Towards Automating Lung-RADS Classification: Insights from Portuguese Radiology Reports

Tarcísio Lima Ferreira[1], **Marcelo Costa Oliveira**[1], **Juliana Simon Petruceli**[1]

[1] Computing Institute, Federal University of Alagoas

{tlf,oliveiramc,jsp}@ic.ufal.br

Abstract. *Early lung cancer detection requires efficient nodule classification. This study compares automated strategies for assigning Lung-RADS categories to Portuguese CT reports using DL and LLMs. Analyzing 963 reports, we employed NER (BiLSTM-CRF, BioBERTpt) and QA (GPT-4o, Gemini 1.5 Flash, Llama 3 70B) with prompt engineering. Gemini 1.5 Flash achieved the highest evaluation metrics (macro-F1: 0.58; weighted-F1: 0.67). Results demonstrate a scalable, adaptable pathway for automating radiology workflows in non-English clinical settings, enhancing diagnostic efficiency through structured NLP post-processing.*

1. Introduction

Lung cancer ranks as the second most frequently diagnosed cancer among both men and women, following only skin cancer in prevalence. Despite advancements in early detection and treatment, it remains the leading cause of cancer-related mortality, responsible for more deaths in 2023 than breast and prostate cancers combined [Siegel et al. 2023]. In Brazil, lung cancer is the third most common type of cancer among men and the fourth most common among women [Mathias et al. 2020]. This emphasizes lung cancer's substantial contribution to overall cancer mortality, highlighting the need for robust strategies in prevention, early detection, and treatment to address this disease effectively [Siegel et al. 2023].

Lung cancer screening (LCS) is a process designed to detect lung cancer in individuals at risk, particularly those with a history of smoking. It involves annual Low-Dose Computed Tomography (LDCT) scans, careful interpretation of the results, and timely follow-up care to ensure early detection and treatment [Deffebach and Humphrey 2015]. The National Lung Screening Trial (NLST) demonstrated that individuals undergoing annual LDCT scans experienced a consistent reduction of approximately 8-11% in lung cancer mortality [National Lung Screening Trial Research Team 2019].

Multiple professional societies, including the American College of Radiology (ACR) and the Fleischner Society, have published guidelines for managing patients with pulmonary nodules detected during lung cancer screening. The guidelines are an important tool for screening programs aimed at reducing unnecessary follow-up exams and guiding optimal patient management. Besides, the guidelines offer more flexibility in follow-up intervals and provide tailored recommendations based on individual risk factors, thus enhancing the ability of radiologists, clinicians, and patients to make well-informed decisions [MacMahon et al. 2017].

One of these guidelines, the Lung CT Screening Reporting & Data System (Lung-RADS®), is a standardized classification system for lung nodules detected in imag-

ing exams such as Computed Tomography (CT) scans. Lung-RADS assesses the risk of malignancy (cancer) in these nodules and guides subsequent management decisions. Lung-RADS is a classification system based on specific nodule characteristics that guides follow-up decisions, ranging from LDCT to a Positron Emission Tomography-Computed Tomography (PET-CT) or biopsy. The higher the risk of malignancy, the higher the Lung-RADS index [of Radiology 2022].

Determining follow-up examinations according to the Lung-RADS guidelines for lung cancer screening CT is a straightforward process for individuals enrolled in a lung cancer screening program. However, extracting and organizing relevant clinical information in a structured format, as required by the Lung-RADS criteria, presents significant challenges. This difficulty arises because clinical data are often recorded in free-text format within chest CT reports, making conversion to structured data time-consuming and prone to information loss [Kreimeyer et al. 2017]. The linguistic complexities of languages such as Portuguese, including acronyms, negations, and intricate grammatical structures, further complicate this task and hinder accurate interpretation. Additionally, cultural differences and varied descriptive styles can introduce inconsistencies in the data [da Rocha et al. 2023].

Natural Language Processing (NLP) and Named Entity Recognition (NER) are essential for managing large-scale clinical data, enabling the identification and categorization of entities in medical records [Pandey et al. 2020, Li et al. 2022]. However, research on information extraction (IE) for lung nodules remains heavily focused on English and Chinese [Zheng et al. 2021, Hu et al. 2024a, Hu et al. 2024b]. This focus, coupled with a lack of public datasets in other languages limits the development of IE methods for diverse radiology contexts.

While rule-based NLP systems were initially successful for lung nodule classification [Gershanik et al. 2011, Nobel et al. 2020, Zheng et al. 2021], linguistic variability led to a shift toward deep learning (DL) models like BiLSTM-CRF and CNNs [Fei et al. 2022, da Rocha et al. 2023, Ferreira et al. 2023]. The emergence of Transformer-based models (e.g., BERT, BioBERT) [Vaswani et al. 2023, Devlin et al. 2019, Lee et al. 2019, Alsentzer et al. 2019, Gu et al. 2021] and, more recently, Large Language Models (LLMs) like GPT-4 and MedPaLM-2, has pushed performance toward near-human levels on medical benchmarks [OpenAI et al. 2024, Chowdhery et al. 2022, Singhal et al. 2023, Wu et al. 2023, Bedi et al. 2024].

In this context, this study evaluates the effectiveness of DL and LLMs in extracting lung nodule features from Portuguese CT reports using NER and Question Answering (QA). These extracted insights feed a rule-based algorithm for automated Lung-RADS classification.

The main contributions of this work are as follows:

- **Comparative Analysis of Automated Strategies**: A comprehensive evaluation of different computational approaches for Lung-RADS classification, determining the most effective methods for interpreting CT report data.
- **Scalable Workflow for Multilingual Clinical Environments**: The development of a reproducible framework to automate radiology workflows, specifically tailored for non-English speaking healthcare settings where such tools are currently

scarce.

- **Annotated Portuguese Corpus for Chest CT Scans**: Creation of a specialized dataset of chest CT reports in Portuguese, providing a resource for future research in medical NLP.

2. Methods

2.1. Lung-RADS

The Lung-RADS (ACR) framework standardizes pulmonary nodule assessment to enhance diagnostic accuracy and clinical consistency [Beyer et al. 2017]. This study utilizes the Lung-RADS 2022 categories, summarized below:

Category 0 (Incomplete): Incomplete findings; requires additional imaging or prior CT for comparison.
Category 1 (Negative): No nodules or benign features (e.g., fat-containing); 12-month follow-up.
Category 2 (Benign Appearance): Likely benign features or size; 12-month follow-up.
Category 3 (Probably Benign): Low malignancy risk; 6-month follow-up.
Category 4 (Suspicious): Includes **4A** (3-month follow-up), **4B** (biopsy or PET/CT recommended), and **4X** (highest suspicion due to spiculation or growth).

Full guidelines and management protocols are available at the official ACR documentation[1].

2.2. Dataset Annotation

The 963 chest high-dose CT reports in Portuguese were collected from January 1, 2022, to April 3, 2023, at the University Hospital of Alagoas. The research was approved by the ethics and research committee of the Federal University of Alagoas with the number: 74747817.4.0000.5013. After obtaining consent from all patients in this study, 963 chest CT reports, irrespective of the clinical indication, were included. All reports were anonymized and underwent a cleaning pipeline, including special character removal and whitespace normalization. Cleaned reports were then imported into the Doccano platform [Nakayama et al. 2018] for annotation.

To establish a robust gold standard, the corpus was manually annotated with six semantic entities essential for Lung-RADS 2022 classification [of Radiology 2022]: **Finding** (nodule presence), **Size** (maximum diameter), **Attenuation** (density), **Calcification** (morphological patterns), **Edges** (margins), and **Localization** (anatomical site). We adopted the Inside-Outside-Beginning (IOB) tagging scheme to accurately delimit entity boundaries. As a result, the annotation tool generated a JSON file containing the labeling information for all reports. Next, we split each report and its labeling information into sentences and tokenize them using the BERT tokenizer. Finally, we padded each sequence of integers representing a report and its labeling information to a fixed size. This step was necessary because models like BERT require a specific input sequence length.

The dataset was partitioned into a 70% training set and a 30% test set (289 reports) for the NER and QA tasks. For the IE task, the same 30% CT reports (289) were utilized

[1] https://edge.sitecorecloud.io/americancoldf5f-acrorgf92a-productioncb02-3650/media/ACR/Files/RADS/Lung-RADS/Lung-RADS-2022.pdf

to perform the Question-answering procedure. The IE task utilized 8 targeted questions based on Lung-RADS criteria (Table 1). A senior thoracic radiologist (15+ years of experience) provided the ground-truth answers for the 289-report test set. All reports in this subset contained pulmonary nodules, with the following Lung-RADS distribution: Category 0 (47 cases), Category 1 (182 cases), Category 2 (37 cases), Category 3 (2 cases), Category 4A (11 cases), Category 4B (3 cases), and Category 4X (7 cases).

Table 1. Questions and statistics for annotated answers ($n = 289$).

No.	Question	Type	Positives/Stats
1	Report ID	Num.	-
2	Solid or soft tissue attenuation?	Bool.	33
3	Ground-glass attenuation?	Bool.	5
4	Spiculated or irregular borders?	Bool.	8
5	Is the nodule calcified?	Bool.	119
6	Nodule location	Cat.	RUL: 39, RML: 27, RLL: 41, LUL: 38, LLL: 32, Others: 28
7	Nodule size (mean $\pm$ SD)	Num.	5.40 ± 4.20 mm

Notes: RUL/LUL: Right/Left Upper Lobe; RML: Right Middle Lobe; RLL/LLL: Right/Left Lower Lobe.

2.3. Information Extraction Techniques

To extract information from the pulmonary nodules described in chest CT reports, we employed two natural language processing techniques. First, NER was employed to detect and categorize rigid designators (e.g., nodule size, location, and attenuation) into predefined semantic types, following the methodology described by Li et al. [Li et al. 2022]. Complementarily, Question Answering (QA) was used to interpret the intent of clinical descriptions and to retrieve structured information from free-text reports [Hirschman and Gaizauskas 2001].

2.4. Models for Named Entity Recognition

For the NER task, we compared a BiLSTM-CRF architecture with BioBERTpt, a transformer-based model fine-tuned with Portuguese clinical narratives and abstracts from PubMed and SciELO. The BiLSTM-CRF hyperparameter configuration followed the experimental setup described in [Ferreira et al. 2023], using word embeddings and LSTM units of size 50, a learning rate of 0.01, and a dropout rate of 0.1.

BioBERTpt was fine-tuned using the HuggingFace Transformers library (PyTorch v2.0.1). Following the recommendations in [Schneider et al. 2020], we employed the AdamW optimizer with a weight decay of 0.01, a learning rate of 3×10^{-5}, and a linear scheduler with 10% warm-up. Fine-tuning task was conducted on an NVIDIA RTX 3060 12GB GPU with a batch size of 4 for 10 epochs. The BiLSTM-CRF and BioBERTpt models aim to predict entity tags in IOB format for each token in the input sequence. Each input sequence corresponded to a chest CT report, with entity tags representing one of six named entities related to the nodule characteristics.

The automated Lung-RADS classification pipeline using the BiLSTM-CRF and BioBERTpt models with the NER technique is illustrated in Figure 1. Raw CT reports undergo preprocessing (cleaning, tokenization, and IOB tagging via Doccano) before being fed to the NER models. These models predict tags for six distinct entity classes related

to nodule characteristics. In the post-processing phase, the identified entities are populated into a Question Table (Table 1)), which serves as input to a rule-based algorithm that assigns the final Lung-RADS category.

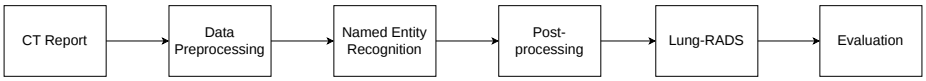

Figure 1. Methodology scheme applied in NER task for Lung-RADS classification.

2.5. Post-processing Named Entity Recognition Extraction

Following the NER stage, a systematic post-processing pipeline was implemented to transform raw entity tags into a structured representation. We used Regular Expressions (Regex) to parse the IOB-tagged outputs from both BiLSTM-CRF and BioBERTpt, mapping identified entities (e.g., size, attenuation, and location) to a standardized Question Table (Table 1). This process involved consolidating overlapping entities and normalizing numerical values (e.g., converting '1.5 cm' to '15 mm'). The structured data then served as input for a deterministic rule-based algorithm, which applied the Lung-RADS criteria to assign the final malignancy risk category.

2.6. Models for Question Answering

To evaluate the potential of LLMs models, we experimented with the following models: GPT-4o (gpt-4o-2024-05-13), Gemini 1.5 Flash (gemini-1.5-flash), and Llama 3 70B . To ensure reproducibility and minimize randomness in response generation, the temperature was set to 0 for all models. This is important in QA tasks, where the accuracy of the information extracted is crucial. The models were accessed through their respective APIs (OpenAI, Google AI, and Together AI). Although GPT-4o represents a high-parameter proprietary baseline, Llama 3 70B, and Gemini 1.5 Flash were included as more cost-effective and efficient alternatives for large-scale radiology workflow automation.

2.7. Prompt Engineering

Based on the work of Danqing Hu et al. [Hu et al. 2024a], we designed the prompts used for QA. Hu et al. used zero-shot (ZS) learning in their input prompts. In addition to ZS learning, we also employed few-shot (FS) learning [Brown et al. 2020]. Our prompt templates consist of three parts: (1) Original CT report; (2) IE instructions and an unfilled Question table; and (3) Additional requirements for the IE task. In this work, the LLMs were instructed to respond with "No" as the default answer for Questions that do not have corresponding information in the given CT report. We provide annotated reports with completed tables to enhance LLMs' task comprehension and result accuracy.

For the FS approach, we implemented a dynamic context selection strategy based on semantic similarity. We used the (paraphrase-multilingual-mpnet-base-v2) model [Reimers and Gurevych 2020] to compute embeddings for the training and test reports. The cosine similarity was then used to measure the similarity between the test reports and the training reports. This step facilitated the identification of similar training examples for FS learning, enhancing task comprehension.

The testing framework used one prompt for ZS learning and another prompt for FS learning. For FS learning, the prompt template was created with five examples, each

combining CT reports with the prompt template to generate answers to questions. This combined prompt is submitted via API to the LLMs, and their responses are obtained. A new request is made for each CT report, preventing previous requests from influencing the QA results. Besides, the LLMs' responses are requested in JSON format to facilitate post-processing of the results. The responses from these language models do not always consist solely of the completed Question table. Therefore, any additional text is disregarded as irrelevant to the analysis. The focus is exclusively on extracting the content into the Question table.

The end-to-end QA pipeline for automated Lung-RADS classification is synthesized in Figure 2. CT reports are concatenated with ZS or similarity-based FS prompts to guide the LLM in extracting nodule-specific characteristics. Although the initial model output is generated as free text, it is constrained by a JSON schema to ensure structural integrity. This post-processing step converts the response into a set of key-value pairs (e.g.,"calcification": "Yes") that populate the Question Table 1. Finally, this structured representation serves as input for a deterministic rule-based algorithm that assigns the definitive Lung-RADS category.

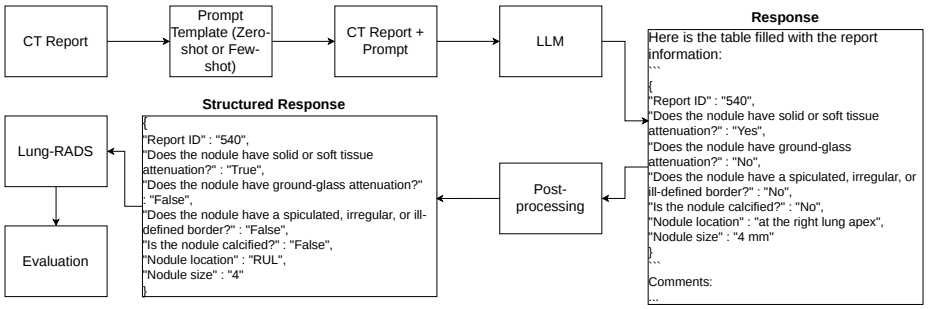

Figure 2. Methodology scheme applied in QA task for Lung-RADS classification.

2.8. Post-processing and Output Normalization

Although LLMs were instructed to generate structured JSON outputs, inconsistencies such as conversational fillers or incomplete tables required a robust post-processing pipeline. Any text outside the schema boundaries was programmatically removed. To ensure a binary state for the rule-based classifier, blank responses or labels such as "Not informed" were assigned the default "No" value. For the categorical attributes in Table 1 (Questions 2–5), responses were converted into Boolean types.

2.9. Lung-RADS Classification: Radiologist Analysis

After information on lung nodules was obtained during the QA and NER tasks, a rule-based algorithm was used to assign the Lung-RADS index to lung nodules described in chest CT reports. To validate the effectiveness of this approach, a rigorous evaluation process was implemented, which included the random selection of 30 test reports for independent review by a radiologist. These 30 reports were chosen due to the limited number of representative examples for the various Lung-RADS categories within the full set of 289 test reports. During this review, the radiologist was asked to:

- Assess the Lung-RADS index assigned by the rule-based algorithm and indicate whether he agreed or disagreed with the generated classification;

- Provide a detailed justification for his evaluation.

This methodological approach enables a comprehensive assessment of the AI tool's performance by comparing machine-generated classifications with expert human interpretations. By soliciting specific rationales for agreement or disagreement, we can identify potential systematic biases or limitations in the AI's Lung-RADS index attribution process.

This iterative refinement process involved:

- Examining the specific cases of misclassification
- Identifying potential sources of algorithmic bias
- Modifying the existing rule set to improve diagnostic accuracy

2.10. Evaluation

To evaluate the effectiveness of BiLSTM-CRF and BioBERTpt in the NER task, as well as the effectiveness of LLMs in the QA task, we used precision, recall, and F1-score as evaluation metrics. For the QA task, we applied the Friedman test [Rainio et al. 2024] followed by the Nemenyi post-hoc test [Rainio et al. 2024] to compare the F1-scores of Gemini 1.5 Flash, GPT-4o, and Llama 3 70B models in ZS and FS learning settings, in order to assess statistically significant differences in effectiveness. Additionally, to evaluate the effectiveness of BiLSTM-CRF, BioBERTpt, and LLMs combined with a rule-based algorithm for Lung-RADS classification, we also employed precision, recall, and F1-score metrics (macro and weighted).

3. Results

3.1. Information Extraction Effectiveness

The comparative effectiveness for NER and QA tasks is summarized in Table 2. For the DL models (BiLSTM-CRF and BioBERTpt), the metrics reflect the accuracy of entity delimitation for IOB-tagged entities. For LLMs, metrics assess their ability to retrieve structured clinical data via prompted questions.

Table 2. Evaluation Metrics for Models.

Model	Task	Precision	Recall	F1-Score
BiLSTM-CRF	NER	0.85	0.89	0.87
BioBERTpt	NER	0.81	0.89	0.85
Gemini 1.5-Flash ZS	QA	0.80	0.79	0.79
Gemini 1.5-Flash FS	QA	0.84	0.82	0.82
GPT-4o ZS	QA	0.80	0.78	0.78
GPT-4o FS	QA	0.84	0.84	0.84
Llama 3 70B ZS	QA	0.79	0.77	0.77
Llama 3 70B FS	QA	0.88	0.88	0.88

The BiLSTM-CRF model achieved a higher F1-score (0.87) than BioBERTpt (0.85), primarily due to higher precision in identifying nodules' characteristics. Regarding LLMs, the transition from ZS to FS learning yielded significant improvements across all architectures. Llama 3 70B exhibited the most substantial gain, with an increase in F1-score of 0.11 (0.77 to 0.88).

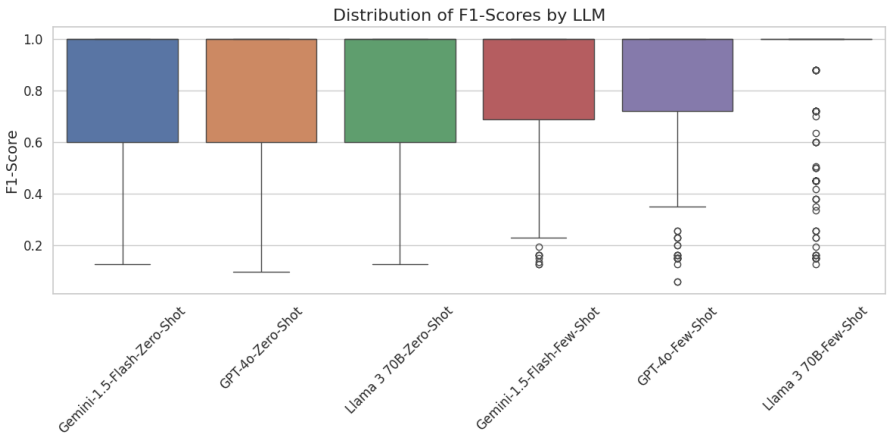

Figure 3. Boxplot with F1-Score by LLM.

Although FS learning generally resulted in higher mean F1-scores, it also led to a greater number of outliers, suggesting increased variability in effectiveness in finding the correct answer to the Questions Table 1 among different reports.

3.2. Statistical Significance and Nemenyi Analysis

To assess whether these differences were statistically significant, we performed the Friedman test, which yielded a $p-value < 0.0001$, indicating significant differences between the models. We then applied the Nemenyi post-hoc test to identify which pairs of models differed significantly. The results of the Nemenyi test, shown in Figure 4, reveal statistically significant differences $(p - value < 0.05)$ between the following model pairs:

- Llama 3 70B FS vs. Gemini 1.5 Flash ZS
- Llama 3 70B FS vs. GPT-4o ZS
- Llama 3 70B FS vs. Llama 3 70B ZS
- GPT-4o FS vs. GPT-4o ZS
- GPT-4o FS vs. Llama 3 70B ZS
- Gemini 1.5 Flash FS vs. GPT-4o ZS
- Gemini 1.5 Flash FS vs. Llama 3 70B ZS

Given the statistically significant differences identified between models using ZS and FS learning techniques, as confirmed by the Friedman test ($p-value < 0.0001$) and the Nemenyi post-hoc analysis (Figure 4), the Lung-RADS classification step for LLMs was based on the FS learning results.

3.3. Lung-RADS Classification Accuracy

Table 3 shows the precision, recall, and F1-score for all evaluated models combined with the rule-based algorithm for Lung-RADS classification.

In general, transformer-based architectures outperformed the BiLSTM-CRF model in most categories. The LLM-based approaches (Gemini 1.5 Flash, GPT-4o, and Llama 3 70B) achieved substantially higher F1 Scores in both the low-risk (categories 0–2) and high-risk (4A–4X) groups. Gemini 1.5 Flash obtained the highest macro-average F1-score (0.58), followed by Llama 3 70B (0.52) and Gpt-4o (0.39), whereas BioBERTpt

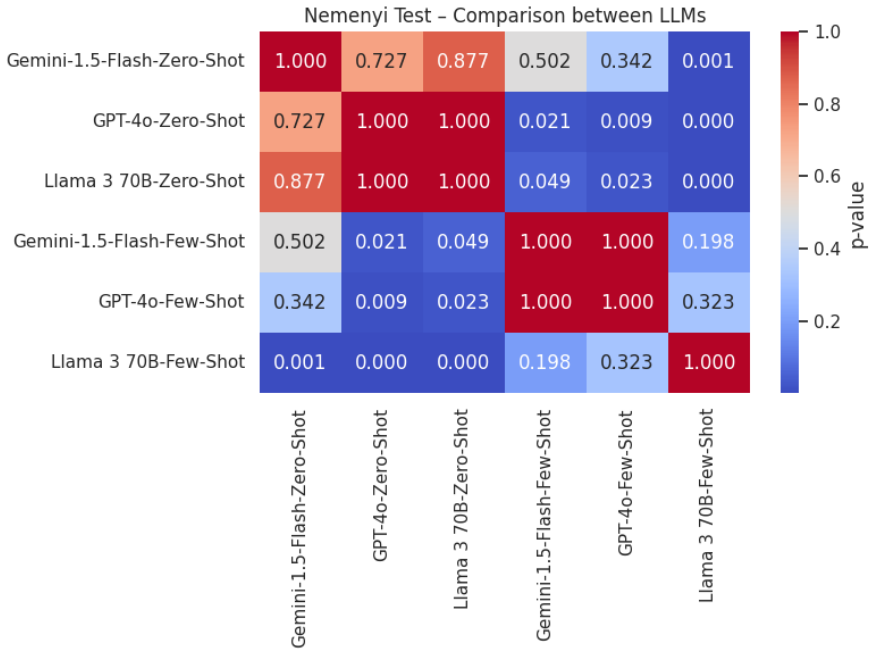

Figure 4. P-value Nemenyi test between LLMs.

Table 3. Lung-RADS Classification Results

Lung-RADS Category		**0**	**1**	**2**	**3**	**4A**	**4B**	**4X**	**Macro**	**Weighted**
Examples		47	182	37	2	11	3	7		
BiLSTM-CRF	Precision	0.13	0.64	0.07	0.00	0.20	0.00	0.00	0.15	0.44
	Recall	0.11	0.80	0.03	0.00	0.09	0.00	0.00	0.15	0.53
	F1-score	0.12	0.71	0.04	0.00	0.13	0.00	0.00	0.14	0.48
BioBERTpt	Precision	0.09	0.64	0.12	0.00	0.20	0.00	0.00	0.15	0.44
	Recall	0.09	0.77	0.05	0.00	0.09	0.00	0.00	0.14	0.51
	F1-score	0.09	0.70	0.07	0.00	0.13	0.00	0.00	0.14	0.47
Gemini 1.5 Flash	Precision	0.39	0.81	0.39	0.17	0.67	0.67	1.00	0.58	0.68
	Recall	0.19	0.79	0.65	0.50	0.73	0.67	0.86	0.58	0.68
	F1-score	0.26	0.80	0.49	0.25	0.70	0.67	0.92	0.58	0.67
GPT-4o	Precision	0.33	0.80	0.53	0.00	0.60	0.00	0.57	0.40	0.66
	Recall	0.38	0.77	0.57	0.00	0.27	0.00	0.57	0.37	0.65
	F1-score	0.35	0.79	0.55	0.00	0.38	0.00	0.57	0.38	0.65
Llama 3 70B	Precision	0.52	0.81	0.53	0.00	0.71	0.67	1.00	0.60	0.72
	Recall	0.34	0.91	0.54	0.00	0.45	0.67	0.43	0.48	0.73
	F1-score	0.41	0.86	0.53	0.00	0.56	0.67	0.60	0.52	0.72

and BiLSTM-CRF achieved considerably lower macro-F1 scores (0.14). These results indicate that generative LLMs, when combined with structured post-processing, provide superior generalization across heterogeneous Lung-RADS categories.

Effectiveness differences were particularly pronounced in high-risk categories. The BiLSTM-CRF model failed to correctly identify any instances of categories 3, 4B, and 4X, yielding an F1-score of 0.00. This limitation is likely associated with its reduced ability to capture complex semantic cues related to irregular margins and suspicious morphological descriptors, which are critical for 4X classification. BioBERTpt

achieved moderate improvements but remained substantially below LLM effectiveness in most high-risk categories.

In contrast, LLM-based models exhibited more balanced behavior across categories. Llama 3 70B achieved high effectiveness in categories 1, 4B, and 4X, while GPT-4o showed competitive effectiveness only in non-relevant high-risk classes. Furthermore, Gemini 1.5 Flash demonstrated high effectiveness across a diverse range of categories, covering both non-risk and high-risk classes such as 1, 4A, 4B, and 4X. These findings suggest that large generative models may better integrate multiple textual cues required for risk stratification.

It is important to note that class imbalance likely influenced macro-level performance. Categories such as 3 (n=2) and 4B (n=3) contained very few samples, which may have amplified performance variance and penalized macro-F1 scores. Future studies with larger, more balanced datasets are necessary to confirm robustness across rare but clinically critical categories. Additionally, it is important to acknowledge that the dataset was collected from a single hospital, which limits the demographic and clinical variability of the sample. As a result, the findings may not fully generalize to different patient populations, institutional protocols, or reporting practices across other healthcare settings. Multi-center evaluations are therefore necessary to assess the external validity and broader applicability of the proposed approach.

4. Conclusion

In this study, we compared two current NLP approaches, deep learning–based named entity recognition and LLM-based question answering with prompt engineering, to extract lung nodule characteristics from chest CT reports written in Portuguese to automate the Lung-RADS classification. Among the evaluated methods, Llama 3 70B achieved the best overall results, suggesting that LLMs can reliably capture clinically relevant attributes needed for Lung-RADS scoring in Portuguese narratives. To our knowledge, this is the first study to apply LLMs to Lung-RADS scoring in Portuguese, and it provides evidence that LLM-centered pipelines can accelerate the development of more adaptable, maintainable systems for report structuring and risk categorization. Future work should assess generalization across institutions and reporting styles, and validate the approach under prospective or real-world deployment conditions.

References

Alsentzer et al. (2019). Publicly available clinical bert embeddings.

Bedi, Jain and Shah (2024). Evaluating the clinical benefits of llms. *Nature Medicine*.

Beyer et al. (2017). Automatic Lung-RADS™ classification with a natural language processing system. *J Thorac Dis*, 9(9):3114–3122.

Brown et al. (2020). Language models are few-shot learners.

Chowdhery et al. (2022). Palm: Scaling language modeling with pathways.

da Rocha et al. (2023). Natural language processing to extract information from portuguese-language medical records. *Data*, 8(1).

Deffebach and Humphrey (2015). Lung cancer screening. *Surg Clin North Am*, 95(5):967–978.

Devlin et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Fei et al. (2022). Quality management of pulmonary nodule radiology reports based on natural language processing. *Bioengineering (Basel)*, 9(6).

Ferreira, Oliveira and De Almeida Vieira (2023). Lung-rads + ai: A tool for quantifying the risk of lung cancer in computed tomography reports. In *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 292–297.

Gershanik, Lacson and Khorasani (2011). Critical finding capture in the impression section of radiology reports. *AMIA Annu Symp Proc*, 2011:465–469.

Gu et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Hirschman and Gaizauskas (2001). Natural language question answering: The view from here. *Natural Language Engineering*, 7:275 – 300.

Hu et al. (2024a). Zero-shot information extraction from radiological reports using chatgpt. *International Journal of Medical Informatics*, 183:105321.

Hu et al. (2024b). Improving large language models for clinical named entity recognition via prompt engineering.

Kreimeyer et al. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73:14–29.

Lee et al. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Li et al. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

MacMahon et al. (2017). Guidelines for management of incidental pulmonary nodules detected on CT images: From the fleischner society 2017. *Radiology*, 284(1):228–243.

Mathias et al. (2020). Lung cancer in brazil. *Journal of Thoracic Oncology*, 15(2):170–175.

Nakayama et al. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

National Lung Screening Trial Research Team (2019). Lung cancer incidence and mortality with extended follow-up in the national lung screening trial. *J Thorac Oncol*, 14(10):1732–1742.

Nobel et al. (2020). Natural language processing in dutch free text radiology reports: Challenges in a small language area staging pulmonary oncology. *Journal of Digital Imaging*, 33(4):1002–1008.

of Radiology (2022). Lung-rads® v2022. `https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads`. Accessed: 2023-05-01.

OpenAI et al. (2024). Gpt-4 technical report.

Pandey et al. (2020). Extraction of radiographic findings from unstructured thoracoabdominal computed tomography reports using convolutional neural network based natural language processing. *PLoS One*, 15(7):e0236827.

Rainio, Teuho and Klén (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

Reimers and Gurevych (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Schneider et al. (2020). BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In Rumshisky et al., editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Siegel et al. (2023). Cancer statistics, 2023. *CA Cancer J Clin*, 73(1):17–48.

Singhal et al. (2023). Towards expert-level medical question answering with large language models.

Vaswani et al. (2023). Attention is all you need.

Wu et al. (2023). Pmc-llama: Towards building open-source language models for medicine.

Zheng et al. (2021). Natural language processing to identify pulmonary nodules and extract nodule characteristics from radiology reports. *Chest*, 160(5):1902–1914.