IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 30, 2020, accepted October 3, 2020, date of publication October 7, 2020, date of current version October 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029429

Extreme Multi-Label ICD Classification: Sensitivity to Hospital Service and Time

ALBERTO BLANCO, ALICIA PÉREZ, AND ARANTZA CASILLAS
HiTZ Center-Ixa, University of the Basque Country (UPV/EHU), 20080 Donostia, Spain

Corresponding author: Alberto Blanco (alberto.blanco@ehu.eus)

This work was supported in part by the Spanish Ministry of Science and Technology under Grant PAT-MED PID2019-106942RB-C31, in part by the Basque Government under Grant Elkartek KK-2019/00045 and Grant IXA IT-1343-19, and in part by the Predoctoral under Grant PRE-2019-1-0158.

ABSTRACT This work deals with clinical text mining for automatic classification of Electronic Health Records (EHRs) with respect to the International Classification of Diseases (ICD). ICD is the international standard for the identification of diseases and health conditions in EHRs and the foundation for reporting health statistics. Machine learning-based techniques have proven robust to infer classification models from EHRs. Since each EHR tends to involve multiple diseases, multi-label classification is required. The concern in this work is the versatility of the models inferred and their ability to generalise in two ways: as time goes ahead and across hospital services or health specialties. Indeed, in this work, we show the capabilities of a Bidirectional Recurrent Neural Network (RNN) with GRU units and ELMo embeddings on two corpora (a corpus comprising a set of EHRs within the Basque Health System, namely Osakidetza, and the well-known MIMIC-III corpus). To delve into and assess the versatility of the models, we focus on their resilience across hospital admissions taken over two different years and also across six distinct hospital services. In addition, we paid attention to the classification performance to estimate ICD codes of different granularity (e.g. with or without essential modifiers). Our best results are 39.55% and 47.28% F-Score for the Osakidetza and MIMIC-III datasets respectively, with the original main label-sets. Regarding the models evaluated per specialty, the most remarkable results are 57.00% and 72.74% F-Score, in the Cardiology and Nephrology medical services respectively.

INDEX TERMS Extreme multi-label classification, electronic health records, international classification of diseases, classification across-time, classification across hospital-services.

I. INTRODUCTION

Natural Language Processing (NLP) is gaining relevance within the clinical documentation services to cope with extensive information conveyed by Electronic Health Records (EHRs). Healthcare data is getting increasingly larger and complex to process [1], but evidence shows its usefulness in such different sectors as Adverse Drug Reaction extraction [2], [3] and identification of complex symptoms, assessed in several cohorts of patients in hemodialysis [4], as well as relevant symptoms in patients with schizophrenia [5], and breast cancer [6], and the creation of phenotypes to characterise patients [7], [8].

Facilitating access to information is crucial for accurate clinical documentation. International Classification of Diseases (ICD) [9] is a standard used to classify diagnosis and procedures within EHRs. These codes are used to quantify vital statistics, for surveillance, to seek cohorts of patients with similar diagnoses in downstream studies and also as a standardised information exchange method between hospitals. The thorough and accurate coding of EHRs affects critical clinical information extraction and also other industries such as insurance billing [10]–[12].

Nowadays EHRs are manually encoded by healthcare professionals specially trained to cope with complex ICD nuances. Note that ICD-10 is arranged in 24 chapters or branches of medicine and comprises nearly 70 thousand codes for diseases (ICD-10-CM) and as many for procedures (ICD-10-PCS). The ICD versions evolve rapidly e.g. from the 9th version to the 10th the number of codes increased five-fold and, what is more, the code structure changed from a maximum of 5 characters to 7; the alpha-numeric coding

The associate editor coordinating the review of this manuscript and approving it for publication was Nilanjan Dey.

structure was also modified. The chapter is encoded in this structure and would reflect a coarse-grained classification of the EHR into branches referred to as *ICD chapter*. Besides, in this coding we typically find a sub-set of characters (often 3 of them) that are referred to as the *main ICD class*, and the remaining characters comprise what is referred to as non-essential modifiers (e.g. laterality and severity). Altogether these are referred to as '*fully-specified ICD class*'. Additionally, manual coding is time-consuming. As an example, let us focus on a well-known collection of EHRs, i.e. MIMIC-III [13], a set of nearly 55, 000 EHRs made available by the MIT Lab for Computational Physiology from the Beth Israel Deaconess Medical Centre. Each EHR has, on average, 1, 947 words that are carefully read by professionals to annotate the EHR with all the ICDs found. On average, each of these EHRs was assigned 11.5 different ICDs (also referred to as *label cardinality*) and, altogether, the set shows 6, 527 different ICDs (also known as the size of the label-set). Given that manually encoding is time-consuming and requires specialised professionals, several health systems took the decision to restrict coding to just the main cause of admission, leading to a relevant loss of valuable information. It is well-known that EHRs show much variety in terms of (often non-standard) linguistic forms, semantics and syntax, and a linguistic style prone to the economy of language [14], [15].

There is evidence that NLP can aid human coders in a decision support system with the human in the loop [16]. The task of automatically assigning multiple codes to a given document is referred to as *multi-label classification*. Automatic multi-label classification of EHRs is, however, far from the scope of current machine learning approaches given that high accuracy is a must. There are numerous **challenges** inherent in this task, not only from the inference perspective but also due to the assessment of the quality of the predicted label-set. The size of the output label-set (e.g. 6, 527 in MIMIC-III) is extremely high for the input evidence (e.g. 55, 000 EHRs in MIMIC-III). While automatic inference can find patterns, pattern repetition in selecting an unknown number (11.6 labels on average in MIMIC-III, with 9.0 as the median value and a standard deviation of 6.3) of ICD codes from a set of 6, 527 lead to 51,980 unique label-sets. Note that there are few pieces of evidence for the output on top of the variability in the input. This highly-variable classification task with thousands of possible labels is referred to as *extreme multi-label classification (XMC)* [17]. Quantitative assessment of multi-label models is still one of the stumbling blocks in the inference since model optimisation (fine-tuning) rests on evaluation approaches and this is not a trivial issue. (We shall discuss issues that might emerge in the assessment in Section III-B).

In this work, we assess a multi-label classification approach in the task of EHR multi-class classification in documents written in Spanish. The set of EHRs is comparable in many aspects to those in MIMIC-III (as we will present in Section IV-A). Often, [18], [19] related work bound the size of the label-set in such a way that the data-set ensures a minimum number of documents per label (in an attempt to ensure minimum repetition per pattern). In this work we assessed the system exhaustively.

First, to assess the resiliency of the learning approach proposed, i.e. Bidirectional Recurrent Neural Network [20] with GRU units [21] and ELMo embeddings [22] (BiGru ELMo), the label-set was not restricted and all the labels available in the data (i.e. 2, 554 labels) were considered and next, the system was assessed with the top 110 and top 16. Second, we assessed the robustness of the system across time. Needless to say, as time goes by, personnel in a hospital might have changed their EHR writing or encoding style: in these circumstances, the system should be adapted. The question also arises here of how often should we re-train the model and also whether previous EHRs are either beneficial or harmful for current EHR coding. The motivation is to assess whether a predictive model inferred with data from a given year can help to predict EHRs from future years. Also, we want to evaluate if non-overlapping data from two consecutive years help to predict EHRs from the later year of the two. In other words, we wish to assess/evaluate if training with data from successive years can generate synergies or, whether the best option is to re-train the system frequently to keep it updated. Third, we did not only assess the system across time but also across use-cases in different hospital services. One of our concerns had to do with data scarcity. We wondered if a general system trained with EHRs from discharge reports from several hospital services (e.g. cardiology, psychiatry etc.) is able to cooperate and make the system capture accurate syntax and semantic nuances, or if, by contrast, accurately encoding EHRs from a given service was bound to train the system with EHRs from that service, while EHRs from other services could lead to lexical explosion and maybe distort the outcome. Through these experiments we tried to shed light on the following three **research questions**: 1) the ability of BiGru ELMo to cope with infrequent and frequent labels, 2) the robustness of the model across time and, 3) across hospital services. Briefly, the novelty of the work resides in the aforementioned research questions. To this end, we apply a state-of-the-art multi-label classification model to a Spanish EHR dataset that can be segmented by year and medical service. The segmentation of data allows checking the robustness of the models against lexical variation due to variations among medical specialties and across-time. We also assessed the model in both coarse-grained (the ability of the model to situate the EHR within a chapter of the ICD) and fine-grained code assessment (referring to the granularity mentioned on page 183534). The finer the granularity, the bigger the size of the label-set.

II. RELATED WORK

Since 2000 CLEF has organised different laboratories in the field of multilingual access evaluation, in particular since 2016 in the automatic assign of ICD codes. In 2016 [23] the task consisted of extracting causes of death from French narratives as coded in the International Classification of Diseases

ICD-10. In 2017 [24] the task goal was to automatically assign ICD-10 codes to English and French death certificates. In 2018 [18] the task focused on French, Hungarian and Italian texts. In 2019 [25] the task explored the automatic assignment of ICD-10 codes to non-technical summaries of animal experimentation in German. The tasks carried out from 2016 - 2018 are focused on the codification of lines (diagnosis) instead of on the codification of whole EHRs. On average, each diagnosis has between 2.06 to 12.38 tokens and between 1.20 to 1.37 codes.

Approaches based on regular expressions or transducers either manually created [26], [27] or automatically inferred from data [28] were used in previous works when it comes to mapping Diagnostic Terms (DT) expressed in natural language into standard DTs within the ICD and, hence, assigning the corresponding ICD. The difference between translating non-standard expressions to a standard form and assigning ICDs to a given full EHR is substantial. The entire EHR in our task has on average ∼1,000 words per document, while the input non-standard DT tends to have around 5 words. In the EHR, the language is likewise, non-standard, although, the DTs are not explicitly informed. Moreover, implicit evidence, such as analytics and current treatment, might yield an ICD. Besides, while the correspondence between the non-standard DT and the ICD codes is 1 to 1, in the EHRs, 1 short phrase could trigger n ICD codes being the correspondences m to n.

The so-called *binary relevance approach* [29] is a simple approach to tackle multi-label classification that comprises as many binary classifiers as classes involved. Each classifier would determine the absence or presence of one class. The drawback of this simplistic approach is that the classes are assumed to be independent, hence, dependencies among ICD codes would be disregarded. Nevertheless, some diagnostics are incompatible (and should not be predicted together), while others tend to co-occur. Accordingly, we opted for a model that considers the label-dependencies.

Rios and Kavuluru [30] explored the use of Convolutional Neural Networks (CNNs) for automatic ICD coding. They stated that when many codes occur infrequently, the Deep Learning (DL) models' performance is inhibited. They proposed a neural transfer learning strategy, supplementing EHR data with PubMed indexed biomedical research abstracts. For the source task, they trained a CNN to predict 1.6M Medical Subject Headings (MeSH) using PubMed indexed biomedical abstracts, whereas, for the target task, they trained a CNN on 71,463 EHRs to predict ICD diagnosis codes. Our approach is also based on the idea of transfer learning, as the ELMo embeddings are derived from a bidirectional LSTM trained with a coupled language model (LM) objective on a large text corpus, including, but not restricted to biomedical texts (i.e. pharmaceutical or medical articles from Wikipedia). They got, respectively, a micro- and macro-F-Score of 56.8 and 28.6, considering $1,231$ truncated ICD-9 labels with $5,303$ average words per instance from $71,463$ instances.

Gangavarapu *et al.* [19] employed the MIMIC data-set. It is usual to exploit the discharge summaries (i.e. the clinical report prepared by the physician after a hospital stay), but in this case, they leverage the nursing notes. One drawback is that the nursing notes present excessive redundant information, due to the anomalous and evolutive data of the patient. This issue was addressed with a fuzzy similarity-based data cleansing approach; The authors applied vector space and topic modelling to extract the rich patient-specific information available in unstructured clinical data. This can be crucial in countries where structured EHR adoption is not widespread [31], [32]. The authors worked with $223,556$ nursing notes of 357.8 words on average, predicting 19 ICD-9 code group labels, and achieved a maximum F1-score (weighted-) of 69.81 across all the tested models.

Most ICD codes appear only in a few samples, that is, the ICD distribution presents a long-tail, which is, precisely, a feature of extreme multi-label classification. Babbar and Schölkopf [33] posed the tail-label detection task in XMC as a robust learning problem, taking into account the worst-case perturbation scenarios. This viewpoint is motivated by a key observation: from the training set to test set, there is a significant change in the distribution of the features of instances belonging to the tail-labels. This is a typical case when classifying EHRs with ICD codes, especially, across time or clinical services [34] since physicians from different medical specialties refer to the same medical concepts in diverse forms. The converse also happens: the same string is employed to refer to different concepts (this happens often with abbreviations) across clinical services.

In an attempt to tackle the scalability issue of state-of-the-art Deep Learning-based methods to extremely large label-sets [35], a hierarchical structure based on Probability Label Trees generated with balanced k-means recursively, and multi-label attention was proposed by You *et al.* [36]. Similarly, Gargiulo *et al.* [37] presented a methodology named Hierarchical Label-Set Expansion (HLSE), used to regularise the data labels, based on the hierarchical structure of the MeSH label-set. Data scarceness and large lexical variability and vocabularies are major concerns in the ICD multi-label classification tasks. Deng *et al.* [38] presented a processing pipeline built upon CNNs and an autoencoder with logistic regression. They applied the combination of embeddings from different sources and proved the positive influence of semantic enrichment to counter the aforementioned strains. The contextual ELMo embeddings can overcome these limitations of the standard embeddings [39]. Cheng *et al.* [40] recognised that some complex semantic problems in the real world require the association of more objects with related labels but also that as data complexity increases, the class imbalance issues become increasingly prominent. A well-known strategy to deal with imbalance between classes is to use label correlations, but their work proposed an alternative approach. They first introduced the classification margin and expanded the original label-space among labels, taking into account the label-density. The BiGru model can handle all

the labels at once thanks to the final dense layer with the Sigmoid activation function, and thus, is able to capture and model the label dependencies. Chalkidis *et al.* [41] compared various neural methods on the EURLEX57K dataset (with 4,271 labels) and concluded that the best results rely on the Recurrent Neural Network with GRU units, but also that it is the most computationally expensive method. Chang *et al.* [42] leveraged the pre-trained language representation model BERT, extending it to the XMC problem to deal with the difficulty of capturing dependencies or correlations among labels and the tractability to scale to the extreme label setting because of the Softmax bottleneck scaling linearly with the output space. Their so-called X-BERT utilises both the label and input text to build label representations. This induces semantic label clusters to better model label dependencies, which can also be applied to the ICD classification task, as all the labels have an associated text, the standard-term description.

The **motivation** and novelty of this work resides in exploring the behaviour of the classifiers under novel circumstances through the characteristics of our task. It conveys a multi-label text classification problem with great lexical variability, especially in the set of EHRs in Spanish, as the dataset can be segmented by year and medical service. To that end, and following the insights of the related works, we developed a model based on Recurrent Neural Networks with Bidirectional Recurrent layers and GRU units leveraging the ELMo contextual embeddings. These models were proven robust and capable of learning from scarce samples, as is the case of ICD coding. The gaps found in previous works, and which we do cover in this article, are related to the capability of such state-of-the-art models to keep a strong performance facing lexical variation inherent to the biomedical domain, but extended to variations over time (i.e. attempting to make predictions across years) and different sub-domains (i.e. across various clinical services). This way, we assess the sensitivity of these models to different factors (time and health services) and, thus, pay attention to their usability.

III. METHODS

A. MULTI-LABEL CLASSIFICATION APPROACH

Having explored previous works, for our task we opted for a Recurrent Neural Network with a Bidirectional layer with GRU units (referred to as BiGru from now onward). The architecture of the model is shown in Figure 1, and formally, is explained in (1), with the bidirectional layer processing the sequences of text in both directions, forward and reverse. Accordingly, it generates forward ($\overrightarrow{h^{(t)}}$) and backward ($\overleftarrow{h^{(t)}}$) hidden states, which are later combined into $h^{(t)}$. Here t is the time-step and T the total number of time-steps ($1 \leq t \leq T$). The parameters to be determined in the inference stage given the EHRs are, on the one hand, the weight matrices, W and V, and, on the other hand, the bias term b. A non-linear activation function, the Sigmoid (σ), is chosen to compute the current hidden-states taking, as input, the weighted sum of previous

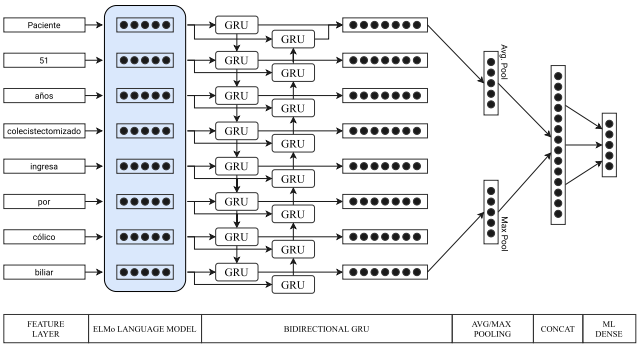

FIGURE 1. Architecture of the BiGru ELMo model: a Bidirectional Recurrent Layer with GRU units powered by ELMo embeddings with Pooling and a final multi-label dense layer.

hidden-states ($h^{(t-1)}$) and current input ($x^{(t)}$) with the weights given by W and V. Finally, both hidden states are combined as a mere concatenation of each matrix.

$$\begin{aligned}\overrightarrow{h^{(t)}} &= \sigma(\overrightarrow{W}x^{(t)} + \overrightarrow{V}\,\overrightarrow{h}^{(t-1)} + \overrightarrow{b}) \\ \overleftarrow{h^{(t)}} &= \sigma(\overleftarrow{W}x^{(t)} + \overleftarrow{V}\,\overleftarrow{h}^{(t-1)} + \overleftarrow{b}) \\ h^{(t)} &= [\overrightarrow{h}^{(t)} \| \overleftarrow{h}^{(t)}]\end{aligned} \tag{1}$$

The output of the bidirectional layer, $h^{(t)}$, is fed to several Pooling [43] layers. To be precise, in our work, an average and a max-pooling layer were used, as in (2).

$$h_{max} = \max_{1 \leq t \leq T} h^{(t)} \text{ and } h_{avg} = \frac{\sum_{t=1}^{T} h^{(t)}}{\left\| h^{(t)} \right\|} \tag{2}$$

The output of both pooling layers (h_{max} and h_{avg}) is again concatenated into $h = [h_{max} \| h_{avg}] \in \mathbb{R}^{2T}$ and passed into a final dense layer, which is responsible of computing the probability estimation of the labels i.e. ICD codes.

The strength of RNN with GRU unit and ELMo is clear in this extreme multi-label scenario as described in what follows. This BiGru model can cope with the multi-label problem since the final dense layer is able to capture and model the label dependencies in contrast with the binary-relevance approach. In fact, by virtue of the Sigmoid activation function, it produces a probability estimation for each label that is not mutually exclusive [17]. Thus, this BiGru model is suitable to cope with the multi-label classification of EHRs through ICDs and address dependencies between diagnoses. Moreover, bearing in mind that EHRs are long documents, with even thousands of words per document, the ability to capture long-term dependencies in the text, as an RNN with GRU does, becomes imperative. Furthermore, in an attempt to cope with lexical variability in the input EHRs, we turned to ELMo embeddings. In general, the word embedding is a technique to transform a word, and therefore a document, into a dense vector and, by extension, into a matrix. The text from a clinical record is fed into an embedding layer, and the output is a matrix representing the document, with one row-vector per word. Each word is referred to as time-step in the formulation of the model. ELMo embeddings [22]

capture contextual information and, according to the authors, these word representations model i) complex characteristics of word use (i.e. syntax and semantics) and ii) how these uses vary across linguistic contexts (i.e. to model polysemy). As a result, in our task, ELMo embeddings can help towards i) the detection of some essential nuances of medical records such as the negation of symptoms and, ii) robustness to variations of the author (i.e. each physician expresses in a particular way) and sub-domain (i.e. a reference to the same medical concepts across several clinical services). We also need embeddings, as they can cope with different linguistic contexts, to deal with the various clinical services, time-frames, and their lexical subtleties. These are the principal reasons why we opted for BiGru ELMo as an appropriate choice to deal with high lexical variability within EHRs and multi-label classification with respect to the ICD.

B. MULTI-LABEL ASSESSMENT CRITERIA

We have resorted to well-known metrics such as Precision, Recall and F-Score, but to adapt them correctly to the multi-label scenario it is necessary to compute averages since these metrics are well suited to mono-label tasks. There are several common averages, i.e. micro-, macro- and weighted-average-, each of them penalising certain aspects more severely than other [44], [45]. For example, a macro-average will compute the metric independently for each class (i.e. a confusion matrix for each ICD) and then compute each metric and take the average of the metric (hence it will treat all classes equally). By contrast, a micro-average will aggregate the contributions of all classes in a single confusion matrix and then compute the average (hence, the performance over more populated classes dominate). Thus, in a situation with imbalanced classes is easier to get higher metric values with the micro-average provided that the dominant class is accurately predicted (with the micro-avg result being almost insensitive to the hits or fails over less populated classes). The weighted-average is a balanced solution which takes into account the support (i.e. frequency) of each label to weight their contribution to the final metric value. For that reason, in this work, we give the weighted-average version of the Precision, Recall and F-Score metrics. Nevertheless, all these approximations and variations come with benefits and disadvantages: there is not a general optimum approach, and the best-suited evaluation will depend on the task and objectives of the work. Note, however, that averages are taken per code and not, strictly, per document. A challenge in ML classification is to on decide the number of codes to assign to each document, as this is variable (on average, EHRs within MIMIC receive 11.6 codes but the deviation is 6.3, quite high).

IV. EXPERIMENTAL FRAMEWORK

A. CORPORA

Here we describe the corpora and provide two perspectives: the input (text from EHRs) and the output (ICDs). The data we had available for this work comprised two separate but analogous data-sets with EHRs, written in Spanish, from the Basque public health system (Osakidetza). Specifically, both sets, denoted as Osa1 and Osa2, comprise discharge summaries from hospitals. Table 1 provides quantitative details of each data-set and the union of both sets (denoted as Osa1+2) leading to a total of 27, 040 unique EHRs.

TABLE 1. Quantitative description of the input (EHRs).

	data-set			
	MIMIC	Osa1	Osa2	Osa1+2
Samples	55,172	13,574	13,466	27,040
Vocab. size	137,207	236,109	255,394	379,477
Words/doc	$1,399 \pm 721$	843 ± 401	885 ± 428	864 ± 415

Regarding the input (EHRs), both Osa1 and Osa2 data-sets are significantly smaller than MIMIC; indeed, there are nearly twice as many samples in MIMIC as in Osa1+2. However, the size of the vocabulary (the number of unique words) of Osa1+2 (379, 477) is approximately three times larger than the vocabulary of MIMIC (137, 207). This means that the lexical variability is notably higher in the Spanish set, even though there are more documents and, besides, the length of the documents is much higher in MIMIC. To enable the drawing of conclusions from the following experiments, we must acknowledge the distributions of the features of the different sets from the experimental setup. To that end, as the classifiers are only fed with the text from the clinical records, we explore the vocabulary and Out-of-Vocabulary (OOV) words. The number of unique words in Osa1 is 89,840, the number of unique words in Osa2 is 94,764, and the number of OOV words in Osa2 with respect to Osa1 is 42,249, which is the 44.6% of the vocabulary. The vocabulary we are dealing with is large, but what is more, we can observe that the amount of disjoint vocabulary between sets is also quite high, leading to the demand for robust classifiers.

Regarding the output, i.e. the **label-sets**, the aim is to predict the set of ICDs in their fully-specified form. However, failing non-essential modifiers might be considered not as bad as failing the chapter of the ICD. Accordingly, we assessed the performance taking into account three different granularities of the ICD codes, from fine- to coarse-grained:

- *"Full-code"* level preserves the original code e.g. "`I13.10`" (this stands for *Hypertensive heart and chronic kidney disease without heart failure, with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease*);
- *"Main"* level drops non-essential modifiers, keeping just the first three characters e.g. "`I13`" (*Hypertensive heart and chronic kidney disease*)
- *"Chapter"* level keeps only the first character, that is, the chapter of the ICD e.g. "`I`" (*Diseases of the circulatory system*).

Trying to predict rare codes is really challenging for inferred systems and often previous works pruned the

TABLE 2. Size of label-set ($|\mathcal{C}|$) taking different granularity of labels for each corpus. Three levels of granularity were assessed: "Full" stands for "fully specified ICD code", "Main" for the essential modifiers, "Chapter" for the ICD chapter. "Full$_{fi\%}$" is the subset of Full in which the ICD codes were seen in at least *fi*% documents (∼1% and ∼5% of the documents respectively).

	$\|\mathcal{C}\|$			
	MIMIC	Osa1	Osa2	Osa1+2
Full	6,918	2,554	2,554	2,554
Main	941	870	870	870
Chapter	12	24	24	24
Full$_{1\%}$	110	110	110	110
Full$_{5\%}$	16	16	16	16

label-set according to a minimum frequency threshold [46], [47]. In an attempt to make comparisons with respect to previous works, we created a sub-set of instances restricted by frequency. $|C_{train_r}|$ denotes the size of the label-set restricted to most prevalent ICDs following repetition boundaries shown in previous works. Note that, $|C_{train_r}| = |C_{train}|$ means that no restriction was applied. We experimented with a subset from the full label-set in which the occurrences of the codes were above a threshold f. In our case, we considered two thresholds leading to two label-sets, denoted as Full$_{1\%}$ and Full$_{5\%}$, which incorporated a code whenever it appeared in at least ∼1% and ∼5% of the documents respectively. Quantitative details of the label-set in our data-sets are given in Table 2. The table reveals substantial sparsity: just 110 out of 2, 554 ICDs appear at least in 1% of the EHRs (i.e. diseases diagnosed around a hundred times in a set of around 10, 000 EHRs). Note that for all the experiments carried out, the train and test partitions are obtained with the iterative stratification algorithm, with a 70/30 split.

B. PERFORMANCE BY LABEL GRANULARITY

Table 3 shows the experimental results of the model on each corpus (Osa1, Osa2, Osa1+2, and MIMIC). Regarding the label-set reduction, the assessment was made on each of the aforementioned label sub-sets (Full, Full$_{1\%}$, and Full$_{5\%}$ presented in Table 2). Additionally, Table 3 shows the assessment of the computer-aided coding system to various levels of granularity.

Note that increasing the size of the label-set from 110 to 2, 500 (an increment of 1 to 22), as we could expect, was detrimental to the F-Score (from 37.87 to 20.43). Nevertheless, the decrease was not as dramatic as 22 to 1 and the same applies to the results of just 16 prevalent labels. This insight suggests that the model is able to learn from rare cases in EHRs (as one-shot learning strategy aims to) and is able to make predictions of ICDs with little prevalence.

Experiments carried out with the entire label-set (with above 2, 550 different ICDs) show reasonable performance in terms of averaged scores. Although it is difficult to make a fair comparison because no standardised set of experiments has been popularised on any ICD code-based multi-label classification data set, there are some reference works with which we can validate the performance of our models.

TABLE 3. Performance of the system over different ICD code-lengths or granularity (F: Full, M: Main, C: Chap) for all specialties together. P denotes Precision, R Recall and F the F-Score.

Train	Eval	EHRs	grain	$\|C_{train_r}\|$	P	R	F
Osa1	Osa1	13,574	F	2,553	30.61	16.77	20.43
		11,370		110	53.41	31.54	37.87
		8,732		16	70.97	50.44	57.08
		13,574	M	870	44.59	34.36	37.50
		9,844		19	76.10	59.98	66.31
		13,574	C	24	78.42	64.54	69.90
		13,513		14	82.00	66.26	72.51
Osa2	Osa2	13,466	F	2,553	24.49	15.93	18.56
		11,635		110	48.78	26.85	33.02
		9,323		16	67.80	45.99	53.77
		13,466	M	870	40.28	32.28	34.99
		10,142		19	71.86	58.06	63.73
		13,466	C	24	75.46	61.64	66.59
		13,392		14	78.82	66.10	68.27
Osa1+2	Osa1+2	27,040	F	2,554	27.64	20.52	22.48
		22,907		110	53.30	33.79	39.89
		18,098		16	70.14	53.52	59.84
		27,040	M	870	44.97	36.92	39.55
		19,925		19	73.56	63.46	67.52
		27,040	C	24	79.96	66.18	71.77
		26,905		14	80.88	67.46	72.85
MIMIC	MIMIC	55,172	F	6,902	30.72	34.87	31.43
		55,172		110	50.29	53.12	51.26
		55,172		16	67.40	63.81	65.42
		53,090	M	941	46.75	49.55	47.28
		45,189		19	69.47	66.61	67.96
		48,120	C	12	79.12	77.85	78.44

Dermouche *et al.* [48] obtained 75.0 micro F-Score and 35.0 macro F-Score with a Support Vector Machine (SVM) model, and 74.0 micro F-Score and 38.0 macro F-Score with a Latent Dirichlet Allocation (LDA) model, but taking into account just 252 codes from the MIMIC dataset, and computing the F-Score retrieving the correct class among the 10 most probable classes returned by the model. Duarte *et al.* [47] achieved 27.04 macro F-Score with their best model based on Hierarchical GRUs considering the Full codes, 40.50 considering main codes and 62.91 considering chapter codes. The number of labels was 1,418, 611 and 19 respectively. In brief, taking as the baseline the aforementioned state of the art approaches, our approach is ahead in several aspects. In light of the results attained with MIMIC-III, the model was proven competitive with respect to previous works in a fully automatic classification scenario entailing fully-specified ICD codes. Having validated the results on a well-known corpus, we have extended the study to the corpus from Osakidetza (a set of EHRs in Spanish).

Our system was assessed with two non-overlapping sets of EHRs from two different years from Osakidetza, named Osa1 and Osa2. As shown in Tables 1 and 2, both Osa1 and Osa2 have a similar number of input EHR texts (about 13, 500 documents). The size of the label-set is the same (2, 550 in round figures), as is the label cardinality (on average, 5.8 labels per document). However, the F-Score differs 2 absolute points in the most extreme case with all the labels (from 20.43 to 18.56). Nevertheless, as expected due to

TABLE 4. Behaviour of current model on unseen current and future EHRs, for all the specialties together, and with granularity Full. P denotes Precision, R Recall and F the F-score.

Train	Test	EHRs	$\|C_{train}\|$	$\|C_{train_r}\|$	$\|C_{test}\|$	P	R	F
Osa1	Osa1	13,574	2,553	2.553	2,023	30.61	16.77	20.43
Osa1	Osa2	13,574	2,553	2,553	2,052	21.10	8.95	11.61
Osa2	Osa2	13,466	2,553	2,553	2,052	24.49	15.93	18.56
Osa1+2	Osa2	27,040	2,554	2,553	2,023	26.22	19.36	21.52
Osa1+2	Osa1+2	27,040	2,554	2,554	2,554	27.64	20.52	22.48

TABLE 5. Osa1+2 generalist model trained on the $\text{Full}_{1\%}$ label-set but re-evaluated per specialty and also applying the 'specialty labels' label-set modification.

Specialty	EHRs	$\|C_{train}\|$	$\|C_{train_r}\|$	$\|C_{test}\|$	C_{modif}	P	R	F
Pneumology	5,199	2,057	110	107	All	50.44	37.17	40.92
				17	Spec	55.62	45.33	48.39
Cardiology	4,731	1,470	110	105	All	54.00	39.84	43.66
				23	Spec	61.53	49.61	53.22
Digestive	3,990	1,896	110	108	All	50.46	25.80	32.39
				6	Spec	27.17	20.12	23.02
Psychiatry	1,724	697	110	46	All	43.25	30.33	28.95
				7	Spec	54.75	47.06	42.14
Hematology	1,350	1,088	110	98	All	55.77	28.77	36.76
				4	Spec	56.52	20.29	29.86
Nephrology	841	785	110	94	All	59.97	39.87	46.43
				7	Spec	82.95	64.88	72.74

the higher number of available samples, the best performance is attained with the union of both data-sets, even though the union conveys ∼500 labels more on the test set. With the union, the results from Osa1 are improved by another two points, leading to an F-Score of 22.48 on the full label-set.

Bearing a computer-aided ICD classification system in mind, we assessed the model in three scenarios with increasing details in the predicted label, ranging from the ability of the model to predict the fully-specified ICD code (denoted as F (Full) in Table 3), the main class without non-essential modifiers (denoted as M (Main) in Table 3) or the ICD chapter (denoted as C (Chap) in Table 3). Given an EHR, the model is shown to be effective in the assignment of the chapters of the ICD, restricting the use to just those chapters, which can be useful, as discussed in Section V. As the label gets more and more specific, the task gets more complex. Restricting the granularity of the model impacts upon the size of the label-set: while there are thousands of fully-specified ICDs, there are just 24 chapters and 870 main labels. Note that with the Osa1+2 data-set, the F-Score with the Full granularity and the label-set reduced from 2,554 to 110 labels is 39.89, almost the same as the 39.55 F-Score obtained with the non-reduced 870 Main labels.

C. SENSITIVITY OF THE MODEL TO LEXICAL VARIANTS ACROSS TIME

Would a system learn consistently from EHRs issued in one year how to classify EHRs in the future? How much does data addition boost the performance of the system? These experiments would suggest that the lexical features within EHRs (possibly clinical personal, clinical specialisations, etc.) changed over time. Also, note that the number of samples can influence the results, especially when the concatenation of both data-sets are used for the training of the model.

Table 4 shows the ability of each model to classify current and forthcoming EHRs. To this end, the model was trained with current and past EHRs and assessed with either current events or events from subsequent years unclassified at that moment. The aim is to test the sensitivity to different time-frames. As we could expect, training the models with EHRs issued in the same year as those in the evaluation is beneficial, and even more so, if the training set is completed with EHRs even from previous years. As we can see in the rows with Test = Osa2 when the training data is from both years, the F-Score raises from 18.56 to 22.48, increasing the performance by ∼4 points, and it is ∼2 times higher than when training only with EHRs from a previous year (11.61 to 22.48).

D. SENSITIVITY OF THE MODEL TO LEXICAL VARIANTS BY HOSPITAL SERVICE

The full data-set comprises discharge reports issued by different hospital services: e.g. Cardiology, Digestive, Neurology, etc. While decreasing the amount of EHRs tends to be detrimental to the effectiveness of the inference process, restricting the service may also reduce the lexical variability in the input and, eventually, might benefit the predictive ability. In other words, we aim to respond to the following question: how sensitive are the generalist model and the Specialty Models are to relevant codes belonging to the specialty in which the patient was admitted?. These results are shown in Tables 5 and 6.

TABLE 6. Specialty Models trained on subsets of EHRs from the Osa1+2 data-set by specialty, with the same $\text{Full}_{1\%}$ label-set and also applying the 'specialty labels' label-set modifications as in Table 5 to enable comparison.

Specialty	EHRs	$\lvert C_{train}\rvert$	$\lvert C_{train_r}\rvert$	$\lvert C_{test}\rvert$	C_{modif}	P	R	F
Pneumology	5,199	2,057	110	110	All	44.05	25.84	30.49
			17	17	Spec	61.50	41.54	45.57
Cardiology	4,731	1,470	110	110	All	50.12	32.17	37.90
			23	23	Spec	66.83	52.69	57.00
Digestive	3,990	1,896	110	110	All	64.01	14.54	21.28
			6	6	Spec	73.01	45.59	52.38
Psychiatry	1,724	697	110	110	All	30.09	24.92	25.40
			7	7	Spec	70.36	48.04	51.74
Hematology	1,350	1,088	110	110	All	66.93	26.18	35.42
			4	4	Spec	53.33	55.00	43.74
Nephrology	841	785	110	110	All	61.45	31.87	39.60
			7	7	Spec	67.56	72.27	66.81

To deal with this question, we begin with Table 5, where we present the performance of a generalist model (from Table 3, trained with EHRs from all the medical services) evaluated over the different subsets of EHRs by medical service. On the other hand, we also trained a model specifically for each service (namely, Specialty Models) limiting the training data to the EHRs issued in that service. The results are shown in Table 6.

Individual models were created, one per clinical service. Each model was trained receiving only discharge reports from a single service. While this reduces the documents accessed by each model, if the language boundaries are subject to lexical nuances particular to individual services, the models might show good performance, particularly in predicting ICD codes from the service with which they were trained. Note, however, that even though the EHRs were restricted to a single service, they convey both codes from the specialty (associated with the cause of admission in that service) as well as other codes regarding the general status of the patient and other findings (e.g. ex-smoker and type-2 diabetes).

For that reason, we consider two alternative evaluations based on different label-set modifications (as shown in the "C_{modif}" column of the Tables 5 and 6). These are i) All: Keep all the labels that appear across the EHRs from the subset of the given specialty. ii) Spec: Consists in taking into account only the "specialty labels", that is, keep the labels of the given specialty. For example, for Cardiology, you will keep only those labels that appear in Chapter IX - Diseases of the circulatory system of the ICD, due to it being the most-related chapter to Cardiology.

To help the reader interpret the results, note that while the records within Pneumology service convey 2, 057 different ICDs (see $|C_{train}|$ column from Table 5), the sub-set of ICDs within the ICD chapters related to Pneumology are 126, and those among the 110 most frequent codes are only 17 (see $|C_{test}|$ column with the 'Spec' $|C_{modif}|$ from Table 5).

First of all, in most clinical services, when evaluating with the "specialty labels", performance improves. In some specialties, such as Nephrology, the increase is considerable (26.3 F-Score points). One reason is that the difference in the number of labels between both label-set modifications is notorious in every medical service (i.e. from 107 to 17 in Pneumology, or from 105 to 23 in Cardiology...). Despite this fact, in some specialties, better performance is achieved with all the labels than only with the specialty labels, such as in the Digestive case (i.e. 32.39 F-Score with all the 108 labels, and 23.02 with only the 6 digestive-related labels). However, the most remarkable aspect is that, regarding all the labels, (the 'All' rows on the C_{modif} column from Tables 5 and 6) for all medical services, the generalist model achieves a better performance in comparison with the Specialty Models. Nonetheless, what behaviour takes place when considering only the 'specialty labels'?

It can be observed (in Tables 5 and 6) that when the aim is to classify the EHRs of a given specialty according to the ICD codes of that specialty, it is worth training the Specialty Models. In four of the six specialties, the results improve in terms of F-Score. Besides, note that for the medical services which perform better with the generalist model (Pneumology and Nephrology) the gain is only around 3 and 6 points respectively. However, the mean improvement obtained with the Specialty Models for the other specialties is about 14 points, presenting some notable increases, such as the ∼30 points improvement (from 21.28 to 52.38) in the Digestive specialty. Figure 2 combines the experiments with the generalist model evaluated per specialty and each of the Specialty Models (i.e. the results from Tables 5 and 6). The picture shows that assessing all the labels, the generalist model is more suitable, without modification, while when considering only the "specialty labels", the Specialty Models line (in light blue), is, in most medical services, superior to the generalist model.

V. DISCUSSION

The experimental setup consists of a popular clinical multi-label dataset, namely, MIMIC-III, used to validate our approximation and compare with previous works, along with some in-house datasets (Osa1 and Osa2), that presents the advantage that has the data segmented by year and medical specialty.

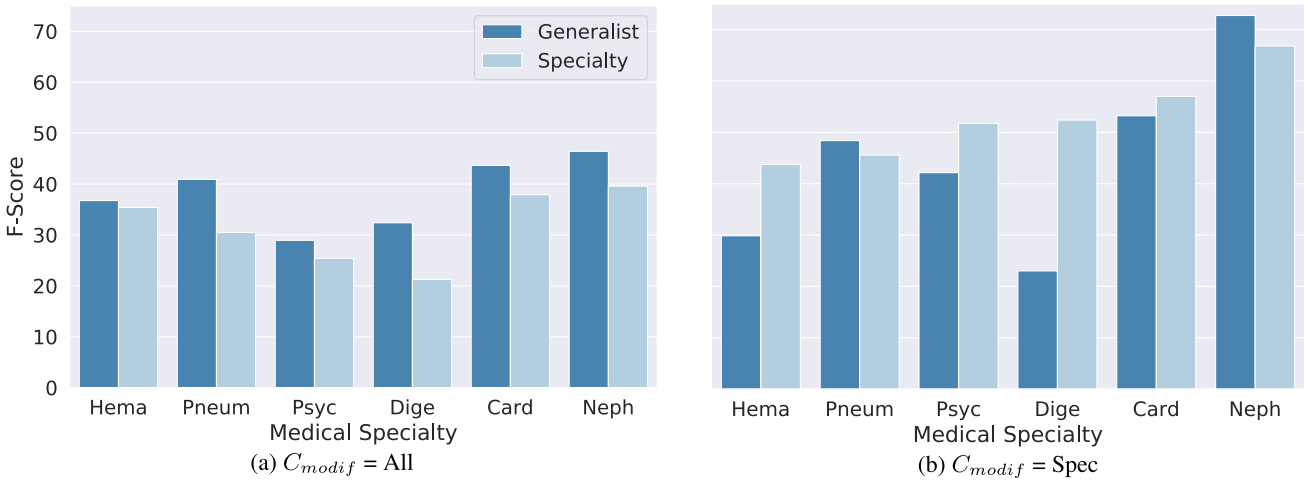

FIGURE 2. F-Scores of models trained with the Osa1+2 data-set on the $\text{Full}_{1\%}$ label-set, but re-evaluated per specialty (i.e. generalist model, dark blue) and trained on subsets of EHRs (i.e. Specialty Models, light blue).

Compared to the related works, the input EHR is not restricted to a diagnostic phrase of few words as in [49] or a short note as in [48]. Our EHRs comprise full notes with 864 ± 415 words on average it is close to a full clinical history of MIMIC-III, entailing several notes for a given patient, with $1,399 \pm 721$ words per history on average. Due to the complex ICD system in conjunction with long medical texts, nowadays, numerous healthcare professionals specially trained are devoted to manual EHR coding. Emerging techniques in NLP are bridging the gap between manual and automatic ICD coding through clinical decision support systems.

In an attempt to assess the usability of the models in practice, we should bear in mind whether the errors produced by the system are minor, due to confusion between non-essential modifiers (i.e. the last characters from the Full codes) yet correctly guessing the main class, or major, failing even the chapter of the ICD. While confusion between non-essential modifiers is counted as a failure, in practice, the system can help the human-coder position in the right branch of the ICD; by contrast, confusion between chapters would require an extra effort on the part of the human expert. In this scenario, the model would guide the expert to a Chapter (branch) of the ICD in which to select the Full code.

The assessment was carried out enabling the model dispense with the non-essential modifiers (and focusing on the main ICD and the chapter). Results showed that the potential of the model increased substantially from the fully- automated scenario to that of the computer-aided classification. A computer-aided ICD classification system can help the human expert to access the chapters of the ICD involved in each record very accurately (with a Precision of 80.88 and a coverage-recall of 67.46) as shown in Table 3. If we turn to a finer-grained classification in a situation in which the system would act automatically and would code the fully-specified label, the system would attain Precision of 27.64 and the Recall decays to 20.52. Depending on the Recall required, the system would demand a more active role from the human, while a significant percentage of the labels would have been correctly assigned. Although we have explored several levels of granularity, namely, Chapter, Main and Full granularity, we have focused on the complete ICD, as it is of great importance for applications such as insurance billing or other clinical information extraction tasks.

Often, previous works discarded learning ICDs that had little prevalence in the set or which only focused on a set of nearly a hundred labels [34], [48], [50]. In our case, we assessed them all, but as we could expect, prevalent ICDs are predicted more accurately than the average prediction quality. Table 3 disclosed that increasing training instances significantly benefitted the predictive ability of the model. The restriction to 110 and 16 most prevalent labels was selected in an attempt to make fair comparisons with previous works. Nevertheless, increasing the number of labels and decreasing the performance does not show a linear relationship, but rather the performance drop is less than expected. This implies that the model can learn from infrequent occurrences and can predict uncommon ICDs.

An **analysis** of the results shows that a generalist model trained overall services achieves, on average, an F-Score of 22.48 for the full set of 2,554 labels (see Table 3). Regarding the experiments to assess the robustness across-time, adding more years (i.e. more EHRs) to the training set benefits performance, as expected. Nevertheless, it is interesting to note that although the models are robust enough to correctly classify some EHRs from future years trained solely on data from past years, there is a negative effect on performance. Therefore, our recommendation is, whenever possible, to continue retraining the models with the new data as it becomes available since the improvement is not negligible.

In what concerns the experiments with the different medical services, one conclusion is that when evaluating the labels without modification by service (that is, all the labels that

appear in the EHRs), the best results are obtained with the generalist model, meaning that the lexical reduction did not overcome the label-set variability. Nevertheless, the most significant insight gained is derived from Table 6, which shows that it is when the Specialty Models are trained on the specialty label-sets that they do better than the generalist model. It is true that this comes with the extra cost of training several models, one for each medical service, and is limited to more restricted specialty-related label-sets. However, we feel that for certain applications, such as intra-specialty pharmacovigilance services in hospitals, these costs could be offset by the associated advantages.

Regarding the evaluation, we believe that there are aspects that do not get reflected in the most widely used metrics (such as Precision, Recall, F-Score, MAP, MRR…). Specifically, the number of codes associated with a document is relevant: therefore, if the prediction yields either a much lower, similar or much higher number of codes than the actual number of codes, this should be penalised/acknowledged accordingly. We feel that further multifaceted metrics should be developed to gain a deeper insight into extreme multi-label classification.

There is room for improvement by exploring other neural approaches e.g. models based on the Transformer architecture (BERT, BioBert, …). Nevertheless, transformers pose challenges in the training process [51] due to the high computational burden and data needed. To remedy this, and inspired by the fine-tuning strategy, we feel that an initial generalist model could be trained; this could be further fine-tuned with new data from subsequent years or new medical services.

VI. CONCLUSION

This work deals with an extreme multi-label classification task on clinical texts. The aim is to assign, to each EHR, the corresponding diagnoses as in the ICD. Each EHR tends to convey 5.8 ± 3.4 ICDs (out of about 2, 500 distinct diagnoses in our study).

Having demonstrated the ability of the approach to be both a fully-automatic and computer-aided multi-label classification, we assessed the resilience of the model to natural variations in order to address omission in previous works. The concern is about the behaviour of a model trained with some EHRs when it comes to classifying EHRs later on (e.g. texts possibly written by different experts). We put our focus on variations in two aspects: **across-time** and through **hospital services**.

Regarding the resilience of the system to time-related variations, the results showed that the datasets from different (non-overlapping and consecutive) years are similar in difficulty, as the results with Osa1 and Osa2 are reasonably comparable, with 20.43 and 18.56 F-Score, respectively. Also, adding more samples is always useful, as this gives best results, as previously seen when the train set is the union Osa1+2 (i.e. 22.48 F-Score when training and testing with Osa1+2). A key insight is that although the datasets are similarly difficult when trying to predict future EHRs with data from previous years, the performance decline significantly (i.e. from 20.43 when training and testing with Osa to 11.61 when training with Osa1 but testing with Osa2 future samples).

With respect to the ability of the system to classify EHRs across medical specialties, our experiments showed that when the predictions are made over non-modified label-sets by specialty, the best option is the generalist model, which benefits from the greater number of EHRs in the training set. However, the approach which achieves the most favourable results on specialty labels is to train Specialty Models with the specialty label sub-sets. Although this carries an extra cost, it can be useful for the development of tools for application in specific medical services in hospitals.

We feel that there is still a gap in the literature that could be exploited for **future work**: namely, knowledge-driven reinforcement learning exploiting the hierarchical structure of the ICD to gain accuracy in different granularity levels. Previous works tried to incorporate the hierarchy [52], [53]. Clinical entity recognition could help to recognise relevant information such as disorders or findings, laterality, severity or body-part. This information is, somehow, included in the hierarchical representation of the ICD and could drive the code generation. Within our framework, the hierarchical boundaries could be modelled as embedded graphs. This approach, however, is outside of the scope of this work.

REFERENCES

[1] P. Mukherjee and A. Mukherjee, "Advanced processing techniques and secure architecture for sensor networks in ubiquitous healthcare systems," in *Sensors for Health Monitoring*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 3–29.

[2] A. Banerji, K. H. Lai, Y. Li, R. R. Saff, C. A. Camargo, K. G. Blumenthal, and L. Zhou, "Natural language processing combined with ICD-9-CM codes as a novel method to study the epidemiology of allergic drug reactions," *J. Allergy Clin. Immunol., Pract.*, vol. 8, no. 3, pp. 1032–1038, 2020.

[3] S. Santiso, A. Perez, and A. Casillas, "Exploring joint AB-LSTM with embedded lemmas for adverse drug reaction discovery," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 5, pp. 2148–2155, Sep. 2019.

[4] L. Chan, K. Beers, A. A. Yau, K. Chauhan, Á. Duffy, K. Chaudhary, N. Debnath, A. Saha, P. Pattharanitima, J. Cho, P. Kotanko, A. Federman, S. G. Coca, T. Van Vleck, and G. N. Nadkarni, "Natural language processing of electronic health records is superior to billing codes to identify symptom burden in hemodialysis patients," *Kidney Int.*, vol. 97, no. 2, pp. 383–392, Feb. 2020.

[5] D. Chandran, D. A. Robbins, C.-K. Chang, H. Shetty, J. Sanyal, J. Downs, M. Fok, M. Ball, R. Jackson, R. Stewart, H. Cohen, J. M. Vermeulen, F. Schirmbeck, L. de Haan, and R. Hayes, "Use of natural language processing to identify obsessive compulsive symptoms in patients with schizophrenia, schizoaffective disorder or bipolar disorder," *Sci. Rep.*, vol. 9, no. 1, pp. 1–7, Dec. 2019.

[6] A. Bhattacherjee, S. Roy, S. Paul, P. Roy, N. Kausar, and N. Dey, "Classification approach for breast cancer detection using back propagation neural network: A study," in *Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications*. Hershey, PA, USA: IGI Global, 2020, pp. 1410–1421.

[7] N. C. Ernecoff, K. L. Wessell, L. C. Hanson, A. M. Lee, C. M. Shea, S. B. Dusetzina, M. Weinberger, and A. V. Bennett, "Electronic health record phenotypes for identifying patients with late-stage disease: A method for research and clinical application," *J. Gen. Internal Med.*, vol. 34, no. 12, pp. 2818–2823, Dec. 2019.

[8] H. Alzoubi, R. Alzubi, N. Ramzan, D. West, T. Al-Hadhrami, and M. Alazab, "A review of automatic phenotyping approaches using electronic health records," *Electronics*, vol. 8, no. 11, p. 1235, Oct. 2019.

[9] *International Statistical Classification of Diseases and Related Health Problems*, World Health Org., Geneva, Switzerland, 2004, vol. 1.
[10] K. J. O'Malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton, "Measuring diagnoses: ICD code accuracy," *Health Services Res.*, vol. 40, no. 5, pp. 1620–1639, Oct. 2005.
[11] R. H. Napier, L. S. Bruelheide, E. T. K. Demann, and R. H. Haug, "Insurance billing and coding," *Dental Clinics North Amer.*, vol. 52, no. 3, pp. 507–527, Jul. 2008.
[12] M. Lau, J. L. Prenner, A. J. Brucker, and B. L. VanderBeek, "Accuracy of billing codes used in the therapeutic care of diabetic retinopathy," *JAMA Ophthalmol.*, vol. 135, no. 7, pp. 791–794, 2017.
[13] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, Dec. 2016, Art. no. 160035.
[14] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than English: Opportunities and challenges," *J. Biomed. Semantics*, vol. 9, no. 1, p. 12, Dec. 2018.
[15] H. Dalianis, *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer, 2018, doi: 10.1007/978-3-319-78503-5.
[16] D. Zikos and N. DeLellis, "CDSS-RM: A clinical decision support system reference model," *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 137, Dec. 2018.
[17] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 115–124.
[18] A. Névéol, A. Robert, F. Grippo, C. Morgand, C. Orsi, L. Pelikan, L. Ramadier, G. Rey, and P. Zweigenbaum, "CLEF eHealth 2018 multilingual information extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian," in *Proc. CLEF (Working Notes)*, 2018, pp. 1–18.
[19] T. Gangavarapu, A. Jayasimha, G. S. Krishnan, and S. Kamath, "Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes," *Knowl.-Based Syst.*, vol. 190, Feb. 2020, Art. no. 105321.
[20] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling," 2016, *arXiv:1611.06639*. [Online]. Available: http://arxiv.org/abs/1611.06639
[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555
[22] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365
[23] A. Névéol, K. B. Cohen, C. Grouin, T. Hamon, T. Lavergne, L. Kelly, L. Goeuriot, G. Rey, A. Robert, X. Tannier, and P. Zweigenbaum, "Clinical information extraction at the CLEF eHealth evaluation lab 2016," in *Proc. CEUR Workshop*, vol. 1609, 2016, p. 28.
[24] A. Névéol, A. Robert, R. Anderson, K. B. Cohen, C. Grouin, T. Lavergne, G. Rey, C. Rondet, and P. Zweigenbaum, "CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French," in *Proc. CLEF (Working Notes)*, 2017, pp. 1–17.
[25] A. Dörendahl, N. Leich, B. Hummel, G. Schönfelder, and B. Grune, "Overview of the CLEF eHealth 2019 multilingual information extraction," in *Proc. CEUR-WS*, 2019, pp. 1–9.
[26] L. Zhou, C. Cheng, D. Ou, and H. Huang, "Construction of a semi-automatic ICD-10 coding system," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–12, Dec. 2020.
[27] A. Casillas, A. D. de Ilarraza, K. Gojenola, M. Oronoz, and A. Pérez, "First approaches on Spanish medical record classification using diagnostic term to class transduction," in *Proc. 10th Int. Workshop Finite State Methods Natural Lang. Process.*, 2012, pp. 60–64.
[28] A. Pérez, A. Atutxa, A. Casillas, K. Gojenola, and Á. Sellart, "Inferred joint multigram models for medical term normalization according to ICD," *Int. J. Med. Informat.*, vol. 110, pp. 111–117, Feb. 2018.
[29] O. Luaces, J. Díez, J. Barranquero, J. J. D. Coz, and A. Bahamonde, "Binary relevance efficacy for multi-label classification," *Prog. Artif. Intell.*, vol. 1, no. 4, pp. 303–313, 2012.
[30] A. Rios and R. Kavuluru, "Neural transfer learning for assigning diagnosis codes to EMRs," *Artif. Intell. Med.*, vol. 96, pp. 116–122, May 2019.
[31] C. J. Murray, A. D. Lopez, R. Black, R. Ahuja, S. M. Ali, A. Baquim, L. Dandona, E. Dantzer, V. Das, U. Dhingra, and A. Dutta, "Population health metrics research consortium gold standard verbal autopsy validation study: Design, implementation, and development of analysis datasets," *Population Health Metrics*, vol. 9, no. 1, p. 27, Dec. 2011.
[32] W. B. A. Karaa, A. S. Ashour, D. B. Sassi, P. Roy, N. Kausar, and N. Dey, "MEDLINE text mining: An enhancement genetic algorithm based approach for document clustering," in *Applications of Intelligent Optimization in Biology and Medicine: Current Trends and Open Problems*. Cham, Switzerland: Springer, 2016, pp. 267–287.
[33] R. Babbar and B. Schölkopf, "Data scarcity, robustness and extreme multi-label classification," *Mach. Learn.*, vol. 108, nos. 8–9, pp. 1329–1351, Sep. 2019.
[34] J. Pérez, A. Pérez, A. Casillas, and K. Gojenola, "Cardiology record multi-label classification using latent Dirichlet allocation," *Comput. Methods Programs Biomed.*, vol. 164, pp. 111–119, Oct. 2018.
[35] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu, "AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5812–5822.
[36] R. You, Z. Zhang, S. Dai, and S. Zhu, "HAXMLNet: Hierarchical attention network for extreme multi-label text classification," 2019, *arXiv:1904.12578*. [Online]. Available: http://arxiv.org/abs/1904.12578
[37] F. Gargiulo, S. Silvestri, M. Ciampi, and G. De Pietro, "Deep neural network for hierarchical extreme multi-label text classification," *Appl. Soft Comput.*, vol. 79, pp. 125–138, Jun. 2019.
[38] Y. Deng, A. Sander, L. Faulstich, and K. Denecke, "Towards automatic encoding of medical procedures using convolutional neural networks and autoencoders," *Artif. Intell. Med.*, vol. 93, pp. 29–42, Jan. 2019.
[39] A. Blanco, O. Perez-de-Viñaspre, A. Pérez, and A. Casillas, "Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity," *Comput. Methods Programs Biomed.*, vol. 188, May 2020, Art. no. 105264.
[40] Y. Cheng, K. Qian, Y. Wang, and D. Zhao, "Missing multi-label learning with non-equilibrium based on classification margin," *Appl. Soft Comput.*, vol. 86, Jan. 2020, Art. no. 105924.
[41] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Extreme multi-label legal text classification: A case study in EU legislation," 2019, *arXiv:1905.10892*. [Online]. Available: http://arxiv.org/abs/1905.10892
[42] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. Dhillon, "X-BERT: Extreme multi-label text classification with BERT," 2019, *arXiv:1905.02331*. [Online]. Available: http://arxiv.org/abs/1905.02331
[43] Y.-T. Zhou and R. Chellappa, "Computation of optical flow using a neural network," in *Proc. IEEE Int. Conf. Neural Netw.*, Jul. 1988, pp. 71–78.
[44] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Oxford, U.K.: Butterworth-Heinemann, 1979.
[45] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, MA, USA: Cambridge Univ. Press, 2008.
[46] M. Apidianaki, S. Mohammad, J. May, E. Shutova, S. Bethard, and M. Carpuat, "Proceedings of the 12th international workshop on semantic evaluation," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 1–18.
[47] F. Duarte, B. Martins, C. S. Pinto, and M. J. Silva, "Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text," *J. Biomed. Informat.*, vol. 80, pp. 64–77, Apr. 2018.
[48] M. Dermouche, J. Velcin, R. Flicoteaux, S. Chevret, and N. Taright, "Supervised topic models for diagnosis code assignment to discharge summaries," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Cham, Switzerland: Springer, 2016, pp. 485–497.
[49] A. Atutxa, A. Casillas, N. Ezeiza, I. Goenaga, V. Fresno, K. Gojenola, R. Martinez, M. Oronoz, and O. P. D. Viñaspre, "IxaMed at CLEF eHealth 2018 task 1: ICD10 coding with a sequence-to-sequence approach," in *Proc. CLEF Online Work. Notes. CEUR-WS*, 2018, pp. 1–9.
[50] P. Nigam, "Applying deep learning to ICD-9 multi-label classification from medical records," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2016.
[51] L. Liu, X. Liu, J. Gao, W. Chen, and J. Han, "Understanding the difficulty of training transformers," 2020, *arXiv:2004.08249*. [Online]. Available: http://arxiv.org/abs/2004.08249
[52] J. C. Ferrao, F. Janela, M. D. Oliveira, and H. M. G. Martins, "Using structured EHR data and SVM to support ICD-9-CM coding," in *Proc. IEEE Int. Conf. Healthcare Informat.*, Sep. 2013, pp. 511–516.
[53] L. Cao, D. Gu, Y. Ni, and G. Xie, "Automatic ICD code assignment based on ICD's hierarchy structure for Chinese electronic medical records," *AMIA Summits Transl. Sci. Proc.*, vol. 2019, p. 417, May 2019.

ALBERTO BLANCO received the B.S. and M.S. degrees in computer engineering and computational engineering and intelligent systems from the University of The Basque Country (UPV/EHU), in 2018 and 2019, respectively. He is currently pursuing the Ph.D. degree with the Department of Electricity and Electronics.

His research activity is being carried out within Ixa Group, HiTZ: Basque Center for Language Technology. His research interests include deep learning, natural language processing, clinical text mining, and multi-label classification.

ALICIA PÉREZ received the B.S. and M.S. degrees in physics engineering from the University of The Basque Country (UPV/EHU), and the Ph.D. degree in computational linguistics in 2010. Since 2011, she has been an Assistant Professor with the Computer Languages and Systems Department. Her research activity is being carried out within Ixa Group, HiTZ: Basque Center for Language Technology. Her research interests include natural language processing and understanding, clinical text mining, and artificial intelligence.

ARANTZA CASILLAS received the B.S. degree in computer science from Deusto University, and the Ph.D. degree in computational linguistics in 2000. Since 2001, she has been with the Science and Technology Faculty (UPV/EHU), Electricity and Electronics Department, where she is currently an Associate Professor. She is also the main Researcher of the Medical and Legal Domains Section, HiTZ: Basque Center for Language Technology. Her research interest includes artificial intelligence, natural language processing and understanding, clinical text mining, and clinical information retrieval.