



UNIVERSITÀ  
DEGLI STUDI  
FIRENZE

# FLORE

## Repository istituzionale dell'Università degli Studi di Firenze

### **Classification of cancer pathology reports: a large-scale comparative study**

Questa è la Versione finale referata (Post print/Accepted manuscript) della seguente pubblicazione:

*Original Citation:*

Classification of cancer pathology reports: a large-scale comparative study / Martina, Stefano; Ventura, Leonardo; Frasconi, Paolo. - In: IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS. - ISSN 2168-2194. - STAMPA. - 24:(2020), pp. 3085-3094. [10.1109/JBHI.2020.3005016]

*Availability:*

The webpage <https://hdl.handle.net/2158/1226419> of the repository was last updated on 2021-02-22T15:13:29Z

*Published version:*

DOI: 10.1109/JBHI.2020.3005016

*Terms of use:*

Open Access

La pubblicazione è resa disponibile sotto le norme e i termini della licenza di deposito, secondo quanto stabilito dalla Policy per l'accesso aperto dell'Università degli Studi di Firenze (<https://www.sba.unifi.it/upload/policy-oa-2016-1.pdf>)

*Publisher copyright claim:*

La data sopra indicata si riferisce all'ultimo aggiornamento della scheda del Repository FloRe - The above-mentioned date refers to the last update of the record in the Institutional Repository FloRe

(Article begins on next page)

# Classification of cancer pathology reports: a large-scale comparative study

Stefano Martina, Leonardo Ventura, and Paolo Frasconi

**Abstract**—We report about the application of state-of-the-art deep learning techniques to the automatic and interpretable assignment of ICD-O3 topography and morphology codes to free-text cancer reports. We present results on a large dataset (more than 80 000 labeled and 1 500 000 unlabeled anonymized reports written in Italian and collected from hospitals in Tuscany over more than a decade) and with a large number of classes (134 morphological classes and 61 topographical classes). We compare alternative architectures in terms of prediction accuracy and interpretability and show that our best model achieves a multiclass accuracy of 90.3% on topography site assignment and 84.8% on morphology type assignment. We found that in this context hierarchical models are not better than flat models and that an element-wise maximum aggregator is slightly better than attentive models on site classification. Moreover, the maximum aggregator offers a way to interpret the classification process.

**Index Terms**—Artificial Intelligence, Attention models, Deep Learning, Hierarchical models, Interpretable models, Machine Learning, Oncology, Recurrent Neural Networks.

## I. INTRODUCTION

Cancer is a major concern worldwide, as it decreases the quality of life and leads to premature mortality. In addition, it is one of the most complex and difficult-to-treat diseases, with significant social implications, both in terms of mortality rate and in terms of costs associated with treatment and disability [1]–[4]. Measuring the burden of disease is one of the main concerns of public healthcare operators. Suitable measures are necessary to describe the general state of population's health, to establish public health goals and to compare the national health status and performance of health systems across countries. Furthermore, such studies are needed to assess the allocation of health care and health research resources across disease categories and to evaluate the potential costs and benefits of public health interventions [5].

Cancer registries emerged during the last few decades as a strategic tool to quantify the impact of the disease and to provide analytical data to healthcare operators and decision makers. Cancer registries use administrative and clinical data

sources in order to identify all the new cancer diagnoses in a specific area and time period and collect incidence records that provide details on the diagnosis and the outcome of treatments. Mining cancer registry datasets can help towards the development of global surveillance programs [6] and can provide important insights such as survivability [7]. Although data analysis software would best operate on structured representations of the reports, pathologists normally enter data items as free text in the local country language. This requires intelligent algorithms for medical document information extraction, retrieval, and classification, an area that has received significant attention in the last few years (see, e.g., [8] for a recent account and [9] for the specific case of cancer).

The study of intelligent algorithms is also motivated by the inherent slowness of the cancer registration process, which is partially based on manual revision, and which also requires the interpretation of pathological reports written as free text [10]–[12]. In practice, significant delays in data production and publication may occur. This weakens data relevance for the purpose of assessing compliance with updated regional recommended integrated case pathways, as well as for public health purposes. Improving automated methods to generate a list of putative incident cases and to automatically estimate process indicators is thus an opportunity to perform an up-to-date evaluation of cancer-care quality. In particular, machine learning techniques like the ones presented in this paper could overcome the delay in cancer case definition by the cancer registry and pave the way towards powerful tools for obtaining indicators automatically and in a timely fashion.

In our specific context, pathology reports can be classified according to codes defined in the International Classification of Diseases for Oncology, third edition (ICD-O3) system [13], a specialization of the ICD for the cancer domain which is internationally adopted as the standard classification for topography and morphology [12]. The development of text analysis tools specifically devoted to the automatic classification of incidence records according to ICD-O3 codes has been addressed in a number of previous papers (see Section II below). Some works have either focused on reasonably large datasets but using simple linear classifiers based on bag-of-words representations of text [14], [15]. Most other works have applied recent state-of-the-art deep learning techniques [16], [17] but using smaller datasets and restricted to a partial set of tumors. A remarkable exception is [18] that applies convolutional networks to a large dataset. Additionally, the use of deep learning techniques usually requires accurate domain-

Paolo Frasconi was supported in part by the Italian Ministry of Education, University, and Research under Grant 2017TWNMH2.

Stefano Martina is with University of Florence (email: stefano.martina@unifi.it).

Leonardo Ventura is with Institute for cancer research, prevention and clinical network (ISPRO), Florence (email: l.ventura@ispro.toscana.it).

Paolo Frasconi is with University of Florence (email: paolo.frasconi@unifi.it).

specific word vectors (embeddings of words in a vector space) that can be derived from word co-occurrences in large corpora of unlabeled text [19]–[21]. Large medical corpora are easily available for English (e.g. PubMed) but not necessarily for other languages.

To the best of our knowledge, the present work is the first to report results on a large dataset ( $> 80,000$  labeled reports for supervised learning and  $> 1,500,000$  unlabeled reports for pretraining word vectors), with a large number of both topography and morphology classes, and comparing several alternative state-of-the-art deep learning techniques, namely Gated Recurrent Unit (GRU) Recurrent Neural Network (RNN) [22], with and without attention [23], Bidirectional Encoder Representations from Transformers (BERT) [21] and Convolutional Neural Network (CNN). In particular, we are interested in evaluating on real data the effectiveness of attention models, comparing them with a simpler form based on max aggregation. We also report an extensive study on the interpretability of the trained classifiers. Our results confirm that recent deep learning techniques are very effective on this task, with attentive GRUs reaching a multiclass accuracy of 90.3% on topography (61 classes) and 84.8% on morphology (134 classes) but (1) hierarchical models does not achieve better accuracy than using flat models, (2) the improvement over a simple support vector machine classifier on bag-of-words is modest, (3) a simpler aggregator of hidden representations taking element-wise maximum over time improves slightly over (flat and hierarchical) attention models for topography prediction while a flat attention model is better for morphology task, and (4) the improvements of flat models over hierarchical is stronger for difficult to learn rare classes. We additionally show that the element-wise maximum aggregator offers a new alternative strategy for interpreting prediction results.

## II. RELATED WORKS

Early works for ICD-O3 code assignment were structured on rule-based systems, where the code was assigned by creating a set of handcrafted text search queries and combining results by standard Boolean operators [24]. In order to prevent spurious matches, rules need to be very specific, making it very difficult to achieve a sufficiently high recall on future (unseen) cases.

Also more recent works employ rule-based approaches. Coden et al. [25] implemented a knowledge representation model that they populate processing cancer pathology reports with Natural Language Processing (NLP) techniques. They performed categorization of classes using rules based on syntactic structure. They also experimented, without satisfactory results, machine learning methods. They validate the model using a small corpus of 302 pathology reports related to colon cancer obtaining an  $F1$  score of 0.82 for primary tumor classification and 0.65 for metastatic tumor. Nguyen et al. [26] developed a rule based system evaluated on a set of 221 pathology reports with 66 full site classes (site plus sub-site) and 94 type classes. They obtained an  $F1$  score of 0.61 and 0.64 respectively for site and type.

A number of studies reporting on the application of machine learning to this problem have been published during the last

decade. Direct comparisons among these works are impossible due to the (not surprising) lack of standard publicly available datasets and the presence of heterogeneous details in the settings. Still, we highlight the main differences among them in order to provide some background. In [14], the authors employed support vector machine (SVM) and Naive Bayes classifiers on a small dataset of 5 121 French pathology reports and a reduced number of target classes (26 topographic classes and 18 morphological classes), reporting an accuracy of 72.6% on topography and 86.4% on morphology with SVM. A much larger dataset of 56 426 English reports from the Kentucky Cancer Registry was later employed in [15], where linear classifiers (SVM, Naive Bayes, and logistic regression) were also compared but only on the topography task and using 57, 42, and 14 classes (determined considering classes that have at least respectively 50, 100, and 1000 examples). The authors reported a micro-averaged  $F1$  measure of 90% on 57 classes using SVM with both unigrams and bigrams. Still, the bag-of-words representations used by these linear classifiers do not consider word order and are unable to capture similarities and relations among words (which are all represented by orthogonal vectors). Deep learning techniques are known to overcome these limitations but were not applied to this problem until very recently. In [17], a CNN architecture fed by word vectors pretrained on PubMed was applied to a small corpus of 942 breast and lung cancer reports in English with 12 topography classes; the authors demonstrate the superiority of this approach compared to linear classifiers with significant increases in both micro and macro  $F1$  measures. In [16], the same research group also experimented on the same dataset using RNNs with hierarchical attention [27], obtaining further improvements over the CNN architecture. Also the same research group implemented in [18] two CNN-based multitask learning techniques and trained them on a big dataset of 95 231 pathology reports (71 223 unique tumors) from the Louisiana Tumor Registry. The models were trained on five tasks: topology main site (65 classes), laterality (4 classes), behavior (3 classes), morphology type (63 classes), and morphology grade (5 classes). They reached a micro and macro  $F1$  score of respectively 0.944 and 0.592 for site prediction and respectively 0.811 and 0.656 for type prediction.

Recent works investigated the interpretability of supervised machine learning models. In [28], a novel technique called *LIME* explains the prediction of any classifier or regressor by locally approximating it.

## III. MATERIALS AND METHODS

### A. Dataset

We collected a set of 1 592 385 anonymized anatomopathological exam results from Tuscany region cancer registry in the period 1990-2014 for which we obtained the approval from the institutional ethics committee<sup>1</sup>. About 6% of these records refer to a positive tumor diagnosis and have topographical and morphological ICD-O3 labels, determined by tumor registry

<sup>1</sup>CEAV 14081\_oss 27/11/2018

experts. Other reports are associated with non-cancerous tissues and with unlabeled records. When multiple pathological records for the same patient existed for the same tumor, cancer registry experts selected the most informative report in order to assign the ICD-O3 code to that tumor case, leaving a set of 94 524 labeled reports. In our dataset each labeled report correspond to the primary report for a single tumor case, thus the classification was performed at report level.

The histological exam records consist of three free-text fields (not all of them always filled-in) reporting tissue macroscopy, diagnosis, and, in some cases, the patient's anamnesis. We found that field semantics was not always used consistently and that the amount of provided details varied significantly from extremely synthetic to very detailed assessments of the morphology and the diagnosis. Field length ranged from 0 to 1 368 words, with lower, middle and upper quartiles respectively 34, 62 and 134. For these reasons, we merged the three text fields into a single text document. We did a case normalization converting the letters to uppercase and we kept punctuation. We finally removed duplicates (records that have the exact same text) and reports labeled with extremely rare ICD-O3 codes (1048 samples that do not appear in either training, validation, and test sets). In the end we obtained a dataset suitable for supervised learning consisting of 85 170 labeled records ( $\sim 99\%$  of them in the period 2004-2012). We further split the records in sentences when using hierarchical models. For this purpose we employed the *spaCy* sentence segmentation tool<sup>2</sup>.

After preprocessing, our documents had on average length of 105 words and contained on average 13 sentences (detailed distributions are reported in Figure 3 of Appendix I). These statistics indicate that reports tend to be much shorter than in other studies. For example, in [18] the average number of words and sentences per document were 1290 and 117, respectively<sup>3</sup>. It is also the case that language is often synthetic, rich in keywords, and poor in verbs (three sample reports are shown in Figure 1).

ICD-O3 codes describe both topography (tumor site) and morphology. A topographical ICD-O3 code is structured as *Cmm.s* where *mm* represent the main site and *s* the subsite. For example, *C50.2* is the code for the upper-inner quadrant (2) of breast (50). A morphological ICD-O3 code is structured as *tttt/b* where *tttt* represent the cell type and *b* the tumor behavior (benign, uncertain, in-situ, malignant primary site, malignant metastatic site). For example, *8140/3* is the code for an adenocarcinoma (*adeno 8140; carcinoma 3*). We defined two associated multi-class classification tasks (1) main tumor site prediction (topography) and (2) type prediction (morphology). The topography task only considers the first part of the topographical ICD-O3 code, before the dot without the sub-site. The morphology task only considers the first part of the morphological ICD-O3 code, before the slash without the behavior. As shown in Figure 4 (Appendix I), our dataset is highly unbalanced, with many of the 71 topographical and

435 morphological classes found in the data being very rare. In an attempt to reduce bias in the estimated performance (particularly for the macro F1 measure, see below), we removed classes with less than five records in the test set, resulting in 61 topographical and 134 morphological classes. Even after these removals, our tasks have no less classes than in previous works (the most comprehensive previous study [18] has 65 topographical classes and 63 histological classes).

In order to provide an evaluation that does not neglect possible dataset shift issues over time (for example due to style changes or to the evolving of oncology knowledge) we split train, validation, and test data using a temporal criterion (based on record insertion date): we used the most recent 20% of data as the test set (17 015 records for site and 16 719 for type, from March 2012 to March 2014), a similar amount of the remaining most recent records as the validation set (17 007 for site and 16 787 for type, from December 2010 to March 2012), and the rest as the training set (50 875 for site and 49 436 for type, before December 2010). Note that many other previous studies have used a k-fold cross validation strategy, which is perhaps unavoidable when dealing with small datasets.

## B. Plain models

In our setting, a dataset  $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}$  consists of variable length sequence vectors  $\mathbf{x}^{(i)}$ . For  $t = 1, \dots, T^{(i)}$ ,  $x_t^{(i)}$  is the  $t$ -th word in the  $i$ -th document and  $y^{(i)} \in \{1, \dots, K\}$  is the associated target class. To simplify the notation in the subsequent text, the superscripts are not used unless necessary. Sequences are denoted in boldface. The GRU-based sequence classifiers<sup>4</sup> used in this work compute their predictions  $f(\mathbf{x})$  as follows:

$$e_t = E(x_t; \theta^e), \quad (1)$$

$$h_t^f = F(e_t, h_{t-1}^f; \theta^f), \quad (2)$$

$$h_t^r = R(e_t, h_{t+1}^r; \theta^r), \quad (3)$$

$$u_t = G(h_t; \theta^h), \quad (4)$$

$$\phi = A(\mathbf{u}; \theta^a), \quad (5)$$

$$f(\mathbf{x}) = g(\phi; \theta^c). \quad (6)$$

$E$  is an embedding function mapping words into  $p$ -dimensional real vectors where embedding parameters  $\theta^e$  can be either pretrained and adjustable or fixed, see Section III-E below. Functions  $F$  and  $R$  correspond to (forward and reverse) dynamics that can be described in terms of several (possibly layered) recurrent cells. Each vector  $h_t = h_t^f \oplus h_t^r$  (the concatenation of  $h_t^f$  and  $h_t^r$ ) can be interpreted as latent representations of the information contained at position  $t$  in the document.  $G$  is an additional Multilayer Perceptron (MLP) (with sigmoidal output units) mapping each latent vector into a vector  $u_t$  that can be seen as contextualized representation of the word at position  $t$ .  $A$  is an aggregation function that creates a single  $d$ -dimensional representation vector for the entire sequence and  $g$  is a classification layer with softmax. The parameters  $\theta^f$ ,  $\theta^r$ ,  $\theta^h$ , and  $\theta^a$  (if present) are determined

<sup>2</sup><https://spacy.io/>

<sup>3</sup>Note, however, than in that study about 76% of the documents consisted of a single report, and the rest of concatenated reports (two reports in 17.7% cases, three in 4.2% cases, and four or more in 2.1% cases).

<sup>4</sup>In a set of preliminary experiments we found that Long Short-Term Memory (LSTM) did not improve over GRU.



by minimizing a loss function  $\mathcal{L}$  (categorical cross-entropy in our case) on training data. Three possible choices for the aggregator function are described below.

1) *Concatenation*:  $\phi = (h_T^f, h_1^r)$ . In this model, called **GRU** in the following,  $G$  is the identity function and we simply take the extreme latent representations; in principle, these may be sufficient since they depend on the whole sequence due to bidirectional dynamics. However, note that this approach may require long-term dependencies to be effectively learned;

2) *Attention mechanism*:  $\phi = \sum_t a_t(\mathbf{u}; \theta^a) u_t$ . In this model, called **ATT** in the following, (scalar) attention weight [23] are computed as

$$c_t = C(\mathbf{u}; \theta^a),$$

$$a_t(\mathbf{u}; \theta^a) = \frac{e^{\langle c, c_t \rangle}}{\sum_{\tau=1}^T e^{\langle c, c_\tau \rangle}},$$

where  $C$  is a single layer that maps the representation  $u_t$  of the word to a hidden representation  $c_t$ . Then, the importance of the word is measured as a similarity with a context vector  $c$  that is learned with the model and can be seen as an embedded representation of a high level query as in memory networks [29];

3) *Max pooling over time*:  $\phi_j = \max_t u_{j,t}$ . In this model [30], [31] (called **MAX** in the following) the sequence of representation vectors is treated as a bag and we apply a form of multi-instance learning: each “feature”  $\phi_j$  will be positive if at least one of  $u_{j,1}, \dots, u_{j,T}$  is positive (see also [32]). The resulting classifier will find it easy to create decision rules predicting a document as belonging to a certain class if a given set of contextualized word representations are present and another given set of contextualized word representations are absent in the sequence. Note that this aggregator can also be interpreted as a kind of hard attention mechanism where attention concentrates completely on a single time step but the attended time step is different for each feature  $\phi_j$ . As detailed in Section III-C, a new model interpretation strategy can be derived when using this aggregator.

### C. Interpretable model

An interpretable model can be used to assist the manual classification process routinely performed in tumor registries and to explain the proposed automatic classification for further human inspection. To this end, the plain model (Eqs. 1–6) can be modified as follows:

$$e_t = E(x_t; \theta^e), \quad (7)$$

$$h_t^f = F(e_t, h_{t-1}^f; \theta^f), \quad (8)$$

$$h_t^r = R(e_t, h_{t+1}^r; \theta^r), \quad (9)$$

$$u_t = G(h_t; \theta^h), \quad (10)$$

$$f(\mathbf{x}) = A(\mathbf{u}; \theta^a), \quad (11)$$

where  $E, F, R, G$  and  $A$  are defined as in Section III-B and the size of  $u_t$  is forced to equal the number of classes so that each component  $u_{j,t}$  of  $u_t$  will be associated with the importance of words around position  $t$  for class  $j$ . This information can be

used to interpret the model decision. Preliminary experiments showed that the interpretation using the attention aggregator was not satisfactory. Therefore in the experiments we only report the interpretable model with the max aggregator that we call **MAXi**. More details on the preliminary experiments are reported in [33]. Besides accuracy, we are also interested in the average agreement between **MAXi** and **MAX** (i.e. the fidelity of the interpretable classifier, see Appendix III for a definition).

### D. Hierarchical models

The last two models in Section III-B can be extended in a hierarchical fashion, as suggested in [27]. In this case, data  $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}$  consist of variable length sequences of sequence vectors  $\mathbf{x}^{(i)}$ , where, for  $s = 1, \dots, S^{(i)}$  and  $t = 1, \dots, T_s^{(i)}$ ,  $x_{s,t}^{(i)}$  is the  $t$ -th word of the  $s$ -th sentence in the  $i$ -th document, and  $y^{(i)} \in \{1, \dots, K\}$  is the associated target class. The prediction  $f(\mathbf{x})$  is calculated as:

$$e_{s,t} = E(x_{s,t}; \theta^e), \quad (12)$$

$$h_{s,t}^f = F(e_{s,t}, h_{s,t-1}^f; \theta^f), \quad (13)$$

$$h_{s,t}^r = R(e_{s,t}, h_{s,t+1}^r; \theta^r), \quad (14)$$

$$u_{s,t} = G(h_{s,t}; \theta^h), \quad (15)$$

$$\phi_s = A(\mathbf{u}_s; \theta^a), \quad (16)$$

$$\bar{h}_s^f = \bar{F}(\phi_s, \bar{h}_{s-1}^f; \bar{\theta}^f), \quad (17)$$

$$\bar{h}_s^r = \bar{R}(\phi_s, \bar{h}_{s+1}^r; \bar{\theta}^r), \quad (18)$$

$$\bar{\phi} = \bar{A}(\bar{\mathbf{h}}; \bar{\theta}^a), \quad (19)$$

$$f(\mathbf{x}) = g(\bar{\phi}; \theta^c). \quad (20)$$

As in the plain model,  $E$  is an embedding function,  $F$  and  $R$  correspond to forward and reverse dynamics that process word representations,  $h_{s,t} = h_{s,t}^f \oplus h_{s,t}^r$  is the latent representation of the information contained at position  $t$  of the  $s$ -th sentence,  $u_{s,t}$  is the contextualized representation of the word at position  $t$  of the  $s$ -th sentence, and  $A$  is an aggregation function that creates a single representation for the sentence. Furthermore,  $\bar{F}$  and  $\bar{R}$  correspond to forward and reverse dynamics that process sentence representations, and  $\bar{A}$  is the aggregation function that creates a single representation for the entire document.  $\bar{h}_s = \bar{h}_s^f \oplus \bar{h}_s^r$  can be interpreted as the latent representation of the information contained in the sentence  $s$  for the document. We call **MAXh** and **ATTTh** the hierarchical versions of **MAX** and **ATT**, respectively.

### E. Word Vectors

Most algorithms for obtaining word vectors are based on co-occurrences in large text corpora. Co-occurrence can be measured either at the word-document level (e.g. using latent semantic analysis) or at the word-word level (e.g. using word2vec [19] or Global Vectors (GloVe) [20]). It is a common practice to use pre-compiled libraries of word vectors trained on several billion tokens extracted from various sources such as Wikipedia, the English Gigaword 5, Common Crawl, or Twitter. These libraries are generally conceived for general purpose applications and are only available for the English

language. Reports in cancer registries, however, are normally written in the local language and make extensive usage of a very specific domain terminology. In fact they can be considered *sublanguages* with a specific vocabulary usage and with peculiar sentence construction rules that differ from the normal construction rules [34].

Another approach is to employ a Language Model (LM) that models language as a sequence of characters instead of words. In particular, in the *Flair* framework [35], the internal states of a trained character level LM are used to produce *contextual string* word embeddings.

## F. Baselines

1) *Linear classifiers*: The classic approach is to employ bag-of-words representations of textual documents. Vector representations of documents are easily derived from bags-of-words either by using indicator vectors or taking into account the number of occurrences of each word using the Term-Frequency Inverse-Document-Frequency (TF-IDF) [36]. In those representations, frequent and non-specific terms receive a lower weight.

Bag-of-words representations (including those employing bigrams or trigrams) enable the application of linear text classifiers, such as Naive Bayes (NB), Support Vector Machine (SVM) [37], or boosted tree classifiers [38]. Those representations suffer two fundamental problems: first, the relative order of terms in the documents is lost, making it impossible to take advantage of the syntactic structure of the sentences; second, distinct words have an orthogonal representation even when they are semantically close. Word vectors can be used to address the second limitation and also allow us to take advantage of unlabeled data, which can be typically be obtained in large amounts and with little cost.

2) *CNN*: Convolutional Neural Network (CNN) can be successfully employed in the context of sentence classification [31]. The CNN model that we trained in our work is a slight variant of the architecture in [31]. The original architecture produces features maps applying convolutional filters on the sequence of word vectors followed by a max pooling and the classification. We used three convolutional layers with filter size of 3, 4 and 5. Moreover we added a linear layer between the word vectors and the convolutional layers. We fine-tuned hyperparameters for the output size of the linear layer and the number of convolutional filters. The input size of the linear layer is the same as the word vector size.

3) *BERT*: BERT [21] is a recent model that represents the state of the art in many NLP related tasks [39]–[42]. It is a bi-directional pre-training model backboneed by the Transformer Encoder [43]. It is an attention-based technique that learns context-dependent word representation on large unlabeled corpora, and then the model is fine tuned end-to-end on specific labeled tasks. During pre-training, the model is trained on unlabeled data over two different tasks. In Masked Language Model (MLM) some tokens are masked and the model is trained to predict those token based on the context. In Next Sentence Prediction (NSP) the model

is trained to understand the relationship between sentences predicting if two sentences are actually consecutive or if they were randomly replaced (with 50% probability).

In our work we pre-trained BERT using the same set of 1.5 million unlabeled records that we used to train word embeddings (see Section IV for details). Then we fine tuned BERT with the specific topography and morphology prediction tasks.

## G. Hyperparameters

All deep models (*GRU*, *MAX*, *ATT*, *MAXi*, *MAXh*, and *ATTTh*) were trained by minimizing the categorical cross entropy loss with Adam [44] with an initial learning rate of 0.001 and minibatches of 32 samples. The remaining hyperparameters (including  $C$  for SVM) were obtained by grid search using the validation accuracy as the objective (see Appendix II for optimal values and details on the hyperparameter space). In particular, we tuned hyperparameters in (1) - (6) and (12) - (20), which control the structure of the model.

$\xi^e$  is associated with the embedding layer  $E$  and in our case refers to GloVe hyperparameters [20]. With an intrinsic evaluation, we found that the better configuration was 60 for the vector size, 15 for the window size, and 50 iterations. We constructed sets of couples of related words, i.e. 11, 12, 11, 7 and 92 couples for respectively the benign-malignant, benign-tissue, malignant-tissue, morphology-site and singular-plural relations. For example, *fibroma*, *fibrosarcoma* and *lipoma*, *liposarcoma* for the benign-malignant relation and *fibroma*, *connective* and *lipoma adipose* for the cancer-tissue relation. We then used those sets to evaluate if the semantic relations are captured by linear substructures in the space of the embeddings, e.g. we measure if  $E(\text{fibrosarcoma}) - E(\text{fibroma}) + E(\text{lipoma}) \approx E(\text{liposarcoma})$  for the benign-malignant relation and  $E(\text{fibroma}) - E(\text{connective}) + E(\text{adipose}) \approx E(\text{lipoma})$  for the cancer-tissue relation. We confirmed the parameters with an extrinsic evaluation on the best model by grid search in the space of  $[2, \dots, 20]$  for window size and  $[40, \dots, 300]$  for vector dimension.

$\xi^f$ ,  $\xi^r$ ,  $\bar{\xi}^f$ , and  $\bar{\xi}^r$  define the number of GRU layers ( $\xi_{(l)}$ ) and the number of unit per each layer ( $\xi_{(d)}$ ) respectively for  $F$ ,  $R$ ,  $\bar{F}$ , and  $\bar{R}$ .  $G$  is a MLP,  $\xi^h$  controls the number of layers ( $\xi_{(l)}^h$ ) and their size ( $\xi_{(d)}^h$ ). Regarding  $F$ ,  $R$ , and  $G$ , we decided to have all the stacked layer with the same size to limit the hyperparameters space.  $\xi^a$  and  $\bar{\xi}^a$  control the kind of aggregating function of  $A$  and  $\bar{A}$  respectively and, in case of *attention*, it controls the size of the attention layer ( $\xi_{(d)}^a$ ). Finally,  $\xi^c$  controls the data-dependent output size of  $g$ .

## IV. RESULTS

In the experiments reported below word vectors were computed by GloVe [20] trained on our set of 1.5 millions unlabeled records. In a set of preliminary experiments, we also compared the best model that we obtained using GloVe embeddings against the same model trained using Flair embeddings obtained using a LM trained on the same unlabeled records. Although Flair has the potential advantage of robustness with respect to typos and spelling variants, extrinsic results on

the topography and the morphology tasks did not show any advantages over GloVe. For example test-set accuracy attained on topography by the best model, *MAX*, were slightly worse with Flair embeddings (89.9%) than with GloVe embeddings (90.3%) (the latter is reported in Table I).

TABLE I

TOPOGRAPHY SITE PREDICTION (61 CLASSES), SIGNIFICANCE AGAINST *MAX* (\*:  $p < 10^{-2}$ ; \*\*:  $p < 10^{-3}$ ; \*\*\*:  $p < 10^{-4}$ )

	Accuracy	Top 3 Acc.	Top 5 Acc.	MacroF1
<i>SVM</i>	89.7**	95.9***	96.8***	60.0
<i>CNN</i>	89.2***	96.0***	97.6***	55.3***
<i>GRU</i>	89.9*	96.5	97.7***	58.3**
<i>BERT</i>	89.9*	96.3*	97.8*	56.6*
<i>MAXi</i>	88.0***	95.4***	96.2***	46.1***
<i>MAXh</i>	89.9*	96.2***	97.8*	58.8*
<i>ATTh</i>	89.9	96.3**	97.7**	58.0**
<i>MAX</i>	<b>90.3</b>	<b>96.6</b>	<b>98.1</b>	<b>61.9</b>
<i>ATT</i>	90.1	96.2***	97.6***	60.0

TABLE II

MORPHOLOGY TYPE PREDICTION (134 CLASSES), SIGNIFICANCE AGAINST *MAX* (\*:  $p < 10^{-2}$ ; \*\*:  $p < 10^{-3}$ ; \*\*\*:  $p < 10^{-4}$ )

	Accuracy	Top 3 Acc.	Top 5 Acc.	Macro F1
<i>SVM</i>	82.4***	94.0***	95.6***	53.7**
<i>CNN</i>	83.3***	94.4***	96.7	55.0**
<i>GRU</i>	83.3***	94.6*	96.6*	55.2**
<i>BERT</i>	84.3	93.2***	94.9***	51.1***
<i>MAXi</i>	73.4***	91.0***	93.6***	31.3***
<i>MAXh</i>	83.7***	94.4***	96.4***	54.5*
<i>ATTh</i>	83.7***	94.4***	96.2***	57.5
<i>MAX</i>	84.6	<b>95.0</b>	<b>96.9</b>	59.2
<i>ATT</i>	<b>84.8</b>	94.9	<b>96.9</b>	<b>61.3</b>

TABLE III

MACRO F1 MEASURE BY GROUPS OF CLASS FREQUENCY, SIGNIFIC. AGAINST *MAX* (\*:  $p < 10^{-2}$ ; \*\*:  $p < 10^{-3}$ ; \*\*\*:  $p < 10^{-4}$ )

	Topography			Morphology		
	easy (4 cls)	avg. (18 cls)	hard (39 cls)	easy (5 cls)	avg. (18 cls)	hard (111 cls)
<i>SVM</i>	95.7*	<b>86.9</b>	50.9	90.5	68.6	48.4*
<i>CNN</i>	95.6	71.0**	43.1***	91.7*	70.5	49.2**
<i>GRU</i>	<b>96.1</b>	72.2	48.0*	91.4	71.6	49.7**
<i>BERT</i>	95.7	73.2	44.9*	<b>92.9</b>	<b>74.4</b>	43.9***
<i>MAXi</i>	95.0	66.6	31.4***	87.1	41.9**	25.1***
<i>MAXh</i>	95.8	72.4	48.8*	92.7	71.8	48.8*
<i>ATTh</i>	96.0	73.1	47.1**	91.9	72.3	52.6
<i>MAX</i>	96.0	73.3	<b>53.1</b>	92.7	72.3	53.8
<i>ATT</i>	96.0	73.1	50.3	92.8	72.3	<b>56.7</b>

In Table I and Table II we summarize the results of different models on test data in terms of multiclass accuracy (or, equivalently, micro-averaged F1 measure), top- $\ell$  accuracy (if the correct class appears within the top  $\ell$  predictions) for  $\ell = 3$  and  $\ell = 5$ , and macro-averaged F1 measure (see Appendix III for definitions). Significance (each method against *MAX*) is reported with asterisks in the tables and was assessed with a one-sided McNemar test [45] for accuracy and with a one-sided macro T-test [46] for F1 score.

Collecting results for all the models (for a single hyperparameters configuration and excluding the training of word vectors and *BERT*) required approximately 11 hours on a

GeForce RTX 2080 Ti GPU<sup>5</sup>. In Table III we report F1 score averaged on different subsets of classes. We consider a class *easy* if it has more than 1000 examples in the test set, *average* if it has between 100 and 1000 examples, and *hard* if it has less than 100 examples.

In the case of topography, when focusing on the performance on classes with many examples, all models tend to perform similarly, with even the interpretable model attaining high F1 scores. The advantage of recurrent networks over bag-of-word representations is more pronounced when focusing on rare classes. One possible explanation is that the representation learned by recurrent networks is shared across all classes, leveraging the advantage of multi-task learning [47] in this case. We also note that in no case hierarchical attention models outperform flat attention models and max-pooling performs the best on rare classes. In the case of morphology, differences among different models are more pronounced, with BERT being very effective for densely populated classes (but not for rare classes). Again hierarchical attention does not outperform flat attention. This result differs from the ones reported in [16] but the datasets are very different in terms of number of examples and number of classes. Differences in the writing style of pathologist trained and practicing in different countries could also impact the relative performance of different models. In this respect, our documents contain on average fewer sentences (see Figure 3 in Appendix I), offering less structure to be exploited by the richer hierarchical models.

The interpretable classifier *MAXi* can be used to explain prediction by highlighting which portions of the text contribute to which classes. Its average agreement with *MAX* was 91.8% on topography and 78.3% on morphology. In Figure 1, we show three examples (topography task) where terms are underlined by class-specific colors and with intensities proportional to the importance  $u_{j,t}$  of word in position  $t$  for class  $j$  (see (10)): high if  $u_{j,t} > 0.8$ , medium if  $u_{j,t} \in [0.3, 0.8)$ , low if  $u_{j,t} \in [0.1, 0.3)$ , not highlighted if  $u_{j,t} < 0.1$ . We consider class  $j$  to be relevant to the document if at least one word has  $u_{j,t} \geq 0.1$ .

The first report was correctly classified and the two most relevant words are *prostatico* (*prostatic*), *prostata* (*prostate*), followed by *PSA* (Prostate-Specific Antigene) and *Gleason* score, that are two common exams in prostate cancer cases [48]. For the second report, the model proposes three codes: *18*, *20* and *21*, suggesting that *intestinal tubular adenoma* and *pedunculated polypus* are terms associated with class colon, *polypus* associated with colon and rectum, and *anal orifice* associated with rectum and anus. Note that the ground truth for this record was rectum, while the text explicitly mentions that the fragments have been extracted at 20 cm from the anal orifice (the human rectum is approximately 12 cm long and the anal canal 3-5 cm [49]). The third report is an even more complex case where the model proposes codes *34*, attached to *plurial effusion* and *lung thickening*, but interestingly also underlines the immunohistochemical results, as the pattern *CK7+* *CK20-* commonly indicates a diagnosis of lung origin for

<sup>5</sup>The source code for the experiments is available at the following address: <https://github.com/trianam/cancerReportsClassification>

Class	Relevant classes	Document text with highlighted words	English translation (by the authors)
61	61 (PROSTATE GLAND)	DISOMOGENICITA' / DIFFUSE . PSA NON PERVENUTO . ADENOCARCINOMA PROSTATICO A GRADO DI DIFFERENZIAZIONE MEDIO - BASSO (GLEASON 3 + 4 ) NEI PRELIEVI DI CUI AI NN . 2 E 3 . AGOBIOPSIA DELLA PROSTATA : 1 ) 1 PRELIEVO LL DX . 2 ) 2 PRELIEVI ML DX . 3 ) 2 PRELIEVI M DX . 4 ) 1 PRELIEVO M SX . 5 ) 2 PRELIEVI ML SX . 6 ) 1 PRELIEVO LL SX . 7 ) 1 PRELIEVO TRANSIZIONALE SX . 8 ) 1 PRELIEVO TRANSIZIONALE DX .	DIFFUSE DISHOMOGENEITY . PSA NOT RECEIVED . PROSTATIC ADENOCARCINOMA OF INTERMEDIATE - LOW GRADE OF DIFFERENTIATION ( GLEASON 3 + 4 ) IN SAMPLES AT N . 2 AND 3 . NEEDLE BIOPSY OF THE PROSTATE : 1 ) 1 RIGHT LL SAMPLE . 2 ) 2 RIGHT ML SAMPLES . 3 ) 2 RIGHT M SAMPLES . 4 ) 1 LEFT M SAMPLE . 5 ) 2 LEFT ML SAMPLES . 6 ) 1 LEFT LL SAMPLE . 7 ) 1 LEFT TRANSITIONAL SAMPLE . 8 ) 1 RIGHT TRANSITIONAL SAMPLE .
20	18 (COLON) 20 (RECTUM) 21 (ANUS AND ANAL CANAL)	ISOLATI FRAMMENTI RIFERIBILI AD ADENOMA TUBULARE INTESTINALE DI ALTO GRADO . FRAMMENTI ( NR . 2 ) DI POLIPO PEDUNCOLATO A 20 CM DALL' ORIFIZIO ANALE . ( ESEGUITA COLORAZIONE EMATOSSILINA - EOSINA ) .	ISOLATED FRAGMENTS ATTRIBUTABLES TO HIGH DEGREE INTESTINAL TUBULAR ADENOMA . FRAGMENTS ( NR . 2 ) OF PEDUNCULATED POLYPUS AT 20 CM FROM THE ANAL ORIFICE . ( PERFORMED HEMATOXYLIN - EOSIN COLORING ) .
34	34 (BRONCHUS AND LUNG) 56 (OVARY) 67 (BLADDER) 80 (UNKNOWN PRIMARY SITE)	VERSAMENTO PLEURICO SX DI N . D . D . E ADDENSAMENTI POLMONARI DI N . D . D . , NODULI PARETE ADDOMINALE . INFILTRAZIONE CANCERIGNA DEGLI STROMI CONNETTIVO - ADIPOSI . IMMUNOISTOCHEMICA : CK7 + , CK20 - , TTF - 1 - , PROTEINA S - 100 - . LESIONE DI CM 2 , 0 X 1 , 3 X 0 , 7 . 1 - 2 ) SEZIONI SERIATE .	LEFT PLEURAL EFFUSION OF UNKNOWN ORIGIN AND LUNG THICKENING OF UNKNOWN ORIGIN , ABDOMINAL WALL NODULES . CANCEROUS INFILTRATION OF THE CONNECTIVE - ADIPOSE STROMA . IMMUNOHISTOCHEMICAL : CK7 + , CK20 - , TTF - 1 - , PROTEIN S - 100 - . 2 CM LESION , 0 X 1 , 3 X 0 , 7 . 1 - 2 ) SERIAL SECTIONS .

Fig. 1. Three sample reports annotated by the interpretable model (underline intensity proportional to class importance).

metastatic adenocarcinoma [50]. Also, immunohistochemistry is a common approach in the diagnosis of tumors of uncertain origin [51]. This can be the reason for the underlying with code 80 of the immunoistochemical part. It is interesting to note that *pleuric* is suggested to be related to ovarian cancer, in fact the pleural cavity constitutes the most frequent site for extra-abdominal metastasis in ovarian carcinoma [52].

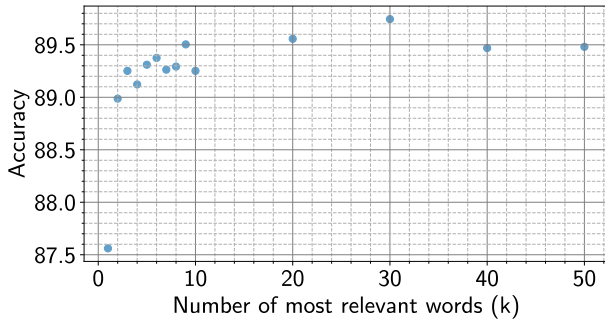


Fig. 2. Training of a plain GRU model on topography task on datasets created using **MAXi** to distill the most relevant  $k$  words.

To quantify the effectiveness of the interpretable model, we designed an experiment where a set of datasets is created taking only the most relevant  $k$  words based on the value of  $u_t$  in (10) of **MAXi** for the topography site prediction. In Figure 2, we plot the accuracy obtained training a plain GRU model on those reduced datasets, for increasing values of  $k$ . Accuracy is high even when selecting only a few words, suggesting that the interpretable model is effective in distilling the most relevant terms, and that the information contained in texts tends to be concentrated in a small number of terms.

## V. CONCLUSIONS

We compared different algorithms on a large scale dataset of more than 80 000 labeled records from the Tuscany region tumor registry collected between 1990 and 2014. Results confirm the viability of automated assignment of ICD-O3 codes, with an accuracy of 90.3% on topography (61 classes) and 84.8% on morphology (134 classes). Top-5 accuracies (fraction of test documents whose correct label is among the top five model's prediction) were 98.1% and 96.9% for topography and morphology, respectively. The latter rates decreased only to 96.2% (topography) and 93.6% (morphology) when using an interpretable model that highlights the most important terms in the text.

In this specific context we did not obtain significant improvements using hierarchical attention methods, compared to a simple max pooling aggregation. The difference between deep learning models and more traditional approaches based on bag-of-words with SVM is significant but not as pronounced as in the results reported in other studies. We also found that a large window size (15 words) and relatively small dimensionality (60) works better for construction of word vectors, while other works in biomedical field [53] found better results with smaller window size larger word vector dimensionality. These differences can be explained, at least in part, with the specificity of the corpus used in this study, where reports tend to be short, synthetic, rich in discriminant keywords, and often lacking verb phrases. As shown in Figure 2, few words are sufficient to achieve good accuracy.

SVM perform well on topography class that are sufficiently well represented in the dataset. Also, we found that hierarchical models are not better than flat models and that a



simple max aggregation achieves the best results in most cases. Interestingly, hierarchical models are outperformed by flat attention or flat max pooling for the more difficult classes (those with less than 100 training examples). Rare classes remain however challenging for all current methods and as discussed in Section III-A our study, like all previous similar studies in this area, do not even consider extremely rare classes. In this respect, future work may consider the use of metalearning techniques capable of operating in the few-shot learning setting [54]–[56] in order to include more classes and to improve prediction accuracy on the underrepresented ones. Results in this study are limited to a specific (but large) Italian dataset and might be compared in the future against results obtained on cancer reports written in other languages.

## REFERENCES

- [1] R. Sullivan, J. Peppercorn, *et al.*, “Delivering affordable cancer care in high-income countries,” *The Lancet Oncology*, vol. 12, pp. 933–980, Sept. 2011.
- [2] B. Stewart and C. P. Wild, eds., *World Cancer Report 2014*. International Agency for Research on Cancer, WHO, Feb. 2014.
- [3] C. E. DeSantis, C. C. Lin, *et al.*, “Cancer treatment and survivorship statistics, 2014,” *CA: A Cancer Journal for Clinicians*, vol. 64, pp. 252–271, July 2014.
- [4] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2016: Cancer Statistics, 2016,” *CA: A Cancer Journal for Clinicians*, vol. 66, pp. 7–30, Jan. 2016.
- [5] M. Brown, J. Lipscomb, and C. Snyder, “The burden of illness of cancer: economic cost and quality of life,” *Annual Review of Public Health*, vol. 22, pp. 91–113, 2001.
- [6] G. Tourassi, “Deep learning enabled national cancer surveillance,” in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 3982–3983, Dec. 2017.
- [7] D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” *Artificial Intelligence in Medicine*, vol. 34, pp. 113–127, June 2005.
- [8] G. Mujtaba, L. Shuib, *et al.*, “Clinical text classification research trends: Systematic literature review and open issues,” *Expert Systems with Applications*, vol. 116, pp. 494–520, Feb. 2019.
- [9] W.-w. Yim, M. Yetisgen, W. P. Harris, and S. W. Kwan, “Natural Language Processing in Oncology: A Review,” *JAMA Oncology*, vol. 2, p. 797, June 2016.
- [10] O. M. Jensen, *Cancer registration: principles and methods*, vol. 95, ch. 5 Data sources and reporting, pp. 35–48. IARC, 1991.
- [11] M. Colombet, S. Antoni, and J. Ferlay, *Cancer incidence in five continents*, vol. 11, ch. 6 Data Processing. Lyon: International Agency for Research on Cancer, 2017.
- [12] S. Ferretti, A. Giacomini, *et al.*, *Cancer Registration Handbook*. AIRTUM, January 2008.  
<https://www.registri-tumori.it/cms/publicazioni/cancer-registrations-handbook-2010>.
- [13] A. Fritz, C. Percy, A. Jack, K. Shanmugaratnam, L. Sobin, D. M. Parkin, and S. Whelan, eds., *International classification of diseases for oncology*. Geneva: World Health Organization, 3 ed., 2000.
- [14] V. Jouhet, G. Defosse, A. Burgun, P. Le Beux, P. Levillain, P. Ingrand, and V. Claveau, “Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer,” *Methods of Information in Medicine*, vol. 51, pp. 242–251, July 2011.
- [15] R. Kavuluru, I. Hands, E. B. Durbin, and L. Witt, “Automatic extraction of ICD-O-3 primary sites from cancer pathology reports,” in *Clinical Research Informatics AMIA symposium*, 2013.
- [16] S. Gao, M. T. Young, J. X. Qiu, H.-J. Yoon, J. B. Christian, P. A. Fearn, G. D. Tourassi, and A. Ramanathan, “Hierarchical attention networks for information extraction from cancer pathology reports,” *Journal of the American Medical Informatics Association*, vol. 25, pp. 321–330, Mar. 2018.
- [17] J. X. Qiu, H.-J. Yoon, P. A. Fearn, and G. D. Tourassi, “Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 244–251, Jan. 2018.
- [18] M. Alawad, S. Gao, J. X. Qiu, H. J. Yoon, J. Blair Christian, L. Penberthy, B. Mumphy, X.-C. Wu, L. Coyle, and G. Tourassi, “Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks,” *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 89–98, 2020.
- [19] T. Mikolov, W.-t. Yih, and G. Zweig, “Linguistic Regularities in Continuous Space Word Representations,” in *Hlt-naacl*, vol. 13, pp. 746–751, 2013.
- [20] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global Vectors for Word Representation,” in *EMNLP*, vol. 14, pp. 1532–1543, 2014.
- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [22] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” in *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, 2014. arXiv:1409.1259.
- [23] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, 2015.
- [24] E. Crocetti, C. Sacchetti, A. Caldarella, and E. Paci, “Automatic coding of pathologic cancer variables by the search of strings of text in the pathology reports. The experience of the Tuscany Cancer Registry,” *Epidemiologia e prevenzione*, vol. 29, no. 1, pp. 57–60, 2004.
- [25] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, and P. C. De Groen, “Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model,” *Journal of biomedical informatics*, vol. 42, no. 5, pp. 937–949, 2009.
- [26] A. N. Nguyen, J. Moore, J. O’Dwyer, and S. Colquist, “Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports,” in *American Medical Informatics Association Annual Symposium*, 2015.
- [27] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical Attention Networks for Document Classification,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 1480–1489, June 2016.
- [28] M. T. Ribeiro, S. Singh, and 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*, pp. 1135–1144, 2016.
- [29] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, “End-to-end memory networks,” in *Advances in Neural Information Processing Systems 28*, pp. 2440–2448, 2015.
- [30] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa, “Natural language processing (almost) from scratch,” *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.
- [31] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pp. 1746–1751, 2014.
- [32] A. Tibo, P. Frasconi, and M. Jaeger, “A network architecture for multi-multi-instance learning,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–752, Springer, 2017.
- [33] S. Martina, *Classification of cancer pathology reports with Deep Learning methods*. PhD thesis, University of Florence, 2020.
- [34] P. Spyns, “Natural language processing in medicine: an overview,” *Methods of information in medicine*, vol. 35, no. 04/05, pp. 285–301, 1996.
- [35] A. Akbik, D. Blythe, and R. Vollgraf, “Contextual string embeddings for sequence labeling,” in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649, 2018.
- [36] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [37] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273–297, Sep 1995.
- [38] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 2016.
- [39] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, “Semeval-2019 task 3: Emocontext contextual emotion detection in text,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 39–48, 2019.

- [40] D. Hu, “An introductory survey on attention mechanisms in nlp problems,” in *Proceedings of SAI Intelligent Systems Conference*, pp. 432–448, Springer, 2019.
- [41] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: pre-trained biomedical language representation model for biomedical text mining,” *arXiv preprint arXiv:1901.08746*, 2019.
- [42] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, “Unsupervised word embeddings capture latent knowledge from materials science literature,” *Nature*, vol. 571, no. 7763, p. 95, 2019.
- [43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [44] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [45] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural Computation*, vol. 10, pp. 1895–1923, Oct 1998.
- [46] Y. Yang and X. Liu, “A re-examination of text categorization methods,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42–49, 1999.
- [47] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
- [48] F. Brimo, R. Montironi, L. Egevad, A. Erbersdobler, D. W. Lin, J. B. Nelson, M. A. Rubin, T. van der Kwast, M. Amin, and J. I. Epstein, “Contemporary grading for prostate cancer: Implications for patient care,” *European Urology*, vol. 63, no. 5, pp. 892 – 901, 2013.
- [49] F. Greene, C. Compton, A. J. C. on Cancer, A. Fritz, J. Shah, and D. Winchester, *Ajcc Cancer Staging Atlas*. Springer New York, 2006.
- [50] S. Kummar, M. Fogarasi, A. Canova, A. Mota, and T. Ciesielski, “Cytokeratin 7 and 20 staining for the diagnosis of lung and colorectal adenocarcinoma,” *British journal of cancer*, vol. 86, no. 12, p. 1884, 2002.
- [51] J. Duraiyan, R. Govindarajan, K. Kaliyappan, and M. Palanisamy, “Applications of immunohistochemistry,” *Journal of pharmacy & bioallied sciences*, vol. 4, no. Suppl 2, p. S307, 2012.
- [52] J. M. Porcel, J. P. Diaz, and D. S. Chi, “Clinical implications of pleural effusions in ovarian cancer,” *Respirology*, vol. 17, no. 7, pp. 1060–1067, 2012.
- [53] B. Chiu, G. K. O. Crichton, A. Korhonen, and S. Pyysalo, “How to train good word embeddings for biomedical NLP,” in *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pp. 166–174, 2016.
- [54] J. Snell, K. Swersky, and R. S. Zemel, “Prototypical networks for few-shot learning,” in *Advances in Neural Information Processing Systems 30*, pp. 4077–4087, 2017.
- [55] S. Ravi and H. Larochelle, “Optimization as a model for few-shot learning,” in *5th International Conference on Learning Representations, ICLR 2017*, 2017.
- [56] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, “Matching networks for one shot learning,” in *Advances in Neural Information Processing Systems 29*, pp. 3630–3638, 2016.

## APPENDIX I DATASET STATISTICS

We report in Figure 3 and in Figure 4 some distributions of the dataset used in this study.

## APPENDIX II HYPERPARAMETER OPTIMIZATION

We report here domains and optimal values (underlined) for the hyperparameters of the models used in our experiments.

In **MAX** we used the *max* aggregation function in the plain model of Section III-D. The hyperparameters space was:

$$\begin{aligned}\xi_{(l)}^f &= \xi_{(l)}^r \in [1, 2], & \xi_{(l)}^h &\in [1, 2, 4], \\ \xi_{(d)}^f &= \xi_{(d)}^r \in [2, 4, 8, 16, 32, 64, \underline{128}, 256, 512], \\ \xi_{(d)}^h &\in [2, 4, 8, 16, 32, 64, 128, 256, \underline{512}, 1024, 2048],\end{aligned}$$

for the *topography* site task, and:

$$\begin{aligned}\xi_{(l)}^f &= \xi_{(l)}^r \in [1], & \xi_{(l)}^h &\in [1, 2, 4], \\ \xi_{(d)}^f &= \xi_{(d)}^r \in [2, 4, 8, 16, 32, 64, \underline{128}, 256, 512], \\ \xi_{(d)}^h &\in [2, 4, 8, 16, 32, 64, \underline{128}, 256, 512, 1024, 2048],\end{aligned}$$

for the *morphology* type task.

In **ATT** we used the *attention* aggregation function in the plain model. The hyperparameters space was:

$$\begin{aligned}\xi_{(l)}^f &= \xi_{(l)}^r \in [1], & \xi_{(l)}^h &\in [0, 1], \\ \xi_{(d)}^f &= \xi_{(d)}^r \in [64, \underline{128}, 256], & \xi_{(d)}^h &\in [256, \underline{512}, 1024], \\ \xi_{(d)}^a &\in [128, \underline{256}, 512, 1024],\end{aligned}$$

for the site, and:

$$\begin{aligned}\xi_{(l)}^f &= \xi_{(l)}^r \in [1], & \xi_{(l)}^h &\in [0, 1], \\ \xi_{(d)}^f &= \xi_{(d)}^r \in [64, 128, \underline{256}], & \xi_{(d)}^h &\in [64, \underline{128}, 256], \\ \xi_{(d)}^a &\in [128, \underline{256}, 512, 1024],\end{aligned}$$

for the morphology.

In **MAXh** we used the *max* aggregation in the hierarchical model of Section III-D. The hyperparameters space was:

$$\begin{aligned}\xi_{(l)}^f &= \xi_{(l)}^r = \bar{\xi}_{(l)}^f = \bar{\xi}_{(l)}^r \in [1], & \xi_{(l)}^h &\in [0, 1, 2, 4], \\ \xi_{(d)}^f &= \xi_{(d)}^r = \bar{\xi}_{(d)}^f = \bar{\xi}_{(d)}^r \in [32, \underline{64}, 128, 256], \\ \xi_{(d)}^h &\in [256, 512, \underline{1024}, 2048],\end{aligned}$$

for the topography, and:

$$\begin{aligned}\xi_{(l)}^f &= \xi_{(l)}^r = \bar{\xi}_{(l)}^f = \bar{\xi}_{(l)}^r \in [1], & \xi_{(l)}^h &\in [0, 1, 2, 4], \\ \xi_{(d)}^f &= \xi_{(d)}^r = \bar{\xi}_{(d)}^f = \bar{\xi}_{(d)}^r \in [32, \underline{64}, 128, 256], \\ \xi_{(d)}^h &\in [256, 512, \underline{1024}, 2048],\end{aligned}$$

for the morphology.

In **ATT**h we used the *attention* aggregation in the hierarchical model. The hyperparameters space was:

$$\begin{aligned}\xi_{(l)}^f &= \xi_{(l)}^r = \bar{\xi}_{(l)}^f = \bar{\xi}_{(l)}^r \in [1], & \xi_{(l)}^h &\in [0, 1, 2, 4], \\ \xi_{(d)}^f &= \xi_{(d)}^r = \bar{\xi}_{(d)}^f = \bar{\xi}_{(d)}^r \in [32, 64, \underline{128}, 256], \\ \xi_{(d)}^h &\in [256, 512, \underline{1024}, 2048], \\ \xi_{(d)}^a &= \bar{\xi}_{(d)}^a \in [64, \underline{128}, 256, 512],\end{aligned}$$

for the topography, and:

$$\begin{aligned}\xi_{(l)}^f &= \xi_{(l)}^r = \bar{\xi}_{(l)}^f = \bar{\xi}_{(l)}^r \in [1], & \xi_{(l)}^h &\in [0, 1, 2, 4], \\ \xi_{(d)}^f &= \xi_{(d)}^r = \bar{\xi}_{(d)}^f = \bar{\xi}_{(d)}^r \in [32, \underline{64}, 128, 256], \\ \xi_{(d)}^h &\in [256, 512, \underline{1024}, 2048], \\ \xi_{(d)}^a &= \bar{\xi}_{(d)}^a \in [64, \underline{128}, 256, 512],\end{aligned}$$

for the morphology.

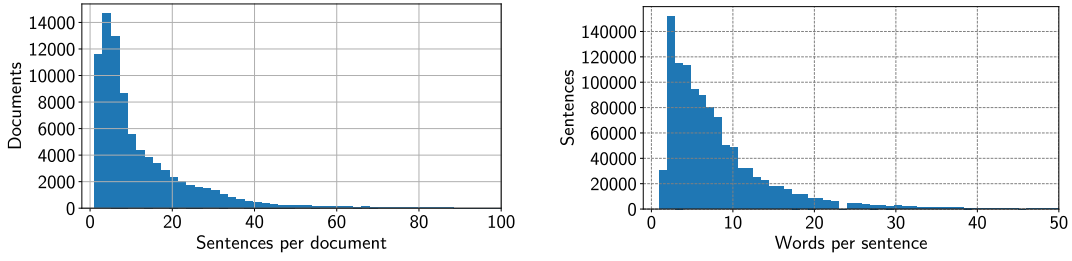


Fig. 3. Distribution of the number of sentences per document (left) and the number of words per sentence (right).

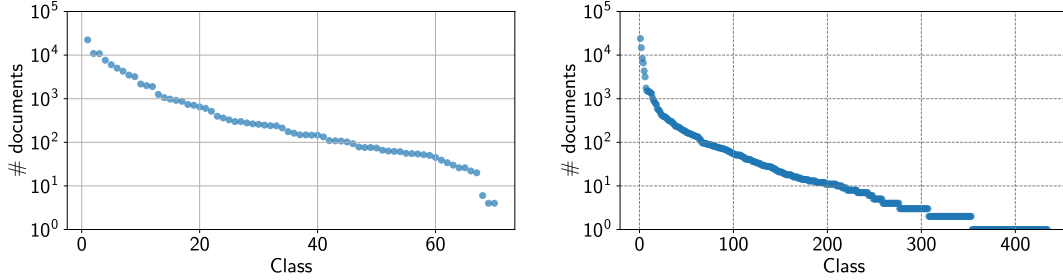


Fig. 4. Class distributions for topography (left) and morphology (right)

In **MAXi** we used the *max* aggregation in the plain model. Also we set the model to be interpretable. The hyperparameters space was:

$$\begin{aligned} \xi_{(l)}^f &= \xi_{(l)}^r \in [1, 2, 4], & \xi_{(l)}^h &\in [1, 2, 4], \\ \xi_{(d)}^f &= \xi_{(d)}^r \in [2, 4, 8, 16, 32, 64, \underline{128}, 256, 512], \\ \xi_{(d)}^h &\in [2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048], \end{aligned}$$

for the topography, and:

$$\begin{aligned} \xi_{(l)}^f &= \xi_{(l)}^r \in [1, 2, 4], & \xi_{(l)}^h &\in [1], \\ \xi_{(d)}^f &= \xi_{(d)}^r \in [64, 128, \underline{256}, 512], & \xi_{(d)}^h &\in [], \end{aligned}$$

for the morphology. Note that, in this setting, the size of the last layer of  $G$  must be equal to the output size of the model (and the softmax is applied directly after the aggregation  $A$ , without any layer). Thus,  $\xi_{(d)}^h$  refers only to the layers before the last one, if they exist.

Regarding **GRU**, we searched in a space of  $[1, 2, 4]$  number of layers of dimension in  $[128, 256, 512, 1024]$ . We found that the best configuration was using 2 layers of dimension 256.

### APPENDIX III PERFORMANCE MEASURES

We report in the following precise definitions of our performance measures.

- The multiclass accuracy is defined as

$$A \doteq \frac{1}{m} \sum_{i=1}^m \mathbb{1} \left\{ y^{(i)} = \arg \max_{j=1, \dots, K} f_j(x^{(i)}) \right\},$$

where  $\mathbb{1}\{\}$  denotes the indicator function and  $m$  is the number of test points (recall that  $f(x)$  denotes the vector of conditional probabilities assigned to each of the  $K$  classes). It is equivalent to micro-averaged F1 measure for mutually exclusive classes.

- The top- $\ell$  accuracy is defined as

$$A_\ell \doteq \frac{1}{m} \sum_{i=1}^m \mathbb{1} \left\{ y^{(i)} \in T_\ell \left( f(x^{(i)}) \right) \right\},$$

where  $T_\ell(a)$  denotes the operator that given array  $a = [a_1, \dots, a_K]$  as input returns the set  $\{\pi_1, \dots, \pi_\ell\}$  being  $\pi_1, \dots, \pi_K$  the permutation sequence that sorts  $a$  in descending order

- The macro-averaged F1 measure is defined as

$$F_1^M \doteq \frac{1}{K} \sum_{k=1}^K \frac{2P_k R_k}{P_k + R_k}$$

where

$$P_k = \frac{\sum_{i=1}^m \mathbb{1} \left\{ y^{(i)} = k \right\} \mathbb{1} \left\{ y^{(i)} = \arg \max_{j=1, \dots, K} f_j(x^{(i)}) \right\}}{\sum_{i=1}^m \mathbb{1} \left\{ k = \arg \max_{j=1, \dots, K} f_j(x^{(i)}) \right\}}$$

is the precision for class  $k$  and

$$R_k = \frac{\sum_{i=1}^m \mathbb{1} \left\{ y^{(i)} = k \right\} \mathbb{1} \left\{ y^{(i)} = \arg \max_{j=1, \dots, K} f_j(x^{(i)}) \right\}}{\sum_{i=1}^m \mathbb{1} \left\{ k = y^{(i)} \right\}}$$

is the recall for class  $k$ ;

- The fidelity is defined as

$$F \doteq \frac{1}{m} \sum_{i=1}^m \mathbb{1} \left\{ \arg \max_{j=1, \dots, K} f_j(x^{(i)}) = \arg \max_{j=1, \dots, K} g_j(x^{(i)}) \right\},$$

where  $f(x)$  and  $g(x)$  denote the vectors of conditional probabilities assigned to each of the  $K$  classes by the two models (**MAX** and **MAXi** in the paper).