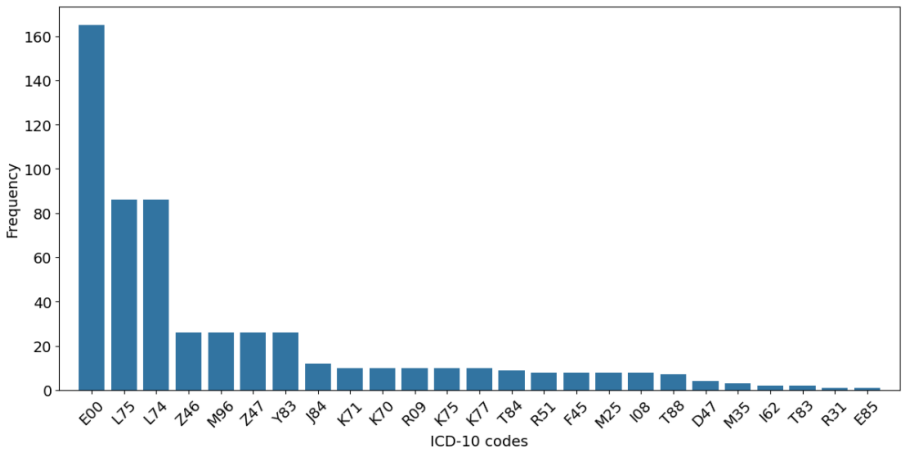

Figure 2: Frequency distribution of the 25 ICD-10 codes excluded from the system due to the absence of mention-level annotations, sorted in descending order of occurrence.

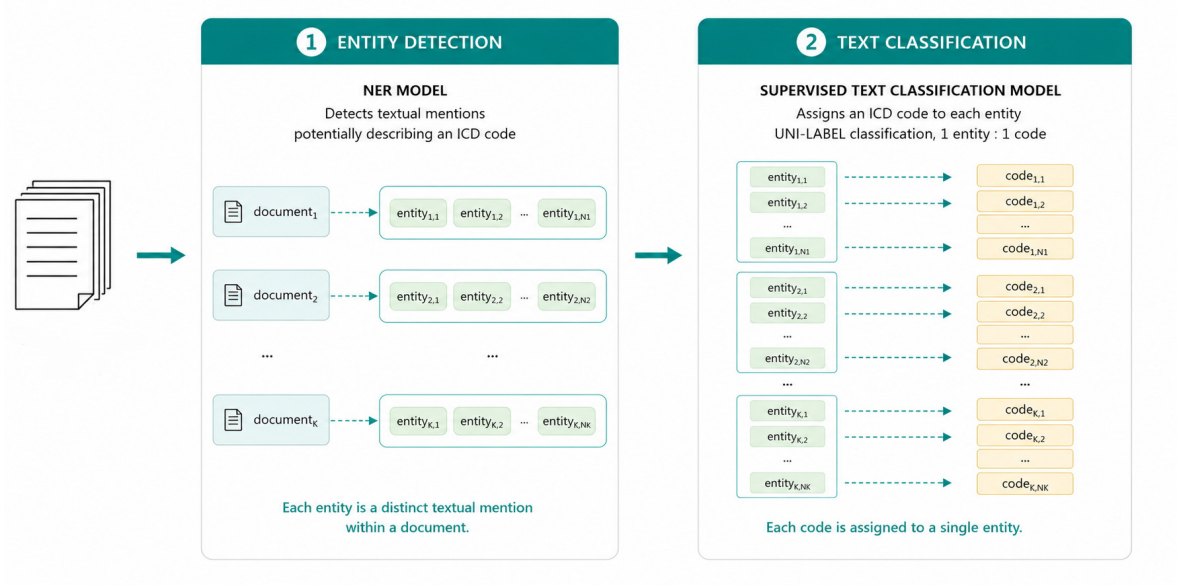

Figure 3: Flow diagram of the proposed system.

only the documents annotated at the mention level, which, as shown in Table 1, constitute a subset of the complete training set. However, we decided to adopt this approach because it has proven to be both effective and robust in the state of the art for automatic clinical coding [26]. It also allows us to simplify the XMTC problem by transforming it from a multi-label task into a single-label one; that is, assigning one code per entity instead of multiple codes per text.

3.2.1. Named Entity Recognition

The first phase of the system consists of a Named Entity Recognition (NER) model that identifies text spans potentially associated with ICD-10 codes. We adopt the W^{2}NER architecture proposed in [27], which reframes NER as a pairwise word-relation classification problem rather than a sequential tagging task. Specifically, it constructs a two-dimensional grid over all word pairs and predicts relations between them, like Next-Neighboring-Word (NNW), which links adjacent tokens within an entity, and Tail-Head-Word (THW), which connects the last token of an entity to its first token and assigns the entity label. The model uses BERT embeddings followed by a BiLSTM to capture contextual information, and applies 2D convolutions over the word-pair grid to learn structural patterns. Entities are then reconstructed by linking words based on the predicted relations. According to the original article,

W^{2}NER achieves state-of-the-art results on 14 benchmark NER datasets in both English and Chinese.

Using this grid-based NER framework, we trained and evaluated two multilingual pre-trained models (bert-base-multilingual-cased[1] [28] and xlm-roberta-base[2] [29]), as well as one model pre-trained on the Greek language, bert-base-greek-uncased-v1[3] [30]. The bert-base-multilingual-cased model has around 110M parameters and was pretrained on the top 104 languages with Wikipedia data. It served as the baseline system in last year's shared task and achieved the second-best overall performance across the three subtasks [18]. The xlm-roberta-base model contains approximately 278M parameters and was trained on large-scale multilingual corpora covering over 100 languages. It was also adopted by the team ranked first in Recall in two of the subtasks of the previous edition of the task [20]. Finally, the bert-base-greek-uncased-v1 model, with roughly 110M parameters and trained primarily on Greek-language corpora including Wikipedia and OSCAR [31], was the best-performing model in last year's shared task across all three subtasks [19].

Regarding the hyperparameters, the training configuration was fixed to a batch size of 8, a learning rate of $1e^{-3}$, and a maximum sequence length of 512 tokens per chunk. We applied 5-fold stratified cross-validation to determine the optimal number of training epochs. With a maximum of 20 epochs, the best F1-score was consistently reached at around 15 epochs on average across folds for xlm-roberta-base and bert-base-greek-uncased-v1, whereas bert-base-multilingual-cased achieved its best performance at an average of 11 epochs. Accordingly, these values were selected as the final number of epochs for fine-tuning each respective model.

Moreover, as the proposed system consists of a two-stage pipeline, the performance of the second stage depends on the output of the first. Consequently, code instances whose entities are not detected during the NER phase cannot be predicted in the subsequent phase. To assess the performance of this initial stage, we conducted several validation tests on the NER component, obtaining an F1-score of approximately 0.7805 when considering the strict boundaries of each entity.

3.2.2. Text classification

After detecting ICD-10 code mentions, we fine-tuned a text classification model to assign a code to each mention. In this phase, we evaluated the three models used in Phase 1, as well as an additional model pre-trained on the Greek language, GreekDeBERTa-base[4]. This model was used by the team that ranked third in the first two subtasks of the previous edition of the shared task [20].

The training hyperparameters were fixed to a batch size of 8, a learning rate of 5×10^{-5}, and a weight decay of 0.001. As in the previous phase, we also performed 5-fold stratified cross-validation to select the optimal number of training epochs. With a maximum of 10 epochs, the multilingual models (xlm-roberta-base and bert-base-multilingual-cased) consistently achieved their best F1-score at around 3 epochs on average, while the Greek-specific models (bert-base-greek-uncased-v1 and GreekDeBERTa-base) peaked at approximately 8 epochs. These values were therefore used as the final epoch settings for fine-tuning each model.

4. Experiments and results

The evaluation framework of ELCardioCC 2026 shared-task assesses predictions at the group level, where each group consists of a set of equivalent or interchangeable ICD-10 codes, using Precision, Recall and F1-score. A prediction is considered correct if at least one code from the corresponding group is correctly identified, while retrieving multiple codes from the same group does not yield additional reward. Conversely, if no code from a group is predicted, the system receives no credit for that group.

In our setup, we chose to form singleton groups, where each ICD-10 code is treated as its own group. This design decision simplifies the prediction task by avoiding assumptions about equivalence or

[1] https://huggingface.co/google-bert/bert-base-multilingual-cased
[2] https://huggingface.co/FacebookAI/xlm-roberta-base
[3] https://huggingface.co/nlpaueb/bert-base-greek-uncased-v1
[4] https://huggingface.co/AI-team-UoA/GreekDeBERTa-base

NER Model	Text Classification Model	Precision	Recall	F1-score
xlm-roberta-base	xlm-roberta-base	0.7786 ± 0.0111	0.9104 ± 0.0055	0.8394 ± 0.0055
xlm-roberta-base	bert-base-multilingual-cased	0.7796 ± 0.0133	0.9136 ± 0.0052	0.8412 ± 0.0088
xlm-roberta-base	bert-base-greek-uncased-v1	0.7814 ± 0.0125	0.9166 ± 0.0032	0.8436 ± 0.0078
xlm-roberta-base	GreekDeBERTa-base	0.6880 ± 0.0144	0.7030 ± 0.0056	0.6954 ± 0.0084
bert-base-multilingual-cased	xlm-roberta-base	0.7896 ± 0.0149	0.9084 ± 0.0086	0.8448 ± 0.0082
bert-base-multilingual-cased	bert-base-multilingual-cased	0.7890 ± 0.0174	0.9100 ± 0.0051	0.8448 ± 0.0102
bert-base-multilingual-cased	bert-base-greek-uncased-v1	0.7902 ± 0.0147	0.9126 ± 0.0049	0.8470 ± 0.0078
bert-base-multilingual-cased	GreekDeBERTa-base	0.6986 ± 0.0172	0.6956 ± 0.0069	0.6972 ± 0.0118
bert-base-greek-uncased-v1	xlm-roberta-base	0.8046 ± 0.0123	0.9114 ± 0.0042	**0.8546 ± 0.0057**
bert-base-greek-uncased-v1	bert-base-multilingual-cased	0.8046 ± 0.0141	0.9130 ± 0.0034	**0.8556 ± 0.0091**
bert-base-greek-uncased-v1	bert-base-greek-uncased-v1	0.8066 ± 0.0128	0.9172 ± 0.0016	**0.8584 ± 0.0080**
bert-base-greek-uncased-v1	GreekDeBERTa-base	0.7108 ± 0.0168	0.6998 ± 0.0058	0.7050 ± 0.0082

Table 2
Performance results of each model combination on the development set. Averages across 5 folds are reported, with standard deviation included. The three systems highlighted in bold achieved the best results across folds and were selected for the final submission to the task.

Run	NER Model	Text Classification Model	Precision	Recall	F1-score
1	bert-base-greek-uncased-v1	bert-base-greek-uncased-v1	**0.8187**	0.8416	**0.8300**
2	bert-base-greek-uncased-v1	xlm-roberta-base	0.8176	**0.8421**	0.8297
3	bert-base-greek-uncased-v1	bert-base-multilingual-cased	0.8173	0.8417	0.8293

Table 3
Precision, recall and F1-score results of the three submitted systems on the test set. For each metric, the best results are highlighted in bold. Results are reported to four decimals.

interchangeability between codes. It also avoids redundancy, as the evaluation only rewards identifying at least one correct code per group and does not provide additional benefit for retrieving multiple codes within the same group.

To ensure a robust evaluation, we applied 5-fold stratified cross-validation and computed the average F1-score across folds for each model combination. The results on the development set, including standard deviations, are reported in Table 2. As can be observed, the systems generally achieve higher Recall than Precision. An interesting observation is the consistently lower performance of GreekDeBERTa-base compared to the BERT-based models. Although we did not investigate this further, we hypothesize that differences in pretraining data or tokenization may make GreekDeBERTa-base less well suited to the specialized medical terminology required for clinical coding.

Based on these experiments, we selected the three best-performing systems for submission, marked in bold in Table 2. All submitted systems used the Greek-specific model bert-base-greek-uncased-v1 for entity extraction. In the text classification phase, the top-performing system employed bert-base-greek-uncased-v1, followed by systems using bert-base-multilingual-cased and xlm-roberta-base, in descending order of performance.

Table 3 shows the results achieved on the test set by the three submitted systems. The best F1-score is 0.8300 and is obtained by the system that employs the model pre-trained on the Greek language (bert-base-greek-uncased-v1) in both phases. As can be observed in the results table, all three systems obtain very similar results across the three metrics, with only minor variations, as all three use the same pre-trained model in the NER phase. An interesting observation is that Precision and Recall are much more balanced than in the development set (see Table 2), where differences of up to 0.134 between metrics were observed. The lower Recall could be due to the fact that our approach only used the mention-level annotated subset of the corpus and, as noted in Section 3.1, 25 codes, representing 3.21% of the total occurrences on the training set, could not be included.

Moreover, Table 4 reports the overall ranking for the task across all participating teams. Our best result (LSI_UNED Run 1) achieved seventh place out of 20 submissions and third place out of seven participating teams. The baseline F1-score provided by the organizers is 0.762264.

Team	Precision	Recall	F1-score
stanimeros	0.850979	0.882995	0.866691
Georgios_1	0.865845	0.861783	0.863809
stanimeros	0.868745	0.853885	0.861251
stanimeros	0.876722	0.843096	0.859581
stanimeros	0.819434	0.881092	0.849145
stanimeros	0.831037	0.867525	0.848889
LSI_UNED Run 1	**0.818709**	**0.841595**	**0.829994**
LSI_UNED Run 2	**0.817621**	**0.842046**	**0.829654**
LSI_UNED Run 3	**0.817259**	**0.841673**	**0.829286**
alikibliona	0.818347	0.824626	0.821474
ELCardioCC baseline	0.659173	0.903579	0.762264
Pakakisd	0.809282	0.720000	0.762035
fernandogd97	0.791516	0.734028	0.761689
fernandogd97	0.786439	0.727364	0.755749
fernandogd97	0.791878	0.714192	0.751032
Pakakisd	0.869471	0.654655	0.746924
Pakakisd	0.883249	0.557310	0.683406
Pakakisd	0.793691	0.561570	0.657752
fernandogd97	0.972806	0.057031	0.107745
fernandogd97	0.972806	0.057031	0.107745

Table 4
Precision, recall and F1-score of the participating systems on the test set, sorted by F1-score. Results are reported to six decimals.

5. Conclusions

This work presents the participation of the LSI_UNED team in the ELCardioCC 2026 shared-task, located in the BioASQ Workshop of the Conference and Labs of the Evaluation Forum (CLEF). The task is focused on ICD-10 clinical coding on a corpus of 3,000 discharge letters written in Greek. We addressed the task by adopting a two-phase pipeline and effectively transforming a multi-label problem into a single-label formulation. This design choice reduced complexity and enabled a more straightforward mapping between text mentions and candidate codes, although it also introduced constraints related to the availability of mention-level annotations, which limited the portion of the corpus that could be used for training the models.

Despite these limitations, the proposed systems achieved competitive results, with a best F1-score of 0.8300 on the test set. Overall, performance analysis showed that Precision and Recall are relatively well balanced, with Recall being slightly higher.

Future work will focus on several directions. First, it would be interesting to conduct hyperparameter optimization to identify the optimal combination of hyperparameters. Second, evaluating fully multi-label classification approaches would also be a promising direction. Finally, we intend to evaluate more recent approaches, including Large Language Models (LLMs), to assess their ability to capture domain-specific knowledge, particularly with respect to medical terminology and the ICD-10 coding system.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the EDHER-MED Project under grant PID2022-136522OB-C21 as well as by the Universidad Nacional de Educación a Distancia (UNED), Spain within project SICAMESP (2023-VICE-0029).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check.

References

[1] D. Dimitriadis, V. Patsiou, E. Stoikopoulou, A. Toumpas, A. Bekiaridou, A. Samaras, G. Giannakoulas, G. Tsoumakas, Overview of ElCardioCC Task on Clinical Coding in Cardiology at BioASQ 2026, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[2] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[3] H. Dong, M. Falis, W. Whiteley, B. Alex, J. Matterson, S. Ji, J. Chen, H. Wu, Automated clinical coding: what, why, and where we are?, npj Digit. Medicine 5 (2022). URL: https://doi.org/10.1038/s41746-022-00705-7.

[4] K. J. O'malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, C. M. Ashton, Measuring diagnoses: Icd code accuracy, Health services research 40 (2005) 1620–1639.

[5] M. Q. Stearns, C. Price, K. A. Spackman, A. Y. Wang, SNOMED clinical terms: overview of the development process and project status, in: AMIA 2001, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 3-7, 2001, AMIA, 2001. URL: https://knowledge.amia.org/amia-55142-a2001a-1.597057/.

[6] A. M. Association, Physicians' Current Procedural Terminology: CPT, The Association, 1991.

[7] O. WHO, International classification of diseases, WHO [Internet] (1992).

[8] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3 (2016) 1–9.

[9] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al., Mimic-iv, a freely accessible electronic health record dataset, Scientific data 10 (2023) 1.

[10] A. Névéol, K. B. Cohen, C. Grouin, T. Hamon, T. Lavergne, L. Kelly, L. Goeuriot, G. Rey, A. Robert, X. Tannier, P. Zweigenbaum, Clinical information extraction at the CLEF ehealth evaluation lab 2016, in: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016, CEUR Workshop Proceedings, CEUR-WS.org, 2016, pp. 28–42. URL: https://ceur-ws.org/Vol-1609/16090028.pdf.

[11] A. Névéol, A. Robert, R. Anderson, K. B. Cohen, C. Grouin, T. Lavergne, G. Rey, C. Rondet, P. Zweigenbaum, CLEF ehealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in english and french, in: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017, CEUR Workshop Proceedings, CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-1866/invited_paper_6.pdf.

[12] A. Névéol, A. Robert, F. Grippo, C. Morgand, C. Orsi, L. Pelikan, L. Ramadier, G. Rey, P. Zweigenbaum, Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian., in: CLEF (Working Notes), CEUR-WS, 2018, pp. 1–18.

[13] M. Chizhikova, P. López-Úbeda, J. Collado-Montañez, T. Martín-Noguerol, M. C. Díaz-Galiano, A. Luna, L. A. U. López, M. T. M. Valdivia, CARES: A corpus for classification of spanish radiological

reports, Comput. Biol. Medicine 154 (2023) 106581. URL: https://doi.org/10.1016/j.compbiomed.2023.106581.

[14] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF ehealth 2020, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, CEUR Workshop Proceedings, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_263.pdf.

[15] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020, CEUR Workshop Proceedings, CEUR-WS.org, 2020, pp. 303–323. URL: https://ceur-ws.org/Vol-2664/cantemist_overview.pdf.

[16] M. L. Neves, D. Butzke, A. Dörendahl, N. Leich, B. Hummel, G. Schönfelder, B. Grune, Overview of the CLEF ehealth 2019 multilingual information extraction, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, CEUR Workshop Proceedings, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_251.pdf.

[17] M. Kittner, M. Lamping, D. T. Rieke, J. Götze, B. Bajwa, I. Jelas, G. Rüter, H. Hautow, M. Sänger, M. Habibi, et al., Annotation and initial evaluation of a large annotated german oncological corpus, JAMIA open 4 (2021) ooab025.

[18] D. Dimitriadis, V. Patsiou, E. Stoikopoulou, A. Toumpas, A. Kipouros, A. Bekiaridou, K. Barmpagiannos, A. Vasilopoulou, A. Barmpagiannos, A. Samaras, D. Papadopoulos, G. Giannakoulas, G. Tsoumakas, Overview of ElCardioCC Task on Clinical Coding in Cardiology at BioASQ 2025, in: Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, 2025, pp. 51–64. URL: https://ceur-ws.org/Vol-4038/paper_4.pdf.

[19] Y. Liu, S. Zhu, Lyx_dmiip_fdu at bioasq 2025: Utilizing bert embeddings for biomedical text mining, in: CLEF, 2025.

[20] B. Velichkov, A. Datseris, S. Vassileva, S. Boytcheva, Enigma@ elcardiocc: bridging ner and icd-10 entity linking-a hybrid method for greek clinical narratives, in: CLEF, 2025.

[21] B. Huang, Clinical entity recognition and linking in greek discharge letters using multilingual-llm-based multi-stage system, in: CLEF, 2025.

[22] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, 2021, pp. 4228–4238.

[23] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, Z. Lu, Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval, Bioinformatics 39 (2023).

[24] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, arXiv preprint arXiv:2402.03216 4 (2024).

[25] M. Almagro, R. Martínez, V. Fresno, S. Montalvo, ICD-10 coding of spanish electronic discharge summaries: An extreme classification problem, IEEE Access 8 (2020) 100073–100083. URL: https://doi.org/10.1109/ACCESS.2020.2997241.

[26] A. Ramirez-Arrabe, J. Martinez-Romo, A. Duque, An end-to-end system for explainable clinical coding across languages and diverse medical data sources, BMC Medical Informatics and Decision Making (2026). URL: https://doi.org/10.1186/s12911-026-03473-6.

[27] J. Li, H. Fei, J. Liu, S. Wu, M. Zhang, C. Teng, D. Ji, F. Li, Unified named entity recognition as word-word relation classification, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press, 2022, pp. 10965–10973. URL: https:

//doi.org/10.1609/aaai.v36i10.21344.

[28] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423.

[29] 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, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747.

[30] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, I. Androutsopoulos, GREEK-BERT: the greeks visiting sesame street, in: C. D. Spyropoulos, I. Varlamis, I. Androutsopoulos, P. Malakasiotis (Eds.), SETN 2020: 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, September 2-4, 2020, ACM, 2020, pp. 110–117. URL: https://doi.org/10.1145/3411408.3411440. doi:10.1145/3411408.3411440.

[31] P. J. Ortiz Suárez, B. Sagot, L. Romary, Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures, in: 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Leibniz-Institut f"ur Deutsche Sprache, Mannheim, 2019, pp. 9 – 16. URL: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.