

Multi-Criteria Optimization of Language Models for HFpEF Symptom Detection in Spanish Electronic Health Records

Jacinto Mata, Victoria Pachón, Ana Manovel, Manuel J. Maña, Manuel de la Villa

Submitted to: Journal of Medical Internet Research
on: April 23, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Multi-Criteria Optimization of Language Models for HFpEF Symptom Detection in Spanish Electronic Health Records

Jacinto Mata¹ PhD; Victoria Pachón¹ PhD; Ana Manovel² MD; Manuel J. Maña¹ PhD; Manuel de la Villa¹

¹ I2C Research Group Universidad de Huelva Huelva ES

² Cardiology Department, Juan Ramón Jiménez University Hospital Multidisciplinary Amyloidosis Unit Huelva (MAUH) Hospital Juan Ramón Jiménez Huelva ES

Corresponding Author:

Jacinto Mata PhD

I2C Research Group
Universidad de Huelva
Escuela Técnica Superior de Ingeniería
Campus El Carmen
Huelva
ES

Abstract

Background: Heart failure with preserved ejection fraction (HFpEF) is a major clinical manifestation of cardiac amyloidosis (CA), a condition frequently underdiagnosed due to its nonspecific symptomatology. Electronic health records (EHRs) offer a promising avenue for supporting early symptom detection through natural language processing (NLP). However, identifying relevant clinical cues within unstructured narratives, particularly in Spanish, remains a significant challenge owing to the scarcity of annotated corpora and domain-specific models.

Objective: This study proposes and evaluates a Transformer-based NLP framework for the automated detection of HFpEF-related symptoms in Spanish EHRs. Our objective is to assess the feasibility of leveraging clinical narratives to support early identification of heart failure phenotypes suggestive of cardiac amyloidosis.

Methods: A novel corpus was developed from over 15,000 Spanish clinical documents, manually annotated and validated by cardiology experts. Several Transformer architectures were benchmarked, including general-purpose, biomedical-specialized, and long-document models (Longformers), under three optimization strategies tailored to clinically relevant metrics: area under the curve (AUC), F1-score, and sensitivity.

Results: All models achieved strong performance (AUC > 0.940). The best-performing model, Longformer Biomedical-Clinical, reached an AUC of 0.987 and F1-score of 0.985. Sensitivity-optimized models achieved false negative rates below 3%, a critical threshold for clinical applicability.

Conclusions: Transformer-based models can robustly detect HFpEF-related symptoms from unstructured Spanish clinical texts, even in the presence of class imbalance and complex language patterns. Our findings highlight the importance of combining domain-specific pretraining, long-context modeling, and tailored optimization to enhance the performance of NLP systems in high-impact clinical applications.

(JMIR Preprints 23/04/2025:76433)

DOI: <https://doi.org/10.2196/preprints.76433>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to the public.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/preprint/76433>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/76433>



Original Manuscript

Original Paper

Jacinto Mata^{a,*}, Victoria Pachón^a, Ana Manovel^{b,c}, Manuel J. Maña^a and Manuel de la Villa^a

^a I²C Research Group, University of Huelva, Campus El Carmen s/n, Huelva, Spain

^b Multidisciplinary Amyloidosis Unit Huelva (MAUH), Huelva, Spain

^c Cardiology Department, Juan Ramón Jiménez University Hospital, Huelva, Spain

*Corresponding Author: Escuela Técnica Superior de Ingeniería, Campus El Carmen, University of Huelva (Spain); +34687862089; mata@uhu.es

Multi-Criteria Optimization of Language Models for HFpEF Symptom Detection in Spanish Electronic Health Records

Abstract

Background: Heart failure with preserved ejection fraction (HFpEF) is a major clinical manifestation of cardiac amyloidosis (CA), a condition frequently underdiagnosed due to its nonspecific symptomatology. Electronic health records (EHRs) offer a promising avenue for supporting early symptom detection through natural language processing (NLP). However, identifying relevant clinical cues within unstructured narratives, particularly in Spanish, remains a significant challenge owing to the scarcity of annotated corpora and domain-specific models.

Objective: This study proposes and evaluates a Transformer-based NLP framework for the automated detection of HFpEF-related symptoms in Spanish EHRs. Our objective is to assess the feasibility of leveraging clinical narratives to support early identification of heart failure phenotypes suggestive of cardiac amyloidosis.

Methods: A novel corpus was developed from over 15,000 Spanish clinical documents, manually annotated and validated by cardiology experts. Several Transformer architectures were benchmarked, including general-purpose, biomedical-specialized, and long-document models (Longformers), under three optimization strategies tailored to clinically relevant metrics: area under the curve (AUC), F1-score, and sensitivity.

Results: All models achieved strong performance (AUC > 0.940). The best-performing model, *Longformer Biomedical-Clinical*, reached an AUC of 0.987 and F1-score of 0.985. Sensitivity-optimized models achieved false negative rates below 3%, a critical threshold for clinical applicability.

Conclusions: Transformer-based models can robustly detect HFpEF-related symptoms from unstructured Spanish clinical texts, even in the presence of class imbalance and complex language patterns. Our findings highlight the importance of combining domain-specific pretraining, long-context modeling, and tailored optimization to enhance the performance of NLP systems in high-impact clinical applications.

Keywords: Natural language processing; Transformer models; Electronic health records; Symptom detection; Heart failure with preserved ejection fraction; Cardiac amyloidosis; Spanish clinical texts

Introduction

Heart failure (HF) represents a major public health challenge, with its prevalence rising as a result of population aging and improved survival among individuals with cardiovascular conditions. Among its phenotypes, heart failure with preserved ejection fraction (HFpEF) accounts for nearly 50% of cases and is characterized by congestion and dyspnea in the absence of marked systolic dysfunction. Despite its clinical relevance, HFpEF remains diagnostically challenging due to its pathophysiological and etiological heterogeneity [1].

A key underlying cause of HFpEF is cardiac amyloidosis, an infiltrative disorder characterized by extracellular deposits of misfolded transthyretin protein that impair myocardial structure and function. The most prevalent variant is the wild-type (wt) form of transthyretin cardiac amyloidosis (ATTR-CM), which predominantly affects elderly individuals and leads to progressive cardiac dysfunction and worsening HF symptoms. Unlike other HFpEF etiologies, ATTR-CM is amenable to disease-modifying therapies that can slow progression, reduce hospitalizations, and improve survival outcomes. Nonetheless, early diagnosis remains difficult due to low clinical suspicion, nonspecific echocardiographic findings, and historically limited awareness within the medical community. Therapeutic efficacy is maximized when treatment is initiated early, underscoring the importance of timely and optimized detection strategies in HF populations [2].

In this context, artificial intelligence (AI) has emerged as a promising approach for the early identification of disease within large-scale clinical datasets [3]. Specifically, natural language processing (NLP) and deep learning models can analyze EHRs to identify, and extract mentions of HFpEF and associated terminology, facilitating the development of high-quality annotated corpora [4]. Such corpora are essential for training AI models capable of detecting textual patterns indicative of ATTR-CM, enabling the prioritization of high-risk patients. Applying these technologies to the structured identification of HF mentions in clinical narratives can enhance diagnostic accuracy, support data-driven decision-making, and enable timely interventions that improve prognosis and quality of life.

This study aims to develop and validate an AI-based system for the automated detection of HF mentions in Spanish EHRs, with the goal of supporting early detection of potential ATTR-CM cases, enhancing diagnostic efficiency, and mitigating underdiagnosis. As outlined in [5], several research challenges remain in clinical NLP. This study addresses three of them: applying deep learning to clinical text classification, overcoming language-related barriers, and leveraging transfer learning for narrative clinical report classification.

The main contributions of this work are: (1) the construction of a manually annotated corpus of Spanish EHRs for HFpEF symptom detection, validated by cardiology specialists and intended for public release; (2) a comparative evaluation of general, biomedical, and long-document Transformer models for clinical text classification; (3) an analysis of optimization strategies tailored to AUC, F1-score, and sensitivity; and (4) a demonstration of the feasibility of using NLP models to support the early detection of HFpEF symptoms relevant to cardiac amyloidosis.

Related Work

Transformer-based models have shown increasing effectiveness in extracting clinically relevant information from unstructured electronic health records (EHRs), particularly within the context of heart failure (HF). Adejumo et al. [6] employed ClinicalBERT to extract NYHA classifications and heart failure symptoms from clinical notes, achieving AUROCs greater than 0.98 and identifying

functional status in 83% more cases than explicit documentation alone. Similarly, Liu et al. [7] used ClinicalBERT embeddings within a predictive framework to identify early HF onset, outperforming traditional methods in a large-scale real-world dataset.

In the context of clinical trials, Martí-Castellote et al. [8] developed a hybrid NLP approach integrating Clinical Longformer and GPT-4o to adjudicate HF-related hospitalizations. Their system reproduced expert adjudications with 83% concordance, substantially reducing manual effort and demonstrating the scalability and reliability of Transformer-based models in high-stakes research. Houssein et al. [11] benchmarked five Transformer architectures (BERT, BioBERT, RoBERTa, XLNet, and BioClinicalBERT) on the i2b2 dataset for risk factor extraction in heart disease, achieving state-of-the-art performance with a micro F1-score of 94.26%. These results reinforce the value of domain-specific Transformer models in clinical NLP applications. Notably, Fan et al. [12] implemented a Transformer-based clustering framework to identify seven HF subtypes in a cohort of more than 379,000 patients. These subgroups, some of which were independent of LVEF, highlight the capacity of deep learning approaches to uncover novel and clinically meaningful HF phenotypes, which is particularly relevant given the heterogeneity of HFpEF.

Despite these advances, most of the existing research focuses on English-language EHRs, frequently relying on structured data or datasets with broad cardiovascular endpoints. To date, only a limited number of studies have addressed the automated detection of HFpEF-specific symptoms and none, to our knowledge, have done so in Spanish-language clinical narratives. Moreover, existing models are rarely optimized for clinically critical metrics such as sensitivity, nor do they account for the long and complex structure of real-world clinical narratives. This study addresses these gaps by introducing a manually annotated Spanish-language corpus for HFpEF symptom detection, evaluating multiple Transformer architectures, including models adapted for long documents, and implementing task-specific optimization strategies aligned with diagnostic priorities. This represents the first demonstration of the feasibility and clinical applicability of Transformer-based NLP for detecting HFpEF in Spanish EHRs, with direct implications for improving early recognition of cardiac amyloidosis.

Methods

The Corpus

Cardiac amyloidosis may present with a broad spectrum of clinical manifestations. This study focuses on HFpEF, one of the most clinically significant manifestations, to evaluate the performance of language models in detecting its presence or absence within clinical documents. Accordingly, the experiments have been conducted using a dataset specifically annotated for HFpEF. The full corpus, comprising fifteen sets of manually annotated clinical documents covering all cardiac amyloidosis symptoms, will be introduced in future publications. This corpus will be made publicly available to the scientific community.

Description

The corpus was compiled from the electronic health records (EHR) of 262 patients. The records of 173 patients who underwent testing for cardiac amyloidosis were selected and supplemented with the records of 89 patients who did not undergo these tests, aiming to mitigate overfitting in the models. These records were obtained from the Cardiology, Internal Medicine, Neurology, Neurophysiology, and Traumatology units at 'Juan Ramón Jiménez' Hospital in Huelva, Spain. Following a detailed analysis to determine the most relevant document types for this study, the following clinical records were selected: *anamnesis*, *consultation reports*, *discharge reports*, and *progress notes*.

To accommodate language model context limitations, documents were segmented into clinically relevant sections: *current illness*, *clinical assessment*, *reason for consultation*, *medical history*, and *complementary tests*. For progress notes, only the *evolution*, *clinical assessment*, and *complementary test* sections were retained to prioritize relevant content and reduce redundancy.

Table 1. Number of documents by type.

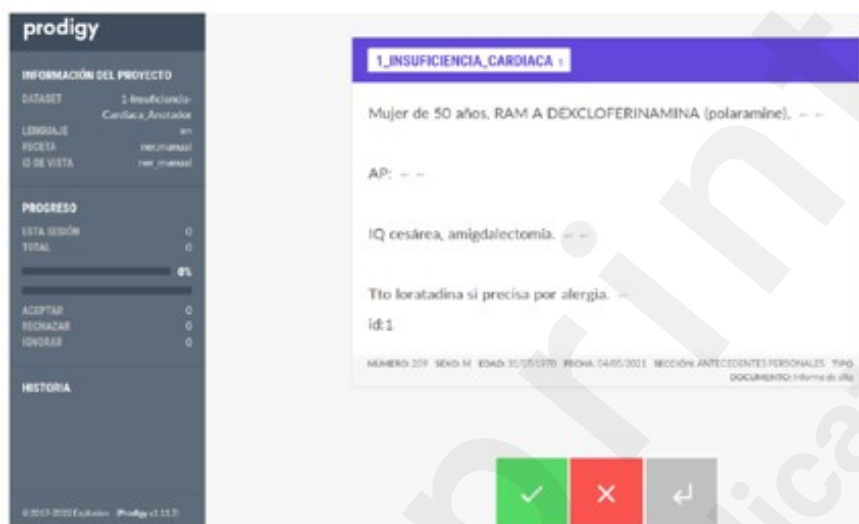
Document Type	Num. Documents
Anamnesis	5,040
Consultation Reports	5,235
Discharge Reports	3,970
Progress Notes	16,122
Total	30,367
Total (after cleaning process)	15,304

Table 1 provides the total number of documents per type, along with the overall corpus size. A data cleaning process was performed to eliminate duplicate and incomplete documents (e.g., those containing minimal text or blank sections) and to correct errors identified during document extraction from EHRs. A sample of the dataset is presented in the Appendix.

Labeling process

This section outlines the corpus labeling process, with particular emphasis on the collection related to HFpEF, as this dataset was used for the experiments presented in this study. The annotation process followed a rigorous methodology appropriate for this type of task. Given the dataset's size, its manual annotation required six months, highlighting the complexity of this phase of the study.

Figure 1. Prodigy interface for dataset annotation.



Based on prior evaluations of annotation tools [13], Prodigy [14] was selected due to its adaptability to the project's requirements and its ease of use. To enhance the annotators' efficiency, improvements were made to the tool's functionality, and a user-friendly interface was designed to streamline the annotation process. Figure 1 illustrates the interface developed for this task.

To ensure consistency and standardization in the annotation process, an annotation guideline was developed under the supervision of a cardiology specialist. This guideline provides a detailed framework on relevant linguistic elements (terms, acronyms, expressions, among others) that annotators must consider determining the presence (positive case) or absence (negative case) of the symptom in the analyzed documents. Table 2 presents an excerpt from the annotation guide, illustrating some of the expressions used to indicate heart failure. Manual labeling was carried out by two specialists in medical documentation. To minimize annotation bias, the task was performed independently by each annotator, without mutual interaction. Discrepancies were resolved by the cardiology specialist who authored the annotation guidelines, ensuring consistency and validity.

Table 2. Excerpt from the annotation guide with some of the expressions used to label the dataset. (*English translation is included to facilitate reading)

Identification of the expression	
Expressions indicating the presence of the symptom	Expressions indicating the absence of the symptom
Ingreso por ICC con FEVI preservada	Sin signos de IC
I Cardíaca con FE conservada	ICC sin especificar si preservada o reducida
Insuficiencia cardíaca diastólica	Insuficiencia cardíaca con FEVI reducida
IC con función sistólica conservada o preservada	IC con FE disminuida

CF III EAP Edema agudo de pulmón NYHA III	IC sistólica
Admitted for CHF with preserved LVEF Heart failure with preserved EF Diastolic heart failure HF with preserved or conserved systolic function Class III APE Acute pulmonary edema NYHA II	No signs of HF CHF unspecified if preserved or reduced Heart failure with reduced LVEF HF with decreased EF Systolic HF

Statistics

The dataset comprises 15,304 clinically labeled documents with binary annotation. The documents exhibit several linguistic characteristics that increase the complexity of classification for language models. Notable challenges include the frequent use of acronyms and abbreviations, semantic variability across equivalent expressions, and acronym polysemy.

Additionally, the dataset exhibits a high-class imbalance, with 1,140 documents (7.5%) labeled as positive (Label '1') and 14,164 documents (92.5%) labeled as negative (Label '0'), posing a challenge for the classification task.

Experimentation framework

The Transformer architecture has become a cornerstone in NLP, offering high efficiency and versatility across a wide range of tasks. Its ability to model long-range dependencies through attention mechanisms has revolutionized the field, driving significant advancements in tasks such as text classification [15, 16, 17]. This architecture is distinguished not only by its accuracy but also by its scalability and adaptability across diverse domains and datasets of varying sizes. Transfer learning allows for domain-specific adaptation through fine-tuning, as applied here to the healthcare domain [18]. By leveraging pre-trained models that have already learned general language representations from large, diverse corpora, transfer learning facilitates their specialization for more domain-specific datasets through additional training. This adaptability has established Transformers as a powerful tool for advancing NLP in specialized fields.

A key limitation of Transformer models is their inherent constraint in processing large texts. Their architecture is typically designed to handle a maximum of 512 tokens, posing challenges for tasks that require longer sequences, such as document-level analysis or summarization of extensive texts. To address this limitation, the Longformer architecture [19] was introduced, specifically designed to efficiently handle longer text sequences. Longformer incorporates sparse attention mechanisms, allowing it to extend the context window while preserving computational efficiency, thus mitigating a core limitation of traditional Transformer models. This innovation broadens the applicability of Transformer models to tasks involving lengthy documents, such as EHRs. These considerations motivated the inclusion of three model categories in this study: general-purpose Transformers, domain-adapted clinical models, and Longformers. This categorization allows for a comprehensive comparison of performance across different levels of domain specialization and sequence-length

flexibility. General-purpose models serve as a baseline, while fine-tuned clinical models assess the added value of domain-specific adaptation. Longformers, by contrast, address the limitations of standard sequence lengths and offer insight into the role of extended context modeling in clinical NLP.

In clinical contexts, maximizing recall for the positive class (Recall-1) is often prioritized to capture as many at-risk cases as possible. However, this strategy can lead to an increased number of false positives, elevating alert rates and potentially burdening healthcare systems. To address this trade-off, we incorporated additional evaluation metrics, specifically the area under the ROC curve (AUC) and the F1-score, to support a more balanced classification approach. Each metric provides a distinct perspective: AUC assesses the model's overall ability to discriminate between positive and negative cases, while F1-score balances precision and sensitivity, helping to reduce both false positives and false negatives.

Models description

As described in the previous section, this study evaluated three categories of Transformer-based models: two general-purpose, four domain-adapted clinical models, and two Longformers. General-purpose models serve as a baseline to assess performance in the absence of domain adaptation, whereas fine-tuned clinical models illustrate the effects of transfer learning on domain-specific performance. Longformer models, designed to efficiently process extended text sequences, were included to address tasks requiring long-range contextual understanding, such as detailed medical reports or full-length patient narratives.

1. *BETO*. The Spanish-BERT model [20] follows the same architecture as BERT-Base and was trained exclusively on Spanish corpora, including Wikipedia and OPUS data. Specifically, we used the *bert-base-spanish-wwm-cased* model.
2. *RoBERTa*. A robustly optimized BERT pre-training approach, RoBERTa is an improved version of BERT that incorporates modifications to key hyperparameters [21]. Several Spanish-language RoBERTa models are available. We used the version pre-trained on corpora from the National Library of Spain (BNE) [22]. Specifically, experiments were conducted using the base version, *roberta-base-bne*.
3. *RoBERTa-biomedical* and *RoBERTa-biomedical-clinical* [23] are monolingual Spanish RoBERTa-based models, trained on a large biomedical and clinical corpus of over 1 billion tokens. These models have demonstrated strong performance in prior Spanish-language biomedical and clinical NLP benchmarks. The specific versions used in this study were *roberta-base-biomedical-es* and *roberta-base-biomedical-clinical-es*.
4. *bsc-bio* and *bsc-bio-ehr* [24] are recently developed Spanish-language models, derived from *roberta-base-biomedical-es* and *roberta-base-biomedical-clinical-es*, respectively, and trained on expanded corpora to improve performance. Specifically, *bsc-bio* is a pretrained model for processing biomedical and clinical texts, suitable for tasks such as literature analysis, information extraction, and interpretation of medical guidelines. In contrast, *bsc-bio-ehr* is specifically adapted for processing electronic health records and clinical notes.
5. *Long Transformer RoBERTa*. This model is a Longformer-based adaptation of RoBERTa for the Spanish language [22]. It combines sliding-window (local) and global attention mechanisms, enabling linear scalability with sequence length and facilitating the processing of documents with thousands of tokens. For this study, we used the model *longformer-base-4096-bne-es*.

6. *Long Transformer Biomedical-Clinical*. This model extends the RoBERTa-based architecture through Longformer adaptations tailored for Spanish biomedical and clinical texts [24]. Initialized from the *roberta-base-biomedical-clinical-es* checkpoint, it underwent further fine-tuning using masked language modeling (MLM), specifically targeting long biomedical and clinical documents. We used the model *longformer-base-4096-biomedical-clinical-es* in our experiments. This model, explicitly designed for processing extended biomedical and clinical texts in Spanish, leverages domain-specific pretraining to achieve strong performance in NLP tasks within the healthcare domain, including the classification of clinical notes.

Evaluation metrics

The models were evaluated using standard performance metrics commonly employed in binary classification tasks for clinical documents. Specifically, the following metrics were considered:

Precision. The proportion of correctly predicted positive clinical documents relative to the total number of documents classified as positive. It is computed as:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \quad (1)$$

Recall (Sensitivity). The proportion of correctly predicted positive clinical documents relative to the total number of documents classified as positive. It is computed as:

$$\text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \quad (2)$$

F1-score. The harmonic mean of precision and recall, providing a single metric that balances both aspects of classification performance. It is computed as:

$$f1\text{-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

Specificity. The proportion of correctly predicted negative clinical documents relative to the total number of actual negative documents. It is computed as:

$$\text{specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \quad (4)$$

Area under the ROC curve (AUC). The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is a performance metric used to evaluate binary classification models. It measures the model's ability to distinguish between positive and negative classes, independent of the decision threshold. The ROC curve plots the True Positive Rate (TPR, or sensitivity) against the False Positive Rate (FPR, or 1-specificity) across different threshold values.

For all metrics, a *positive document* indicates that it contains the symptom, while a *negative document* indicates that the symptom is not found within the text of the document.

Experimental methodology

Building robust models for detecting complex symptoms, such as those associated with heart failure in clinical contexts, requires fine-tuning of pretrained models and optimizing both hyperparameters and training strategies for the target task. To this end, the experimental framework was structured into two phases: an initial hyperparameter search using a reduced dataset, and a subsequent full training and evaluation phase. During Phase 1, multiple hyperparameter configurations were systematically evaluated for each target metric, with the goal of optimizing models to maximize AUC, F1-score, or recall for the positive class. Phase 2 applied the optimal hyperparameter configurations identified previously to train each model on the full dataset and evaluate performance on the test set, enabling the selection of the best-performing configuration per strategy.

Phase 1: Hyperparameter tuning. To limit computational demands, a further reduced subset of the training set was used during the hyperparameter search. This subset included 2,000 randomly sampled instances obtained through undersampling, comprising 1,500 Class 0 and 500 Class 1 examples. Table 3 summarizes the hyperparameter search space used in our experimentation.

A comprehensive grid search was conducted to evaluate all possible combinations of defined parameter values. This approach allows for systematic exploration of the hyperparameter space to identify the optimal configuration for each target performance metric. Optimization was carried out independently for each predefined strategy: (a) maximizing AUC, (b) maximizing F1-score, and (c) maximizing recall (sensitivity) for the positive class.

Table 3. Hyperparameter search space.

Hyperparameter	Values
Number of epochs	10 (using early stopping)
Batch size (BS)	[8, 16, 32]
Learning rate (LR)	[2e-5, 3e-5, 5e-5]
Weight decay (WD)	[0.1, 0.01, 0.001]
Optimizer	[adamw_hf, adamw_torch, adafactor]

Tables 4, 5, and 6 present the models and their optimal configurations for each optimization strategy, detailing key parameters such as batch size, learning rate, weight decay, and optimizer. This systematic, metric-driven optimization process enables the identification of the ideal configuration for each model based on the target metric, thereby enhancing model performance according to the specific requirements of the symptom classification task. As shown in the tables, the optimal hyperparameter values vary depending on the optimization strategy employed. For instance, the bio model achieves better performance in AUC optimization when trained with a batch size of 16. However, its performance improves for F1-score and sensitivity optimization when batch sizes of 32 and 8 are used, respectively. By conducting an exhaustive search for these values, the final models attained optimal performance for each metric.

Table 4. Best hyperparameter values for the AUC optimization strategy.

Model	Batch size (BS)	Learning rate (LR)	Weight decay (WD)	Optimizer
BETO	8	5e-05	0.001	adamw_torch
RoBERTa	8	2e-05	0.01	adafactor

RoBERTa-biomedical	16	2e-05	0.01	adamw_torch
RoBERTa-biomedical-clinical	16	5e-05	0.001	adamw_torch
bsc-bio	16	5e-05	0.001	adafactor
bsc-bio-ehr	32	3e-05	0.01	adamw_hf
Long Transformer RoBERTa	8	3e-05	0.01	adamw_hf
Long Transformer Biomedical-clinical	8	5e-05	0.01	adamw_hf

Table 5. Best hyperparameter values for the F1-score optimization strategy.

Model	Batch size (BS)	Learning rate (LR)	Weight decay (WD)	Optimizer
BETO	8	5e-05	0.001	adamw_torch
RoBERTa	16	5e-05	0.01	adafactor
RoBERTa-biomedical	16	5e-05	0.01	adamw_torch
RoBERTa-biomedical-clinical	8	3e-05	0.1	adamw_hf
bsc-bio	32	5e-05	0.1	adafactor
bsc-bio-ehr	8	2e-05	0.01	adamw_torch
Long Transformer RoBERTa	8	3e-05	0.01	adamw_hf
Long Transformer Biomedical-clinical	8	3e-05	0.001	adamw_hf

Table 6. Best hyperparameter values for the Sensitivity optimization strategy.

Model	Batch size (BS)	Learning rate (LR)	Weight decay (WD)	Optimizer
BETO	8	5e-05	0.1	adamw_torch
RoBERTa	8	2e-05	0.01	adafactor
RoBERTa-biomedical	16	2e-05	0.001	adamw_torch
RoBERTa-biomedical-clinical	32	5e-05	0.001	adamw_hf
bsc-bio	8	5e-05	0.1	adamw_hf
bsc-bio-ehr	8	2e-05	0.1	adamw_hf
Long Transformer RoBERTa	8	2e-05	0.1	adamw_torch
Long Transformer Biomedical-clinical	8	5e-05	0.001	adafactor

Phase 2: Final training and evaluation. Using the optimal hyperparameters identified in Phase 1, the pretrained models were subsequently fine-tuned, each according to a specific optimization strategy (AUC, F1-score, or Sensitivity). Given the pronounced class imbalance in the original training dataset, with a much larger proportion of majority class (label '0') samples than minority class (label '1'), a reduced training set was created using undersampling. All minority class examples (1,176) were retained, and three times that number were randomly sampled from the majority class (3,528), yielding a training dataset of 4,704 instances. To enable model tuning without compromising final evaluation, 20% of the reduced training set was reserved as a validation dataset. Undersampling helps mitigate class imbalance by ensuring a more equitable representation of classes. This technique

yields a manageable dataset size while maintaining a class distribution appropriate for the task, especially for detecting clinical symptoms in both classes. To preserve model generalizability, the test dataset was kept intact and fully isolated from the training process, enabling objective performance evaluation across the full dataset. Although undersampling can lead to the exclusion of some majority class instances, this trade-off was deemed acceptable to ensure an efficient and clinically relevant training process.

Results

This section presents and analyzes the results obtained on the test dataset for the models trained under the three previously described optimization strategies.

The dataset was divided into training (80%), validation (20%), and test (20%) subsets to support model development and evaluation. Table 7 shows the number of documents in each dataset partition. As previously described, an undersampling strategy was applied to the training and validation sets to mitigate class imbalance.

Table 7. Size and label distribution in the training, valid and test datasets.

Dataset	# Docs	Label 0	Label 1
Original	15,304	13,853	1,451
Train	3,764	2,823	941
Valid	940	705	235
Test	3,032	2,757	275

To improve model robustness and ensure greater statistical validity, two strategies were implemented in the construction of the test dataset:

- Ensuring patient-level separation between the training and test datasets. That is, test documents originated from patients not included during model training. This ensured that the test dataset contained entirely unseen data for the model. Satisfactory results under this condition would indicate strong generalization capability in detecting HFpEF symptoms.
- Preserving the original class distribution in the test dataset (91% class '0', 9% class '1'). This strategy allows evaluation under realistic conditions reflecting the natural class imbalance between negative and positive HFpEF cases.

Table 8. Results for the AUC optimization strategy.

Model	Sens.	Spec.	F1	AUC
BETO	0.920	0.960	0.959	0.940
RoBERTa	0.938	0.985	0.983	0.962
RoBERTa-biomedical	0.960	0.972	0.972	0.966
RoBERTa-biomedical-clinical	0.967	0.961	0.964	0.964
bsc-bio	0.945	0.980	0.977	0.963
bsc-bio-ehr	0.953	0.980	0.978	0.966
Long Transformer RoBERTa	0.960	0.952	0.957	0.956

Long Transformer Biomedical-clinical	0.956	0.987	0.985	0.971
--------------------------------------	-------	--------------	--------------	--------------

Table 9. Results for the F1-score optimization strategy.

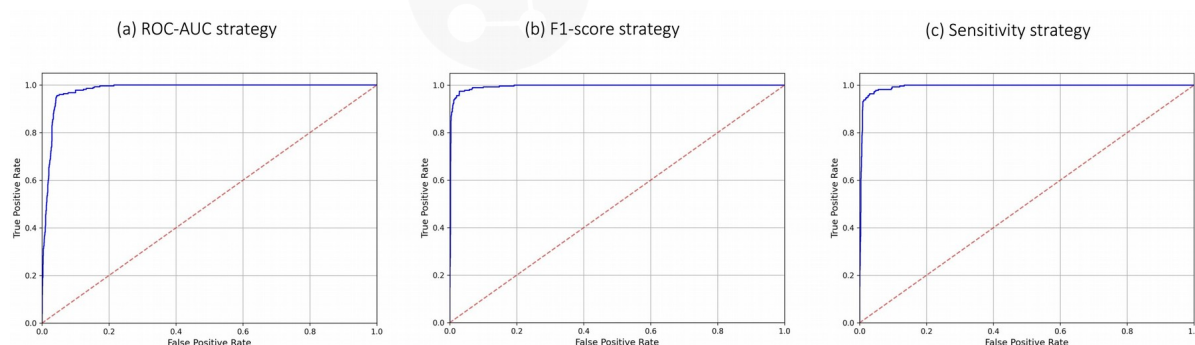
Model	Sens.	Spec.	F1	AUC
BETO	0.949	0.935	0.942	0.942
RoBERTa	0.967	0.946	0.952	0.957
RoBERTa-biomedical	0.985	0.849	0.885	0.917
RoBERTa-biomedical-clinical	0.967	0.923	0.936	0.945
bsc-bio	0.956	0.971	0.971	0.963
bsc-bio-ehr	0.956	0.971	0.971	0.964
Long Transformer RoBERTa	0.920	0.987	0.981	0.954
Long Transformer Biomedical-clinical	0.987	0.987	0.985	0.987

Table 10. Results for the Sensitivity optimization strategy.

Model	Sens.	Spec.	F1	AUC
BETO	0.938	0.955	0.957	0.947
RoBERTa	0.938	0.986	0.983	0.962
RoBERTa-biomedical	0.953	0.934	0.942	0.943
RoBERTa-biomedical-clinical	0.967	0.925	0.937	0.946
bsc-bio	0.985	0.943	0.952	0.964
bsc-bio-ehr	0.960	0.978	0.977	0.969
Long Transformer RoBERTa	0.979	0.956	0.968	0.968
Long Transformer Biomedical-clinical	0.971	0.922	0.935	0.946

Tables 8, 9, and 10 summarize the performance of the six selected language models trained under the three optimization strategies. The evaluation metrics indicate robust and consistent performance across models. A key observation is the consistently high AUC values across models, all surpassing the 0.80 threshold considered clinically significant [25]. Figure 2 displays the ROC curves of the models that achieved the best performance in terms of AUC.

Figure 2. ROC curves obtained by the classifiers on the test dataset using the three hyperparameter optimization strategies. The first two curves correspond to the AUC and F1-score optimization strategies and were generated by the *Long Transformer Biomedical-clinical* model. The curve representing the Sensitivity strategy corresponds to the *bsc-bio-ehr* model.

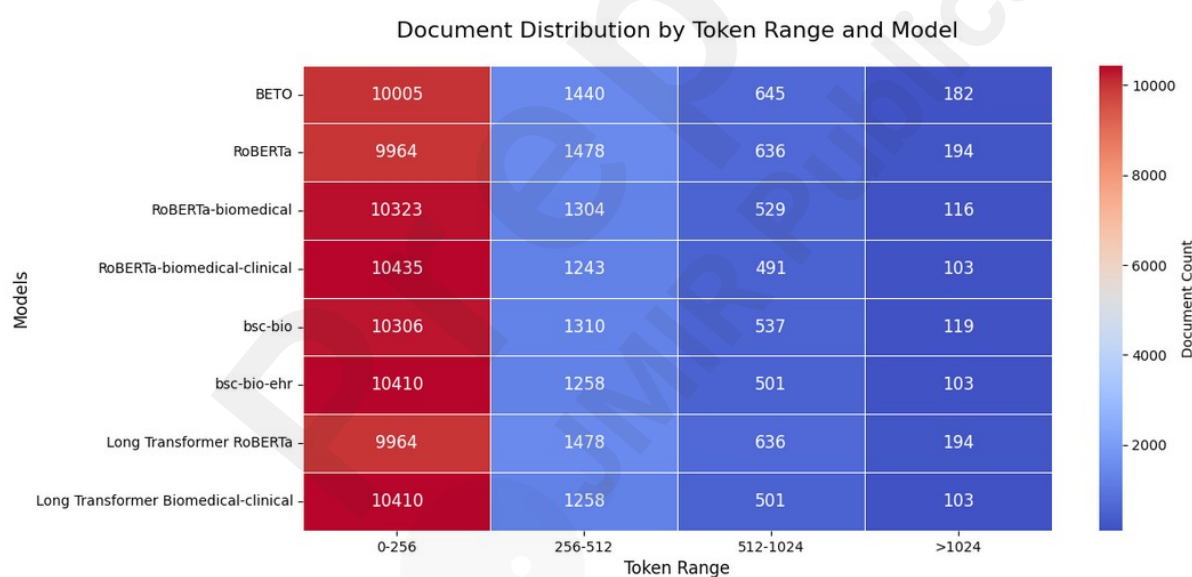


Overall, the *Long Transformer Biomedical-clinical* model achieved the highest performance,

highlighting that domain-adapted pretrained models generally outperform general-purpose counterparts in specialized clinical tasks. Moreover, models capable of processing extended input sequences exhibit superior suitability for clinical domains, where documents often exceed the standard 512-token limitation of conventional Transformer architectures.

As shown in Figure 3, analysis of token distributions across models revealed that the majority of clinical documents (approximately 85%–95%, depending on the tokenizer) contain 512 tokens or fewer. However, a non-negligible subset of documents (5%–15%) exceeds this limit, underscoring the presence of longer narratives in the dataset. In standard Transformer-based models, these longer documents were truncated during processing due to sequence length limitations. Although this limitation may result in the loss of relevant contextual information, the models still achieved robust performance, indicating their effectiveness even when operating on truncated inputs. To better handle longer documents, Longformer-based models were employed, enabling a maximum input length of 1,024 tokens. This extended capacity allows for more complete document representations while preserving computational efficiency. While only a small fraction of the dataset (1%–2%) exceeded 1,024 tokens, the increased sequence length enhanced contextual coverage for longer documents compared to standard Transformer models. The inclusion of Longformers in the experimental framework highlights their value in handling tasks involving lengthy and complex texts, particularly in clinical NLP, where capturing contextual nuances is crucial for effective classification.

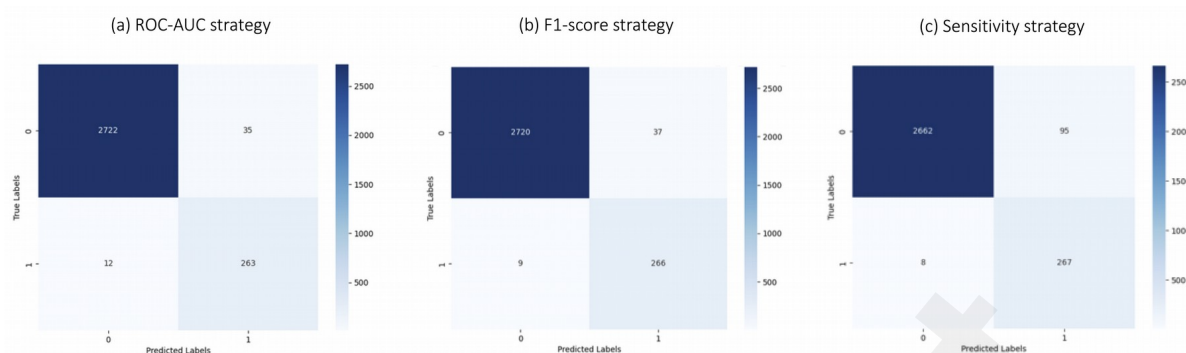
Figure 3. Document distribution by token count based on each model's tokenizer.



In diagnostic applications, minimizing the false negative (FN) rate is critical to improving the model's ability to correctly identify positive cases and facilitate early disease detection. Accordingly, sensitivity was prioritized in one experimental condition by training a model specifically optimized for this metric. This approach increased the correct classification rate of positive cases, thereby reducing the likelihood of missing patients in need of further clinical evaluation. To evaluate the impact of each optimization strategy on model performance, confusion matrices were analyzed on the test dataset. Figure 4 shows the confusion matrices of the best-performing models under each optimization strategy.

Figure 4. Confusion matrices obtained by the classifiers on the test dataset using the three hyperparameter optimization strategies. The first two matrices, corresponding to the *AUC* and *F1-score* optimization

strategies, were obtained by model *Long Transformer Biomedical-clinical*, while the confusion matrix for the *Sensitivity* strategy was obtained by model *Long Transformer RoBERTa*.



The trained Transformer-based models exhibited high sensitivity across all three optimization strategies, consistently achieving low false-negative rates. In particular, classification accuracy for the positive class exceeded 96% in all cases, highlighting the robustness of both the annotated corpus and the experimental design. These findings underscore not only the models' effectiveness in detecting positive cases, but also the consistency and reliability of the adopted methodology. Confusion matrix analysis reveals subtle performance differences depending on the optimization strategy applied. Notably, the sensitivity-optimized strategy reduced false negatives to just 8, resulting in a 97% accuracy rate for positive case detection. However, this improvement came at the cost of increased false positives (95), illustrating the inherent trade-off between sensitivity and specificity. In contrast, the AUC and F1-score strategies, which aim to balance accuracy across classes, significantly reduced false positives (35 and 37, respectively). Simultaneously, these strategies maintained relatively low false-negative counts (12 and 9, respectively).

Discussion

It should be noted that direct comparisons with related studies are not feasible, as each employs different datasets. Nevertheless, the results outlined above demonstrate the validity of the experimental design and the robustness of the trained models under the proposed optimization strategies, as evidenced by the evaluation metrics. Additionally, the involvement of cardiology specialists provides expert validation and qualitative support for these findings.

Error analysis

A qualitative error analysis was performed from a clinical perspective, focusing on both false negative and false positive cases. Special attention was given to false negatives to identify the reasons why the models predicted the absence of the symptom in a document (label '0') when, in reality, it had been classified as positive, meaning that the document contained information indicating the presence of the symptom. We selected the misclassified documents from the three models with the lowest false negative rates (see Figure 4).

Table 11. Example of a misclassified document (false negative). Document manually labeled as positive and classified as negative by all three classifiers. (*Automatic English translation is included to facilitate reading)

Document

Ecg: en fa a unos 135-140 lpm, e isquemia subendocardica lateral.cr 1.6, probnp 6869, tnt 20, iones normales, hb 10.7g, inr tp 2.8.bq planta: cr 1.40mg, ggtp 84u , urea 68mg, iones normales, hierro 24 mcg, ab 3.3g, mg 1.52 mg.pco2 63, ph 7.33, bicarbonato 33.orina normal.rx de torax: cardiomegalia, signos de hipertension pulmonar precapilar y poscapilar con infiltrados de edema intersticioalveolar bilaterales.

ECG: Atrial fibrillation at approximately 135-140 bpm, with lateral subendocardial ischemia. CR 1.6, ProBNP 6869, TnT 20, normal ions, Hb 10.7 g, INR/TP 2.8. Ward: CR 1.40 mg, GGTP 84 U, Urea 68 mg, normal ions, Iron 24 mcg, AB 3.3 g, Mg 1.52 mg. PCO2 63, pH 7.33, Bicarbonate 33. Urinalysis: Normal. Chest X-ray: Cardiomegaly, signs of pre- and post-capillary pulmonary hypertension, with bilateral interstitial-alveolar edema infiltrates.

Table 11 presents a false negative example, in which all models predicted the case as negative, despite human annotation indicating the presence of the symptom. However, a review of the annotation guidelines revealed that the annotator's interpretation relied on clinical reasoning that, while valid, extended beyond the defined annotation criteria. The text includes terms such as *"cardiomegalia, signos de hipertensión pulmonar precapilar y poscapilar con infiltrados de edema intersticial alveolar bilaterales"* (*"cardiomegaly, signs of pre- and post-capillary pulmonary hypertension with bilateral interstitial-alveolar edema infiltrates"*), which can be associated with heart failure with preserved left ventricular ejection fraction (HFpEF). However, the guidelines did not explicitly state that these findings alone were sufficient for a positive classification without additional diagnostic information or clinical context. In this regard, the model's prediction strictly adhered to the established criteria, whereas the human annotation reflected a broader interpretation based on implicit clinical knowledge. This example highlights the inherent complexity of symptom detection and the subtle distinction between rule-based annotation and clinically contextual judgment. Despite these occasional discrepancies, the models have demonstrated a high level of consistency in applying the predefined criteria. Future revisions of the annotation process could explore potential adjustments to better align guideline application with the expected clinical reasoning for this task.

Table 12. Example of a misclassified document (false positive). Document manually labeled as negative and classified as positive by all three classifiers. (*Automatic English translation is included to facilitate reading)

Document
Hoy tiene prevista tercera infusión del fármaco. peso 49.5 kg según lo recogido hoy. clínicamente bien, salvo edemas en ambos maléolos, creo que en relación al patisiran pues no existe daño renal ni signos de icc.ef. beg. acr con cor rítmico, sin soplos audibles, con bmv. abdomen sin masas ni megalias. mmii cion edemas que dejan fôve a nivel maleolar. no signos de tvp.plan:- próxima cita 10 de julio- controles analíticos los de nefrología por ahora (niveles de tacrolimus en rango en último control).
Today, the third infusion of the drug is scheduled. Weight: 49.5 kg as recorded today. Clinically well, except for edema in both malleoli, which I believe is related to patisiran, as there is no renal damage or signs of heart failure. Physical examination: Well-nourished and in good general condition. Peripheral circulation: rhythmic heart sounds, no audible murmurs, with good bilateral breath sounds. Abdomen: no masses or organ enlargement. Lower limbs: with edema that leaves a pit at the malleolar level. No signs of deep vein thrombosis. Plan: Next appointment: July 10. Laboratory tests: nephrology monitoring for now (tacrolimus levels within range in the last test).

Conversely, Table 12 shows an example of a false positive error. In this case, the models incorrectly predicted the presence of HFpEF symptoms in a document explicitly indicating their absence. Although the key phrase *"no existe daño renal ni signos de ICC"* ("no kidney damage or signs of CHF") clearly indicated the negative class, the models failed to interpret the negation correctly in this context. Such errors are a well-known challenge in clinical NLP, where negation is often expressed using diverse and context-dependent linguistic patterns. Notably, all three models produced the same misclassification, indicating that the challenge in interpreting negation patterns is model-independent. Despite these specific instances, the overall performance of the models has been consistent across most evaluated cases. However, future improvements could focus on refining negation handling, either through adjustments in language representation or specific preprocessing strategies that enhance the accurate interpretation of negative structures.

Clinical implications of misclassification errors

From a clinical standpoint, accurately identifying patients with suspected HFpEF is critical for ensuring appropriate management. Classification errors may result in the omission of key diagnostic procedures or delays in initiating disease-modifying therapies. In particular, false negatives represent a major concern, as they may leave early-stage patients undiagnosed and untreated. Conversely, false positives may prompt unnecessary diagnostic procedures, increasing both clinical workload and healthcare costs. Therefore, tuning strategies that prioritize sensitivity, while preserving acceptable specificity, are essential for the effective deployment of these models in clinical settings.

Conclusions

This study proposed an experimental framework for optimizing Transformer-based language models to automatically detect clinical documents suggesting the presence of heart failure with preserved ejection fraction (HFpEF). Our findings underscore the importance of hyperparameter tuning, as optimal configurations directly influence model performance according to the prioritized evaluation metric. Specifically, to minimize false negatives, optimization should prioritize model sensitivity. This approach allows model performance to be adapted to varying clinical and operational requirements, maintaining a suitable balance between sensitivity and specificity depending on the context of use.

It has been demonstrated that using of domain-specific pretrained language models significantly improves adaptability and transferability for specialized clinical task. This advantage lies in the ability of context-aware models to better capture linguistic patterns, domain-specific terminology, and semantic relationships, thereby enhancing task-specific performance. Moreover, the capacity to process longer sequences during fine-tuning and inference is especially beneficial in the clinical domain. Medical documents are often lengthy and rich in interrelated concepts, which may be lost when contextual processing is constrained. Therefore, employing models designed for extended input sequences is essential for comprehensive and accurate clinical text analysis.

Having validated our experimental approach and confirmed model reliability for HFpEF detection, future work will focus on developing and fine-tuning models capable of comprehensively identifying all symptoms associated with cardiac amyloidosis. To this end, the top-performing models from this study will be retrained on manually annotated corpora to incorporate the full range of clinically relevant manifestations. This process will improve the system's ability to detect textual patterns indicative of the disease with increased precision and sensitivity. The ultimate aim is to implement an

automated screening system based on EHR analysis to support the early identification of patients with suspected cardiac amyloidosis. This approach may optimize diagnostic workflows, improve clinical decision-making, and enable more personalized treatment strategies for a condition that remains difficult to detect at early stages.



Conflicts of Interest

None declared.

Data availability

Data will be made available on request.

Ethical Approval

This study was approved by the Andalusian Biomedical Research Ethics Coordinating Committee under protocol version 1 dated 10/02/2020, with internal protocol code: 2382-N-20.

Acknowledgments

We thank María Isabel Rengel, IT Specialist from the Team IT SAS Huelva, for her valuable support in extracting data from clinical records and curating clinical episodes. Her work in identifying relevant terms related to potential signs of amyloidosis has been essential to the quality of our dataset.

Funding

This project was funded by the Institute of Health Carlos III, Ministry of Science, Innovation and Universities, Spanish Government, grant number PI20/01485.

Appendix

Sample of documents extracted from the dataset. (*Automatic English translation is included to facilitate reading)

Document	Label
Consulta por urgencias hace 48hrs por sensación de palpitaciones y disnea en reposo así como tos escasamente productiva de una semana de evolución. Refiere dolor torácico atípico tipo pleurítico que aumenta con la inspiración. Afebril hasta el momento. No otra sintomatología por aparatos ni sistemas. -Tabaquismo activo de 5 cig/día. Impresiona de mala adherencia al tratamiento y cumplimiento errático. En urgencias se describe con buen estado general. Hallazgo de acidosis respiratoria en GSV (pH: 7.27 y pCO₂: 88 mmHg), que empeora tras la administración de aerosoles, por lo que pasa a observación y se inicia soporte VNI, manteniéndose dependiente de la VNI por trabajo respiratorio y empeoramiento de la hipercapnia tras su intento de retirada.	0
<i>Emergency consultation 48 hours ago due to a sensation of palpitations and dyspnea at rest, as well as a mildly productive cough lasting one week. Reports atypical pleuritic chest pain that worsens with inspiration. Afebrile so far. No other symptoms affecting other systems. Active smoker, 5 cigarettes/day. Appears to have poor adherence to treatment and erratic compliance. In the emergency department, described as being in good general condition. Finding of respiratory acidosis in arterial blood gas (pH: 7.27 and pCO₂: 88 mmHg), which worsened after aerosol administration, leading to observation and initiation of non-invasive ventilation (NIV), remaining dependent on NIV due to respiratory effort and worsening hypercapnia after attempted withdrawal.</i>	0
Mujer de 78 años refiere aumento de disnea progresivo desde hace 2 meses, acentuado en los últimos 5 días, NYHA III. Opresión torácica sin contejo vegetativo tanto en esfuerzo como con algunos episodios de reposo. Refiere ortopnea con episodios de DPN. Palpitaciones. Refiere edemas ocasionales en MMII. Refiere episodios de mareos frecuentes con caídas en relación con su cervicopatía y otros nocturnos de perfil ortostático sin síncope. En ecg realizado en MAP, se observa TSV a 120 lpm, En Urgencias, ecg en ritmo sinusal a 80 lpm. Se realiza placa de tórax con signos sugestivos de descompensación de origen cardíaco (pulmón congestivo, con signos de líquido intersticial en cisura y borramiento senos costofrénicos), con Pro BNP 2500 y Hb 7, 8. Se procede al sondaje vesical de la paciente y se administra 1 ampolla de furosemida. Al llegar a planta, se realiza transfusión de 1 Concentrado de hematíes + Ampolla de furosemida.	1
<i>A 78-year-old woman reports a progressive increase in dyspnea over the past two months, worsening in the last five days, classified as NYHA III. She experiences chest tightness without autonomic symptoms, both with exertion and during some episodes at rest. Reports orthopnea with episodes of paroxysmal nocturnal dyspnea (PND). Palpitations. Occasional lower limb edema. Frequent dizziness episodes with falls related to her cervical pathology, as well as nocturnal episodes with an orthostatic profile, without syncope. An ECG performed by the primary care physician (PCP) shows supraventricular tachycardia (SVT) at 120 bpm. In the emergency department, the ECG shows sinus rhythm at 80 bpm. A chest X-ray reveals signs suggestive of cardiac decompensation (congestive lungs, interstitial fluid in the fissure, and blurring of the costophrenic angles), with Pro BNP of 2500 and hemoglobin (Hb) of 7.8. A urinary catheter is placed, and one ampoule of furosemide is administered. Upon admission to the ward, the patient receives a transfusion of one unit of packed red blood cells plus another ampoule of furosemide.</i>	1

References

1. Borlaug BA, Sharma K, Shah SJ, Ho JE. Heart failure with preserved ejection fraction: JACC scientific statement. *J Am Coll Cardiol*; 2023 (81):1810-1834. doi: <https://doi.org/10.1016/j.jacc.2023.01.049>
2. Yamamoto H, Yokochi T. Transthyretin cardiac amyloidosis: an update on diagnosis and treatment. *ESC Heart Fail*; 2019 (6):1128-1139. doi: <https://doi.org/10.1002/ehf2.12597>
3. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*; 2023 (388):1201-1208. doi: <https://doi.org/10.1056/NEJMra2302038>
4. Turchioe MR, Volodarskiy A, Pathak J, Wright DN, Tcheng JE, Slotwiner D. Systematic review of current natural language processing methods and applications in cardiology. *Heart*; 2022 (108):909-916. doi: <https://doi.org/10.1136/heartjnl-2021-319769>
5. Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical text classification research trends: systematic literature review and open issues. *Expert Syst Appl*; 2019 (116):494-520. doi: <https://doi.org/10.1016/j.eswa.2018.09.034>
6. Adejumo P, Thangaraj PM, Dhingra LS, Aminorroaya A, Zhou X, Brandt C, et al. Natural language processing of clinical documentation to assess functional status in patients with heart failure. *JAMA Netw Open*. 2024;7(11):e2443925. doi: <https://doi.org/10.1001/jamanetworkopen.2024.43925>
7. Liu Y, Tan Z, Zhang Z, Wang S, Guo J, Liu H, Chen T, Bian J. A natural language processing-based approach for early detection of heart failure onset using electronic health records. *medRxiv*. Preprint posted online April 6, 2025. doi: <https://doi.org/10.1101/2025.04.04.25325211>
8. Marti-Castellote PM, Reeder C, Claggett B, Singh P, Lau ES, Khurshid S, et al. Natural language processing to adjudicate heart failure hospitalizations in global clinical trials. *Circ Heart Fail*. 2025;18(1):e012514. doi: <https://doi.org/10.1161/CIRCHEARTFAILURE.124.012514>
9. Ahmad FS, Luo Y, Wehbe RM, Thomas JD, Shah SJ. Advances in machine learning approaches to heart failure with preserved ejection fraction. *Heart Fail Clin*. 2022;18(2):287–300. doi: <https://doi.org/10.1016/j.hfc.2021.12.002>
10. Houssein EH, Mohamed RE, Ali AA. Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. *Sci Rep*. 2023;13:7173. doi: <https://doi.org/10.1038/s41598-023-34294-6>
11. Houssein EH, Mohamed RE, Hu G, et al. Adapting transformer-based language models for heart disease detection and risk factors extraction. *J Big Data*. 2024;11(47). doi: <https://doi.org/10.1186/s40537-024-00903-y>
12. Fan Z, Mamouei M, Li Y, Rao S, Rahimi K. Identification of heart failure subtypes using transformer-based deep learning modelling: a population-based study of 379,108 individuals. *EBioMedicine*. 2025;114:105657. doi: <https://doi.org/10.1016/j.ebiom.2025.105657>
13. Neves M, Ševa J. An extensive review of tools for manual annotation of documents. *Brief Bioinform*; 2019 (22):146-163. doi: <https://doi.org/10.1093/bib/bbz130>
14. Honnibal M, Montani I, et al. Prodigy: an annotation tool for AI, machine learning & NLP. Available from: <https://prodi.gy>
15. Fields J, Chovanec K, Madiraju P. A survey of text classification with transformers: how wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*; 2024 (12):6518-6531. doi: <https://doi.org/10.1109/ACCESS.2024.3349952>
16. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning-based text classification: a comprehensive review. *ACM Comput Surv*; 2022 (54):62. doi: <https://doi.org/10.1145/3439726>
17. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in

- natural language processing via large pre-trained language models: a survey. *ACM Comput Surv*; 2024 (56):30. doi: <https://doi.org/10.1145/3605943>
18. Nerella S, Bandyopadhyay S, Zhang J, Contreras M, Siegel S, Bumin A, et al. Transformers and large language models in healthcare: a review. *Artif Intell Med*; 2024 (154):102900. doi: <https://doi.org/10.1016/j.artmed.2024.102900>
 19. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. *arXiv*, 2020. doi: <https://doi.org/10.48550/arXiv.2004.05150>
 20. Cañete J, Chaperon G, Fuentes R, Ho JH, Kang H, Pérez J. Spanish pre-trained BERT model and evaluation data. In: *PML4DC at ICLR*; 2020. doi: <https://doi.org/10.48550/arXiv.2308.02976>
 21. Liu Y. Roberta: a robustly optimized BERT pretraining approach. *arXiv*, 2019. doi: <https://doi.org/10.48550/arXiv.1907.11692>
 22. Gutiérrez Fandiño A, Armengol Estapé J, Pàmies M, Llop Palao J, Silveira Ocampo J, et al. MarIA: Spanish language models. *Proces Leng Nat*; 2022 (68):39-46. doi: <https://doi.org/10.26342/2022-68-3>
 23. Carrino CP, Armengol-Estapé J, Gutiérrez-Fandiño A, Llop-Palao J, Pàmies M, et al. Biomedical and clinical language models for Spanish: on the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv*, 2021. doi: <https://doi.org/10.48550/arXiv.2109.03570>
 24. Carrino CP, Llop J, Pàmies M, Gutiérrez-Fandiño A, Armengol-Estapé J, et al. Pretrained biomedical language models for clinical NLP in Spanish. In: *Proc 21st Workshop on Biomedical Language Processing*; 2022. p.193-199. doi: <https://doi.org/10.18653/v1/2022.bionlp-1.19>
 25. Çorbacioğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: a guide to interpreting the area under the curve value. *Turk J Emerg Med*. 2023;23(4):195–198. doi: https://doi.org/10.4103/tjem.tjem_182_23