Computers in Biology and Medicine 182 (2024) 109127

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Quantifying decision support level of explainable automatic classification of diagnoses in Spanish medical records

Nuria Lebeña [a,*], Alicia Pérez [b], Arantza Casillas [a]

[a] HiTZ Center - Ixa, Department of Electricity and Electronics, University of the Basque Country (UPV/EHU), Barrio Sarriena 2, Leioa 48940, Spain
[b] HiTZ Center - Ixa, Department of Computer Languages and Systems, University of the Basque Country (UPV/EHU), Rafael Moreno "Pitxitxi" 2/3, Bilbao 48013, Spain

ARTICLE INFO

Keywords:
Clinical natural language processing
Electronic health records in Spanish
International classification of diseases
Transformers
Explainability in large language models

ABSTRACT

Background and Objective: In the realm of automatic Electronic Health Records (EHR) classification according to the International Classification of Diseases (ICD) there is a notable gap of non-black box approaches and more in Spanish, which is also frequently ignored in clinical language classification. An additional gap in explainability pertains to the lack of standardized metrics for evaluating the degree of explainability offered by distinct techniques.

Methods: We address the classification of Spanish electronic health records, using methods to explain the predictions and improve the decision support level. We also propose Leberage a novel metric to quantify the decision support level of the explainable predictions.

We aim to assess the explanatory ability derived from three model-independent methods based on different theoretical frameworks: SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Integrated Gradients (IG). We develop a system based on longformers that can process long documents and then use the explainability methods to extract the relevant segments of text in the EHR that motivated each ICD. We then measure the outcome of the different explainability methods by implementing a novel metric.

Results: Our results beat those that carry out the same task by 7%. In terms of explainability degree LIME appears as a stronger technique compared to IG and SHAP.

Discussion: Our research reveals that the explored techniques are useful for explaining the output of black box models as the longformer. In addition, the proposed metric emerges as a good choice to quantify the contribution of explainability techniques.

1. Introduction

Automating Electronic Health Records (EHR) classification according to the International Classification of Diseases (ICD) is becoming a critical task in the healthcare environment. The ICD classification, which is a laborious manual process, is crucial for enhancing healthcare systems' efficiency freeing healthcare workers from this task and allowing them to focus on tasks more closely related to the patient. This study delves into the clinical decision support for the classification of EHRs in Spanish, aiding the decision support by using evidence-based diagnostic prediction. Due to limited resources, Spanish is often disregarded in the area of automated clinical language processing, yet it is essential as clinical coding is mandatory not only in Spain but also in other Spanish-speaking countries.

Large Language Models (LLMs), such as longformers, are frequently seen as black boxes as how the decision is made is hardly understandable and thus, criticized for their opacity. Under the European Union's "General Data Protection Regulation" and the "European General Data Protection Regulation" users have a legal right to an explainable argumentation on the logic involved in the decisions adopted by computer systems [1,2]. Moreover, the first law written on AI from the European Parliament challenges the widespread usage of "black box" models and highlights the need for transparency across all general-purpose AI systems [3]. This increasing need for transparent and human-oriented models has led to the so-called eXplainable Artificial Intelligence (XAI). XAI aims to respond to society not only with effective AI solutions but also with understandable evidence that motivated the results provided by the AI. Sensitive domains like healthcare are particularly in need of XAI because it relies on complex data sources and the need

* Corresponding author.
E-mail address: nuria.lebena@ehu.eus (N. Lebeña).

https://doi.org/10.1016/j.compbiomed.2024.109127
Received 26 February 2024; Received in revised form 17 August 2024; Accepted 5 September 2024
Available online 12 September 2024

for transparency and liability in decision-making [4,5]. Due to these requirements of the institutions, the concern of generating explainable models is gaining importance and is becoming a core topic of international relevant conferences [6–8] In addition, certain projects run by the Defense Advanced Research Projects Agency (DARPA) [9] as well as other European projects such as Horizon and Antidote are working with XAI techniques [10].

The concept of "*explainability*" itself has different definitions [1]. Our focus is on the "*outcome explanation issue*", that is, a user-centric meaning of XAI in which the user should be able to understand how a model has arrived at a result. For example, in text classification, this could be done by highlighting the relevant parts of the document that motivated the class proposed by the model. This could help the developer improve the system based on the system output, as well as, the users to trust the system's decision [4].

If Spanish EHR classification literature is scarce, those that do so without using black box models are even fewer. In Spanish, explainability has been explored by using convolutional neural networks as a tool for EHR categorization. The precise text passages that contributed to the assignment of each code were highlighted, enhancing transparent AI and, therefore, improving decision support systems for clinicians in charge of ICD coding [11].

In an attempt to gain some sense of explainability using transformer models, in another approach apart from the predicted codes the details about the precise text span that influenced the inference were also extracted [12]. They achieved this by combining two transformers, first one was trained in Medical Entity Recognition (MER) task, and the output served then as input to a transformer trained in medical Named Entity Normalization (MEN) that classified according to ICD-10 the clinical entities identified by the first transformer. While classifying documents, it is important to consider codes that are not specifically specified in the document; this method overlooks these codes, highlighting a gap in the proposed approach [12]. Moreover, the existing explainability methods in Spanish focus on using systems that are self-explainable, that is, the system itself is designed to be able to mark those words that motivated a prediction, therefore it is a characteristic of the system to be self-explanatory. Given the fact that the majority of Deep learning (DL) models lack this quality, i.e., they are not self-explanatory, we focus on explainability strategies that may be applied to any current DL model.

Another notable gap in explainability is the absence of standardized metrics to measure and compare the explainability degree provided by different explainable techniques [13]. The lack in this regard inhibits the capacity to enhance the transparency of AI models as different explainability techniques outputs cannot be compared among different researches [1]. Doxpy is an example of a model-agnostic metric for objective evaluation of explainability in generative AI, inspired by Ordinary Language Philosophy [14]. It bases explanations on Achinstein's theory, where explanations are answers to archetypal questions, implying that a system's explainability increases with its ability to answer these fundamental questions [15]. Similarly, in our work, we developed a metric aimed at assessing the understandability degree of the aforementioned explainability techniques in a classification task. We develop an application that can identify and highlight words and phrases that are pertinent to each prediction.

In this work, we, therefore, put in value the relevance of XAI approaches to aid clinical decision-making for clinical documentation of health reports in Spanish. Diagnoses are extracted and, thus, the EHRs are classified following the ICD standard as in clinical documentation, associating each EHR with all the inherent diagnoses (i.e. ICDs) either explicitly or implicitly stated. The XAI highlights relevant segments of text in the EHR that motivated each ICD (being the ICDs the outcome of the model) as a support system for clinical documentation. That is, humans are aided with XAI in the EHR classification task. We have explored three XAI approaches, being the main **contribution** of our paper to assess, quantitatively, the performance of XAI approaches. We have proposed a novel metric to this end, with which we have evaluated the performance of the three XAI approaches, providing thus, quantitative and qualitative evidence of each of them.

2. Methods

In this paper, we seek to evaluate the explanatory ability extracted by different model-independent techniques that can be applied to any neural network. Our goal is to create a system able to handle long documents whose predictions can subsequently be explained using the aforementioned model-independent explainability techniques. To this end, first, we build a classification model based on transformers able to handle long documents. We also investigate and assess several explainability strategies using a unique metric that aims to broaden the explainability evaluation standard methods.

In this section, first, we describe the methodology used for the classification of large clinical texts. Next, we describe different techniques used to evaluate the explainability of the model and we give insights about Leberage, the developed evaluation metric.

2.1. Long document handling

The maximum input token capacity of traditional LLMs, which is usually restricted at 512 tokens [16], is a limitation as the EHRs employed in our research significantly exceed this limit, averaging 1575 tokens per document. This disparity makes it essential to explore novel approaches that can handle long documents.

To overcome this limitation, a number of innovative approaches have been explored to address the challenge of processing large EHRs. In the work titled "*PLM-ICD: Automatic ICD Coding with Pretrained Language Models*" they employed an approach known as *segment pooling* that consisted of dividing documents into smaller segments that could be processed by Pretrained Language Models (PLMs) such as BioBERT [17], PubMedBERT [18], and RoBERTa-PM [19]. Segments were then combined to represent the entire document overcoming the maximum length restriction of PLMs [20] .

Needless to say, disregarding parts of documents due to processing limitations might yield to missing relevant contextual information. A major breakthrough in processing long sequences was made when Longformer was first introduced [21]. By employing the localized sliding window and global attention mechanisms, Longformers were able to reduce the computational expenses of full self-attention mechanisms from quadratic to linear. A similar approach, known as BigBird, also emerged at the time. This approach reduced the computing expense of full self-attention mechanisms by applying a combination of sparse attention, global attention, and random attention mechanisms to the input sequence [22].

In this research, we need to process long clinical texts, so alternatives such as Longformer or Big-bird emerge as good candidates to carry out our research. An investigation compared the performance of Longformer and BigBird [23]. They proposed Clinical Longformer and Clinical BigBird; Longformer and BigBird further trained with clinical knowledge. The proposed Clinical Longformer outperformed Clinical BigBird and traditional Longformer in several proposed tasks, including clinical document classification, Named Entity Recognition (NER), and question answering, proving to be able to handle long clinical documents. We employed this Clinical Longformer variant as it meets our two requirements showing good results: it can process long documents and is tailored to clinical terminology [23].

2.2. Model-independent explainability techniques

To describe our model's decision-making processes, we use three well-known explainability techniques: SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Integrated Gradients (IG). These approaches have been selected for their model independence, meaning that they can be used to explain different machine-learning models without modifying their internal structure. In the context of our study, these techniques are employed to clarify the decision-making process of the Longformer transformer

model. Specifically, they analyze the outputs generated by the Longformer to determine how different tokens of the input text contribute to the final predictions. They are also widely used in XAI tasks in the clinical domain [24,25], and they are based on different theoretical frameworks, offering an extensive explanation perspective.

SHAP [26], are values based in cooperative game theory and can be used to explain any output derived from a model based on machine learning that uses certain features to infer a prediction. They provide a way to measure how much each feature (for example a word) contributes to the model's decision on the prediction. These values take into account the prior expected prediction of the model $E[f(z)]$ (i.e. the expected prediction of the model for a subset of input features z). The difference between the prior expected prediction with respect to the real output $f(x)$ in which x refers to the full set of input features for a certain prediction, plays a core role. Indeed, the model's expected prediction is computed conditioned on a feature value and, thus, the difference between prior and conditioned expectation is seized. For each prediction, every possible feature subset is examined and determined how much each feature contributes positively or negatively to the prediction.

Deep SHAP [26] is a variant of SHAP devoted to explaining Deep Learning models. It combines two concepts, SHAP and DeepLIFT [27]. In this variant, DeepLIFT attributes the impact of each input to a reference value that is used by SHAP as the prior expected prediction $E[f(z)]$. These input features are considered independent and the model to be linear. Then, SHAP values are propagated from the output layer back to the input layer, recursively calculating SHAP values for each neuron. Therefore, DeepSHAP maintains the coherence of SHAP values while providing understandable explanations for complex deep learning models with the added value that explanations are not merely limited to the inputs but also provided in the inner neurons.

LIME is a technique that approximates any black box machine learning model [28]. To do so, it creates an interpretable model that approximates the original opaque model for a specific prediction. Then, the input is perturbed (e.g. dropping words, altering the word order, etc.), and the effect in the prediction is seized in an attempt to observe how the changes affect the prediction.

The prediction obtained from perturbed inputs is compared to the prediction obtained from the original input and by means of the proximity of the predictions, the relevance of the perturbations is weighted with respect to the original input. By analyzing the weighted importance of each feature, LIME helps to determine which words in the input were essential to get the outcome [28].

IG is a technique used in deep learning to explain the decisions made by a model [29]. It works by computing the relative contributions of each feature in the input to the model's output and measuring how much each feature in the input contributes to the decision made by the model. A key component of IG is the use of a 'baseline' input. The baseline represents an absence of features, for text features, a zero embedding vector. IG works by interpolating between this baseline and the actual input. It essentially creates a series of steps from the baseline to the actual input, and at each step, it computes the gradient of the model's output concerning the input. These gradients are then aggregated across all steps. This aggregation represents the integrated gradient and indicates how each feature contributed to the change from the baseline output to the actual output.

Explainable AI relies heavily on SHAP, LIME, and IG, but each takes a different method to determine how important input features are, therefore, each of them has strengths and weaknesses depending on the task to which it is applied. Moreover, each of these approaches lacks a predefined or fixed range for assigning weights to each feature. While SHAP and IG assign weights at a token level as depicted in Fig. 1, LIME constructs the local model based on full words. To bring together these diverse approaches, we obtain word weights in SHAP and LIME by aggregating the weights of each token as in Fig. 1. As the weight of each word is estimated independently by each method, even if it pertains to the same word, we average the weight of each word, assigning the same weight to equal words. This is the case of the word man in the example in Fig. 1, it has been assigned two different weights in different appearances in the text, therefore we average both. Finally, the weights are normalized to be between 0 and 1 and be consistent across all three techniques, and serve as the basis for comparison using the metric we propose. With this normalization, the contributions of individual words as determined by SHAP, LIME, and IG can be better understood and compared, thus providing a global view of their explanatory capabilities.

2.3. Leberage: A quantitative evaluation of the understandability degree

The task of measuring the degree of explainability may seem ambiguous. Previous work has defined the explainability of generative LLM as the ability of the output (the generated text) to answer certain general and basic questions (e.g., what, how). The degree to which those questions are adequately answered is the degree of explainability of the text [15]. Our focus is not on generating text, but on classifying it, so is fundamental to understand the extent to which the explanations in our model facilitate accurate and informed decision-making. Therefore, we define the degree of explainability as the extent to which an explanation of a prediction aids in decision-making.

To determine this level of explainability, first, our system (Longformer) outputs a set of predicted codes. Next, the previously defined model agnostic explainability techniques (SHAP, LIME, and IG) highlight the relevant words from the text that motivated the predicted code. Each code, at the same time, has a related Diagnostic Term (DT) which is a short definition of the disease that the code represents. For a certain document the predicted codes, their corresponding DTs, and the sets of highlighted words are as follows:

Clinical document fragment

Mujer sufre de hta y diabetes mellitus mellitus, no hábitos tóxicos. En último control se detectó dislipemia.... Debido a la hipertensión arterial control recurrente de hba... Se realiza ecocardiograma de esfuerzo...Implante de 3 stents farmacoactivos...

Predicted code: E119
Diagnostic Term: *Diabetes mellitus insulinodependiente sin mención de complicación*
Highlighted words: *hba, mellitus, diabetes, dislipemia, hta*

Predicted code: I10
Diagnostic Term: *Hipertensión esencial primaria*
Highlighted words: *ecocardiograma, stents, mellitus, tóxicos, hipertensión*

At first glance, we can infer that the system's prediction is correct, since among the highlighted words are those that refer to the diagnosis: *diabetes mellitus* related to code E119 and *hipertensión* related to code I10. As a result, we can also conclude that the example has a very high degree of explainability since the prediction significantly assists in decision-making throughout document coding.

The core idea behind Leberage is to measure the extent to which the highlighted words aid in decision-making quantitatively. In order to achieve this, we look at the semantic relationship between the embeddings of the highlighted words and the definition of the DTs as shown in Fig. 2. In essence, an explanation (a set of highlighted words) will be considered relevant if it has a clear and direct relationship to the associated classification code. The higher that similarity, the higher would be the understandability degree of the explanation. That is the

By token weights

man	history	of	hyperch	olester	olemia	man
0.90	-2.02	0.03	1.32	0.18	0.35	1.78

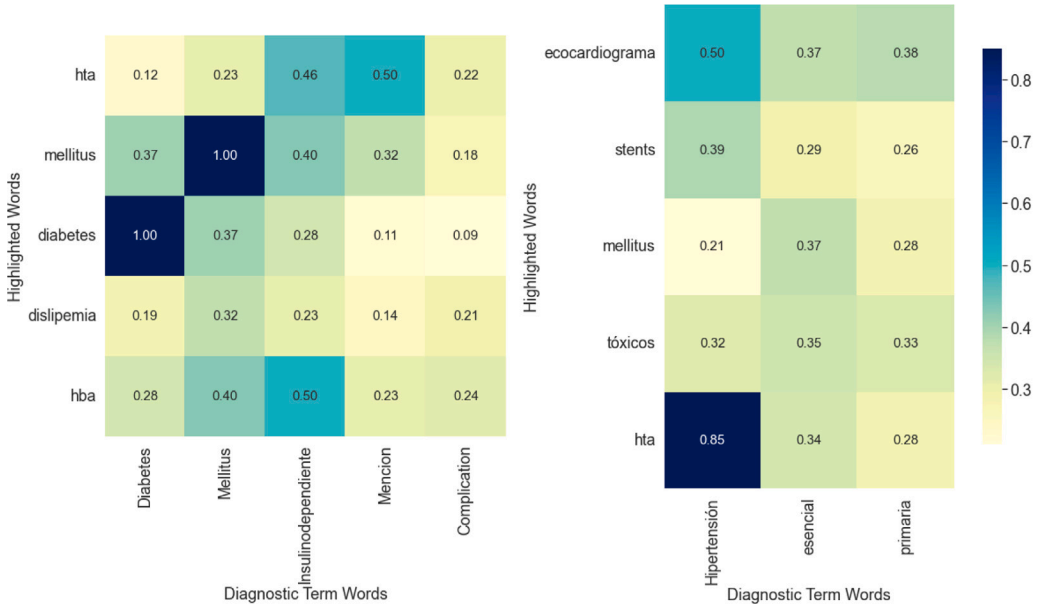
By word weights

man	history	of	hypercholesterolemia
1.34	-2.02	-0.03	1,85

By word normalized weights

man	history	of	hypercholesterolemia
0.87	0	0.51	1

Fig. 1. Example of the token weights processing and normalization. The weights of the tokens that correspond to a word are first added together, and then different weights that relate to the same word are averaged. Finally, the weights are normalized so they range between 0 and 1.



(a) Heatmap of the semantic similarity between the Highlighted words embeddings and Diagnostic Term words embeddings for ICD code E119

(b) Heatmap of the semantic similarity between the Highlighted words embeddings and Diagnostic Term words embeddings for ICD code I10

Fig. 2. Heatmap of the semantic similarity between the Highlighted words embeddings and Diagnostic Term words embeddings. Each cell represents the similarity score, ranging from 0 to 1, where 1 indicates a perfect association and 0 indicates no association.

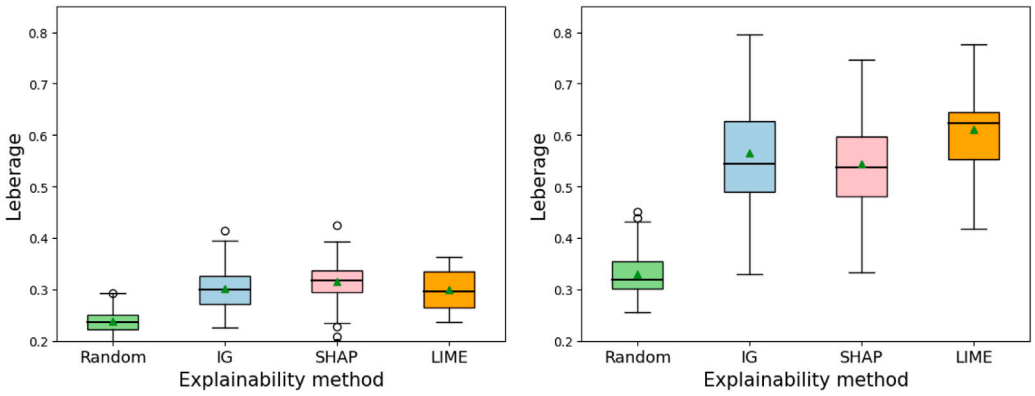
smaller the distance between the embeddings of the highlighted words and the terms of the definition, the higher the understandability degree of the explanation.

Consequently, to determine the degree of explainability ($Leberage(D)$) of a set of predicted codes C_D for a given document D, our method involves extracting the highlighted words W_j^D associated to each code j predicted for document D and DT_j^D, the DT associated to the code j. We then calculate the cosine similarity between these highlighted words and the words in the DT as in Expression (1).

$$CosSim(e(DT_{j,k}^D), e(W_{j,k}^D)) = \frac{e(DT_{j,k}^D) \cdot e(W_{j,k}^D)}{\left\|e(DT_{j,k}^D)\right\| \left\|e(W_{j,k}^D)\right\|} \quad (1)$$

Following the previous example we intuitively saw that the degree of explainability of the codes was very high, reaching 1 to 1 between some highlighted words and the definition terms in many cases. At this point, we propose two alternatives when calculating the average cosine similarity. First, we suggest that the average cosine similarity between all the highlighted terms and the words of the definition should be used to determine a code's explainability level as in Expression (2). However, in the previous example has been noted that in situations when the diagnostic terms and the highlighted words correspond one to one, the remaining correspondences introduce noise into the mean (e.g., hypertension with toxic, or diabetes with dyslipidemia). Therefore, we also propose to calculate the mean as the maximum between each definition word and the diagnostic terms as in Expression (3), that is, for each highlighted word we find the most related word in the DT and calculate the average between them. In this case, the maximum value of the Leberage metric (4) will be given in the case where all relevant words are in the definition of the DT.

$$AvgCosSim(W_j^D, DT_j^D) = \frac{1}{|D|} \sum_{i=1}^{|D|} CosSim(e(DT_{j,k}^D), e(W_{j,k}^D)) \quad (2)$$



(a) Leberage determined for 100 documents using equation 2 for computing the similarity mean

(b) Leberage determined for 100 documents using equation 3 for computing the similarity mean

Fig. 3. Boxplot Comparing Understandability Degrees Across Different Explainability Methods. This figure illustrates the distribution of understandability degrees for four explainability methods: Random, IG, SHAP, and LIME. Each box represents the interquartile range (IQR) of understandability degrees, with the line inside the box denoting the median. The means are indicated by triangles.

$$AvgCosSim(W_j^D, DT_j^D) = \frac{1}{|D|}\sum_{i=1}^{|D|} Max_{k=1}^{|DT_j|} CosSim(e(DT_{j,k}^D), e(W_{j,k}^D)) \tag{3}$$

In both, Expression (2) and Expression (3) $|D|$ represents the full vocabulary of the document D, which we restricted to the top 5 relevant words highlighted by the system.

Then, the understandability of a document, what we call Leberage, is determined by averaging the similarity scores between each predicted label as in Eq. (4).

$$Leberage(D) = \frac{1}{|C_D|}\sum_{j=1}^{|C_D|} AvgCosSim(W_j^D, DT_j^D) \tag{4}$$

To validate this metric, and decide which approach best helps us to measure the degree of explainability of the documents, we analyzed the relevant words associated with the predictions of 100 documents extracted using the three explainability techniques analyzed: Shap, LIME, and IG. We added a third technique which we called Random, which consists of randomly highlighting 5 words. This last technique serves as a basis to measure the contribution of the explainability techniques and to analyze whether the metric helps to isolate the degree of explainability.

The findings of this experiment are displayed in a boxplot in Fig. 3. On the one hand, Fig. 3(a) displays the Leberage determined for 100 documents using Eq. (2), computing the similarity mean. On the other hand, Fig. 3(b) shows the Leberage that was determined using Eq. (3) to find the mean of the similarity for the 100 documents. The boxplots reflect the variations of the Leberage of the predictions of the 100 documents.

As we believed, calculating Leberage as the mean between the distance between the relevant words and the DT words introduces a noise that does not help to isolate the Random approach from the other explainability methods. As depicted in Fig. 3(a) both, the Random approach and the explainability methods are close. However, by computing the average cosine similarity by using the maximum between each definition word and the diagnostic terms, the metric manages to isolate the three techniques with respect to the random one. Thus, we finally defined our Leberage metric as in Expression (5)

$$Leberage(D) = \frac{1}{|C_D|}\sum_{j=1}^{|C_D|}\frac{1}{|D|}\sum_{i=1}^{|D|} Max_{k=1}^{|DT_j|} CosSim(e(DT_{j,k}^D), e(W_{j,k}^D)) \tag{5}$$

3. Materials

This work is assessed in a real framework, employing a set of EHRs from the Basque Health System. This dataset comprises 26 731 discharge summaries (unstructured texts), each labeled with multiple diagnostic terms, following the ICD-10 standard [30]. The ICD-coding process was carried out by clinicians in their daily practice. We faced a multi-label classification problem: each EHR can contain multiple ICD codes that are not mutually exclusive. We accounted for 5 540 distinct ICD codes. We removed codes with fewer than ten occurrences, which reduced the initial 5540 code to 1307 (as shown in Table 1b).

Following the tendency in previous works [31,32] and with the aim of comparing our outcomes with related works, we found of interest to evaluate the performance in different ICD-10 granularity levels. An ICD-10 code can be made up of up to five characters arranged in a hierarchy. The first character is the most general one and designates the **Chapter** in which the diagnose is located within the ICD; the first three characters together encode the diagnostic term without non-essential modifiers (referred to as **Main**); the remaining characters, also known as non-essential modifiers, complete code providing details like severity and laterality of the main disease. A code with all its characters is known as a fully specified and will be referred to as **Full** throughout this paper.

The details of the input and the output, i.e. the documents and the ICD codes associated, respectively, are given in Table 1. It is evident that both the vocabulary and the average word count per document are notorious: 1575, highlighting the need to use LLM, such as longformers, to classify them. In Table 1b we provide the quantitative details of the ICD codes (label-set) divided by the aforementioned hierarchies.

Label imbalance is one of the major challenges in the multi-label classification task. To ensure that all labels are fairly represented in the training, validation, and testing sets, we used multi-label stratified sampling as part of our methodology, making sure that the codes in the training set are represented in test and development (dev) subsets. This sampling strategy led to the division of the dataset into training, validation, and test sets, with partitions of 18,380 documents for training, 4744 for testing, and 3607 for development. We also ensured that no patient occurred in both the training and test sets. We used the training set to fit the model, allowing the algorithm to learn the underlying patterns within the data. The dev subset was used during training to evaluate the model performance while the model was being tuned. Finally, the test set was used as a final evaluation to assess the generalization ability and performance of the model. The text was pre-processed as in relevant antecedents [33], lower-casing texts, and removing non-alphanumeric characters for the experimentation phase.

Table 1
Quantitative description of the EHRs amount, EHR size and vocabulary Table 1a. Size of the label-set taking different granularity levels 1b.

EHRs	26 731
Vocabulary	315 801
Words per EHR	1 575

(a) Quantitative description of the EHRs: EHRs represent the total number of documents, the cardinality of the vocabulary and the average number of words per document.

		Full	Main	Chapter
ICD count		1307	527	22
EHRs per ICD	**Mean**	73	184	3,011
	Max	5,361	5,361	11,378
	Min	3	4	85
	Stdev	240	493	3,163
ICDs per EHR	**Mean**	5.45	5.31	3.60
	Max	266	23	13
	Min	1	1	1
	Stdev	3.21	3.05	1.82

(b)Size of label-set taking different granularity of labels. "Full" stands for fully specified ICD code, "Main" for the essential modifiers and "Chapter" for the ICD chapter. EHRs per ICD indicate the repetition of codes among EHRs, while ICDs per EHR refer to the amount of codes related to an EHR.

Table 2
System performance on all services reports for all specialties combined across various ICD code granularities: Full, Main and Chapter.

	Precision	Recall	F-measure
Full	59.62	45.36	51.42
Main	66.28	52.41	53.47
Chapter	80.95	66.03	73.52

4. Experimental results

In this section, we propose two experiments to evaluate the automatic classification of ICD codes and their contribution to decision making. We seek to develop a model with a good performance in ICD classification and helpful in decision-making. Consequently, in this section, we first give insights about the classifier system built with the longformer and evaluate it using the f-measure, precision, and recall measures. Next, we analyze the contribution to decision-making: we give and comment on an example of the system's output and use Leberage to examine the explainability degree derived from the predictions.

4.1. Predicting overall ability

In this phase we build a model based on Longformer, the aim of the model is to give a clinical document to output the diagnoses of ICD-10 codes related to it. So given a clinical input our model will output a set of different ICD codes. We train our model for 100 epochs using the training data. As the computational requirements of the device grow with the window size, we limit the window size of the long former to 1512 tokens which is the closest to the mean document length of 1575.

Table 2 shows the results attained by the system in terms of Precision, Recall, and F-measure. As expected, at the higher level of the hierarchy labels (**Chapter**), the results improve as there are fewer labels but also less specific. If we focus on the **Main** code, the results decline leading to a F-score deterioration of 20%. However, when adding the complexity of predicting the **Full** code F-score slightly decreases (2%), giving the advantage of getting the completely detailed diagnosis code.

Previous work also makes this granularity distinction, when dealing with the same classification task [31]. If we compare our results we beat that earlier study in 2%, 7%, and 28% respectively in Full, Main, and Chapter granularities. Moreover, comparing our results with a previous work that carried out the same task focusing on explainability [11], our results on the Full granularity level beat those by 7%.

A critical examination of these related works indicates a significant limitation in their methodology — the restriction to 512 tokens. That means that they dropped two-thirds of the information of an average document (according to Table 1 the average document in our data contains 1575 tokens). Needless to say, restricting the text to the first 512 tokens might lead to a significant drop of information that would impact the ability of the model to compute the underlying diagnoses. Given that our health system deals with long documents, we promote the use of systems able to process large volumes of input text.

4.2. Evaluation of the understandability degree

Having compared the overall ability of the model to make accurate predictions in comparison to antecedents, we focused on the ability of the model to assist in decision-making. That is, our next goal is to focus on the ability of the system to motivate the outcomes provided (i.e. ICD codes) about the content of the input document. To this end, we carried out a qualitative experiment using the explainability techniques described in Section 2. The purpose of this experiment is to shed light on the model's decision-making procedure for a particular prediction as a clinical decision support system. Using each of the techniques, our approach is able to identify the most relevant words or tokens that correspond to each predicted label (diagnostic term expressed as an ICD). These words are then highlighted, providing a visual representation of the elements the model considers significant in its predictions.

Fig. 4 provides an example derived from our experiments translated into English to aid the reader. Is structured as follows: on top of the input text (initially given without any color) and on the bottom the ICD codes provided by the system. The codes are colored (e.g. $Z98.61$ in green) by the system and the color pallet is used by the system to highlight the segments of input text that are more relevant to predict the code (e.g. the tokens *Cardiology; losartan; catheterization; stent; aneurysm*) are also highlighted in green as in connection with the ICD $Z98.61$.

The example in Fig. 4 was obtained with the explainability mechanism underlying LIME, explained in Section 2.2, applied to the output of the Longformer. For the input text, the system provided three codes ($I10$, 98.61, and $I25.1$) and also highlighted the words in the report that resulted in relevance for each code. Some of these words might appear more than once in the document and, in an attempt to alleviate the colors and redundancies in the document, we chose to highlight only one occurrence of each relevant word. Notably, all the highlighted terms are directly related to the ICD codes. For instance, $I10$ is strongly linked with hypertension, and the terms coronary artery have a 1 to 1 relation with the DT related to code $I25.1$. In cases where a term is related to two different ICDs, as is the case of the word stent (implant in the coronary artery), the system outputs the word twice, in each case colored according to the color of each ICD. In this case stent is related to code $Z98.61$ as it is implanted during a coronary angioplasty procedure, and to code $I25.1$ (atherosclerotic heart disease) as the stent is implanted due to this disease ($I25.1$), which the patient suffers from.

By selecting the 5 words highlighted by the system for each code in each test document, we are able to determine the Leberage of each explored explainability technique. This provides us with quantitative knowledge about the level of contribution of each technique to the decision support of our system. Following the methodology carried out during the development of the metric we also provide the results of Leberage when using the Random approach that consisted of randomly highlighting 5 words. In Table 3 we show the Leberage results of the test documents for each method.

All three of the explainability strategies perform better than the random approach in terms of the proposed metric. IG and SHAP have very similar results and LIME is the one with the highest understandability degree.

SEX: F
SERVICE: Cardiology
ALLERGIES: Allergic to ASA, skin reaction to ramipril.
PERSONAL MEDICAL HISTORY: hypertension, dyslipidemia.
Medications: Nebivolol 5 mg, losartan 50 mg, fluvastatin 80 mg, omeprazole 20 mg, NTG sublingual as needed, tramadol/paracetamol 375/325 mg.
HISTORY OF PRESENT ILLNESS: The patient has been experiencing retrosternal discomfort for 3 months, exacerbated by nervousness and increased dyspnea. She is admitted for scheduled cardiac catheterization due to clinically negative but electrically positive ischemia stress test.
PHYSICAL EXAMINATION: Blood Pressure: 145/81 mmHg; Heart Rate: 65 bpm. The patient is afebrile, conscious, well-oriented, and shows normal breathing at rest. Heart Sounds: Rhythmic with no murmurs or additional sounds. Abdomen and Lower Extremities: Unremarkable.
HOSPITAL COURSE: Performed via the left radial artery. A significant lesion in the anterior descending artery was found and treated with the placement of a drug-eluting stent stent. A small coronary aneurysm was also noted.
DISCHARGE DIAGNOSIS: Coronary artery disease with a significant lesion in the anterior descending artery.
DISCHARGE INSTRUCTIONS: Continued with Adiro 10 and Vernies. In case of pain, add Plavix (clopidogrel) 75 mg for at least one year. Continue prescribed medications as instructed. Monitor for any chest pain or other symptoms and seek medical attention as necessary.

ICD codes	Diagnostic term
I10	Essential (primary) hypertension
Z98.61	Coronary angioplasty status
I25.1	Atherosclerotic heart disease of native coronary artery

Fig. 4. Example of a clinical note translated into English for better understanding. The system assigns ICD codes automatically (bottom). Besides, the words from the input text connected to each ICD-10 code provided are highlighted by the relevance weight received according to LIME (top).

Table 3
Leberage from testing documents for each of the explainability approaches.

	Random	IG	Shap	Lime
Leberage	32.44	56.45	54.45	61.00

5. Discussion

Our work marks a significant step forward in the explainable automatic classification of EHRs in Spanish, a language that, despite its wide usage, often receives less attention in the realm of automated clinical documentation.

The use of methods able to deal with long texts resulted in a significant improvement in the clinical documentation task. This implies that meaningful information is along the document and cannot be captured limiting the method to the first 250 or 500 tokens as in the antecedents. The F-score in our work outperformed those related works [11,31] in up to 7% and 24% respectively dealing with EHRs classification according to ICD.

With this first set of experiments, we learned that in computer-aided clinical documentation, all the text contains valuable information that should not be disregarded. By taking into account whole documents our system overcomes those that do not take into account complete documents.

Having set a system devoted to multi-label document classification, as in clinical documentation, we assessed its value to aid experts in clinical documentation ICD-10. To this end, we focused on a quantitative means of assessing the explainability, indeed, diverse techniques were compared in our study. The transparency provided by these techniques is essential in medical applications where understanding the rationale behind a diagnosis or decision is as crucial as the decision itself. By emphasizing the most essential phrases and making a black box model explicable, the IG, SHAP, and LIME approaches have proven their capacity to accomplish this aim.

Unlike antecedents [13], which merely focus on extracting explainable predictions and lack a standardized metric to measure and compare the quality of explanations across different models, in our work we seek to evaluate it. With that aim, we proposed Leberage, a metric to assess explainability.

Leberage has proven effective in measuring the degree of explainability when classifying a document. As shown in Table 3, Leberage can quantify the contribution of explainability techniques in contrast to a random methodology. In this case, the Random approach obtains a 32.44% of Leberage while explainability methods get a 54.45% or higher, proving that the metric is valid in quantifying the decision support level of the explanations. To our knowledge, there are no previous works that quantitatively compare different explainability techniques.

An interesting finding about explainability is the direct proportionality between the degree of explainability and the performance in supervised classification. This is particularly relevant in the classification of documents in Spanish, where better classification results also correlate with a higher degree of explainability. Regarding the comparative analysis of explainability strategies, the outcomes for SHAP and IG are very similar. However, there is an increase in explainability when LIME is used.

Our methodology demonstrates robust potential for adaptation across various languages. These employed techniques are based on learning mechanisms that do not inherently prefer one language over another. With appropriate fine-tuning using language-specific corpora, our approach could be effectively applied to clinical text mining in languages other than Spanish.

However, despite our effort to construct a generic metric to measure the quality of the explainability of predictions, it is only applicable in labeling tasks where labels have a direct definition associated with them, as in the case of CIE codes, or in named entity recognition tasks. In other tasks where labels are not associated with definitions, our metric is not applicable, at least without a prior task in which labels are defined. Furthermore, it is far from accurate to regard the explanations offered by the system as the explanations that motivate the predictions are merely words underlined in the text.

6. Conclusion

In this article, we delve into the computer-aided classification of EHRs in accordance with the ICD widely employed in clinical documentation, with a focus on Spanish. EHRs are long documents and, while traditional NLP approaches such as BERT [16] and RoBERTa [19] found difficulties in analyzing entire documents, experimental results show that crucial information is spread along the entire document. Indeed, LLMs resulted in excellent alternatives.

Competitive models, such as LLMs, are often criticized for their opacity. To cope with this issue, we explored different model-agnostic explainability techniques: IG, SHAP, and LIME. Furthermore, we found a lack of measures for categorizing the level of explainability. In this work, we proposed Leberage a novel metric to assess and compare the effectiveness of these techniques. One of the strengths of our work rests on the thorough experimental framework disclosing alternative perspectives of this versatile approach, not only performance but also with the added value provided by explainability, core in clinical decision support initiatives.

In terms of future work, plenty of room remains for additional progress on explainable AI. To improve the use of the system to a wider public, we propose to improve the depth of the explanations provided, by using generative LLMs. This integration can potentially enrich the explanations by adding context and narrative, making the AI's decision-making process more transparent and understandable. Generative LLMs, with their capacity to produce coherent and contextually relevant text, could offer a more detailed backdrop to each explanation, bridging the gap between the technical output of the system and the intuitive understanding of its users. This strategy would not only expand the applicability of our explainability metric but also move towards explaining AI decisions by improving explanation quality to a level more akin to human reasoning.

Our study opens up new possibilities in the realm of medical document classification and paves the way for AI models in healthcare that are more precise, transparent, and trustworthy to users.

CRediT authorship contribution statement

Nuria Lebeña: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Alicia Pérez:** Writing – review & editing, Validation, Resources, Investigation, Funding acquisition. **Arantza Casillas:** Writing – review & editing, Validation, Supervision, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation (EDHIA PID2022-136522OB-C22); by the Basque Government (IXA IT-1570-22 and Predoctoral Grant PRE-2022-1-0069).

Besides, this work was elaborated within the framework of LOTU (TED2021-130398B-C22) funded by MCIN/AEI/10.13039/501100011033, European Commission (FEDER), and by the European Union NextGenerationEU/PRTR.

Appendix. Software and hardware requirements

A.1. Software dependencies

For the implementation of the Longformer model and the application of the explainability techniques SHAP, LIME, and IG, we employed Python 3.9.7 programming language next to the following key libraries:

- Transformers version 4.27.3 - Hugging Face's library, including the Clinical Longformer model.
- SHAP version 0.42.1 a library with DeepShap implementation.
- LIME version 0.2.0.1 a library with LIME implementation
- Captum 0.6.0 that implements IG explainability technique.

Detailed requirements and additional libraries are listed in the requirements.txt file available in our code repository.

A.2. Hardware requirements

The model training and experiments were conducted using an NVIDIA Tesla V100 GPU (32 GB) and 138 GB of CPU.

A.3. Code availability

In an attempt to aid reproducibility, the complete source code is available online at: https://github.com/nuriale207/Quantifying-decision-support-level.

References

[1] A. Rosenfeld, Better metrics for evaluating explainable artificial intelligence, in: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, 2021, pp. 45–50.

[2] European Union, European general data protection regulation, 2018, URL https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en.

[3] D. Mammonas, Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world, European Council Council of the European Union, 2023.

[4] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[5] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (xai): Toward medical, xai. IEEE Trans. Neural Netw. Learn. Syst. (2020).

[6] H. Bouamor, J. Pino, K. Bali, Proceedings of the 2023 conference on empirical methods in natural language processing, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[7] N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T.K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, URL https://aclanthology.org/2022.coling-1.0.

[8] M. Carpuat, M.-C. de Marneffe, I.V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, URL https://aclanthology.org/2022.naacl-main.0.

[9] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—Explainable artificial intelligence, Sci. Robotics 4 (37) (2019) eaay7120.

[10] R. Agerri, I. Alonso, A. Atutxa, A. Berrondo, A. Estarrona, I. Garcia-Ferrero, I. Goenaga, K. Gojenola, M. Oronoz, I. Perez-Tejedor, et al., HiTZ@ antidote: Argumentation-driven explainable artificial intelligence for digital medicine, 2023, arXiv preprint arXiv:2306.06029.

[11] O. Trigueros, A. Blanco, N. Lebeña, A. Casillas, A. Pérez, Explainable ICD multi-label classification of EHRs in spanish with convolutional attention, Int. J. Med. Inf. 157 (2022) 104615.

[12] G. López-García, J.M. Jerez, N. Ribelles, E. Alba, F.J. Veredas, Explainable clinical coding with in-domain adapted transformers, J. Biomed. Inform. 139 (2023) 104323.

[13] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, 2020, arXiv preprint arXiv:2010.00711.

[14] F. Sovrano, F. Vitali, Generating user-centred explanations via illocutionary question answering: From philosophy to interfaces, ACM Trans. Interact. Intell. Syst. 12 (4) (2022) 1–32.

[15] F. Sovrano, F. Vitali, An objective metric for explainable AI: how and why to estimate the degree of explainability, Knowl.-Based Syst. 278 (2023) 110866.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

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

[18] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthc. (HEALTH) 3 (1) (2021) 1–23.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692, 2019, arXiv:1907.11692.

[20] C.-W. Huang, S.-C. Tsai, Y.-N. Chen, PLM-icd: Automatic ICD coding with pretrained language models, 2022, arXiv:2207.05289.

[21] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The long-document transformer, 2020, arXiv preprint arXiv:2004.05150.

[22] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, A. Ahmed, Big bird: Transformers for longer sequences, 2021, arXiv:2007.14062.

[23] Y. Li, R.M. Wehbe, F.S. Ahmad, H. Wang, Y. Luo, Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences, 2022, arXiv preprint arXiv:2201.11838.

[24] R.O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, A.A. Mäkitie, Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP, Sci. Rep. 13 (1) (2023) 8984.

[25] M. Laatifi, S. Douzi, H. Ezzine, C.E. Asry, A. Naya, A. Bouklouze, Y. Zaid, M. Naciri, Explanatory predictive model for COVID-19 severity risk employing machine learning, shapley addition, and LIME, Sci. Rep. 13 (1) (2023) 5481.

[26] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017, arXiv:1705.07874.

[27] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.

[28] M.T. Ribeiro, S. Singh, C. Guestrin, " Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[29] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.

[30] World Health Organization, et al., ICD-10: international statistical classification of diseases and related health problems: tenth revision, World Health Organization, 2004.

[31] A. Blanco, A. Pérez, A. Casillas, Exploiting ICD hierarchy for classification of EHRs in spanish through multi-task transformers, IEEE J. Biomed. Health Informat. 26 (3) (2021) 1374–1383.

[32] F. Duarte, B. Martins, C.S. Pinto, M.J. Silva, Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text, J. Biomed. Informat. 80 (2018) 64–77.

[33] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1101–1111, URL https://aclanthology.org/N18-1100.