IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 April 2024, accepted 8 May 2024, date of publication 16 May 2024, date of current version 23 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402066

APPLIED RESEARCH

Language Models for Hierarchical Classification of Radiology Reports With Attention Mechanisms, BERT, and GPT-4

MATTEO OLIVATO[1], LUCA PUTELLI[1], NICOLA ARICI[1], ALFONSO EMILIO GEREVINI[1], ALBERTO LAVELLI[2], AND IVAN SERINA[1]

[1]Department of Information Engineering, University of Brescia, 25121 Brescia, Italy
[2]Fondazione Bruno Kessler, 38123 Trento, Italy

Corresponding authors: Matteo Olivato (matteo.olivato@unibs.it) and Alfonso Emilio Gerevini (alfonso.gerevini@unibs.it)

ABSTRACT Radiology reports are a valuable source of textual information used to improve clinical care and support research. In recent years, deep learning techniques have been shown to be effective in classifying radiology reports. This article investigates the use of deep learning techniques with attention mechanisms to achieve better performance in the classification of radiology reports. We focus on various Natural Language Processing approaches, such as LSTM with Attention, BERT, and GPT-4, evaluated on a chest tomography report dataset regarding neoplastic diseases collected from an Italian hospital. In particular, we compare the results with a previous machine learning system, showing that models based on attention mechanisms can achieve higher performance. The Attention Mechanism allows us to identify the most relevant bits of text used by the model to make its predictions. We show that our model achieves state-of-the-art results on the hierarchical classification of radiology reports. Moreover, we evaluate the performance of GPT-4 on the classification of these reports in a zero-shot setup through prompt engineering, showing interesting results even with a small context and a non-English language. Our findings suggest that deep learning techniques with attention mechanisms may be successful in the classification of radiology reports even in non-English languages for which it is not possible to leverage on large text corpus.

INDEX TERMS Attention mechanism, BERT, BioBIT, deep learning, GPT-4, large language models, natural language processing, Italian language, prompt engineering, radiology reports, Italian radiology reports, text classification.

I. INTRODUCTION

In recent years, Artificial Intelligence (AI) and, in particular, Machine Learning (ML) have become a fundamental tool to solve many different tasks in complex domains, one for all medical and healthcare domains [1], [2], [3], [4]. The emergence of electronic health records (EHRs) has led to the accumulation of a large number of laboratory tests and narrative clinical texts within hospitals. These rich sources of information hold immense potential for medical research and improving patient care effectiveness and quality. Due to the unstructured nature of textual data, manual analysis is challenging and time-consuming. Consequently, ML and Natural Language Processing (NLP) techniques are employed to extract and organise content from clinical reports, making it readily accessible to radiologists. This study focusses on the classification of Italian chest tomography (CT) reports that follow a diagram proposed by radiologists at *Spedali Civili di Brescia* in Italy. The potential benefits of an accurate classification system for existing and new reports are numerous, covering logistics, healthcare management, follow-up examination frequency monitoring, and case collection for research or educational purposes [5].

Previously, a classification system for chest CT reports was integrated into the software used by the radiologists at Spedali Civili di Brescia for the writing of the report. On completion of the report, radiologists receive a real-time classification from the system, subject to confirmation or

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara.

modification as needed. The results of the initial evaluation were encouraging [6]; however, the performance in the real world was below expectations. Several factors contributed to this discrepancy:

1) The training set was limited in size because it is composed by only 346 reports of a unique radiologist annotated manually.
2) Real-world cases exhibited greater complexity compared to the training and test set used.
3) Radiologists employ varying writing and classification styles, resulting in reduced uniformity in the writing style.

Nowdays, the models that have shown remarkable success in data analysis for text documents [7], [8] are based on Deep Neural Networks (DNNs). Word embedding techniques such as Word2vec [9] and GloVe [10] are widely used to represent words in a vector space, capturing semantic information [11]. Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) networks have long been staples in various tasks of NLP, including text classification and data mining [12], [13]. These solutions process the entire sequence of words, enabling these models to grasp word dependencies and retain context information for the entire document. Moreover, the introduction of attention mechanisms [14], [15] has allowed models to "focus" on the most informative parts of the input. In the realm of classifying clinical texts, LSTM Neural Networks and the Attention Mechanism have achieved significant results even with a limited number of documents [16].

Lately, models fully based on the attention mechanism, such as Transformer-based language models (BERT), have been used with great success in Natural Language Processing (NLP) tasks, including machine translation and text classification, and have established a new standard [17], [18]. These models are composed of multiple encoders that gradually gain knowledge of words and learn semantic associations between different words. Each encoder produces a vector representation of each word and a vector that represents the entire sentence or document. Encoders are made up of several parallel self-attention mechanisms, called *heads*, among other components. For each word, a head computes a probability distribution that reflects the degree of association between the word and all other words in the document.

With the advent of the so-called Large Language Models (LLMs) [19], [20], which are essentially an evolution of the Transformer model, many works have been presented trying to solve specialised NLP tasks via LLMs using a technique called prompting. Prompting is the process of providing instructions (the prompt) to a LLM to guide its output [21], [22], [23]. The performance of an LLM on the resolution of a particular task is closely related to the quality of the given prompt. In particular, a good prompt aims to help the LLM understand what task to be solved for a particular piece of text, what the desired response style, what the correct output format, etc. Therefore, an effective prompt enables LLMs to perform a wide range of tasks [24], [25], [26]. General guidelines for good prompts suggest that the prompt should be clear and specific in instructions and provide enough context and examples. All these quality characteristics in practise are achieved by a trial and error procedure, trying different prompt styles and different contexts or examples. The problem of engineering prompts to be effective for a specific LLM in a particular task is called prompt engineering [23]. Due to the great interest in LLM applications and foundational models in recent years, prompt engineering is becoming an important part of the work related to those models. One of the latest LLMs released is the fourth version of the Generative Pre-Trained Transformer (GPT-4) by OpenAI.[1] It has been shown to be one of the best LLMs for solving many different NLP tasks, such as classification, in a zero-shot manner [22], that is, only by prompting, without the need for labelled training data and a fine-tuning procedure.

Considering our task, we show that Deep Learning (DL) techniques, such as LSTM Neural Networks and the Attention Mechanism, can achieve remarkable results when classifying radiology reports written in Italian, as presented in [6]. Even with a limited lexicon and a small number of documents, these techniques have proven successful, as reported in [27]. Furthermore, in [28], we demonstrate how the attention mechanism can be used to emphasise the most critical parts of a radiology report and how these are strongly associated with a collection of pertinent snippets that were manually labelled by radiologists.

This paper presents a comparison of different hierarchical classification systems that combine NLP and DL solutions for the clinical domain. To test the system, a dataset of more than 5,000 labelled reports and 9,000 unlabelled reports was created, without identifying relevant snippets, simplifying the task of classifying reports. The text classification task presents additional challenges due to the patient's access to reports, which could lead to fuzzy or ambiguous radiology explanations. Furthermore, classification relies solely on text data, excluding pictures or demographic information (e.g., sex or age).

We compare the performance of a system that combines the LSTM and Attention mechanism [27] with BERT [29], and GPT-4 in classifying Italian radiology reports (specifically lung CT reports) obtained from the radiology department of Spedali Civili di Brescia. Finally, we evaluate the performance of GPT-4 on the same classification task using various prompts in a zero-shot setup.

The primary contribution of this work is a novel DL-based (BERT) system for the classification of computed tomography reports that incorporates domain knowledge in a hierarchical manner. Experimental results demonstrate superior performance of the new system over the previous one, with the inclusion of domain information further enhancing performance even with no hand-written annotations by experts.

[1] https://openai.com/gpt-4

II. RELATED WORK

An overview and practical approach to NLP with a specific emphasis on its applications to radiology is treated in different works. The work of Mozayan et al. [30] introduced common steps to perform and the problems commonly found in an NLP pipeline in the medical domain and, in particular, to extract information from radiology reports using automated learning methods. They described a brief history of NLP, its strengths and challenges, and freely available resources and tools to help and guide further studies, paying particular attention to recent developments in the field. Donnely et al. [31] conducted a similar investigation focussing more on recent technical advances in NLP techniques and reporting commonly used terms in NLP.

Analysing solutions and applications of radiology reports, Casey et al. [32] provide a systematic synthesis of more than 150 recent publications on natural language processing techniques applied to radiology. They based the analysis on more than 20 indicators, including radiology characteristics, NLP methodology, performance, and clinical application characteristics. Each work analysed is categorised into one of six clinical application categories, such as: Diagnostic Surveillance, Disease Information and Classification, Quality Compliance, Cohort / Epidemiology, Language Discovery and Knowledge Structure, Technical NLP. Their study shows how DL methods are increasing in the recent literature, but conventional ML methods are still prevalent due to the scarcity of data and, in particular, well-labelled data in healthcare domains. Moreover, the difficulty of performing data augmentation without risks in these contexts limits the extension of a small dataset to train large models. When considering explainability, ML techniques are still more straightforward to understand by physicians than neural networks that are considered black-box solutions.

In [33] the authors study the performance and reliability of four NLP tools to predict stroke phenotypes in radiology reports testing the F_1-score, precision, and recall metrics. They suggest the importance of a deep understanding of the development context of an NLP tool to correctly assess whether it is suitable for the task at hand or whether further training, retraining, or modification is required to adapt it to the target task. Surprisingly, they found that the best tool out-of-the-box for this particular task is still a complete Rule-Based System (RBS) where rules are provided by domain experts (radiologists) and that first solves three simple NLP tasks (named entity recognition, negation detection, and relation extraction) and then classifies reports.

In recent times, a variety of DL techniques have been used to find useful information from medical and clinical documents, from Recurrent Neural Networks, particularly LSTM networks [34], to models based on the Attention Mechanism [14] such as Transformer [17], BERT [18], etc.

Focussing on modern architectures, Miller et al. [35] show the highest performance of a pre-trained biomedical BERT (BioClinicalBERT) to solve multiple classification tasks in patients with acute ischemic stroke from radiology reports of computed tomography (CT) and magnetic resonance imaging (MRI). They considered a dataset of more than 2000 reports from more than 500 individuals and compared their fine-tuned version of BioClinicalBERT with other simpler ML approaches and a RBS. In two particular regression tasks on MRI reports, the RBS slightly outperforms BioClinicalBERT, suggesting the actual validity of rule-based solutions in particular conditions. Furthermore, Yan et al. [36] performed a comprehensive comparison of the BERT approaches for radiology. The authors presented RadBERT, a family of transformed-based models adapted to radiology, pre-trained using more than 4 million reports with more than 2 million unique patients from U.S. Department of Veterans Affairs (VA). They compared four different initialisation schemas (BERT-base, Clinical-BERT, RoBERTa, and BioMed-RoBERTa) to create six variants of RadBERT and fine-tuned them for three NLP tasks in radiology: abnormal sentence classification, report coding, and report summation. These variants achieve higher scores than baselines (BERT-base, BioBERT, Clinical-BERT, BlueBERT and BioMed-RoBERTa) in the four tasks when given only 10% of the training sample, while the RadBERT-BioMed-RoBERTa variant performs best overall.

Fink et al. [37] propose a fully automated scalable data mining and curation pipeline using structured radiology reports (SOR) to build and train NLP models to determine oncologic outcomes in multi-institutional FTOR. The deep NLP model with the best performance, BERT, is trained on data from more than $10,000$ patients and achieved a F_1-score of 0.70 in FTOR to predict TRC based on the descriptions in the findings section, outperforming a conventional NLP model. The authors also found that their NLP models achieve an F_1-score similar to that of normal medical students but not of radiology technology students who outperformed the model. However, models are also prone to the lexical complexity and semantic diversity of the radiological narrative. The study has limitations, including the lack of review to report quality and the use of a German dataset. The authors conclude that such systems may be able to extract clinically relevant oncologic end points from large volumes of longitudinal free-text reports and offer a potential advantage as an automated clinical decision support tool for patients referred for multidisciplinary tumour board assessment.

This paper examines the use of supervised learning considering dataset composed by radiology reports written in Italian. Galbusera et al. [38] investigated the feasibility of using NLP in the Italian language to automate medical image annotation and compared its performance with a DL model trained on manually annotated images. The authors found that their NLP model was able to generate accurate annotations for most radiological findings, even when the reports were not complete or contained errors. However, the model had more difficulty detecting some

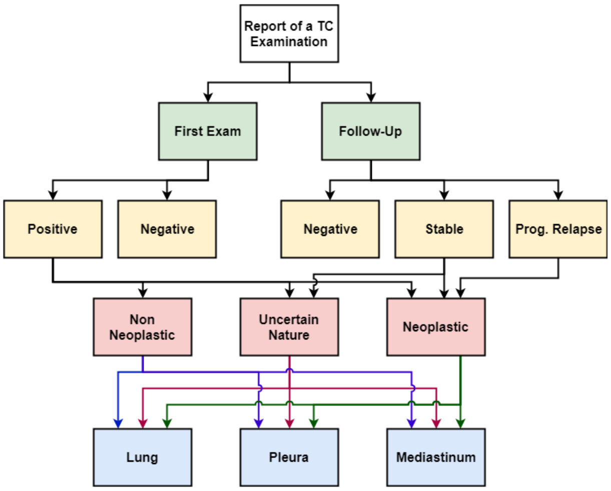


FIGURE 1. The hierarchical classification schema composed by four levels. Radiologists by Spedali Civili di Brescia proposed it for radiology reports related to neoplastic lung injuries.

specific findings, such as retrolisthesis and, to a lesser extent, anterolisthesis, fractures, and SIJ sclerosis. In contrast, the DL model trained on manually annotated images was more accurate in detecting these specific findings. However, the limitations associated with the noisy nature of the NLP predictions and the reports themselves need to be addressed. Fanni et al. [39] proposed an AI-based approach to convert unstructured free-text COVID-19 chest CT reports from an Italian Hospital into structured reports. They trained a deep-learning model using 475 manually structured reports and evaluated its performance on a test set of 400 CT scans. The model achieved a per-statement accuracy of 91.6% and 95.9% depending on whether strict or modified criteria were used. The authors also compared their results with previous studies and found that their approach was more accurate than previous methods and is a promising way to convert unstructured free-text COVID-19 chest CT reports into structured reports. This could have several benefits, including better data mining and communication with providers.

Our previous system [6] was based on the annotation of pertinent excerpts, which were then used to categorise reports with the help of ML methods, as shown in Fig. 1. More information about this system is provided in Section IV-A.

In contrast to the studies that necessitate the development, application, and evaluation of unique algorithms and models (which require a considerable effort), our research looks into the possibility of using pre-trained LLMs without any fine-tuning. In this study, we use prompt engineering designed to get the best results from these pre-set models.

White et al. [23] have created a catalogue of 17 patterns to address common problems that arise when interacting with LLMs. These patterns help manage the data inputted into the model, the structure and style of the output, and any inaccuracies in the content, such as fabricated answers based on unsubstantiated data. Additionally, the catalogue provides strategies to refine the prompt for improved responses, and deals with the communication between the user and the model, as well as the context needed for the model to generate better replies.

Reynolds et al. [22] showed that zero-shot learning (which does not require any examples to be given to the model) can outperform the traditional few-shot learning approach (which requires some examples) when a prompt is used correctly. To do this, they coined the term *meta-prompt*, which encourages the model to generate its own natural language prompt to complete a task. Zhou et al. [21] presented Automatic Prompt Engineer, a system for automatic directive creation and selection. It was designed to enhance the prompt by exploring a range of directive options proposed by an LLM to optimise a chosen scoring function, obtaining remarkable results.

LLMs, in particular ChatGPT and GPT-4, have proven to be very effective in solving specific NLP tasks in general domains. Even in specific domains, such as public health [40], environmental problems [41] or legal rulings and laws [42], LLMs achieve acceptable performance.

In the field of radiology, Liu et al. [43] compared the performance of 32 LLMs from different countries (including ChatGPT, GPT-4, PaLM2 [44], Claude2,[2] and 18 Chinese LLMs), in interpreting radiology reports. The results showed that many Chinese LLMs performed competitively against their global counterparts and that these multilingual and diverse LLMs have the potential to contribute to an improved global healthcare delivery system. They also suggested that there is a relevant scope to expand these LLMs into different medical specialities and develop multimodal LLMs. However, they warned that it is important to consider the ethical implications of deploying these models.

III. CASE STUDY

This section describes the application of machine learning techniques to the classification of radiology reports from an Italian hospital. We focus on chest tomography (CT) reports, specifically seeking for neoplastic lesions. Automatically classifying old and new reports could bring about a number of benefits, such as improved logistics, better health care management, monitoring the frequency of follow-up examinations, and collecting cases for research or teaching. Radiology reports are typically composed of free text that can be organised into standard sections.

This proposed system for classifying reports is based on a schema developed in close collaboration with radiologists from *Spedali Civili di Brescia*. It is designed to emphasise the important aspects of a radiology report and follows the policy used during the evaluation process [27] as shown in Fig. 1, with some modifications. To simplify training and problem description, binary classification tasks are used and different classes are grouped together to create only two classes in each level.

[2] https://www.anthropic.com/index/claude-2

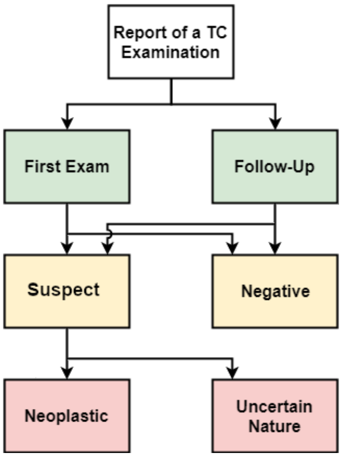


FIGURE 2. The new hierarchical classification schema composed of three levels concordant with radiologists from Spedali Civili di Brescia for radiology reports concerning lung injuries of a neoplastic nature.

Furthermore, the last level (*Lesion Nature*) with its three classes (*Lung*, *Pleura* and *Mediastinum*) is not taken into account anymore. This is because it is difficult to accurately identify the region of the lesion on CT images and the labels provided by different physicians on the same image have excessively high variability.

The new schema shown in Fig. 2 consists of three levels that correspond to the main aspects considered by physicians during the evaluation of a report.

1) **Exam Type**: *First Exam* or *Follow-Up*;
2) **Result**: *Suspect* (grouping Positive, Stable, Prog. Relapse) or *Negative*;
3) **Lesion Nature**: *Neoplastic*, or *Uncertain Nature* (grouping Non Neoplastic and Uncertain Nature). This third level is specified only for the Suspect reports.

The strict relationship between the first three levels, shown by the arrows in the schema of Fig. 2, is formalised by the following rules:

A. The presence of even a suspicion of a neoplastic lesion automatically classifies a *First Exam* as *Suspect*.
B. In certain cases, such as pneumonia or pulmonary embolism, a *First Exam* can be labelled *Suspect* even without suspicion of a neoplastic lesion. These cases are referred to as non-neoplastic positives.
C. On the contrary, since radiological follow-ups are conducted primarily to monitor the conditions of neoplastic patients, if there is no suspicion of a neoplastic lesion, a *Follow-Up* is automatically classified as *Negative*, even in the case of pneumonia or pulmonary embolism.

A. RADIOLOGY REPORTS DATASET

The dataset comprises 5,752 categorised and anonymised CT reports written in Italian. This involves the removal of the names of the patient and the medical personnel. A collection of 9,581 uncategorised and anonymised reports was collected for word representation enhancement (see Section IV-B1) and for pre-training tasks such as for BERT models (see Section IV-C). The reports available to us are text-only; CT images are not included. These reports are the official output of the Radiology Departments of the Spedali Civili di Brescia for communication and documentation purposes. They contain a description (typically verbless) of physician CT scan observations (nodules, lesions, etc.), their comparison with previous visits (e.g., if the size is the same as the previous exam), and any indications that may rule out the presence of specific symptoms or abnormalities (e.g., pleural effusion).

Our reports share the characteristic of non-standard language with other clinical texts, including abbreviations, ungrammatical language, acronyms, and typos. This is because reports are frequently written quickly or dictated to speech recognition software. Additionally, abbreviations and acronyms may be specific to the hospital or department in question.

In terms of absolute percentage of each class in the simplified hierarchical schema in Fig. 2, for the Exam Type level we have a 63.6% of *Follow-Up* and 36.4% of *First Exam*, for the Result level we have 61.1% of *Negative* and 38.9% of *Suspect* and for the Lesion Nature level we have 64.4% of *Uncertain Nature* and 35.6% of *Neoplastic*. Taking into account the percentages of the combined labels according to the previous schema, we obtain the following.

- (*Follow-Up*, *Negative*): 38.7%
- (*First Exam*, *Negative*): 22.35%
- (*Follow-Up*, *Suspect*, *Uncertain Nature*): 17.71%
- (*First Exam*, *Suspect*, *Uncertain Nature*): 7.5%
- (*Follow-Up*, *Suspect*, *Neoplastic*): 7.22%
- (*First Exam*, *Suspect*, *Neoplastic*): 6.5%

This confirms that the dataset is unbalanced and the need for an investigation to find the proper approach able to solve this classification task.

IV. ARCHITECTURES

A. ANNOTATION-BASED METHOD

This section provides an overview of the preceding system, which relied on the annotation of relevant excerpts. The full explanation can be found in [6].

The TextPro suite [45] was used to preprocess reports and extract textual features. Subsequently, a Conditional Random Field algorithm [46] was utilised to automatically identify the most significant snippets, which were indicative of the various classes. The data used to train this algorithm was taken from a collection of reports that had been labeled by hand. Machine learning techniques were used to classify automatic annotations, and a set of heuristic rules defined by radiologists was applied to derive the classification of the report. This section outlines the key aspects of the annotation-based approach.

The annotation process requires substantial effort, and therefore only 346 reports were chosen, categorised, and annotated by an expert physician in the original dataset. This may not be sufficient to capture the complexity of the classification problem to be addressed, which typically requires many more reports annotated by multiple radiologists. Furthermore, in real-world scenarios, some reports are deliberately cryptic and ambiguous, as patients have access to them. Additionally, in the original dataset, a single concept is often annotated without further elaboration, while its characteristics, such as shape or size, are crucial for classification, particularly in more complex real-world cases. For example, the presence of a nodule on a chest tomography is not sufficient to determine whether the patient has a neoplastic lesion. If the nodule margins are clear and rounded, it can be considered benign. In contrast, if the margins are irregular or spiculated or if the nodule expands, a neoplastic lesion cannot be ruled out. Therefore, the annotation of the word "nodule" is not sufficient to describe a specific condition and can be present in both Non-Neoplastic and Neoplastic reports. Furthermore, the ability of the CRF to capture long and complex expressions, such as radiology concepts with all their characteristics, is limited [46]. Although text annotations can effectively highlight the most important sections of a report from the original dataset, they cannot fully capture the meaning of more complex real-world cases.

B. LSTM WITH ATTENTION METHOD

To address the significant challenges outlined in the preceding section, we have developed an innovative LSTM-based model for hierarchical categorisation of radiology reports. Our system, which eliminates the need for manual text annotation, was evaluated using real-world reports obtained from the radiology department of *Spedali Civili di Brescia*. This section delves into the pre-processing phase and the core components of our system.

1) PRE-PROCESSING AND INPUT REPRESENTATION

The pre-processing phase utilises the "it_core_news_sm" model of spaCy,[3] a Python-based NLP tool, and consists of the following steps:

I. **Section segmentation**: A radiology exam may examine multiple body parts, and its report typically contains sections denoted by body part names in uppercase letters. Our custom algorithm extracts the introduction (which may include details about the type of exam), the section related to the chest, and the conclusions.

II. **Sentence segmentation, tokenization, and PoS tagging**: SpaCy splits the report into sentences and each sentence into individual words; this is accomplished using the spaCy Italian model (trained on a news article collection). Part-of-speech (PoS) tags are then assigned to each word.

III. **Length standardisation**: We analyse our dataset to determine the appropriate maximum word count for a report suitable for input into the neural network. 95% of our reports contain less than 450 words, so we set this as the maximum length. If a report exceeds this length, we remove prepositions, articles, and conjunctions to reach the desired length. If it still exceeds the maximum length, we select the first 450 words. This occurs in only 0.7% of the dataset.

To construct the input for our neural network model, each feature must be mapped to a real-valued vector [47]. Each word in our corpus is represented by a 200-dimensional vector obtained by applying the Word2vec [9] algorithm to both classified and unclassified reports in our dataset. Since the reports are in Italian, we cannot use any pre-trained biomedical word embeddings [48].

Similarly, we also include PoS embeddings in our input. Each PoS tag is represented by a 10-dimensional vector obtained by applying Word2vec to the PoS tag sequence.

2) BIDIRECTIONAL LSTM

A RNN is a DL model used to process sequential data, such as sentences in natural language. It is designed to avoid the problems of gradient vanishing through the use of LSTM cells [34], [49]. Given a document of length m, with each term represented by a vector $x_i \in \mathbb{R}^d$ (obtained by combining the word embedding and the Part-of-Speech (PoS) embedding), and the previous LSTM cell's hidden state and cell state (h_{t-1} and c_{t-1} respectively, with h_0 and c_0 initialised as zero vectors), new hidden state h_t and cell state c_t values are computed as follows:

$$\begin{aligned} h_t &= tanh(c_t) \odot o_t \\ c_t &= \hat{c}_t \odot i_t + c_{t-1} \odot f_t \\ \hat{c}_t &= tanh(W_c[h_{t_i}, x_t] + b_c) \\ o_t &= \sigma(W_o[h_{t_i}, x_t] + b_o) \\ i_t &= \sigma(W_i[h_{t_i}, x_t] + b_i) \\ f_t &= \sigma(W_f[h_{t_i}, x_t] + b_f) \end{aligned} \tag{1}$$

The sigmoid activation function σ is used, with $\odot$ representing the element-wise product. The weight matrices W_f, W_i, W_o, W_c and bias vectors b_f, b_i, b_o, b_c are randomly initialised and learnt by the neural network during the training phase. These matrices and vectors are of size $(N + d) \times N$ and $\mathbb{R}^N$ respectively, where N is the LSTM layer size and d is the dimension of the feature vector for each input word. Vectors in square brackets are concatenated on the last axis (columns).

Bidirectional LSTM not only processes the input sequence in the order of the document, but also reverses [50]. Therefore, we can calculate h^r using the same equations as before, but with words in opposite order. After obtaining h_t in document order and h_t^r in reverse order, the output of the t bidirectional LSTM cell h_t^b is the combination of the two:

$$h_t^b = [h_t, h_t^r] \tag{2}$$

3) ATTENTION MECHANISM

LSTM neural networks have difficulty in maintaining connections between words that are far apart [51]. This is especially true for long sequences, where h_m may not be

[3] https://spacy.io

affected by the initial words or may overlook some important words while processing the entire document. The Attention Mechanism [14], [52] is designed to address these issues, taking into account each h_i and computing the weights α_i for the contribution of each word.

$$\alpha_i = softmax(v^T u_i) = \frac{exp(v^T u_i)}{\sum_{k=1}^{n} exp(v^T u_k)}$$
$$u_i = tanh(W_a h_i + b_a) \tag{3}$$

where $W_a \in \mathbb{R}^{N\times N}$, $b_a \in \mathbb{R}^N$ and $v \in \mathbb{R}^N$ are trainable parameters of the attention mechanism. The attention mechanism outputs the document representation, also called the *context vector*:

$$s = \sum_{i=1}^{m} \alpha_i h_i \tag{4}$$

C. BERT

BERT [18] is a Transformer-based architecture [17] that uses multiple encoding layers to analyse a sequence of tokens (words or parts of words) to understand their meaning. Each layer applies multiple self-attention mechanisms (called *heads*) in parallel.

A Transformer consists of two components: a stack of encoders and a stack of decoders. The encoders extract key information from a text, whereas the decoders generate output from the extracted data. BERT, a model for creating vector representations of words, only uses encoders and does not include decoders. The number of encoders in the stack is adjustable, but the original article suggests 12 of them [17].

For a sequence of tokens S of length N, this method produces a matrix $A_{i,j} \in \mathbb{R}^{N\times N}$, where i is the number of the encoding layer and j is the head number. Each token $w \in S$ has a vector $a_w \in A_{i,j}$ that contains the attention weights that indicate how much w is related to the other tokens in S. To determine these weights, the token sequence $X \in \mathbb{R}^{N\times d}$ is projected into three distinct representations, known as key (K), query (Q) and value (V), in each head using three matrices W_k, W_q and W_v.

$$K = X \times W_k$$
$$Q = X \times W_q$$
$$V = X \times W_v \tag{5}$$

The attention weights are determined by taking the scaled dot-product of Q and K, and then applying the softmax function. The new token representation Z is obtained by multiplying the attention weights by V.

$$A = softmax(\frac{Q \times K^{\intercal}}{\sqrt{d}})$$
$$Z = A \times V \tag{6}$$

where d is the length of the input representation of each token.

The multi-head attention mechanism employs multiple independent components, each generating a separate representation. These individual representations are then consolidated and fed into a subsequent layer. As described in [17], the multi-head attention mechanism is followed by a feed-forward layer and residual connections. The output of one encoding layer serves as the input of the next.

Leveraging a vast collection of documents, BERT is trained on two objectives: *Language Modelling*, where BERT predicts a certain percentage (typically 15%) of words based on contextual information, and *Next Sentence Prediction*, a binary classification task where BERT determines whether a pair of sentences belong to the same sentence, that is, whether they were originally parts of the same sentence. For the latter task, BERT introduces two special tokens: [CLS], whose representation is employed for the binary classification task and signifies the entire sequence, and [SEP], which separates the two sentences. By learning these two tasks, BERT develops a meaningful representation of each word and the ability to summarise the most crucial information in a sentence. After training, the model can be adapted using smaller datasets for specific NLP tasks, such as Named Entity Recognition, Text Classification, Sentiment Analysis, etc.

D. GPT-4

The GPT-4 (Generative Pre-trained Transformer 4) is a large language model developed by OpenAI [53]. It is a huge Transformer-based model trained on a massive dataset of text and code that is capable of generating text, translating languages, creating various types of content, and responding to questions in an informative manner [54], [55]. It appears to have up to 100 trillion parameters, which is significantly more than the 175 billion parameters of GPT-3 [56], [57] and the 1.5 billion parameters of GPT-2 [58], [59], [60]. This allows GPT-4 to learn more complex patterns from the data and generate more sophisticated outputs compared to its predecessors. Furthermore, it has been trained on a much larger and more diverse dataset than previous versions, including text from books, articles, code, and other sources in multiple languages. Therefore, it reached a deeper and broader understanding of textual content, allowing it to perform better on a wider range of tasks [54]. GPT-4 consists of a decoder-only architecture based on self-attentions and feed-forward neural networks. Oppositely to BERT, it uses only the decoder part of the Transformer model, essentially taking the sequence of hidden states and generating a sequence of output tokens. In fact, GPT models are specifically designed to generate text using a decoder to predict the next word in a sequence given the previous words in the sequence, in an auto-regressive manner.

GPT-4 was trained using self-supervised and supervised learning techniques, which involved providing it with pairs of input and target sequences. The input sequences were usually text passages or code snippets, whereas the target sequences were typically the continuation of the input sequences. To train GPT-4, similarly to BERT's training, a method called masked language modelling was used. This method involves randomly masking some of the tokens in the input sequence and then training GPT-4 to predict the masked tokens based on the remaining tokens in the sequence.

Recently, GPT-4 has become an intriguing and effective tool for a variety of reasons [61]. It is capable of recognising intricate patterns in the data and producing sophisticated results. The Transformer architecture is ideal for natural language processing tasks, as it is able to detect long-term correlations in the data. Therefore, GPT-4 has a wide range of potential applications, including the following:

- **Text generation**: generating text for a variety of purposes, such as writing blog posts, articles, and even books.
- **Translation**: translating text from one language to another.
- **Code generation**: generating code for a variety of programming languages.
- **Question answering**: answering questions in a comprehensive and informative way.
- **Creative writing**: generating creative content, such as poems, storeys, and scripts.

An important aspect for an effective use of GPT-4, as in the other LLM, is the way the prompt is composed, i.e., how the prompting is performed. The prompt provides GPT-4 with additional information about the desired output, helping generate more accurate and relevant outputs. In fact, the prompt can be used to specify the genre, style, tone, and other aspects of the desired output. Moreover, some OpenAI guidelines have been suggested for effective prompt writing to avoid unwanted outputs or to produce outputs that violate its legal terms [53].

The GPT-4 has the potential to be used for sentence classification, taking advantage of its Question-Answering capabilities. To do this, the sentence classification task must be transformed into a question-answering task. For example, if the goal is to classify a sentence as either positive or negative, the task can be rephrased as: "Is the previous (following) sentence positive or negative?" and GPT-4 can then be used to provide the answer, either "Positive" or "Negative". The answer generated by the model will be the predicted class of the sentence.

This approach to sentence classification has a number of benefits:

- it is highly versatile and can be applied to a wide range of tasks;
- it is highly accurate because LLMs have been demonstrated to be very successful in question answering;
- it is highly efficient, as GPT-4 can generate answers to questions in a short amount of time and without complex pretraining in a zero-shot setup.

Additionally, GPT-4 can be used to classify sentences for more complex tasks, such as Sentiment Analysis, Topic Identification, and Natural Language Inference [24], [25], [26]. This task can be changed into a question-answering task in a manner similar to the example given earlier.

V. METHODOLOGY AND MODELS FOR HIERARCHICAL CLASSIFICATION

Most of the text classification tasks in the literature are flat, which means that they require documents to be associated with one or more labels without any relation between them. On the contrary, our project requires the provision of an appropriate label for each level of the hierarchical schema in Section III, to collect information on different aspects of the radiology report, such as the type of examination, the result of the examination or the location of any neoplastic lesions. To achieve this, the main contribution of our work is the development of an architecture that not only processes the data, but also takes into account the results of other levels, rules, and other domain knowledge set by radiologists. This not only ensures predictions that are consistent with the pre-defined rules but also improves the performance.

Firstly, we discuss the metrics used to evaluate the performance and secondly our three distinct hierarchical techniques and models for categorising a radiology report in accordance with the proposed structure.

A. CLASSIFICATION METRICS

The classification metrics considered in our study are: Accuracy, F_1-score, Recall (also called Sensitivity), Specificity, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). In particular, the most relevant metric considered is F_1-score, which is a good measure in unbalanced classification tasks because it does not favour the majority class as is. Considering True Positive (TP) the number of samples correctly classified as positive, True Negative (TN) the number of samples correctly classified as negative, and False Positive (FP) the number of samples erroneously classified as positive and False Negative (FN) the number of samples erroneously classified as negative, the previous metrics could be described as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{9}$$

$$F_1 = \frac{2TP}{2TP + 2FN + FP} \tag{10}$$

The Receiver Operating Characteristic Curve is the plot of the True Positive Rate (Recall) against the False Positive Rate (Specificity) at each threshold setting. The area under this curve is a value between 0 and 1 that measures the goodness of the classifier, which is considered optimal with a ROC-AUC value of 1.

B. LSTM WITH ATTENTION MODELS

We investigate three different models based on LSTM with an Attention Layer on top. These models vary in the way hierarchical information is processed and combined to

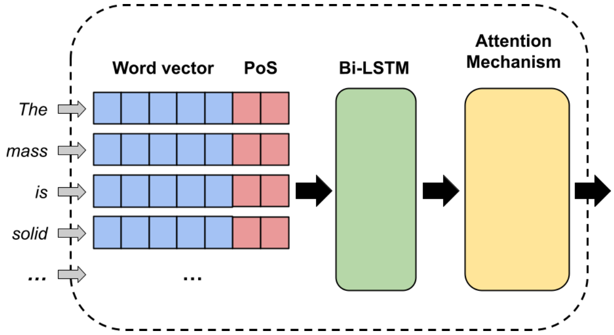


FIGURE 3. A block of our hierarchical classification system. The words have been translated into english to make it easier to understand.

complete the hierarchical classification task. From a practical point of view, splitting samples after a classification level to solve the next classification level can be detrimental to the DL architecture training procedure due to a lack of training data or an unbalanced classification task. Therefore, it is essential to analyse different approaches for an effective hierarchical classification.

1) CLASSIFICATION BLOCK

Our hierarchical classification (illustrated in Fig. 2) is composed of a series of blocks, which are trained with different training sets and combined according to a set of rules established by radiologists during the development of the classification schema (as described in Section III). Each block, shown in Fig. 3, consists of:

- two parallel **embedding layers** that provide the pre-trained embedding representation of each word and PoS-tag in a report, which are then combined into a single input vector;
- a **bidirectional LSTM layer** that transforms the input sequence in a recurrent embedding space;
- an **attention mechanism** that weighs the importance of each word for the classification task and provides the document representation;
- on **output layer** responsible for providing the classification. If it is a binary classifier, it consists of a single neuron with Sigmoid activation. If it is a multiclass classifier, it is composed of *n* neurons, where *n* is equal to the number of classes, and the activation is Softmax.

2) MODEL A

Model A (Fig. 4) is the simplest architecture, combining the two configurations of the Result level into one. It only has three classification blocks, one for each level, such as the following:

- the **Exam Type Block** is used to determine if the report is a *First Exam* or a *Follow-Up* trained on the entire training set.
- the **Result Block** classifies the report as either *Negative* or *Suspect*, combining the two *Negative* categories of the *First Exam* and *Follow-Up*. Even this is trained on the whole training set.
- the **Lesion Nature Block** is a neural network that is trained with the entire training set, excluding *Negative* reports. If the Result block classifies a report as *Suspect*, then the Lesion Nature Block will process it. This block is used to differentiate between *Uncertain* and *Neoplastic*.

3) MODEL B

As shown in the classification schema, Model B (Fig. 5) separates the computation of the Result level between two different classification blocks:

- the **Result (*First Exam*) Block**, which is used to determine if the report is *Suspect* or *Negative*. It is trained with all the *First Exams* in our training set.
- the **Result (*Follow-Up*) Block**, which indicates whether the report is *Suspect* or *Negative*. It is trained with all the *Follow-Up*.

If the report is identified as *First Exam*, it is handled by the Result (*First Exam*) Block. On the other hand, if it is classified as *Follow-Up*, it is processed by the Result (*Follow-Up*) Block. If the report is not *Negative*, another neural network can accurately predict the Lesion Nature level, similar to Model A.

4) MODEL C

Model C (illustrated in Fig. 6) is the most complex of the three models. It follows the guidelines outlined in Section III. The **Exam Type Block** predicts if the report is a *First Exam* or a *Follow-Up*, as with the two previous models. The **Suspect Block** is designed to identify if a report is suspected to contain a neoplastic lesion, according to the rules of Section III. It is trained on a re-labelled version of the dataset, with the labelling modification procedure described as follows:

- If the outcome of both *Follow-Up* and *First Exam* are *Negative*, then the report is deemed to be not *Suspect*.
- If the Result is *Suspect*, we check the classification of the *Neoplastic* level and if it is *Uncertain* or *Neoplastic*, the report is considered *Suspect*.
- If it is *Non Neoplastic*, the report is considered *Negative*.

Following Rule C in Section III, if a *Follow-Up* is classified as *Non-Suspect*, then the result can be automatically set as *Negative*. If Rule A is applied and the *First Exam* is deemed to be *Suspect*, then the outcome is *Positive*. Despite the fact that some *First Exam* are deemed *Suspect* even in the absence of a neoplastic lesion (Rule B), we have created the **Non Neoplastic Positives Block** to identify these particular cases. If the *First Exam* is classified as *Suspect* or *Non-Neoplastic Positive*, then the result is labelled as *Positive*. We used the entire dataset with the appropriate labels for training.

C. BERT METHODOLOGY AND MODELS

To assess the effectiveness of BERT in this intricate setting, we employ BioBIT, an improved biomedical language model for Italian made by Buonocore et al. [62]. This model is

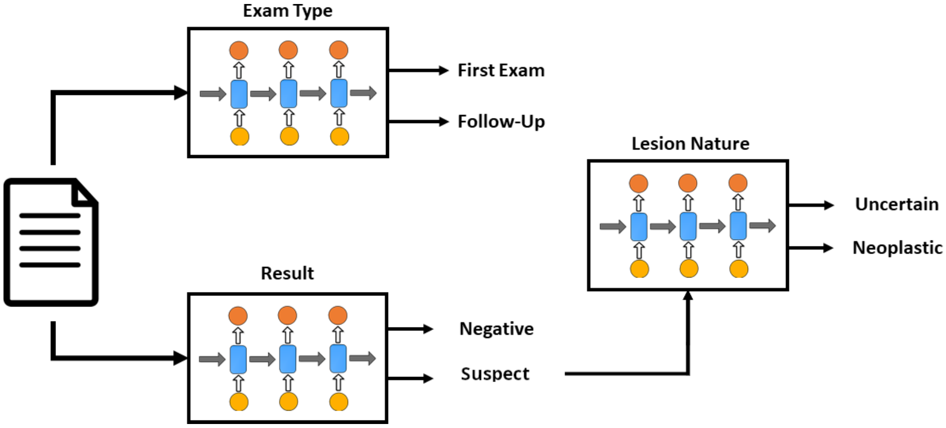


FIGURE 4. Model A (Section V-B2). A single classification block provides the result without any relation with the exam type block. Then the *Suspect* samples are classified by the lesion nature block, which does not have as input information from the exam type block.

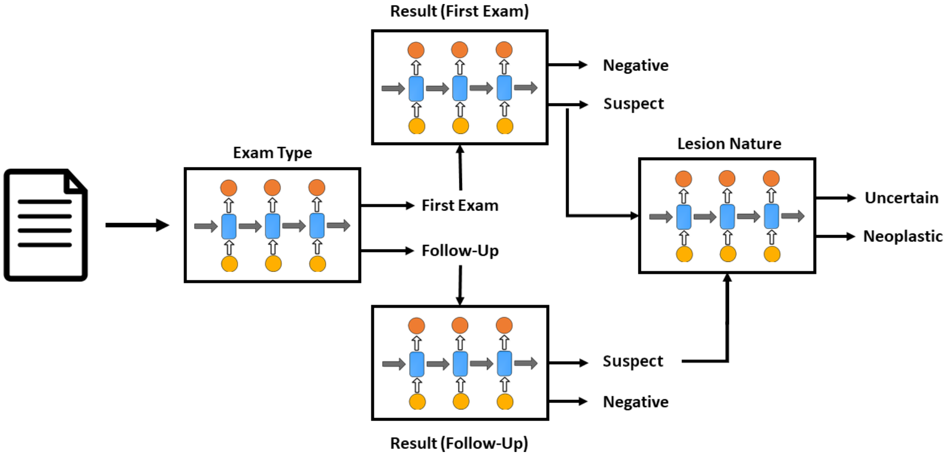


FIGURE 5. Model B (Section V-B3). Each box symbolises a classification block. In this model, if the report is classified as the *First Exam*, a distinct block gives the result for the *First Exam*, otherwise a different one gives the result for *Follow-Up*.

based on the BERT-Base-Italian-XXL-Cased version,[4] a famous BERT solution for the Italian language made by the Bavarian State Library. The authors leveraged machine translation from Google NMT to obtain an Italian biomedical corpus based on the English PubMed abstracts that could be large enough (gigabytes of text) to effectively train BioBIT. In addition, they used the WordPiece tokenizer for compatibility with out-of-vocabulary words in the biomedical corpus. We consider the BioBIT pre-trained weights publicly available on the HuggingFace author's page[5] in compound with its tokeniser and, to adapt this model to our use case, we train it for tasks *Mask Language Model* and *Next Sentence Prediction*, using $9,581$ unclassified reports from the same hospital. Subsequently, we fine-tuned the model using our supervised training set, employing the AdamW [63] optimiser with a learning rate of $2 \cdot 10^{-5}$ and ϵ value $1 \cdot 10^{-8}$, a batch size 8, 4 epochs. Moreover, we use a learning rate scheduler that linearly decreases the learning rate to zero, without using warm-up steps ($warmup\_steps = 0$). We fine-tuned the model to solve binary classification tasks, one for each classification level (Exam Type, Result, Lesion Nature), adding a linear layer with two output neurons on top of the pooled output of BERT and training the whole model without freezing parts, as shown in Figure 7. For each classification level, performance was evaluated using a 10-fold cross-validation approach, resulting in 10 trained models with different training and validation sets. The median and averages of the metrics obtained by the cross-validation procedure are calculated for comparison with other approaches; in particular, we consider the median value as the reference performance because it is robust to outliers (i.e., particularly good or bad folds).

D. GPT-4 PROMPTING METHODS

In Table 1 are shown the Italian prompts (and the corresponding translated versions) used to perform the classification task using GPT-4 through its API.[6] We defined three different

[4] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[5] https://huggingface.co/IVN-RIN/bioBIT

[6] https://chat.openai.com/

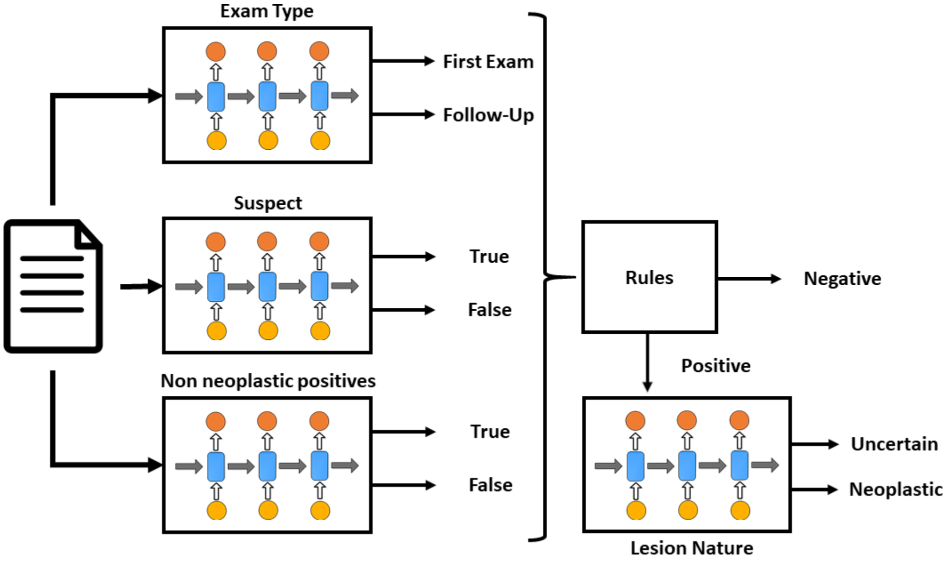


FIGURE 6. Model C (Section V-B4). This model classifies whether a report is a *First Exam* or a *Follow-Up*, if there is a potential neoplastic lesion, and if it is a *Non-Neoplastic Positive*. The overall Result is determined by applying the rules outlined in Section III and the Stable or Prog. Relapse block.

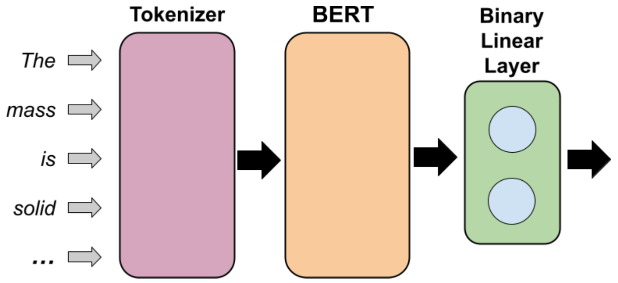


FIGURE 7. The BERT classification model. The words are tokenised before the BERT model processes the sequence, and then a binary linear layer is applied to the output of BERT to solve the classification task.

prompts for the investigation named *Simple*, *Result Refined* and *Experts Group* respectively.

The Simple prompt is a first attempt based on the ‘‘Persona’’ Pattern and following the best practise described in [23]. In prompt engineering, a Persona pattern actually involves taking the LLM to the role of a character to get the desired response. In our case, the prompt asks GPT-4 to play the role of an expert Italian radiologist who has to classify CT reports made by other radiologists. Then we present the different classes, their meaning, and what the labels are associated with each class. Finally, we instruct the LLM on how to present the results and what the output schema is that it has to follow.

The Result Refined prompt is based on the Simple prompt but with some improvements. Taking into account the simplification made in the original hierarchical classification schema (Fig. 1) especially for the Result level, the *Suspect* class in the new schema (Fig. 2) could be difficult to predict or in some way confusing for the LLM due to the grouping of several classes that do not exactly have the same meaning (*Positive*, *Stable*, *Prog. Relapse*). On the other hand, the *Negative* class is practically the same and, generally speaking, is easier to explain and understand. Therefore, in the Result Refined prompt, the description of the *Negative* class and what kind of report should be classified as *Negative* have been refined and extended, with also a better description of the *Suspect* class. Moreover, we add to the output description a statement forcing the model to respond following the classification scheme, even if it does not have enough information or is uncertain.

The last prompt called Experts Group is inspired by the Tree-of-Thought (ToT) pattern [64], [65], [66] for prompt engineering, and in particular the ‘‘Experts Group’’ pattern. The Experts Group pattern is a method for generating text based on the opinions of a group of experts used by the ToT framework. This method can be used to generate text that is more accurate and reliable than the text generated by a single expert. Therefore, the LLM will play the role of a group of Italian expert radiologists who have to classify CT reports made by other radiologists following the new hierarchical classification schema. In particular, we ask the model to continue to think about the classification labels until all three experts agree. Moreover, we maintain the prompt enhancements introduced in the Result Refined prompt for the classification of the Result level.

Although providing examples as an extended context is in the prompting best-practises [22], [23], we do not evaluate or prompt adding positive and negative examples in the context because it could lead GPT-4 to be polarised on them losing generalisation capabilities. In particular, many of our Italian reports present a very cryptic and ambiguous language which could confuse the LLM to focus on particular expressions or unrelevant adjectives instead of considering the whole meaning of the reports.

TABLE 1. Different prompts used to instruct GPT-4 about the classification task to solve with respect to the italian CT reports. On the left side is the Italian prompt used for the classification task, and on the right is the English version to help non-Italian readers. The part of the prompt describing the character to be used by GPT-4 for role-playing is coloured teal, the classification task and its associated labels are coloured purple, and the instructions for generating and formatting the output are coloured blue.

Prompt	*Italian*	*English*
Simple	Sei un medico radiologo italiano che deve classificare il testo di un referto radiologico fatto da altri medici su pazienti che hanno avuto o che sono sospetti di avere un tumore nella zona polmonare (polmone, pleura, mediastino). Le classificazioni da fare sono: 1) TIPO ESAME: se è un referto relativo ad un PRIMO ESAME o ad un controllo di FOLLOW-UP. 2) RISULTATO ESAME: NEGATIVO oppure in tutti gli altri casi è SOSPETTO. 3) NATURA LESIONE: NEOPLASTICO o in tutti gli altri casi è di NATURA DUBBIA. Rispondimi solamente rispettando lo schema: 1: 2: 3:	You are an Italian radiologist who needs to classify the text of a radiological report made by other doctors on patients who have had or are suspected of having a tumor in the lung area (lung, pleura, mediastinum). The classifications to be made are: 1) TYPE OF EXAM: if it is a report related to a First Exam or a FOLLOW-UP check. 2) EXAM RESULT: NEGATIVE or in all other cases it is SUSPECT. 3) NATURE OF THE LESION: NEOPLASTIC or, in all other cases, is of UNCERTAIN NATURE. Please respond only to the scheme: 1: 2: 3:
Result Refined	Sei un medico radiologo italiano che deve classificare il testo di un referto radiologico fatto da altri medici su pazienti che hanno avuto o che sono sospetti di avere un tumore nella zona polmonare (polmone, pleura, mediastino). Le classificazioni da fare, con le relative etichette in maiuscolo, sono: 1) TIPO ESAME: se è un referto relativo ad un PRIMO ESAME, oppure ad un controllo di FOLLOW-UP. 2) RISULTATO ESAME: NEGATIVO se non presenta tumori o masse sospette, oppure in tutti gli altri casi è SOSPETTO. 3) NATURA LESIONE: NEOPLASTICO, oppure in tutti gli altri casi è di NATURA DUBBIA. Rispondi solamente con i valori delle etichette rispettando il seguente schema (anche quando pensi di non avere abbastanza informazioni): 1: 2: 3:	You are an Italian radiologist who needs to classify the text of a radiology report made by other doctors on patients who have had or are suspected of having a tumour in the lung area (lung, pleura, mediastinum). The classifications to be made, with the corresponding labels in uppercase, are: 1) TYPE OF EXAM: whether it is a report related to a First Exam, or a FOLLOW-UP check. 2) EXAM RESULT: NEGATIVE if it does not present tumours or suspicious masses, or in all other cases, it is SUSPECT. 3) NATURE OF THE LESION: NEOPLASTIC, or in all other cases is of UNCERTAIN NATURE. Respond only with the labels values with respect to the following scheme (even when you think you do not have enough information): 1: 2: 3:
Experts Group	Siete un gruppo di medici esperti italiani che deve classificare il testo di un referto radiologico fatto da altri medici su pazienti che hanno avuto o che sono sospetti di avere un tumore nella zona polmonare (polmone, pleura, mediastino). Continua a pensarci fino a che tutti e tre gli esperti non sono d'accordo. Le classificazioni da fare, con le relative etichette in maiuscolo, sono: 1) TIPO ESAME: se è un referto relativo ad un PRIMO ESAME, oppure ad un controllo di FOLLOW-UP. 2) RISULTATO ESAME: NEGATIVO se non presenta tumori o masse sospette, oppure in tutti gli altri casi è SOSPETTO. 3) NATURA LESIONE: NEOPLASTICO, oppure in tutti gli altri casi è di NATURA DUBBIA. Rispondi solamente con i valori delle etichette rispettando il seguente schema (anche quando pensi di non avere abbastanza informazioni): 1: 2: 3:	You are a group of Italian expert doctors who need to classify the text of a radiological report made by other doctors on patients who have had or are suspected of having a tumour in the lung area (lung, pleura, mediastinum). Continue to think about it until all three experts agree. The classifications to be made, with the corresponding labels in uppercase, are: 1) TYPE OF EXAM: whether it is a report related to a First Exam, or a FOLLOW-UP check. 2) EXAM RESULT: NEGATIVE if it does not present tumours or suspicious masses, or in all other cases it is SUSPECT. 3) NATURE OF THE LESION: NEOPLASTIC, or in all other cases, it is of UNCERTAIN NATURE. Respond only with the values of the labels respecting the following scheme (even when you think you do not have enough information): 1: 2: 3:

VI. EXPERIMENTAL RESULTS

In this Section, we present the results of different architectures previously introduced in Section IV following the methodology proposed in Section V. For the LSTM with Attention Models we compare and discuss the performance of three different architectures (Section V-B) to identify the best model that will be compared with the other approaches. As introduced in Section V-D, we evaluate different prompts and in the following section we compare the results of these prompts for each of the three classification levels shown in Fig. 2. Finally, we compare the old method based on annotations, the LSTM with Attention method, and GPT-4 with results from BERT discussing the best overall approach.

A. BEST LSTM WITH ATTENTION MODEL

We evaluated the performance of our three different architectures described in Section V-B based on LSTM with Attention. We implemented them using the Keras [67] library with Tensorflow [68] back-end. To optimise the learning phase and avoid overfitting, we performed a random hyperparameter search [69] and used a validation set of 20% of the training reports for each classification block. We then evaluated the entire model in 10-fold cross-validation using accuracy (Acc.), macro-averaged F_1-score (F_1), macro-averaged recall (Rec.) also called Sensitivity, Specificity (Spec.) and the Area Under the Receiver Operating Characteristic Curve (AUC).

The results of our hierarchical classification models are presented in Table 2. The standard deviation of the 10-fold cross-validation ranged from 0.7% to 2.8%, which was not included in Tables 2 for simplicity. Despite the complexity of the architecture, Model C, illustrated in Fig. 6, outperforms the other models in almost all three classification levels. The classification of the Exam Type at the first level of the classification hierarchy results in the simplest task to solve for all models. They achieve practically the same performance with respect to Accuracy (96.2%) and F_1 (96.0%), but for other metrics (Recall, Specificity, ROC-AUC) Model C shows the highest values (98.2%, 92.7%, 93.6%). Considering the Result level, Model C is the best model to solve this particular task, reaching Accuracy 81.2%, F_1 72.8% and AUC 83.6%, while Model A, illustrated in Fig. 4, is the second-best. However, it is important to note that the results of all three models are very similar on the second level in terms of F_1-score but Model C differs in terms of Accuracy and AUC, focussing more on the positive samples in the class. The third level (Lesion Nature) proves to be the most difficult task to solve in both the metrics

TABLE 2. Performance comparison of our different LSTM with attention models. They are compared and evaluated in terms of accuracy (Acc.), F_1-score (F_1), Recall (Rec.), Specificity (Spec.) and ROC-AUC (AUC). The best and the second-best results are in bold and underlined, respectively. Model C is the best-performing model for every classification level.

	Model A					*Model B*					*Model C*				
Class. Levels	Acc.	F_1	Rec.	Spec.	AUC	Acc.	F_1	Rec.	Spec.	AUC	Acc.	F_1	Rec.	Spec.	AUC
Exam Type	96.2	96.0	98	92.5	93.4	96.2	96.0	97.9	92.4	93.3	**96.2**	**96.0**	**98.2**	**92.7**	**93.6**
Result	77.4	72.7	82.5	73.9	83.2	77.1	72.5	81.8	73.2	79	**81.2**	**72.8**	**83**	**74.1**	**83.6**
Lesion Nature	72.9	70.7	78.4	68.2	73.5	72.8	70.4	77.6	68.1	73.1	**73.2**	**71.2**	**79.2**	**68.8**	**74**

TABLE 3. GPT-4 Prompts results. The Acc. stands for Accuracy, Rec. for Recall (Sensitivity), Spec. for Specificity, AUC for area under the receiver operating characteristic curve. The best and the second-best values are in bold and underlined, respectively. The result refined prompt is the best prompt for GPT-4 overall.

	Simple					*Result Refined*					*Experts Group*				
Class. Levels	Acc.	F_1	Rec.	Spec.	AUC	Acc.	F_1	Rec.	Spec.	AUC	Acc.	F_1	Rec.	Spec.	AUC
Exam Type	91.9	88.5	81	**98.7**	89.9	**92.3**	**89.3**	**83.4**	97.8	**90.6**	92.0	88.9	82.7	97.8	90.3
Result	67.4	65.0	**88.7**	56.4	72.5	**70.9**	**67.5**	88.3	**61.8**	**75**	70.0	66.8	87.9	60.8	74.3
Lesion Nature	**82.9**	64.5	60.7	**92.6**	76.6	81.0	**64.6**	**62.4**	**92.6**	**77.5**	80.2	63.1	61.5	92.4	77

considered because of the reduced number of reports used to solve the task. Being the last level in the classification hierarchy leads to a reduction in the number of reports with respect to the output classification of the blocks in the previous levels. Model C reaches 73.2% Accuracy, 71.2% F_1 and 74 AUC in the Lesion Nature level with small differences compared to the other models. Also, for the last classification level, Model A results in the second-best model.

These results show that Model B, illustrated in Fig. 5, is the worst approach compared to the others. We argue that splitting data after the Exam Type level and classifying the reports in two different groups significantly reduce the number of reports present in the dataset passed to the Result (*First Exam*) and Result (*Follow-Up*) blocks. A poor classification performance for the Result level caused by the dataset splitting impacts also the Lesion Nature level, which is trained with reports classified as *Suspect* from the previous blocks rejoined. In contrast, the Model C architecture, which results in the best model, gives for both the Exam Type, Suspect and Non-Neoplastic Positive blocks the same number of reports resulting in large datasets for blocks training. Moreover, this architecture takes care of different problems in the previous classification schema described in Section III, training specific classification blocks to mitigate them, and finally training the Lesion Nature block. Furthermore, as illustrated in Fig. 8, Model C outperforms the other two models even when only 50% of the number of samples in the training set is used. It shows higher performance than other models at the Result level due to the parallelisation of the related classification tasks and then takes advantage of larger training sets for each subclassification block, as illustrated in Fig. 6. The Neoplastic level (Lesion Nature level) shows similar performance for all three models because its classification block is always downstream of the previous levels in all architectures, having similar small training sets. This is especially noteworthy since biomedical NLP often relies on small datasets [70], [71], [72].

Taking into account the hierarchical classification schema depicted in Fig. 1, the *Uncertain* class poses the greatest challenge, as demonstrated by the inter-annotator agreement study conducted in [6]. In this analysis, a second radiologist was assigned to classify all 68 reports in the original test set. Perfect agreement was observed at the Exam Type level of the classification schema, with a 93% agreement at the Result level and a mere 73% agreement at the *Lesion Nature* level. The majority of disagreements at the third level stemmed from the *Uncertain* class, with 78% of uncertain reports classified as such by one of the two radiologists. Distinguishing between the *Neoplastic* and *Uncertain* classes appears to be the main challenge, with 61% of the disagreements at the final level falling into this category. We deem these reports to harbour the most sensitive information and their language to be the most obscure and enigmatic. Furthermore, we believe that certain cases can be classified as either *Uncertain* or *Neoplastic* depending on the doctor's discretion. Consequently, we opt to merge the *Neoplastic* and *Uncertain* classes into a single *Uncertain* class to improve performance, as illustrated in the revised schema shown in Fig. 2. Radiologists have affirmed that this simplification does not compromise the validity of the classification procedure, which is focused primarily on identifying neoplastic lesions with high precision.

Addressing a crucial aspect, we emphasise the significant impact of data scarcity on prediction accuracy [73], particularly for lower-level categories such as Lesion Nature, which are only present in the 2, 248 non-negative reports. As depicted in Fig. 8, we illustrate the accuracy enhancements of our proposed architectures for the Result and Lesion Nature levels as the dataset size expands. Our models were trained using varying proportions of the training set while employing the same test set. The absence of unclassified reports undoubtedly hinders the performance of our classification system. In fact, while conventional Word2vec models are typically trained on millions of documents [48], our word representation is trained on a mere 10K reports. We firmly believe that even a modest increase in the unclassified report count would translate into better performance.

B. GPT-4 PROMPTS RESULTS

Table 3 compares the classification performance in terms of Accuracy (Acc.), macro-averaged F_1, macro-averaged Recall (Rec.), Specificity (Spec.) and ROC-AUC (AUC) for the

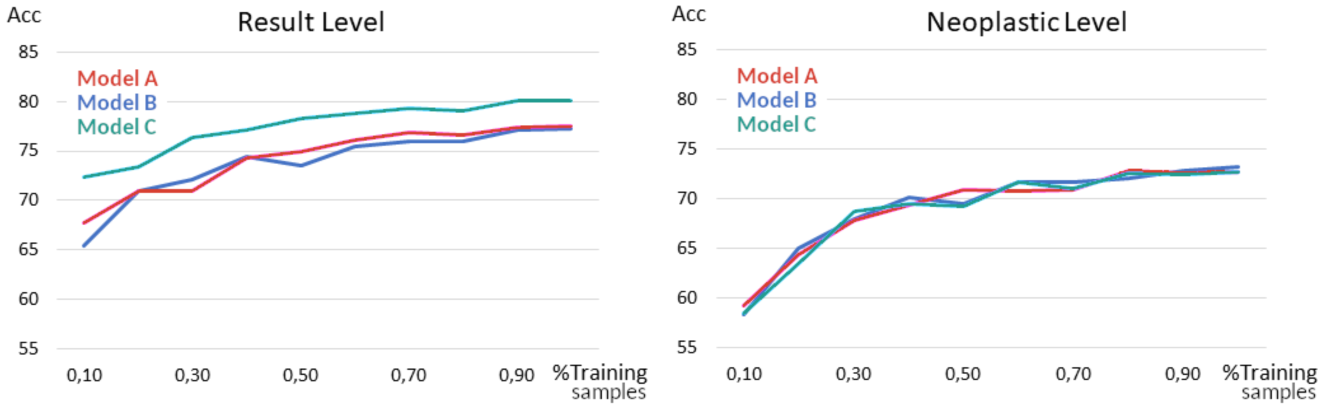


FIGURE 8. Improvement in macro-average accuracy when increasing the size of the training set for models A (red curve), B (blue curve), and C (green curve) in 10-fold cross validation. On the x-axis there is the percentage of the training set used for training the model, while on the y-axis there is the accuracy for both the first exams and follow-ups at the result and lesion nature levels.

different classification levels obtained by the three different prompts described in Section V-D.

The Exam Type classification level is the simplest level to classify for GPT-4 with the different suggested prompts. In this task, in which the model has to classify the report as *First Exam* of *Follow-Up*, the Result Refined prompt shows the best performance, reaching 92.3% Accuracy, 89.3% F_1 and 90.6% AUC, with the other prompts very close to it. Only considering specificity, the Result Refined performs second best on par with the Experts Group behind the Simple prompt.

The Result Refined prompt confirms to be the best even for the classification of the Result level with 70.9% Accuracy, 67.5% F_1 and 75% AUC followed by the Experts Group prompt, which is the second best only for a small margin. On the other hand, Simple prompts show to be not effective on the Result level, losing 2% on average in terms of F_1 and AUC with respect to the others.

In the classification of the Lesion Nature, which is the last classification level, the Simple prompt reaches the best results in terms of Accuracy, 82.9%, and Specificity, 92.6%, followed by the Result Refined prompt with Accuracy 81.0% and the same Specificity value. On the contrary, the Result Refined prompt reaches the best result in terms of F_1 (64.6%) and AUC (77.5%) followed by the Simple prompt for F_1 and by the Experts Group for the ROC-AUC.

Taking into account all three classification levels, the Result Refined is shown to be the best prompt to use with GPT-4 for hierarchical classification in general, and the Experts Group prompt results are shown to be the second-best one.

C. RESULTS COMPARISON

Table 4 shows the performance results achieved by the original Annotation-based system [6], that we consider as our baseline, while Table 5 shows the results of our new approaches; specifically, Table 5 compares the performance of our fine-tuned BERT model, the best LSTM with the Attention model outlined in Section VI-A (Model C), the best prompt-based results introduced in Table 3 (Result Refined).

TABLE 4. Predictive performance of the annotation-based model on different classification levels (class. Levels) of the hierarchy. accuracy is abbreviated with "Acc."; ROC-AUC is abbreviated with "AUC".

	Annotation-based Model				
Class. Levels	Acc.	F_1	Recall	Specificity	AUC
Exam Type	96.0	95.8	95.2	91.9	95.2
Result	75.6	70.8	83.5	71.3	77.4
Lesion Nature	66.3	62.3	73.7	61.6	67.7

With respect to the baseline, the results of the new approaches show better performance under almost all conditions. The LSTM with Attention (Model C) and BERT models outperform the baseline results by a considerable margin, especially for the Result and Lesion Nature levels. The Annotation-based Model is still better than the best GPT-4 prompting technique (Result Refined) in the first two levels (Exam Type and Results), but it shows lower performance in the Lesion Nature level. The GPT-4 prompting technique has difficulty in correctly classifying the Result level, even if the prompt was specifically engineered for that purpose. In fact, the best GPT-4 prompting technique for the last classification level beats the baseline in Accuracy, F_1 and ROC-AUC by 15%, 2% and 10% points, respectively.

Regarding the BERT model, Table 6 shows the result of the 10-fold cross-validation for all relevant metrics. We compute the median, mean, and standard deviation for the Accuracy, F_1, Recall, Specificity and ROC-AUC measures to obtain a robust value for the performance evaluation of each measure. In terms of standard deviation, the Exam Type classification level is deemed the level with less variability between folds. In contrast, the Result level shows the highest variability between fold results, and in particular for the specificity measure. This measure shows the highest variability for all three classification levels as a result of the difficulty in correctly predicting negative samples. When

TABLE 5. Predictive performance of the best model based on LSTM with Attention Mechanism (Model C), BERT, and GPT-4 on the different Classification Levels (class. Levels) of the hierarchy. They are compared and evaluated in terms of accuracy (Acc.), F_1-score (F_1), Recall (Rec.), Specificity (Spec.) and ROC-AUC (AUC). The best and the second-best values are in bold and underlined, respectively. BERT is the best-performing model overall.

	LSTM + Att.					*BERT*					*GPT-4*				
Class. Levels	Acc.	F_1	Rec.	Spec.	AUC	Acc.	F_1	Rec.	Spec.	AUC	Acc.	F_1	Rec.	Spec.	AUC
Exam Type	**96.2**	**96.0**	**98.2**	92.7	**96.3**	**96.2**	95.9	95.7	93.3	95.7	92.3	89.3	83.4	**97.8**	90.6
Result	81.2	72.8	83	74.1	83.6	**87.5**	**84.4**	84.6	**75.9**	**84.6**	70.9	67.5	**88.3**	61.8	75
Lesion Nature	73.2	71.2	79.2	68.8	74	**82.8**	**82.4**	**82.3**	85.6	**82.3**	81.0	64.6	62.4	**92.6**	77.5

TABLE 6. Accuracy (Acc.), F_1, Recall (Rec.), Specificity (Spec.), ROC-AUC (AUC) metrics calculated over the 10 folds for the BERT model, considering the median (Med.), mean ($\overline{\mu}$) and standard deviation ($\overline{\sigma}$). The median is the reference value due to its robustness to outliers.

	Exam Type		*Result*		*Lesion Nature*	
	Med.	$\overline{\mu} \pm \overline{\sigma}$	Med.	$\overline{\mu} \pm \overline{\sigma}$	Med.	$\overline{\mu} \pm \overline{\sigma}$
Acc.	96.2	96.1 ±0.8	87.5	87.2 ±2.1	82.8	82.9 ±1.9
F_1	95.9	95.8 ±0.9	84.4	83.8 ±2.5	82.4	82.5 ±1.9
Rec.	95.7	95.5 ±1.1	84.6	83.4 ±2.7	82.3	82.5 ±1.9
Spec.	93.3	93.2 ±2.4	75.9	74.8 ±5.2	85.6	85.1 ±3.0
AUC	95.7	95.5 ±1.1	84.6	83.4 ±2.7	82.3	82.5 ±1.9

comparing median and mean values, we notice the presence of outlier values in the fold results. In fact, the medians for the Exam Type and Result levels are slightly better than the corresponding means indicating the presence of folds with considerably lower values in performance than in the median case. In contrast, at the level of Lesion Nature level the higher means with respect to the medians in almost all measures indicates the presence of particularly good fold results. For the comparison between BERT and other solutions, we consider the median value for each metrics, to avoid the influence of outlier values in the comparison.

In general, we can say that for all three classification levels, BERT achieves a performance higher than 80% in terms of Accuracy, F_1-score and ROC-AUC, which are the most important performance measures. The level of Lesion Nature shows to be the most difficult classification level. This can be attributed to two primary factors. First, as noted in [27], there is a lack of consensus among physicians on the identification of *Uncertain Nature* cases, often leading to misclassification as *Negative* or *Neoplastic* reports. We hypothesise that these reports contain the most sensitive information, rendering their language particularly ambiguous and enigmatic. Furthermore, certain reports are plausible to be classified as either *Uncertain* or *Negative* depending on the individual physician's judgement. Second, the third level is solely applicable to non-negative reports, thereby limiting the number of reports (fewer than 2000) available for fine-tuning the BERT model.

In general, when comparing the results obtained by BERT with those of the other models, we can observe that BERT almost always performs best in Accuracy, F_1-score and ROC-AUC for all three classification levels. For the Exam Type classification, BERT is only 0.01% less than the best LSTM with the Attention model (Model C) in terms of F_1 and only 0.6% in terms of ROC-AUC, which is practically equivalent to it. The BERT model loses 2.5% in Recall with respect to the LSTM-based one, but still considerably better than GPT-4 on the same metric. Although GPT-4 has the worst performance with respect to accuracy, F_1 and ROC-AUC, it achieves the best specificity value (97.8%), while BERT achieves the second best (93.3%).

For the Result classification task, BERT achieves Accuracy 87.5%, F_1 84.4% and ROC-AUC 84.6%, with a considerable margin in Accuracy (around 6%) and in F_1 (11%) from the second best (LSTM + Att.), and a small margin (1%) in ROC-AUC. On the other hand, GPT-4 shows particularly good performance in Recall, achieving 88.3%, followed by BERT with 84.6%, but it shows very low values on other important metrics.

For the Lesion Nature task, BERT results the best model with a considerable performance gap of 11% and 5% with respect to the second best, in terms of F_1 and ROC-AUC, respectively. LSTM with Attention is the second-best model in terms of F_1 and Recall, while GPT-4 results the second best in terms of Accuracy (with only 1% less than the best) and ROC-AUC. Similarly to the first classification level (Exam Type), GPT-4 shows the best performance in terms of Specificity, achieving a remarkable 92.6% followed by BERT with 85.6%.

The results of our experimental analysis confirm the capability of BERT to provide very good performance, with improvements with respect to the LSTM-based models that are particularly notable for the Result and Lesion Nature tasks, which are the most interesting classification levels. For the Exam Type level, the results of BERT are practically equivalent to those of the LSTM with Attention model and are better than the best GPT-4 prompt, confirming BERT as the best model overall.

Despite the poor results obtained by GPT-4 in the main metrics, its performance is nevertheless noteworthy considering that we did not fine-tune the model, and we did not provide any example in the prompt context for reasons previously discussed in Section V-D. Furthermore, GPT-4 shows good performance in identifying negative samples, due to the high Specificity values for the first and third classification levels, and good Recall for the second level (Result).

VII. CONCLUSION AND FUTURE WORK

Our research demonstrates that a deep learning-based approach can effectively automate the classification of Italian CT reports, potentially leading to significant improvements in efficiency and a reduction in workload within the healthcare domain. This work presents a novel solution that uses a fine-tuned Italian-based BERT model, outperforming our previous approach based on traditional machine learning and manual annotations across all three classification tasks (Exam Type, Result, and Lesion Nature). This achievement highlights the potential of deep learning models for automating complex medical report classification tasks, potentially freeing up valuable time for radiologists and other healthcare professionals. Furthermore, our findings provide valuable information on specific aspects of model performance and open the door to exciting future research directions. In particular, while GPT-4 showed promising results for certain classification levels in a zero-shot setting, more research is needed to comprehensively compare its capabilities with other deep learning models and refine prompt design strategies [22], [23], [64], [66].

For the sake of completeness, we tried to ensemble our models using simple approaches (bagging, voting, etc.) to investigate if performance can be improved, but we did not obtain better results. The BERT model, which performs considerably better than the other models, is penalised by the erroneous predictions made by LSTM with Attention and (or) GPT-4 prompts, which increases the overall error and reduces performance. Therefore, a possible investigation for future work is to study more sophisticated ways of ensembling the different models that we have studied, taking also into account that the GPT-4 prompting technique relies on an API-based prediction, and so its availability is dependent on the network connection and OpenAI server uptime.

Beyond performance improvements, this work paves the way for advances in explainable artificial intelligence (XAI) within the medical domain. Our aim is to explore the integration of XAI techniques to improve the interpretability of model decisions, allowing healthcare professionals to gain deeper insight into the rationale behind classifications and fostering trust in the system. In addition, we plan to investigate the potential for attention mechanisms within deep learning models to identify key sections within CT reports that significantly influence classification [74]. This line of inquiry has the potential not only to improve model performance, but also to provide valuable information about the report sections that are most crucial for accurate diagnoses.

Finally, we aim to evaluate the generalisability of our approach by applying it to the classification of reports from other body parts, such as the abdomen or the encephalon. This exploration may necessitate adaptations or extensions to the current classification schema, which fosters a more comprehensive and generalisable solution.

In conclusion, this work represents a significant step forward in automating medical report classification using deep learning, offering promising avenues for further advancement that can ultimately benefit healthcare professionals and patients alike.

REFERENCES

[1] A. E. Gerevini, R. Maroldi, M. Olivato, L. Putelli, and I. Serina, "Machine learning techniques for prognosis estimation and knowledge discovery from lab test results with application to the COVID-19 emergency," *IEEE Access*, vol. 11, pp. 83905–83933, 2023.

[2] M. Olivato, N. Rossetti, A. E. Gerevini, M. Chiari, L. Putelli, and I. Serina, "Machine learning models for predicting short-long length of stay of COVID-19 patients," *Proc. Comput. Sci.*, vol. 207, pp. 1232–1241, Sep. 2022.

[3] M. Chiari, A. E. Gerevini, M. Olivato, L. Putelli, N. Rossetti, and I. Serina, "An application of recurrent neural networks for estimating the prognosis of COVID-19 patients in northern Italy," in *Proc. 19th Int. Conf. Artif. Intell. Med. (AIME)*, in Lecture Notes in Computer Science, vol. 12721, A. Tucker, P. H. Abreu, J. S. Cardoso, P. P. Rodrigues, and D. Riaño, Eds. Springer, Jun. 2021, pp. 318–328.

[4] T. Mehmood, I. Serina, A. Lavelli, L. Putelli, and A. Gerevini, "On the use of knowledge transfer techniques for biomedical named entity recognition," *Future Internet*, vol. 15, no. 2, p. 79, Feb. 2023.

[5] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[6] A. E. Gerevini, A. Lavelli, A. Maffi, R. Maroldi, A.-L. Minard, I. Serina, and G. Squassina, "Automatic classification of radiological reports for clinical care," *Artif. Intell. Med.*, vol. 91, pp. 72–81, Sep. 2018.

[7] J. Praful Bharadiya, "A comprehensive survey of deep learning techniques natural language processing," *Eur. J. Technol.*, vol. 7, no. 1, pp. 58–66, May 2023.

[8] A. Raj, R. Jindal, A. K. Singh, and A. Pal, "A study of recent advancements in deep learning for natural language processing," in *Proc. IEEE World Conf. Appl. Intell. Comput. (AIC)*, Jul. 2023, pp. 300–306.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst., 27th Annu. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 3111–3119.

[10] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, Doha, Qatar, 2014, pp. 1532–1543.

[11] S. Al-Saqqa and A. Awajan, "The use of Word2vec model in sentiment analysis: A survey," in *Proc. Int. Conf. Artif. Intell., Robot. Control*, Dec. 2019, pp. 39–43.

[12] J. Xiao and Z. Zhou, "Research progress of RNN language model," in *Proc. IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Jun. 2020, pp. 1285–1288.

[13] K. M. Tarwani and S. Edem, "Survey on recurrent neural network in natural language processing," *Int. J. Eng. Trends Technol.*, vol. 48, no. 6, pp. 301–304, Jun. 2017.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.

[15] L. Putelli, A. Gerevini, A. Lavelli, and I. Serina, "Applying self-interaction attention for extracting drug-drug interactions," in *Proc. 18th Int. Conf. Italian Assoc. Artif. Intell.*, in Lecture Notes in Computer Science, vol. 11946, Rende, Italy, M. Alviano, G. Greco, and F. Scarcello, Eds. Springer, Nov. 2019, pp. 445–460.

[16] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, vol. 1, New Orleans, LA, USA, M. A. Walker, H. Ji, and A. Stent, Eds. Association for Computational Linguistics, Jun. 2018, pp. 1101–1111.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., Dec. 2017, pp. 5998–6008.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Minneapolis, MN, USA, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[19] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," 2023, *arXiv:2307.03109*.

[20] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," 2022, *arXiv:2206.07682*.

[21] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2023.

[22] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, Yokohama, Japan, Y. Kitamura, A. Quigley, K. Isbister, and T. Igarashi, Eds. ACM, May 2021, pp. 314:1–314:7.

[23] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with ChatGPT," 2023, *arXiv:2302.11382*.

[24] A. G. Møller, J. A. Dalsgaard, A. Pera, and L. M. Aiello, "The parrot dilemma: Human-labeled vs. LLM-augmented data in classification tasks," 2023, *arXiv:2304.13861*.

[25] J. A. Baktash and M. Dawodi, "Gpt-4: A review on advancements and opportunities in natural language processing," 2023, *arXiv:2305.03195*.

[26] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," 2023, *arXiv:2303.13375*.

[27] L. Putelli, A. E. Gerevini, A. Lavelli, M. Olivato, and I. Serina, "Deep learning for classification of radiology reports with a hierarchical schema," in *Proc. 24th Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, vol. 176, M. Cristani, C. Toro, C. Zanni-Merk, R. J. Howlett, and L. C. Jain, Eds., Sep. 2020, pp. 349–359.

[28] L. Putelli, A. E. Gerevini, A. Lavelli, R. Maroldi, and I. Serina, "Attention-based explanation in a deep learning model for classifying radiology reports," in *Proc. 19th Int. Conf. Artif. Intell. Med. (AIME)*, in Lecture Notes in Computer Science, vol. 12721, A. Tucker, P. H. Abreu, J. S. Cardoso, P. P. Rodrigues, and D. Riaño, Eds. Springer, Jun. 2021, pp. 367–372.

[29] L. Putelli, A. Gerevini, A. Lavelli, T. Mehmood, and I. Serina, "On the behaviour of BERT's attention for the classification of medical reports," in *Proc. CEUR Workshop*, vol. 3277, 2022, pp. 16–30.

[30] A. Mozayan, A. R. Fabbri, M. Maneevese, I. Tocino, and S. Chheang, "Practical guide to natural language processing for radiology," *RadioGraphics*, vol. 41, no. 5, pp. 1446–1453, Sep. 2021.

[31] L. F. Donnelly, R. Grzeszczuk, and C. V. Guimaraes, "Use of natural language processing (NLP) in evaluation of radiology reports: An update on applications and technology advances," in *Seminars in Ultrasound, CT and MRI*, vol. 43. Amsterdam, The Netherlands: Elsevier, 2022, pp. 176–181.

[32] A. Casey, E. Davidson, M. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley, H. Wu, and B. Alex, "A systematic review of natural language processing applied to radiology reports," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, p. 179, Dec. 2021.

[33] A. Casey, E. Davidson, C. Grover, R. Tobin, A. Grivas, H. Zhang, P. Schrempf, A. Q. O'Neil, L. Lee, M. Walsh, F. Pellie, K. Ferguson, V. Cvoro, H. Wu, H. Whalley, G. Mair, W. Whiteley, and B. Alex, "Understanding the performance and reliability of NLP tools: A comparison of four NLP tools predicting stroke phenotypes in radiology reports," *Frontiers Digit. Health*, vol. 5, Sep. 2023, Art. no. 1184919.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[35] M. I. Miller, A. Orfanoudaki, M. Cronin, H. Saglam, I. S. Y. Kim, O. Balogun, M. Tzalidi, K. Vasilopoulos, G. Fanaropoulou, N. M. Fanaropoulou, J. Kalin, M. Hutch, B. R. Prescott, B. Brush, E. J. Benjamin, M. Shin, A. Mian, D. M. Greer, S. M. Smirnakis, and C. J. Ong, "Natural language processing of radiology reports to detect complications of ischemic stroke," *Neurocritical Care*, vol. 37, no. S2, pp. 291–302, Aug. 2022.

[36] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu, "RadBERT: Adapting transformer-based language models to radiology," *Radiol., Artif. Intell.*, vol. 4, no. 4, Jul. 2022, Art. no. e210258.

[37] M. A. Fink, K. Kades, A. Bischoff, M. Moll, M. Schnell, M. Küchler, G. Köhler, J. Sellner, C. P. Heussel, H.-U. Kauczor, H.-P. Schlemmer, K. Maier-Hein, T. F. Weber, and J. Kleesiek, "Deep learning-based assessment of oncologic outcomes from natural language processing of structured radiology reports," *Radiol., Artif. Intell.*, vol. 4, no. 5, Sep. 2022.

[38] F. Galbusera, A. Cina, T. Bassani, M. Panico, and L. M. Sconfienza, "Automatic diagnosis of spinal disorders on radiographic images: Leveraging existing unstructured datasets with natural language processing," *Global Spine J.*, vol. 13, no. 5, pp. 1257–1266, Jun. 2023.

[39] S. C. Fanni, C. Romei, G. Ferrando, F. Volpi, C. A. D'Amore, C. Bedini, S. Ubbiali, S. Valentino, and E. Neri, "Natural language processing to convert unstructured COVID-19 chest-CT reports into structured reports," *Eur. J. Radiol. Open*, vol. 11, Dec. 2023, Art. no. 100512.

[40] S. S. Biswas, "Role of ChatGPT in public health," *Ann. Biomed. Eng.*, vol. 51, no. 5, pp. 868–869, 2023.

[41] J.-J. Zhu, J. Jiang, M. Yang, and Z. J. Ren, "ChatGPT and environmental research," *Environ. Sci. Technol.*, vol. 57, no. 46, pp. 17667–17670, Nov. 2023.

[42] J. H. Choi, K. E. Hickman, A. Monahan, and D. B. Schwarcz, "ChatGPT goes to law school," *J. Legal Educ.*, vol. 71, p. 387, Jan. 2023.

[43] Z. Liu et al., "Evaluating large language models for radiology natural language processing," 2023, *arXiv:2307.13693*.

[44] R. Anil et al., "PaLM 2 technical report," 2023, *arXiv:2305.10403*.

[45] E. Pianta, C. Girardi, and R. Zanoli, "The TextPro tool suite," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, Marrakech, Morocco. European Language Resources Association, May 2008.

[46] V. C. Nguyen, N. Ye, W. S. Lee, and H. L. Chieu, "Conditional random field with high-order dependencies for sequence labeling and segmentation," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 981–1009, 2014.

[47] Y. Li and T. Yang, *Word Embedding for Understanding Natural Language: A Survey*. Cham, Switzerland: Springer, 2018, pp. 83–104.

[48] R. McDonald, G. Brokos, and I. Androutsopoulos, "Deep relevance ranking using enhanced document-query interactions," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, Oct. 2018, pp. 1849–1860.

[49] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000.

[50] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[51] C. Raffel and D. P. W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," 2015, *arXiv:1512.08756*.

[52] R. Kadlec, M. Schmid, O. Bajgar, and J. Kleindienst, "Text understanding with the attention sum reader network," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany. The Association for Computer Linguistics, Aug. 2016.

[53] *GPT-4 System Card*, OpenAI, San Francisco, CA, USA, 2023.

[54] J. Achiam, "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[55] N. Arici, A. E. Gerevini, L. Putelli, I. Serina, and L. Sigalini, "A BERT-based scoring system for workplace safety courses in Italian," in *Proc. 21st Int. Conf. Italian Assoc. Artif. Intell.*, in Lecture Notes in Computer Science, vol. 13796, Udine, Italy, A. Dovier, A. Montanari, and A. Orlandini, Eds. Springer, 2022, pp. 457–471.

[56] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[57] T. B. Brown et al., "GPT-3: Generative pre-trained transformer 3," 2020, *arXiv:2005.14165*.

[58] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI Blog*, Jan. 2018.

[59] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019, *arXiv:1905.14165*.

[60] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "GPT-2: A 1.5B parameter language model," *OpenAI Blog*, Jan. 2019.

[61] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.

[62] T. M. Buonocore, C. Crema, A. Redolfi, R. Bellazzi, and E. Parimbelli, "Localizing in-domain adaptation of transformer-based biomedical language models," *J. Biomed. Informat.*, vol. 144, Aug. 2023, Art. no. 104431.

[63] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Represent (ICLR)*, New Orleans, LA, USA, May 2019.

[64] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," 2023, *arXiv:2305.10601*.

[65] J. Long, "Large language model guided tree-of-thought," 2023, *arXiv:2305.08291*.

[66] H. Sun, A. Hüyük, and M. van der Schaar, "Query-dependent prompt evaluation and optimization with offline inverse RL," 2023, *arXiv:2309.06553*.

[67] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io

[68] M. Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://tensorflow.org

[69] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 281–305, 2012.

[70] H. Hassanzadeh, M. Kholghi, A. N. Nguyen, and K. Chu, "Clinical document classification using labeled and unlabeled data across hospitals," in *Proc. Amer. Med. Inform. Assoc. Annu. Symp. (AMIA)*, San Francisco, CA, USA, Nov. 2018.

[71] B. Shin, F. H. Chokshi, T. Lee, and J. D. Choi, "Classification of radiology reports using neural attention models," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Anchorage, AK, USA, May 2017, pp. 4363–4370.

[72] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, S. Amin, and H. Liu, "A clinical text classification paradigm using weak supervision and deep representation," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, p. 1, Dec. 2019.

[73] T. Mehmood, A. Gerevini, A. Lavelli, and I. Serina, "Multi-task learning applied to biomedical named entity recognition task," in *Proc. 6th Italian Conf. Comput. Linguistics*, vol. 2481, R. Bernardi, R. Navigli, and G. Semeraro, Eds., Bari, Italy, Nov. 2019.

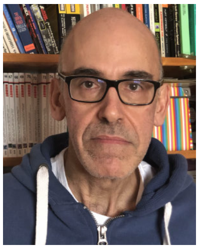
[74] L. Putelli, A. E. Gerevini, A. Lavelli, and I. Serina, "The impact of self-interaction attention on the extraction of drug-drug interactions," in *Proc. 6th Italian Conf. Comput. Linguistics*, vol. 2481, Bari, Italy, R. Bernardi, R. Navigli, and G. Semeraro, Eds., vsss, Nov. 2019.

MATTEO OLIVATO received the Ph.D. degree from the Department of Information Engineering, University of Brescia, in 2022. He is currently a Research Fellow with the Department of Information Engineering, University of Brescia. His research interests include the application of machine learning and deep learning for clinical data, deep learning for prognosis estimation and health applications, and machine learning in the domain of cybersecurity.

LUCA PUTELLI received the Ph.D. degree from the Department of Information Engineering, University of Brescia, in 2021. He is currently a Researcher with the Department of Information Engineering, University of Brescia. His research interests include the application of natural language processing in the biomedical domain, machine learning for prognosis estimation and health applications, the use of transformer-based architectures for the Italian language, and the application of deep learning techniques in the planning and scheduling domain.

NICOLA ARICI is currently pursuing the Ph.D. degree with the Department of Information Engineering, University of Brescia. His research interests include AI, NLP, deep learning, and machine learning, with works on identifying gender bias in BERT and workplace safety courses and ticket assignments. He is developing prompting techniques to exploit recent LLM solutions for research and industrial applications.

ALFONSO EMILIO GEREVINI was a Research Scientist with IRST (Now FBK), Trento, Italy, from 1989 to 1995. In 1995, he joined the University of Brescia, Italy, where he is currently a Full Professor of information processing systems with the Department of Information Engineering. He is the author or coauthor of more than 150 published articles on various aspects of artificial intelligence. He is also involved in a number of research projects in AI, two of which are funded by the EU. His research interests include the fundamental and applied issues of automated planning, knowledge representation and reasoning, machine learning, deep learning, and data mining. He is a fellow of the European Association for Artificial Intelligence (EurAI). He served as a program committee member of the most prestigious AI conferences for many years. He was an Editorial Board Member and an Associate Editor of *Artificial Intelligence* (Elsevier), for several years, and an Editorial Board Member of *Journal of Artificial Intelligence Research* and *Intelligenza Artificiale* (IOS Press).



ALBERTO LAVELLI was a Researcher with the Cognitive and Communication Technologies Division, Istituto per la Ricerca Scientifica e Tecnologica (ITC-irst), from 1988 to 2005. Since 2005, he has been a Senior Researcher with the NLP Research Unit, Fondazione Bruno Kessler, Center for Augmented Intelligence. He has worked on several research projects, including the European project E3C (European Clinical Case Corpus), the project "MelanoBase" in cooperation with the University of Zurich, and the project "Automatic Classification of Radiological Reports for Clinical Care" in cooperation with the University of Brescia and ASST Spedali Civili di Brescia. He is currently working on Horizon Europe projects eCREAM and IDEA4RC. His research interests include machine learning techniques for information extraction from text, information extraction in the biomedical domain, and parsing techniques (both constituency parsing and dependency parsing from a multilingual perspective).

IVAN SERINA received the Ph.D. degree in computer science engineering from the University of Brescia, Italy, in 2000, and a Marie Curie Fellowship in the field of planning and scheduling from the University of Strathclyde, Glasgow, in 2003. He was with the Faculty of Education, Free University of Bozen-Bolzano, from 2008 to 2012. He is currently an Associate Professor of information processing systems with the Department of Information Engineering, University of Brescia. His research interests include the development and the experimental analysis of machine learning, deep learning, efficient automatic domain independent AI planning techniques with innovative applications in several areas, including medicine and health care, predictive maintenance and intelligent manufacturing for Industry 4.0, and AI for e-learning.

Open Access funding provided by 'Università degli Studi di Brescia' within the CRUI CARE Agreement