UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

Dissertação de Mestrado

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

de **Tarcísio Lima Ferreira**

orientado por
Prof. Dr. Marcelo Costa Oliveira
Prof. Dr. Thales Miranda de Almeida Vieira

Universidade Federal de Alagoas
Instituto de Computação
Maceió, Alagoas
13 de Fevereiro de 2025

UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

TOWARDS AUTOMATING LUNG-RADS CLASSIFICATION IN CLINICAL ROUTINE: INSIGHTS FROM PORTUGUESE RADIOLOGY REPORTS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal de Alagoas como requisito parcial para a obtenção do grau de Mestre em Informática.

Tarcísio Lima Ferreira

Orientador: Prof. Dr. Marcelo Costa Oliveira
Coorientador: Prof. Dr. Thales Miranda de Almeida Vieira

Banca Avaliadora:

Prof. Dr. Álvaro Alvares de Carvalho César Sobrinho	UFAL - Instituto de Computação
Prof. Dr. Paulo Mazzoncini de Azevedo Marques	USP - Universidade de São Paulo

Maceió, Alagoas
13 de Fevereiro de 2025

Catalogação na Fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico

Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 - 1767

F383t Ferreira, Tarcísio Lima.

Towards automating lung-rads classification in clinical routine : insights from portuguese radiology reports / Tarcísio Lima Ferreira. – 2025.

63 f. : il.

Orientador: Marcelo Costa Oliveira.

Co-orientador: Thales Miranda de Almeida Vieira.

Dissertação (mestrado em informática) - Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2025.

Bibliografia: f. 55-63.

1. Neoplasias pulmonares. 2. Lung-RADS. 3. Processamento de linguagem natural (Computação). 4. Armazenamento e recuperação da informação. 5. Large language models. I. Título.

CDU: 004:616.24-006.6

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
Av. Lourival Melo Mota, S/N, Tabuleiro do Martins, Maceió - AL, 57.072-970
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO (PROPEP)
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Folha de Aprovação

TARCÍSIO LIMA FERREIRA

TOWARDS AUTOMATING LUNG-RADS CLASSIFICATION IN CLINICAL ROUTINE: INSIGHTS FROM PORTUGUESE RADIOLOGY REPORTS

AUTOMATIZAÇÃO DA CLASSIFICAÇÃO DE LUNG-RADS NA ROTINA CLÍNICA: INSIGHTS DOS RELATÓRIOS DE RADIOLOGIA PORTUGUESA

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 13 de fevereiro de 2025.

Banca Examinadora:

gov.br Documento assinado digitalmente
MARCELO COSTA OLIVEIRA
Data: 13/02/2025 17:40:21-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. MARCELO COSTA OLIVEIRA
UFAL – PPGI- Instituto de Computação
Orientador

gov.br Documento assinado digitalmente
ALVARO ALVARES DE CARVALHO CESAR SOBRIN
Data: 14/02/2025 12:29:59-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. ALVARO ALVARES DE CARVALHO CESAR SOBRINHO
Universidade Federal do Agreste de Pernambuco, UFAPE e membro permanente- PPGI- IC/UFAL
Examinador Interno

gov.br Documento assinado digitalmente
THALES MIRANDA DE ALMEIDA VIEIRA
Data: 14/02/2025 14:57:08-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. THALES MIRANDA DE ALMEIDA VIEIRA
UFAL – PPGI- Instituto de Computação
Coorientador

gov.br Documento assinado digitalmente
PAULO MAZZONCINI DE AZEVEDO MARQUES
Data: 14/02/2025 12:57:17-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. PAULO MAZZONCINI DE AZEVEDO MARQUES
USP– Universidade de São Paulo.
Examinador Externo

Acknowledgments

First, to God for illuminating my entire journey and granting me strength in difficult times.

To my wife, Heloyza Helena, for her love, patience, and support of my studies.

To my advisors throughout the course, Thales and Marcelo. And a special thanks to my advisor Marcelo for the opportunity given, for the patience in teaching and dedication to my progress in studies. To the professors who accepted the invitation to be part of my panel, Álvaro e Paulo.

To all those who helped me directly or indirectly.

Resumo

O câncer de pulmão tem a maior taxa de mortalidade entre todos os tipos de câncer, tanto para homens quanto para mulheres. Estima-se que o câncer de pulmão seja responsável por 21% das mortes por câncer em cada gênero no mundo. Essa estatística alarmante destaca o impacto significativo do câncer de pulmão na mortalidade geral por câncer, sublinhando a necessidade urgente de estratégias eficazes de prevenção, detecção precoce e tratamento para combater essa doença. O rastreamento do câncer de pulmão é um processo projetado para detectar o câncer de pulmão em indivíduos em risco, particularmente aqueles com histórico de tabagismo. Envolve tomografias computadorizadas de baixa dose anuais, interpretação cuidadosa dos resultados e acompanhamento oportuno para garantir a detecção e o tratamento precoces. Várias sociedades profissionais, incluindo a *American College of Radiology* e a Sociedade Fleischner, publicaram diretrizes para o manejo de pacientes com nódulos pulmonares detectados durante o rastreamento de câncer de pulmão. As diretrizes são uma ferramenta importante em programas de rastreamento que visam reduzir a incidência de exames de acompanhamento desnecessários e orientar o manejo ideal do paciente. *Lung Computed Tomography Screening Reporting & Data System* (Lung-RADS) é um sistema de classificação padronizado para nódulos pulmonares detectados em exames de imagem, como tomografias computadorizadas. O Lung-RADS avalia o risco de malignidade (câncer) nesses nódulos e orienta as decisões de manejo subsequentes. Neste contexto, este trabalho visa analisar a eficácia de modelos de aprendizado profundo e *Large Language Model* na extração de características de nódulos pulmonares de laudos de Tomografia Computadorizada em português para permitir a classificação automatizada do Lung-RADS. Este trabalho avaliou a eficácia de BiLSTM-CRF, BioBERTpt, Gemini 1.5 Flash, GPT-4o e Llama 3 70B. Os resultados sugerem que o Gemini 1.5 Flash se destaca como o modelo com maior eficácia, superando os demais em quatro das cinco classificações Lung-RADS no conjunto de teste, com um *F1-score* ponderado de 0,95, destacando sua eficácia na avaliação precisa de nódulos pulmonares em vários cenários de classificação.

***Palavras-chave*: *Câncer de Pulmão*; *Lung-RADS*; *NLP*; *Extração de Informação*; *LLM*.**

Abstract

Lung cancer has the highest mortality rate among all cancer types, affecting both men and women. It is estimated that lung cancer accounts for 21% of cancer deaths in each gender worldwide. This alarming statistic highlights the significant impact of lung cancer on overall cancer mortality, underscoring the urgent need for effective prevention, early detection, and treatment strategies to combat this disease. Lung cancer screening is a process designed to detect lung cancer in at-risk individuals, particularly those with a history of smoking. It involves annual low-dose computed tomography (CT) scans, careful interpretation of results, and timely follow-up to ensure early detection and treatment. Several professional societies, including the American College of Radiology and the Fleischner Society, have published guidelines for the management of patients with pulmonary nodules detected during lung cancer screening. The guidelines are an important tool in screening programs aimed at reducing the incidence of unnecessary follow-up examinations and guiding optimal patient management. The Lung Computed Tomography Screening Reporting & Data System (Lung-RADS) is a standardized classification system for pulmonary nodules detected on imaging examinations, such as CT scans. Lung-RADS assesses the risk of malignancy (cancer) in these nodules and guides subsequent management decisions. In this context, this work aims to analyze the effectiveness of deep learning and large language models in extracting features of pulmonary nodules from Portuguese CT reports to enable automated classification of Lung-RADS. This work evaluated the effectiveness of BiLSTM-CRF, BioBERTpt, Gemini 1.5 Flash, GPT-4o, and Llama 3 70B. The results suggest that the Gemini 1.5 Flash stands out as the most effective model, outperforming the others in four of the five Lung-RADS classifications in the test set, with a weighted F1-score of 0.95, highlighting its effectiveness in accurately assessing lung nodules in various classification scenarios.

***Keywords*: *Lung Cancer*; *Lung-RADS*; *NLP*; *Information Extraction*; *LLM*.**

List of Figures

List of Tables

Acronyms

ACR *American College of Radiology*

API *Application Programming Interface*

BERT *Bidirectional Encoder Representations from Transformers*

BioBERTpt *BioBERT Portuguese*

BiLSTM-CRF *Bidirectional Long Short-Term Memory with Conditional Random Fields*

BRNNs *Bidirectional Recurrent Neural Networks*

CNN *Convolutional Neural Networks*

CoNLL *Computational Natural Language Learning*

CRF *Conditional Random Field*

CTLS *CT Lung Screening*

EHRs *Electronic Health Records*

GGNs *Ground-Glass Nodules*

GPT *Generative Pre-trained Transformer*

GPU *Graphics Processing Unit*

GQA *Grouped Query Attention*

IE *Information Extraction*

IO *Inside-Outside*

IOB *Inside-Outside-Beginning*

IOE *Inside-Outside-Ending*

IOBES *Inside-Outside-Beginning-Ending-Single*

JSON *JavaScript Object Notation*

LCS *Lung Cancer Screening*

LDCT *Low-Dose Computed Tomography*

LLMs *Large Language Models*

LSTM *Long Short-Term Memory*

Lung-RADS *Lung CT Screening Reporting & Data System*

MLM *Masked Language Modeling*

MUC-6 *Message Understanding Conference-6*

NE *Named Entity*

NER *Named Entity Recognition*

NLP *Natural Language Processing*

NLST *National Lung Screening Trial*

PALM *Pathways Language Model*

PET *Positron Emission Tomography*

POS *Part-of-Speech*

QA *Question Answering*

RoPE *Rotary Position Embedding*

SGD *Stochastic Gradient Descent*

SVM *Support Vector Machine*

TF-IDF *Term Frequency-Inverse Document Frequency*

USMLE *United States Medical Licensing Examination*

Contents

Chapter 1

Introduction

Lung cancer ranks as the second most frequently diagnosed cancer among both men and women, following only breast cancer in prevalence. Despite advancements in early detection and treatment, it remains the leading cause of cancer-related mortality, responsible for more deaths in 2020 than breast and prostate cancers combined [Sung et al., 2021]. In that year, approximately 2,206,771 individuals received a lung cancer diagnosis, with 1,796,144 succumbing to the disease [Sung et al., 2021]. In Brazil, lung cancer is the third most common type of cancer among men and the fourth most common among women. Estimates from 2018 show that there were 18,740 new cases of lung cancer in men and 12,530 in women [Mathias et al., 2020]. This statistic 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 [Ahmad et al., 2025]. 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] [Ahmad et al., 2025]. The national lung screening trial (NLST) demonstrated that individuals undergoing annual LDCT scans experienced a consistent reduction in lung cancer mortality. The extended follow-up from the NLST confirmed a lung cancer mortality reduction of 8-11%, emphasizing the long-term benefits of LCS in high-risk populations [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 on computed tomography (CT) exams [ACR, 2022] [MacMahon et al., 2017]. The guidelines are an important tool aimed at reducing the incidence of unnecessary follow-up exams and guiding optimal patient management. In addition, the guidelines offer more flexibility in follow-up intervals and provide tailored recommendations based on individual risk factors, thus enhancing the ability of ra-

diologists, clinicians, and patients to make well-informed decisions [Polanco et al., 2024].

One of these guidelines is the Lung CT Screening Reporting & Data System (Lung-RADS®) published by the ACR, the Lung-RADS is a standardized classification system for lung nodules detected in imaging exams such as LDCT scans [ACR, 2022]. Lung-RADS assesses the risk of malignancy (cancer) in these nodules and guides subsequent management decisions. The Lung-RADS index is based on some characteristics of pulmonary nodules, including the size, shape, growth rate, and other nodule characteristics. The greater the risk of malignancy, the higher the Lung-RADS index. The follow-up examination for a nodule with a lower Lung-RADS index (1-3) is an LDCT in 12 months, while for a nodule with a higher Lung-RADS index (4-5), a Positron Emission Tomography-Computed Tomograph (PET CT) or biopsy is recommended [ACR, 2022].

Determining follow-up examinations according to the Lung-RADS guideline 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 Lung-RADS criteria, presents considerable challenges. Analyzing medical data in Portuguese is a complex and time-consuming task, due to the fact that clinical data is often recorded in a free text format. There are the presence of acronyms, negation adverbs, and grammatical errors. Furthermore, cultural differences and variations in descriptive style can lead to inconsistencies in the data. Finally, human error in data entry is also a possibility [da Rocha et al., 2023].

Natural language processing (NLP) is a branch of artificial intelligence that enables machines to understand, interpret, and generate human language. In the healthcare sector, NLP has greatly improved our capacity to manage and analyze large volumes of textual data, such as medical records and clinical notes [Pandey et al., 2020]. Named entity recognition (NER) is a subfield of NLP that identifies and classifies named entities in text and involves detecting and categorizing named entities (NEs) in text into specified entity classes. These entities can include names of people, organizations, locations, and more [Li et al., 2022]. Several works leverage NER to systematically extract structured information from clinical texts and radiology reports, such as identifying clinical conditions, symptoms, diagnoses, medications, exams, treatment, and pulmonary nodules descriptions [Lopes et al., 2019] [da Rocha et al., 2023] [Fei et al., 2022]. In [Beyer et al., 2017], the authors showed how NLP can assist radiologists by recommending the appropriate Lung-RADS category and identifying reports that lack sufficient information for accurate Lung-RADS classification. Recent study have shown that advancements in NLP provide a promising approach for automatically extracting Lung-RADS malignancy index data from the unstructured text found in radiology reports [Gandomi et al., 2024].

However, previous research on information extraction (IE) has primarily focused on Chinese and English idioms to identify clinical entities, lung nodules, tumors, and their associated characteristics [Zheng et al., 2021] [Hu et al., 2024a] [Hu et al., 2024b], leaving

a gap in understanding this phenomenon in other languages. Additionally, the absence of publicly available datasets containing pulmonary nodule reports further limits the development and evaluation of information extraction methods in this domain, particularly for radiology reports in diverse languages.

1.1 Objective

In this context, this work aims to analyze the effectiveness of strategies for extracting lung nodule characteristics from Portuguese chest CT reports to enable automated Lung-RADS classification. It compares deep learning models and large language models (LLMs), employing natural language processing techniques such as named entity recognition and question answering.

1.2 Work Structure

This work is organized as follows.

- Chapter 2 literature review examining related works that contextualize and support the research;
- Chapter 3 provides the theoretical background;
- Chapter 4 outlines the research methodology;
- Chapter 5 is dedicated to the presentation of experiments conducted and results obtained;
- Chapter 6 summarizes the study's conclusions, highlighting key findings, limitations encountered during the research, future works, and scientific contributions.

Chapter 2

Related Works

2.1 Rule-Based NLP Systems in Medical Applications

Rule-based natural language processing systems rely on manually defined linguistic rules to extract or process information from text. These systems typically use dictionary-based methods and pattern matching to identify and categorize information. Several works have demonstrated the effectiveness of these techniques. For example, Gershanik et al. [Gershanik et al., 2011] introduced iSCOUT, an NLP application to retrieve documents and assess discrepancies between the "findings" and "impressions" sections of radiology reports, discovering inconsistencies more than one-third of the cases. Although achieving a precision of 96.0% and a recall of 80.00% in identifying pulmonary nodules, the NLP application (iSCOUT) remains constrained by the variability and inconsistency of terminology present in the reports. Moreover, variations in language or reporting style can lead to missed findings, despite advanced NLP tools. Nobel et al. [Nobel et al., 2020] developed a rule-based NLP model to classify lung nodule T-stages in Dutch radiology reports, achieving significant accuracy (87.0%) in its evaluation. However, the study faced limitations due to its relatively small dataset, which affects the generalizability of the findings, suggesting the need for a larger dataset for more robust training and validation, especially for machine learning-based approaches. In a study using more than 350,000 CT transcripts, Zheng et al. [Zheng et al., 2021] developed a rule-based algorithm to extract a range of nodule characteristics, achieving high sensitivity (98.6%) and specificity (100.0%) to identify lung nodules.

Although rule-based NLP systems have show significant success in specific medical applications, as evidenced by [Gershanik et al., 2011], [Nobel et al., 2020], and [Zheng et al., 2021], they consistently face limitations related to linguistic variability and the need for manual rule creation. These challenges, particularly in handling the complex and often inconsistent terminology found in medical texts, have motivated the exploration

of more flexible and robust approaches to NLP. A paradigm shift has subsequently occurred in the field, with a strong emphasis on deep learning solutions, particularly those that utilize transformer architectures.

2.2 Machine Learning Approachs

As an intermediary between rule-based NLP systems and deep learning approaches, machine learning methods have been widely explored in radiology-related NLP tasks. Machine learning techniques, such as support vector machines (SVM), logistic regression, and random forests, have been successfully applied to extract structured information from radiology reports. For example, Carrodeguas et al. demonstrated that machine learning models could identify follow-up recommendations in radiology reports, with SVM achieving an F1-score of 0.85, exceeding both a rule-based NLP system and deep learning models in this specific task [Carrodeguas et al., 2019]. Similarly, Zech et al. used natural language processing and machine learning to annotate clinical radiology reports, highlighting the potential of automated methods in structuring large-scale radiological data for downstream applications [Zech et al., 2018].

These approaches demonstrate the effectiveness of machine learning to process medical text while mitigating some of the limitations associated with rigid rule-based systems. However, despite their advantages, machine learning models often require extensive feature engineering and large annotated datasets to achieve high performance, thereby motivating the transition toward deep learning-based methods for more flexible and scalable solutions.

2.3 Deep Learning Approachs

Deep learning methods have emerged as an important pillar in natural language processing, revolutionizing how we analyze and interpret medical texts. The evolution toward deep learning has been driven by the inherent limitations of rule-based systems, particularly their inability to effectively handle the complex, nuanced, and often inconsistent terminology found in medical documentation [Hu et al., 2024b]. For example, Fei et al. [Fei et al., 2022] proposed a Bidirectional Long Short-Term Memory with Conditional Random Fields (BiLSTM-CRF), a model for named entity extraction from Chinese radiology reports. Their approach demonstrated outstanding performance, achieving high accuracy (94.22%), precision (94.56%), recall (93.96%), and F1-score (94.26%). These results highlight the model's effectiveness in handling complex medical text data. However, reliance on unstructured data poses significant challenges for automating data analysis and ensuring standardization.

Naila et al. [da Rocha et al., 2023] proposed a Convolutional Neural Network (CNN) model that uses unstructured data from Portuguese medical records to identify seven entities, including symptoms, diagnoses, medications, conditions, exams, and treatment. They constructed a corpus using 1,200 of 30,000 records. The CNN model for named entity recognition achieved an average precision of 72.72%, an average recall of 56.93%, and an average F1-score of 63.87%.

2.4 BERT and Domain-Specific Transformer Models

Current state-of-the-art NLP systems have predominantly utilized transformer-based architectures, such as Bidirecional Encoder Representations from Transformers (BERT), achieving remarkable performance across a wide range of tasks, including natural language understanding, text classification, question answering, and text generation. These architectures leverage self-attention mechanisms to capture complex relationships within textual data, allowing their dominance in general-purpose and domain-specific applications [Vaswani et al., 2023] [Devlin et al., 2019]. Despite the fact that BERT has set new benchmarks in many NLP tasks, [Sugimoto et al., 2021] demonstrated that BiLSTM-CRF was more effective than BERT and BERT-CRF in extracting relevant information from chest CT reports. In their study, they used BiLSTM-CRF, BERT, and BERT-CRF models. This highlights that despite the general superiority of transformers, certain tasks may benefit from architectures better tailored to structured outputs or sequence tagging. For instance, Fei et al. [Fei et al., 2022] proposed a BiLSTM-CRF model for entity extraction from Chinese radiology reports. Their approach demonstrated outstanding performance, achieving high accuracy (94.22%), precision (94.56%), recall (93.96%), and F1-score (94.26%). These results highlight the model's effectiveness in handling complex medical text data.

Futhermore, based on the BERT architecture, domain-specific models were developed for clinical applications, such as BioBERT [Lee et al., 2019], ClinicalBERT [Alsentzer et al., 2019], and PubMedBERT [Gu et al., 2021]. These models were pre-trained on large biomedical and clinical text datasets. When these domain-specific models were applied to various clinical NLP tasks such as biomedical named entity recognition, biomedical relation extraction, and biomedical question answering, they demonstrated superior performance to the original BERT or BERT models pre-trained on more general text corpora. Building upon the advancements of domain-specific BERT-based models in clinical NLP tasks, recent developments in transformer-based large language models have further expanded the boundaries of language understanding and application.

2.5 Large Language Models in Medical Applications

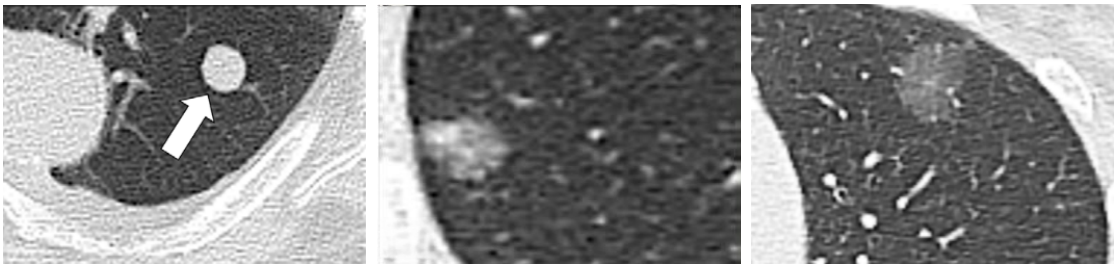
Recent advances in Large Language Models (LLMs) such as Generative Pre-trained Transformer 4 (GPT-4) [OpenAI et al., 2024], LLaMA [Grattafiori et al., 2024], Gemini [Team et al., 2024b], and Pathways Language Model (PaLM) [Chowdhery et al., 2022], contain millions to billions of parameters; they are pre-trained on vast amounts of text data, demonstrate superior capability on a variety of natural language processing tasks [Minaee et al., 2024]. Motivated by these capabilities, several works were developed to adapt general LLMs for question answering for medical use [Singhal et al., 2022] [Singhal et al., 2023]. The LLMs, PaLM, GPT-4, MedPaLM-2 [Singhal et al., 2023], and MedPrompt [Nori et al., 2023] have achieved an accuracy of 86.5% and 90.2%, respectively, against 87.0% when compared to human experts in the United States Medical Licensing Examination [Wu et al., 2023]. As a result, the application of medical LLMs has gained increasing research interest in helping medical professionals due to their ability to process and comprehend complex medical language and their potential to improve efficiency, accuracy, and patient care within the healthcare industry [Bedi et al., 2024].

Chapter 3

Theoretical Background

3.1 Lung Nodules in Medical Images

A pulmonary nodule is defined as a rounded or irregular opacity, well or poorly defined, with a diameter of 3 cm or less. Nodules are classified as small if their largest diameter is 10 mm or less, while micronodules measure under 3 mm. Most nodules smaller than 1 cm are not detectable on chest radiographs and can only be observed on CT scans [Sánchez et al., 2018]. Pulmonary nodules are categorized based on their density into three main types, as shown in Figure 3.1: solid nodules, non-solid nodules, and part-solid nodules. Solid nodules (3.1a), which are the most common type, have a soft-tissue density that obscures the contours of adjacent vessels and airways. Non-solid nodules (3.1b), also known as pure ground-glass nodules (3.1c), are focal areas of increased lung attenuation that do not obscure the underlying parenchymal structures, including airways and vessels. Part-solid nodules, or semi-solid nodules, exhibit both ground-glass and solid soft-tissue components [Hansell et al., 2008].



(a) Solid nodule. (b) Part-solid nodule. (c) Ground-glass nodule.

Figure 3.1: Different lung nodules types in a Chest CT [Bankier et al., 2024].

3.2 Computed Tomography

Computed tomography is a computerized X-ray imaging technique where a narrow x-ray beam is directed at the patient and rapidly rotates around the body. The CT scanner's computer processes these signals to create cross-sectional "slices" known as tomographic images, which provide more detailed insights than standard x-rays. When multiple slices are captured, they can be digitally combined to form a three-dimensional (3D) image, helping clinicians identify structures and potential tumors or abnormalities. Unlike conventional x-rays that use a fixed x-ray source, a CT scanner employs a motorized x-ray tube that revolves around a circular opening called a gantry. During a scan, the patient lies on a bed that slowly advances through the gantry as the x-ray tube rotates around, projecting narrow x-ray beams through the body. Instead of film, CT scanners use specialized digital detectors positioned directly opposite the x-ray source. As the X-rays pass through the patient, the detectors capture them and transmit data to a computer [of Biomedical Imaging and (NIBIB), 2024]. Figure 3.2 shows an overview of a CT scanner operation.

With each full rotation of the x-ray source, the CT computer uses advanced mathematical methods to create a two-dimensional image slice of the body. The thickness of each slice can vary, typically between 1-10 millimeters depending on the machine. Once a slice is complete, the image is stored, and the bed moves slightly forward, allowing the process to repeat until the desired number of slices is obtained. These image slices can be viewed individually or stacked by the computer to form a 3D representation of the patient, revealing bones, organs, and tissues along with any abnormalities. This approach offers significant advantages, including the ability to rotate the 3D model or view slices sequentially, making it easier to pinpoint specific issues [of Biomedical Imaging and (NIBIB), 2024].

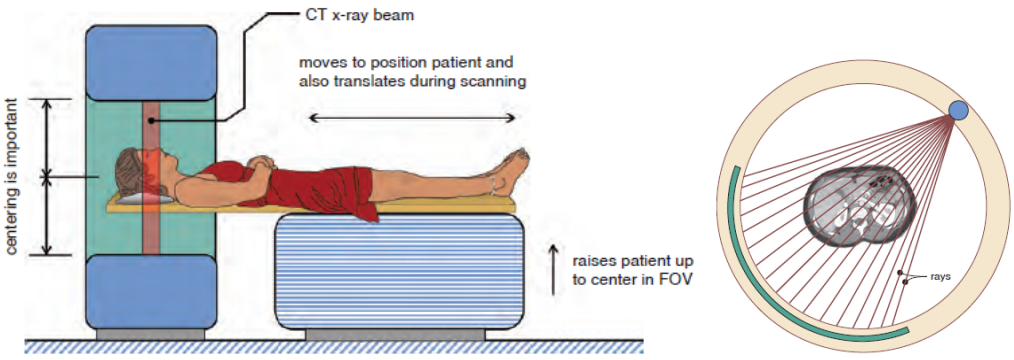


Figure 3.2: CT Schematics [Bushberg et al., 2012].

3.3 Natural Language Processing

Natural language processing is a branch of Artificial Intelligence that is devoted to making computers understand, interpret, and generate human language [Khurana et al., 2022]. In the healthcare sector, NLP has greatly improved our capacity to manage and analyze large volumes of textual data, such as medical records and clinical notes [Pandey et al., 2020]. In radiology, NLP has been used for information retrieval, classification, text extraction, text summarization, question answering, text generation, and named entity recognition [Ewoud et al., 2016] [Luo and Chong, 2020] [Arlene et al., 2021] [Zheng et al., 2021]. Natural language processing employs a range of techniques, broadly categorized as rule-based and machine learning approaches. Rule-based methods rely on predefined rules, often using regular expressions, word matching, and annotation, to select or retrieve text or synonyms. In contrast, machine learning approaches require substantial text datasets for training, validation, and testing. Machine learning methods often use classifiers, such as random forest or statistical techniques that utilize vectorization, such as TF-IDF [Linna and Kahn, 2022].

Deep learning techniques have revolutionized natural language processing with a more effective way to handle NLP problems. With the introduction of word embeddings, words are treated as vectors capturing semantic relationships based on context. This allows documents to be represented as matrices of these vectors, suitable for deep learning architectures like CNN and Long Short-Term Memory (LSTM) networks, including the bidirectional variant (BiLSTM) which captures contextual information from both directions in a text. Further advancements like sequence-to-sequence architectures and attention mechanisms, culminating in the development of pre-trained language models like Transformer, BERT, and GPT, have significantly boosted NLP performance across various tasks, establishing a new standard for the field [Tho, 2022].

Information extraction is the process of automatically identifying and encoding relevant clinical information from unstructured free-text data, such as electronic health records. This task is essential for leveraging free-text data in electronic health records to support clinical decision-making, quality improvement, and research. In NLP, information extraction specifically refers to extracting key concepts, entities, events, and their relationships and attributes from text, enabling the transformation of unstructured data into a structured format that can be more easily analyzed and used for various healthcare applications [Wang et al., 2018]. An information extraction system typically includes one or more of the following components: named entity recognition, relation extraction, and post-processing [Hu et al., 2024a].

3.3.1 Named Entity Recognition

The term "Named Entity" was first recognized as essential for information extraction in the sixth message understanding conference evaluation [Grishman and Sundheim, 1996]. A named entity refers to a word or phrase that identifies elements with shared properties within a dataset. It serves as a rigid designator, or atomic element, representing a member of a semantic class, which can vary depending on the domain of interest. In general domain, person, location, organization, number, date, time, etc. are important entities [Goyal et al., 2018].

Named entity recognition aims to identify mentions of rigid designators within text, associating them with predefined semantic types. NER involves detecting and categorizing named entities in text into specified entity classes. Formally, given a sequence of tokens (small units that can be words, character, punctuation) $s = \langle w_1, w_2, \cdots, w_N \rangle$, NER outputs a list of tuples $\langle I_s, I_e, t \rangle$, each of wich is a named entity mentioned in s. Here, $I_s \in [1, N]$ and $I_e \in [1, N]$ are the start and end indexes of a named entity mention; t is the entity type from a predefined category set [Li et al., 2022]. Figure 3.3 shows an example where NER system recognizes three named entities from the sentence "Michael Jeffrey Jordan was born in Brooklyn, New York.". The tokens w_1 = Michal and w_3 = Jordan were recognized as the named entity "Person". The token w_7 = Brooklyn was recognized as the named entity "Location". The tokens w_9 = New and w_{10} = York were recognized as the named entity "Location".

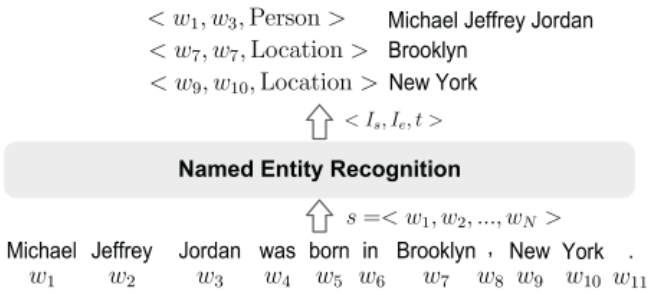


Figure 3.3: Example of NER task [Li et al., 2022].

3.3.2 Question Answering

Question answering (QA) is one of the most important natural language processing tasks, focused on developing systems that can automatically answer questions posed by users in natural language. This involves creating models and algorithms that can understand the intent behind a user's question, retrieve relevant information from various sources such as (databases, documents, or web pages), and formulate a coherent response [Hirschman and Gaizauskas, 2001]. Figure 3.4 shows an example of an QA task in wich

given a text about precipitation, the QA system answers three questions asked by the user.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 3.4: Example of Question-answer pairs for a sample passage in the SQuAD dataset [Rajpurkar et al., 2016].

3.4 Bidirectional Long Short-Term Memory with Conditional Random Fields

Recurrent neural networks were designed with the primary objective of capturing and modeling long-term dependencies or patterns in sequential data, such as text. However, these networks have faced challenges such as the problems of exploding and vanishing gradient, which have hindered their effectiveness in capturing such dependencies [Quinta de Castro et al., 2018]. To deal with these limitations, the LSTM, a recurrent network architecture was employed [Zhang et al., 2018]. The key feature of an LSTM is its memory cell, which stores and propagates information over time. LSTMs utilize a gating mechanism consisting of three main gates: the input, forget, and output. These gates control the flow of information into, out of, and within the memory cell, allowing the LSTM to retain or discard information at different time steps selectively. Figure 3.5 ilustrated the structure of LSTM unit.

At each time step, the cell receives an input x_t along with the previous hidden state h_{t-1}. The forget gate determines how much of the past information from the cell state should be retained or discarded, using a sigmoid activation function to produce values

between 0 and 1. Simultaneously, the input gate regulates how much new information from the current input should be added to the cell state, ensuring that only relevant updates are incorporated. These mechanisms modify the cell state, which serves as the long-term memory of the network.

Once the cell state is updated, the output gate decides how much of this information should contribute to the new hidden state. This process ensures that only essential information is propagated forward while preserving past knowledge in a controlled manner. The final hidden state is obtained by applying a non-linear activation function to the cell state and modulating it through the output gate.

The input at time step t, denoted as x_t, is processed within the LSTM cell and influences different gates that regulate the flow of information.

The forget gate decides how much of the information from the previous cell state, C_{t-1}, should be retained or discarded. It receives x_t and the previous hidden state h_{t-1}, applies a sigmoid activation function (σ), and outputs values between 0 (completely forget) and 1 (fully retain). The corresponding equation is:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3.1}$$

where W_f and b_f are the learned weights and biases of the network.

The input gate (i_t) controls how much new information from x_t should be added to the cell state. It also uses a sigmoid activation function to determine which values should be updated. Additionally, a new candidate cell state, $\tilde{C}_t$, is created using a *tanh* activation function. The equations are:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3.2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3.3}$$

The cell state (C_t), which acts as long-term memory, is updated by combining the forget gate and input gate information. The cell state is updated according to the following equation where $\odot$ represent the element-wise product:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{3.4}$$

This enables the preservation of relevant information over time.

The output gate (o_t) regulates how much of the stored information in the cell state will be used to generate the new hidden state h_t. The final output is obtained by applying a sigmoid activation function and a *tanh* function to the cell state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3.5}$$

$$h_t = o_t \odot \tanh(C_t) \tag{3.6}$$

The hidden state h_t represents the output of the LSTM unit at time step t and is used as an input for the next time step. It stores short-term information and interacts with the cell state to maintain temporal dependencies.

In LSTM cell, several multiplication operations regulate the information flow:

- The output of the **forget gate** f_t multiplies the previous cell state C_{t-1} to decide what should be retained.
- The output of the **input gate** i_t multiplies the candidate cell state $\tilde{C}_t$ to regulate the update of the cell state.
- The output of the **output gate** o_t multiplies the transformed version of C_t to determine the new hidden state.

The activation functions used in the LSTM cell are:

- **Sigmoid** (σ): used in all three gates (forget, input, and output) to restrict values between 0 and 1, controlling the passage of information.
- **Tanh**: used to transform the candidate cell state and regulate the final cell output, keeping values between -1 and 1, ensuring stability in state updates.

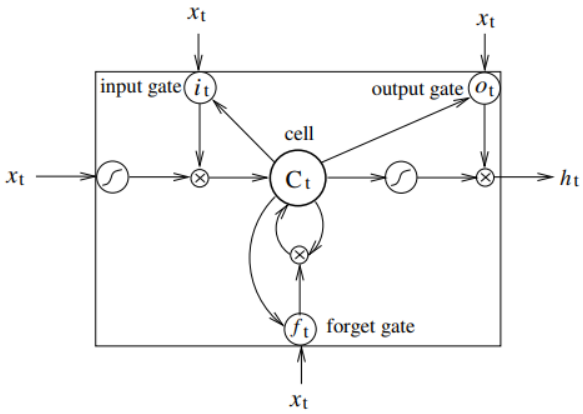


Figure 3.5: Example of LSTM cell [Huang et al., 2015].

Bidirectional recurrent neural networks (BRNNs) can process the input sequence in two passes: one in the forward direction and another in the backward direction. This architecture does this with two separated hidden layers that capture information from their respective directions and subsequently feed it forward to the same output layer. This allows the network to capture information from both past and future contexts, enabling a richer understanding of the sequential data. A bidirectional LSTM is a neural

network comprising LSTM units that function in both forward and reverse directions [Jang et al., 2020].

Figure 3.6 represents a Bidirectional Long Short-Term Memory model applied to a Named Entity Recognition task. This model processes a sentence to identify named entities such as organizations, people, or locations.

In the input layer at the bottom, we see a sequence of words:

- EU, rejects, German, call.

Each of these words is first converted into a numerical representation (word embeddings) before being processed by the LSTM layers. The model consists of two LSTM layers:

- A forward LSTM (arrows pointing left to right) that reads the sentence in normal order;
- A backward LSTM (arrows pointing right to left) that reads the sentence in reverse order.

Each word's representation is thus influenced by both its preceding and following words. This is important in NER, where context from both directions helps determine an entity's classification. In the output layer at the top, each word is assigned an NER label:

- EU → B-ORG (Beginning of an Organization);
- rejects → O (Outside any entity);
- German → B-MISC (Beginning of a Miscellaneous entity);
- call → O (Outside any entity).

These labels indicate which words belong to named entities and their types.

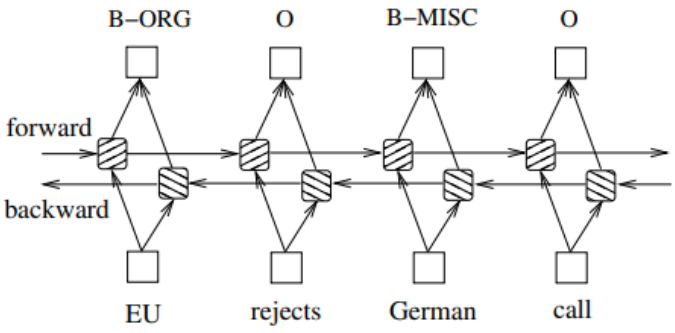


Figure 3.6: Example of BiLSTM Architecture [Huang et al., 2015].

The conditional random field (CRF) is a statistical modeling method that is used to predict sequences of labels based on a given set of observed features. CRFs are particularly useful for structured prediction tasks, where each label in the sequence is not independent but depends on neighboring labels, which makes it possible to capture contextual information. Unlike traditional classifiers that make independent predictions, CRFs consider the correlations between labels, allowing them to predict the optimal label sequence for a given input. This characteristic makes CRFs highly effective in tasks such as NER, Part-of-Speech (POS) tagging, and other sequence labeling applications, as they allow the model to account for label dependencies within the sequence [Zheng et al., 2015].

Figure 3.7 demonstrates how a CRF models sequences for NER by considering both individual word features and the dependencies between labels. The goal of a CRF in this context is to predict the most likely sequence of entity labels for a given input sentence.

The sequence of words: "EU rejects German call" represents a sentence or phrase being processed. The labels at the top are entity labels based on the BIO (Beginning, Inside, Outside) tagging scheme.

The squares connected by horizontal lines at the top represent CRF model. Each square corresponds to a word in the sentence an the connections between them indicate dependencies between neighboring labels, which is a key feature of CRFs.

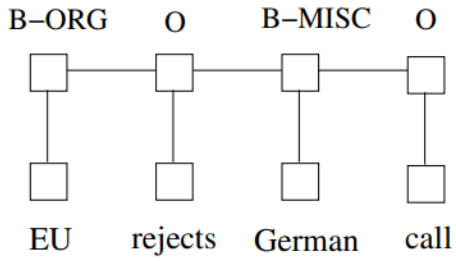


Figure 3.7: Example of CRF Network [Huang et al., 2015].

The BiLSTM model, as shown in the Figure 3.8, leverages the strengths of both a BiLSTM and a CRF.

The input layer (bottom row) contains the input tokens: "EU," "rejects," "German," and "call." Each word is processed individually and mapped to an embedding before being passed to the BiLSTM.

The BiLSTM (middle layer with hatched nodes) consists of two LSTMs: a forward LSTM that processes the sequence from left to right, and a backward LSTM that processes the sequence from right to left. These two LSTMs capture contextual information from both directions.

The outputs of the BiLSTM are then passed to a CRF layer (top row of squares). The CRF makes the final label predictions, such as B-ORG, O, B-MISC, and O. These labels correspond to named entity classes (e.g., organizations, miscellaneous entities, or

non-entity words). The CRF ensures that predictions are made considering dependencies between labels, resulting in valid entity tag sequences.

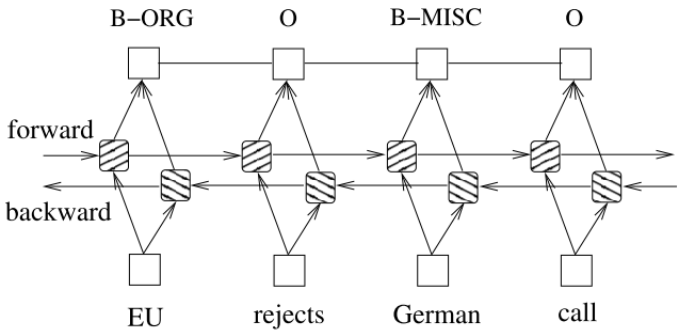


Figure 3.8: BiLSTM-CRF architecture [Huang et al., 2015].

3.5 Transformers

Transformers is a model architecture proposed by [Vaswani et al., 2023], designed to avoid recurrence and instead uses the attention mechanism to capture global dependencies between input and output. The transformer architecture employs stacked self-attention and point-wise, fully conected layer for both the encoder and decoder. Figure 3.9 provides an overview of this architecture, at the bottom there is the input & output embeddings and the positional encode. The encoder is shown in the left half of the figure and the decoder in the right half.

In the Input & Output Embeddings the words are first converted into numerical vectors using an embedding layer. The Positional encoding is added to retain word order information since the Transformer lacks recurrence (like in RNNs).

The encoder in the Transformer architecture consists of $N = 6$ identical layers. Each layer includes two sub-layers: a multihead self-attention mechanism and a position-wise fully connected feedforward network. A residual connection is applied around each of the two sub-layers, followed by layer normalization.

Similarly, the decoder is composed of a stack of $N = 6$ identical layers. However, in addition to the two sublayers in each encoder layer, the decoder includes a third sub-layer that performs multi-head attention over the encoder's output. As in the encoder, residual connections are applied around each sub-layer in the decoder, followed by layer normalization.

The self-attention sublayer in the decoder is modified to prevent positions from attending to subsequent positions. This masking, combined with the fact that output embeddings are offset by one position, ensures that predictions for position i depend only on known outputs at positions less than i.

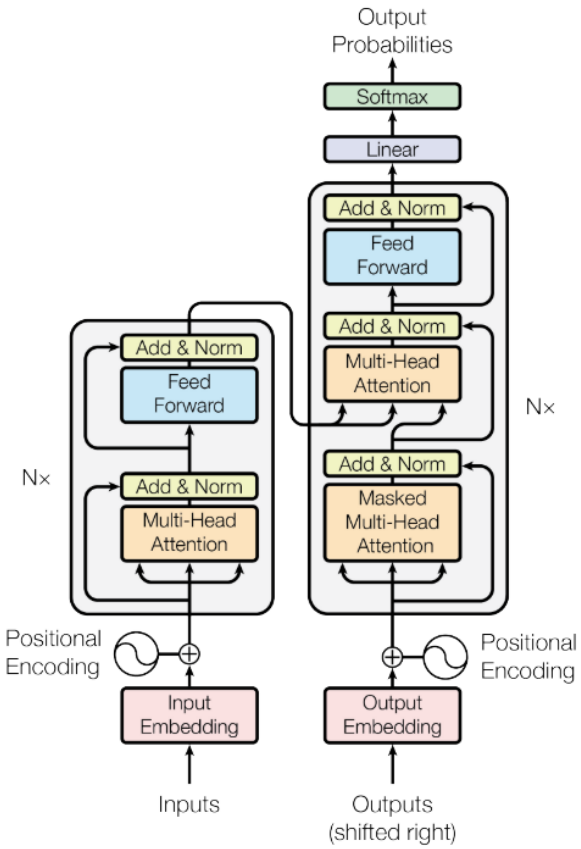


Figure 3.9: The Transformer - model architecture [Vaswani et al., 2023].

The BERT and GPT models are both based on the Transformer architecture, but they use different parts of it: BERT is based solely on the encoder part of the Transformer, on the other hand, GPT is based on the decoder part of the Transformer.

3.6 Large Language Models

Large Language Models are designed to understand, generate, and interact using human language. These models are built using deep learning, specifically through neural network architectures such as transformers, and are trained on vast datasets of text from sources such as books, websites, and other publicly available materials. The large in LLMs typically refers to the substantial number of parameters (billions or even trillions) used in these models. LLMs have a wide range of applications, including chatbots, language translation, content generation, and even programming assistance [Minaee et al., 2024].

3.6.1 Bidirectional Encoder Representations for Transformers

BERT is a language representation model introduced by [Devlin et al., 2019] designed to pre-train deep bidirectional representations from unlabeled text by jointly both left and right contexts in all layers. The model architecture consists of three main modules: (1)

an embedding module that transforms input text into a sequence of embedding vectors, (2) a stack of Transformer encoders that processes these embedding vectors to generate contextual representation vectors, and (3) a fully connected layer that maps the final-layer representations into one-hot vectors.

The pre-training and fine-tuning of the BERT model is shown in Figure 3.10. BERT's pre-training relies on two primary objectives: masked language modeling (MLM) and next sentence prediction (NSP). The MLM objective involves randomly masking certain tokens in the input sequence, requiring the model to predict the original token based solely on the surrounding context. This approach allows BERT to fuse both left and right contexts, enabling the pre-training of a deeply bidirectional Transformer—a key distinction from traditional left-to-right language models. In addition to MLM, BERT employs a next sentence prediction task, designed to jointly pre-train text-pair representations. This task involves predicting whether a given sentence logically follows another, further enhancing the model's capacity to understand relationships between sentences [Devlin et al., 2019].

Fine-tuning involves plugging task-specific inputs and outputs into BERT and training all parameters end-to-end. During this process, sentence A and sentence B from pre-training are analogous to (1) sentence pairs in paraphrasing, (2) hypothesis-premise pairs in entailment, (3) question-passage pairs in question answering, and (4) a degenerate text-null pair in text classification or sequence tagging. At the output stage, token representations are used for token-level tasks, such as sequence tagging or question answering, while the [CLS] token's (the first token of every sequence) representation is used for classification tasks, like entailment or sentiment analysis. Fine-tuning BERT is relatively computationally inexpensive compared to pre-training, making it accessible for many language understanding applications [Devlin et al., 2019].

The input representation of BERT is shown in Figure 3.11 and consists of three main embedding types:

- Token Embeddings (Yellow Boxes): Each word or subword token in the input is converted into a vector representation using a pre-trained embedding matrix. For example, the input consists of two sentences:

  "[CLS] my dog is cute [SEP]"
  "he likes play ##ing [SEP]"

  [SEP] is a separator token used to distinguish segments in tasks like question answering or sentence-pair classification.

- Segment Embeddings (Green Boxes): These embeddings indicate which sentence a token belongs to. "A" represents the first sentence, and "B" represents the second. This helps BERT differentiate between multiple sentences in tasks like next-sentence prediction.

- Position Embeddings (Black Boxes): These embeddings provide information about the position of each token within the input sequence. Since transformers don't inherently understand word order, position embeddings enable the model to learn and utilize this information.

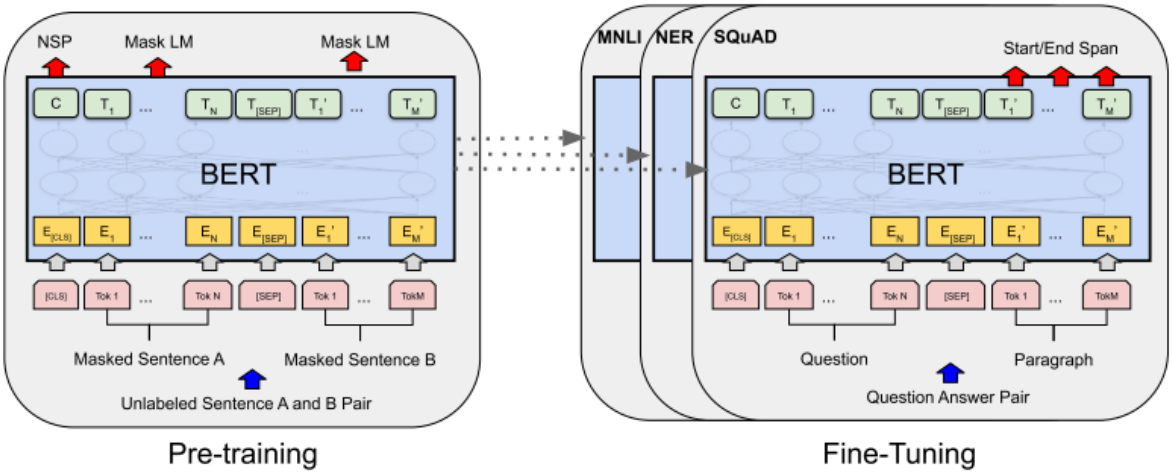


Figure 3.10: Overall pre-training and fine-tuning procedures for BERT [Devlin et al., 2019].

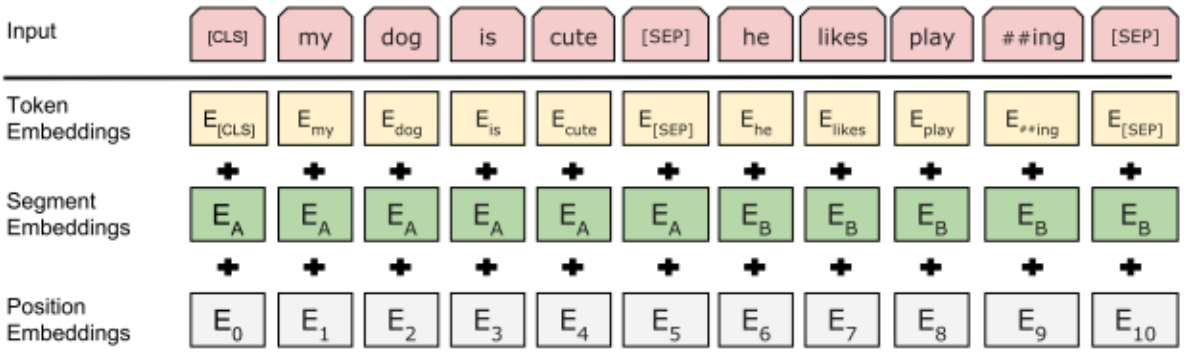


Figure 3.11: BERT input representation [Devlin et al., 2019].

3.6.2 BioBERTpt

BioBERTpt is a language model developed specifically for Portuguese clinical and biomedical text, designed to improve performance on tasks such as named entity recognition in clinical narrative [Schneider et al., 2020]s. It is based on BERT and fine-tuned on Portuguese-language medical texts, including clinical notes from Brazilian hospitals and biomedical scientific papers published in Pubmed and Scielo. BioBERTpt was created to address the limited availability of high-performing NLP tools for clinical data in Portuguese. Three BERT-based models were fine-tuned on these Portuguese-language medical texts and the models were initialized with multilingual BERT weights.

The BioBERTpt achieved better performance than general BERT models in recognizing specific medical entities on the two Portuguese corpora: SemClinBr and CLINpt [Schneider et al., 2020].

3.6.3 Gemini

Gemini is a family of highly capable multimodal models developed by Google. These models are designed to process and understand data in multiple modalities, including text, images, audio, and video, and exhibit advanced reasoning and understanding capabilities. The Gemini 1.0 models are built on an enhanced Transformer decoder architecture [Vaswani et al., 2023] designed to support multimodal inputs and outputs, including text, images, audio, and video. These models are natively multimodal, allowing seamless integration of various data types within a single context. They are capable of processing interleaved sequences of inputs, such as text with accompanying images or audio, while supporting a large context length of 32,768 tokens. To enable efficient handling of such extensive data, the architecture incorporates advanced attention mechanisms, such as multi-query attention, which improve scalability and performance during training and inference [Team et al., 2024a].

Figure 3.12 shows an overview of Gemini 1.0 model. The input sequence consists of text, audio, an image, and video. Each input modality is processed to create feature embeddings, which are then combined into a single sequence. This combined sequence is fed into a transformer, the core of the model. The transformer processes this combined information and generates outputs through two decoders: an image decoder, producing an image, and a text decoder, generating text.

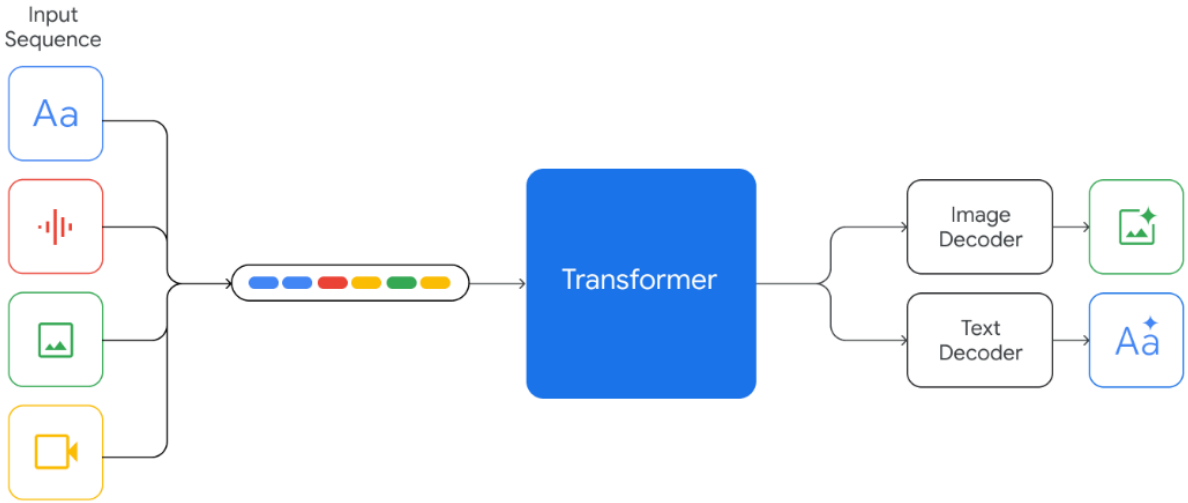


Figure 3.12: Gemini model overview [Team et al., 2024a].

Continuing the development of these models, Google launched Gemini 1.5, which is a family of multimodal large language models designed to advance capabilities in reasoning,

efficiency, and long-context understanding. This model supports mixed-modality inputs, allowing it to process and analyze text, audio, video, and code within a single context seamlessly. A defining feature of Gemini 1.5 is its ability to handle extremely long input sequences, supporting context lengths of up to 10 million tokens, enabling it to work with massive datasets, extended documents, hours of video, or lengthy audio recordings. The family includes two variants: Gemini 1.5 Pro, optimized for peak performance across various benchmarks, and Gemini 1.5 Flash, a lighter, efficiency-focused model that maintains high quality with reduced computational demands. Built on a sparse mixture-of-experts Transformer [Fedus et al., 2022] that builds on Gemini 1.0's [Team et al., 2024a], Gemini 1.5 employs advanced routing mechanisms to efficiently scale its parameter count while keeping computational requirements manageable. It has achieved significant improvements over its predecessors, outperforming state-of-the-art models such as Gemini 1.0 Ultra on benchmarks in areas such as reasoning, multilinguality, and multimodal understanding [Team et al., 2024b].

3.6.4 GPT 4

GPT-4 is a large-scale, multimodal model capable of processing both text and image inputs to produce text outputs. It is based on a Transformer architecture [Vaswani et al., 2023] and was pre-trained to predict the next token in a sequence using vast amounts of publicly available and licensed data. After pre-training, the model underwent fine-tuning with techniques such as Reinforcement Learning from Human Feedback to enhance its alignment with user intents and improve its factuality. However, specific details about its architecture, such as model size, exact hardware, training compute, and dataset construction—have not been disclosed due to competitive consideration [OpenAI et al., 2024].

3.6.5 Llama 3

Llama 3 is a herd of language models developed by Meta AI, designed to natively support multilinguality, coding, reasoning, and tool usage. The largest model in the series is a dense Transformer with 405 billion parameters, capable of processing up to 128,000 tokens in a single context window. The Llama 3 family includes models of varying sizes—8B (Billions of parameters), 70B, and 405B parameters—all trained on a massive and diverse dataset. The 405B model was pre-trained on a corpus of 15 trillion tokens, which significantly surpasses the scale of its predecessors, Llama and Llama 2. The flagship model demonstrate state-of-the-art performance on various benchmarks, rivaling leading language models like GPT-4 [OpenAI et al., 2024] in many tasks. In addition to language capabilities, Llama 3 incorporates safety-focused designs, such as the Llama Guard 3 variant, to ensure secure handling of inputs and outputs. Meta has also conducted promising

multimodal experiments with image, video, and speech processing, though these extensions remain under development [Grattafiori et al., 2024].

The Figure 3.13 shows an overview of the overall architecture and training of Llama 3. The process begins with input text tokens. These are the individual words or sub-word units that form the input sequence. These tokens are then fed as input to the model. Each input token is converted into a vector representation called an embedding. This maps each token to a point in a high-dimensional space, capturing semantic information about the token. The core of the transformer model is formed by stacked layers of identical blocks. The input embeddings are processed through the first layer's Self-Attention and Feedforward Network. The output of this layer is then passed to the next layer. The final layer of the transformer produces a probability distribution over the vocabulary. The token with the highest probability (or sampled from the distribution) is selected as the output text token. This token is then appended to the generated sequence, and the process repeats to generate the next token until a complete sequence is reached.

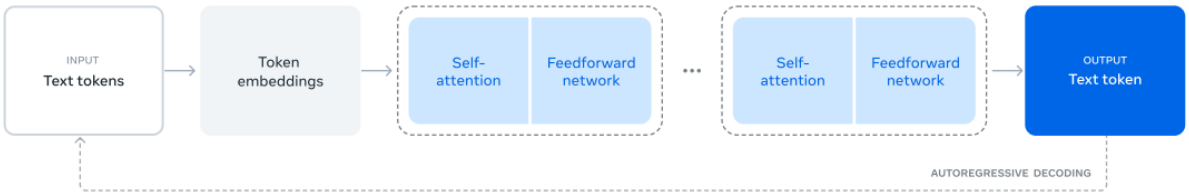


Figure 3.13: Llama 3 model overview [Grattafiori et al., 2024].

The architecture of Llama 3 is based on a standard dense Transformer model [Vaswani et al., 2023], similar to its predecessors, Llama [Touvron et al., 2023a] and Llama 2 [Touvron et al., 2023b], but with key enhancements that drive significant performance improvements. Some small modifications were made compared to LLama 2. In Llama 3, Grouped Query Attention (GQA) [1] was used with 8 key-value heads, as proposed by [Ainslie et al., 2023], to improve inference speed and reduce the size of key-value caches during decoding. Additionally, they employed an attention mask that prevents self-attention between different documents within the same sequence. Although this adjustment had minimal impact during standard pre-training, it proved to be crucial when continuing pre-training on very long sequences. A vocabulary with 128K tokens was used, combining 100K tokens from the tiktoken tokenizer with 28K additional tokens to better support non-English languages. To better support longer contexts, they increased the Rotary Position Embedding (RoPE) [2] base frequency hyperparameter to 500,000, a value shown by [Xiong et al., 2023] to be effective for context lengths up to 32,768. The architecture of Llama 3 405B includes 126 layers, a token representation dimension of 16,384,

[1] GQA emerges as an innovative extension of traditional attention mechanisms, aiming to address several challenges associated with processing long sequences efficiently.

[2] This technique is used to improve the ability of LLMs to handle longer sequences of text than those seen during training by modifying the base value used in the RoPE calculations.

and 128 attention heads.

3.7 Prompt Engineering

Prompt engineering is a technique used to enhance the capabilities of large language models by creating specific instructions or queries known as prompts. These prompts guide the model in producing desired outputs without the need to retrain or modify the underlying model parameters [Chen et al., 2024]. The process involves designing task-specific prompts that can take various forms, ranging from natural language instructions to structured vector representations that help to activate the relevant knowledge within the model. This strategic use of prompts allows the models to perform a variety of tasks, such as question answering, reasoning, and text generation, effectively utilizing their pretrained knowledge base. In addition, prompt engineering facilitates the integration of these models into different applications, enhancing their usability in various domains [Sahoo et al., 2024].

3.7.1 Zero-Shot Learning

In the context of language models, zero-shot learning means that these models can understand and generate responses to tasks they haven't been specifically fine-tuned for, based only on their training on a wide variety of text data. This method takes advantage of the large textual data on which the model was trained to eliminate the need for large task-specific datasets. Instead, carefully designed prompts guide the model to perform these novel tasks [Radford et al., 2019].

Figure 3.14 shows an example of a prompt using zero-shot learning for the task of sentiment analysis to determine whether a text expresses a positive, negative, or neutral sentiment. The instruction given is: "Classify the text as neutral, negative, or positive." This tells the model what it needs to do. The input text is: "I think the vacation is good." This is the text to be analyzed. The expected output format is: "Sentiment:" This indicates the type of answer expected.

3.7.2 Few-Shot Learning

Few-shot learning, refers to a setting where a large language model is given a few input-output examples of task at inference times as conditioning, but no weight updates or fine-tuning are performed. The model learns the task purely from these examples presented in the input context. The model is presented with a small number of examples (typically between 10 to 100) within its context window, which helps it infer the correct pattern and generate the appropriate outputs [Brown et al., 2020].

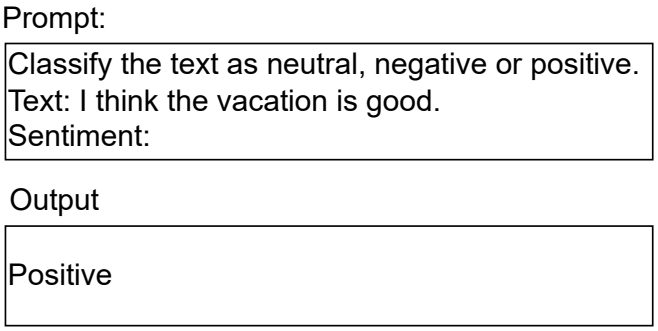


Figure 3.14: Prompt with zero-shot learning.

Figure 3.15 shows an example of a prompt using zero-shot learning to the sentiment analysis. However, in this case, the model is provided with a few examples before being asked to classify a new sentence. The prompt with few-shot learning a few examples, such as: "This is amazing! // Positive," "This is bad! // Negative," and "Wow, that movie was awesome! // Positive." These examples illustrate the relationship between the text and its sentiment, allowing the model to learn which types of phrases correspond to positive or negative sentiment. The new input text is "What a terrible show! //", which is the sentence the model needs to classify. The expected output format is implicit, as the "//" followed by a blank space indicates that the model should generate the appropriate sentiment label.

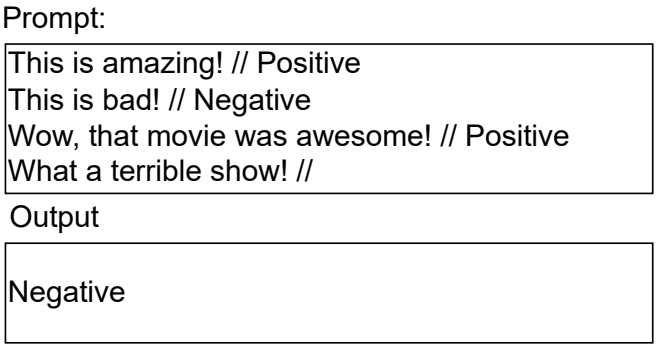


Figure 3.15: Prompt with few-shot learning.

Chapter 4

Materials and Methods

4.1 Lung-RADS

The Lung Imaging Reporting and Data System, established by the American College of Radiology, serves as a standardized reporting framework aimed at optimizing the assessment and management of pulmonary nodules identified via low-dose computed tomography screenings. By systematically categorizing findings from lung cancer screenings, the Lung-RADS classification system enhances diagnostic accuracy, reduces interpretative variability, and facilitates consistent decision-making in clinical practice [Beyer et al., 2017] [Christensen et al., 2024]. HThe following is a summary of each category from the Lung-RADS 2022

1. Category 0: Incomplete
    - Findings: Chest CT is incomplete, part or all of lungs cannot be evaluated, or previous CT is needed for comparison.
    - Management: Additional CT needed; follow-up CT within 1-3 months for suspected inflammatory or infectious processes.
2. Category 1: Negative
    - Findings: No nodules or nodules with benign features (complete, central, popcorn, or concentric ring calcifications or fat-containing).
    - Management: Follow-up in 12 months.
3. Category 2: Benign Appearance or Behavior
    - Findings: Likely benign features; includes specific size and shape nodules.
    - Management: 12-month follow-up.
4. Category 3: Probably Benign

- Findings: Low risk of malignancy; solid or part-solid nodules within specific size ranges.
- Management: 6-month follow-up.

5. Category 4A: Suspicious
   - Findings: Increased suspicion for malignancy; larger nodules or growth seen.
   - Management: 3-month follow-up; PET/CT or additional imaging may be recommended for solid nodule or solid component $\geq$ 8 mm.
6. Category 4B: Very Suspicious
   - Findings: Highly suspicious; larger nodules or solid components.
   - Management: Diagnostic CT with/without contrast, PET/CT imaging may be considered for solid nodule or solid component $\geq$ 8 mm, possible biopsy, or clinical evaluation.
7. Category 4X: Highest Suspicion
   - Findings: Nodules with additional suspicious features (e.g., spiculation or metastasis).
   - Management: Tailored to specific findings, with thorough evaluation and management per clinical guidelines.

The complete Lung-RADS description and the patient management for each Lung-RADS indexis is available at the following url: `https://edge.sitecorecloud.io/americancoldf5f-acrorgf92a-productioncb02-3650/media/ACR/Files/RADS/Lung-RADS/Lung-RADS-2022.pdf`.

4.2 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 patient consent, all chest CT reports, irrespective of the clinical indication were included. It is important to highlight that all patient data was anonymized. Data cleaning was applied to all reports, which included removing special characters, adding spaces between words, and removing emojis in the text. Next, each report was uploaded to the Doccano annotation tool [Nakayama et al., 2018]. The text was labeled with six named entities in Portuguese, corresponding to the characteristics of pulmonary nodules. The data annotation was performed by the author, and the results

were reviewed by a radiologist. The NEs used were "Atenuação" (Attenuation), "Calcificação" (Calcification), "Bordas" (Edges), "Achado" (Finding), "Localização" (Localization) and "Tamanho" (Size). These characteristics were chosen based on Lung-RADS guidelines [ACR, 2022]. Figure 4.1 shows an example of report annotated in Doccano tool. In this repot 5 NEs were identifed: "Finding", "Edges", "Calcification", "Location", and "Size".

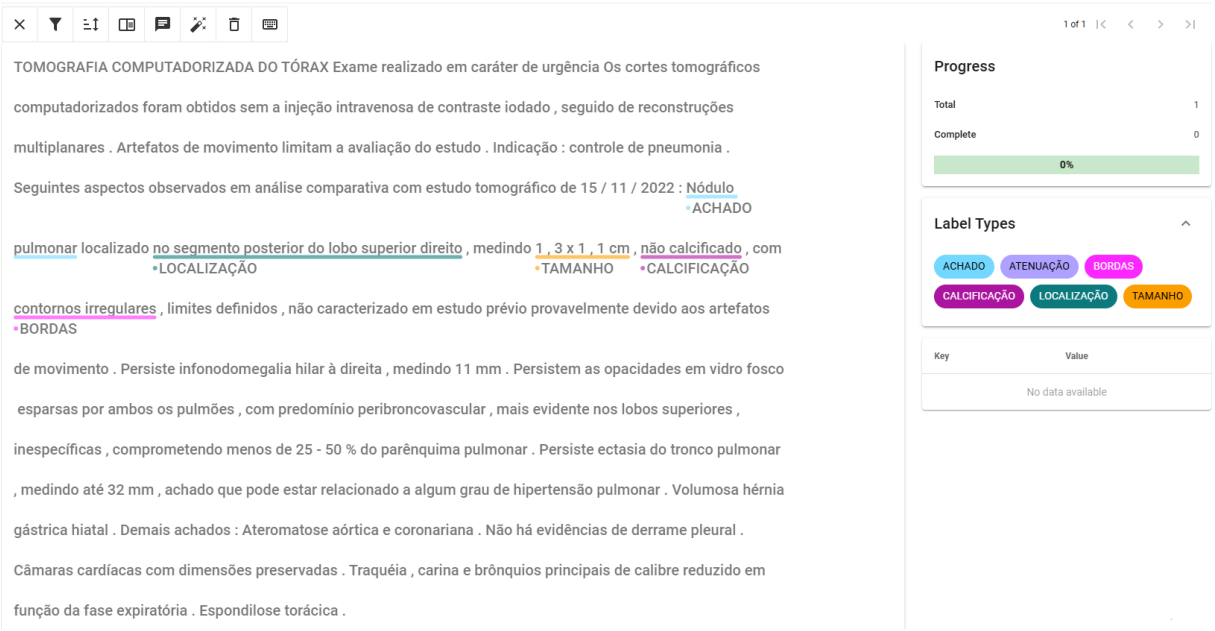


Figure 4.1: Example of a chest CT Report annotated in Doccano [Nakayama et al., 2018].

There are multiple annotation schemes for named entity recognition, such as:

- IO: Each token from the dataset is assigned one of two tags: an inside tag (I) and an outside tag (O). The I tag is for named entities, whereas the O tag is for normal words;
- IOB: IOB is also called BIO scheme. It's famous due to adoption by Conference on Computational Natural Language Learning (CoNLL) . It assigns one of these three tags to a token: the beginning of a known named entity (B), an inside tag (I) and an outside tag (O);
- IOE: This scheme is similar to IOB, but instead of indicating the beginning of named entity (B), it indicates the end (E);
- IOBES: The IOBE in IOBES scheme is a combination of IOB scheme and IOE scheme. It increases the amount of information related to the boundaries of named entities. (S) is a new tag, created for single token named entity.

For the annotation scheme, we adopted the IOB format. This format was adopt because the different values assessed for annotation scheme did not have a considerable im-

Report	***Token***	***Tag***
06	Existem	O
06	pequenos	B-Tamanho
06	nódulos	B-Achado
06	calcificados	B-Calcificação
06	no	B-Localização
06	segmento	I-Localização
06	basal	I-Localização
06	posterior	I-Localização
06	esquerdo	I-Localização
06	,	O
06	o	O
06	maior	O
06	deles	O
06	medindo	B-Tamanho
06	cerca	I-Tamanho
06	de	I-Tamanho
06	0,23	I-Tamanho
06	x	I-Tamanho
06	0,18	I-Tamanho
06	cm	I-Tamanho

Table 4.1: Example of a report annotated with the IOB scheme.

pact in the results of Portuguese NER using LSTM-CRF [Quinta de Castro et al., 2018]. Note that each token is individually annotated, thus resulting in a sequence of tags.

Table 4.1 shows an example of a report and its IOB tags. For the sentence "... There are small calcified nodules in the left posterior basal segment, the largest measuring about 0.23 x 0.18 cm ...", the IOB tags are: "O, B-Finding, B-Finding, B-Calcification, B-Localization, I-Localization, I-Localization, I-Localization, I-Localization, O, O, O, O, B-Size, I-Size, I-Size, I-Size, I-Size, I-Size, I-Size".

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. From the set of 963 reports, the report with the highest number of tokens contained 497 tokens. However, the BERT model requires an input of 512 tokens. Therefore, all texts were padded with zeros at the end of the list to ensure that their input had 512 tokens. This specific token quantity was employed because it aligns with the token limit of the BioBERTpt model [Schneider et al., 2020]. Padding tokens were designated with a distinctive tag: "–PADDING–".

The 963 texts from the CT reports were divided into two proportions: 70% for training

and 30% for testing in the Named Entity Recognition task. This data split was chosen based on the study by [Lopes et al., 2019]. For the IE task, a total of 100 CT reports were utilized to perform the Question-answering procedure. This data split was chosen based on the study by [Hu et al., 2024a].

Based on the Lung-RADS guidelines [ACR, 2022], the author defined eight questions related to pulmonary nodules for the QA task. A thoracic radiologist with 15+ years of experience provided the answers to the questions used for QA. Table 4.2 presents the questions and their corresponding statistics. The analysis of the 100 reports used to perform the QA procedure identified pulmonary nodules in all reports, with Lung-RADS categorization as follows: Category 0 (0 cases), Category 1 (61 cases), Category 2 (25 cases), Category 3 (2 cases), Category 4A (7 cases), Category 4B (0 cases), and Category 4X (5 cases).

Table 4.2: Pulmonary nodules Questions and the statistics of the annotated answers.

No.	Question	Answer type	Answer statistic
1	Report ID	Numerical	-
2	Is the nodule solid	Boolean	10 (Positive)
3	Is the nodule soft tissue, semisolid or subsolid	Boolean	7 (Positive)
4	Is the nodule ground glass	Boolean	2 (Positive)
5	Is the nodule spiculated or irregular	Boolean	4 (Positive)
6	Is the nodule calcified	Boolean	61 (Positive)
7	Nodule location	Categorical	21 (RUL) 10 (RML) 28 (RLL) 18 (LUL) 17 (LLL) 7 (Others)
8	Nodule size	Numerical	5.41 mm ± 3,27

RUL - Right upper lobe RML - Right middle lobe RLL - Right lower lobe LUL - Left upper lobe LLL - Left lower lobe

4.3 Models for Named Entity Recognition

For the named entity recognition task, the effectiveness of the BiLSTM-CRF and BioBERTpt models was compared. To choose the best hyperparameters for BiLSTM-CRF, a grid search was carried out with the following values:

- BiLSTM-CRF;
- Word embedding size = [50, 100, 200];
- LSTM units = [25, 50, 100];

- Batch size = [4, 8, 16];
- Epochs = 10;
- Dropout = 0.1;
- Learning Rate = 0.01.

From all the possible values for each of these parameters for the BiLSTM-CRF model, there are 27 different combinations to evaluate. Following the approach used by [Quinta de Castro et al., 2018] to choose the best set of hyperpameters for the BiLSTM-CRF models, each of these combinations was run 10 times in different data splits and only 3 ephocs were used for this training step. To determine if there are statistical differences between the 27 models, Friedman test [Rainio et al., 2024] [1] was perfomed. After that, the Nemenyi [Demšar, 2006] [2] test was done to determine which ones are significantly different. Based on this analysis, a single model was chosen from the 27 potential models. This selected model was trained for 10 epochs using the 10 data splits and was subsequently evaluated for its performance in the named entity recognition task.

The BERT model was fine-tunned using BioBERTpt [Schneider et al., 2020]. The model weights were initialized using the Transformers library, available on HuggingFace [Wolf et al., 2020], and the model's PyTorch (version 2.0.1) implementation was used. The model was fine-tuned using the following hyperparameter values based in [Schneider et al., 2020]: AdamW optimizer, with a weight decay as 0.01, batch size 4, maximum length as 256, learning rate as 3e-5, maximum epoch as 10, and the linear schedule that decreases the learning rate throughout the epochs with warmup as 0.1.

When the text of the report texts was tokenized by the BERT tokenizer, the resulting number of tokens was greater than 512 tokens, which is the BioBERTpt input limit. To deal with this problem, the reports were divided into 4 parts to ensure that when they were tokenized, each part of the report did not exceed the limit of 512 tokens. Therefore, the value of the maximum length hyperparameter was changed to 512. The GPU utilized in the fine-tuning procedure was an RTX3060 12GB.

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 nodule characteristics.

Figure 4.2 shows an overview of the Lung-RADS classification process using the BiLSTM-CRF and BioBERTpt models with the NER technique. The process begins

[1] The Friedman test is a non-parametric statistical test used to detect differences in treatments across multiple test attempts. In machine learning, the Friedman test is commonly used to compare the performance of multiple models or algorithms across different datasets or tasks

[2] The Nemenyi test is a post-hoc test used after the Friedman test to determine which specific models differ significantly when multiple comparisons are made.

with a CT report and this report undergoes data preprocessing, which includes data cleaning to remove noise, annotation using Doccano to manually label relevant entities, tokenization to break the text into units, and splitting the data into train and test sets. In the NER step, the models learn to identify and classify entities. This involves fine-tuning a pre-trained model (BioBERTpt) and training a BiLSTM-CRF model. Both models, BioBERTpt and BiLSTM-CRF, are used in conjunction for the NER task. Following NER is post-processing, where the extracted information is used for Lung-RADS classification. Finally, the entire process undergoes evaluation to assess its performance.

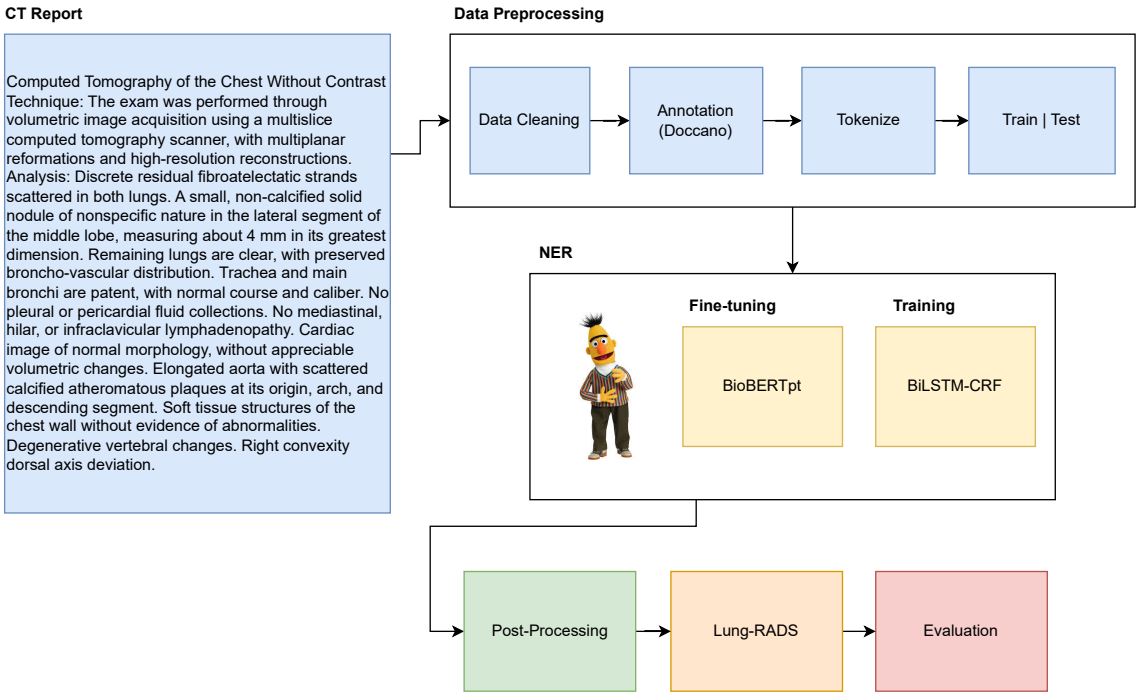


Figure 4.2: Methodology scheme applied in NER task for Lung-RADS classification.

4.4 Post-processing Named Entity Recognition Extraction

Building upon our named entity recognition process, we implemented a systematic post-processing approach to transform the raw entity recognition results into a structured and meaningful representation. Specifically, after the BiLSTM-CRF and BioBERTpt models identified the named entities, a method was developed to populate the question table (Table 4.2). The process involved carefully parsing the outputs from both AI models, cross-referencing the identified entities, and systematically populating the predefined table structure.

4.5 Models for Question Answering

For the LLMs models, we used the specific versions: GPT-4o (gpt-4o-mini-2024-07-18) [OpenAI et al., 2024], Gemini 1.5 Flash (gemini-1.5-flash) [Team et al., 2024b], and the Llama 3 70B [Grattafiori et al., 2024]. In the context of LLMs, the temperature parameter regulates the uncertainty or randomness in the generation process. This parameter typically ranges from 0 (completely deterministic) to 1 or higher (resulting in increasingly random and diverse outputs) [Hu et al., 2024b]. For the GPT-4o, Gemini 1.5 Flash, and Llama 3 models, the temperature was set to 0 to minimize randomness in response generation. Using a lower temperature, we limited the model's tendency to take creative leaps, ensuring more predictable and consistent outputs. This is important in IE tasks, where the accuracy of the information extracted is crucial.

Requests for the three LLM models were made via the API. We used the OpenAI API for GPT-4o, the Together AI API for Llama 3 70B, and the Google AI API for the Gemini 1.5 Flash model. During the tests conducted in this study, the cost of GPT-4o was US$0.075 per 1 million tokens for input and US$0.60 per 1 million tokens for output. For Llama 3 70B, the cost was approximately US$0.88 per 1 million tokens for both input and output. Gemini 1.5 Flash model has no cost for processing incoming and outgoing tokens. Depending on the volume of text processed, the cost of these paid models may hinder their use in real-world systems.

To evaluate the consistency of the responses of these LLMs, the approach of [Hu et al., 2024a] [Saxena et al., 2024] was used. This approach consist in repeatedly presenting the same query to the models and observing variations in their responses. The same query was presented three times for each LLM, and the consistency and inconsistency percentage was calculated for each question in the Question table.

The consistency percentage is given by:

$$C_p = \frac{n_c}{n_t} \cdot 100 \tag{4.1}$$

The inconsistency percentage is given by:

$$C_n = \frac{n_{ic}}{nt} \cdot 100 \tag{4.2}$$

were, C_p is the consistency percentage, C_n is the inconsistency percentage, n_c is the total number of consistent responses, n_ic is the total number of inconsistent responses, and n_t is the total number of question in the data.

4.6 Prompt Engineering

Based on the work of Danqing Hu et al. [Hu et al., 2024a], we designed the prompts used for QA. Danqing Hu et al. employed zero-shot learning in their input prompts. In adittion to zero-shot learning, we also employed few-shot learning technique. The 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. To improve LLM task comprehension, annotated reports with completed tables were provided. Two prompt templates, detailed in the appendices 6.2, were used for zero-shot and few-shot learning.

To calculate the similarity of the test reports with the training reports, the cosine similarity was used, which is given by the following equation.

$$\text{cosine similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4.3}$$

where:

- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of $\mathbf{A}$ and $\mathbf{B}$.
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the magnitudes (or Euclidean norms) of $\mathbf{A}$ and $\mathbf{B}$, respectively.
- n is the number of dimensions of the vectors.
- A_i and B_i are the i-th components of vectors $\mathbf{A}$ and $\mathbf{B}$.

From our database of 963 reports, only those that contained answers to all the questions in Table 4.2 were selected. After the filtering process, the dataset was narrowed down to 300 reports. One hundred reports were used for testing, while the remaining 200 were used as examples for few-shot learning. The testing framework utilized two prompts for zero-shot learning. For few-shot learning, it leveraged two prompt templates, each instantiated with five and ten examples, resulting in four few-shot prompts (Template 1 with 5 examples, Template 1 with 10 examples, Template 2 with 5 examples, and Template 2 with 10 examples).

The CT reports were combined with prompt templates 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 IE results. Additionally, 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 it is irrelevant to the analysis. The focus is exclusively on extracting the content in the form of the Question table.

Figure 4.3 details the methodology used for a QA task for Lung-RADS classification, focusing on extracting structured data from CT scan reports. The process begins with the original CT report. This report is combined with a specific prompt, a question designed to guide the LLM in extracting relevant information about lung nodule to Lung-RADS classification. The prompt can be zero-shot or few-shot. The LLM processes the combined report and prompt, generating a free-text response. This response is then converted through post-processing into a structured format, typically a JSON table. This table contains key-value pairs representing specific Lung-RADS criteria (e.g., "Is the nodule solid?") and their corresponding values (e.g., "Yes"). The final output is a structured response, the JSON table ready for analysis and use. The structured data are then evaluated against a gold standard to evaluate the effectiveness of LLMs for the Lung-RADS classification.

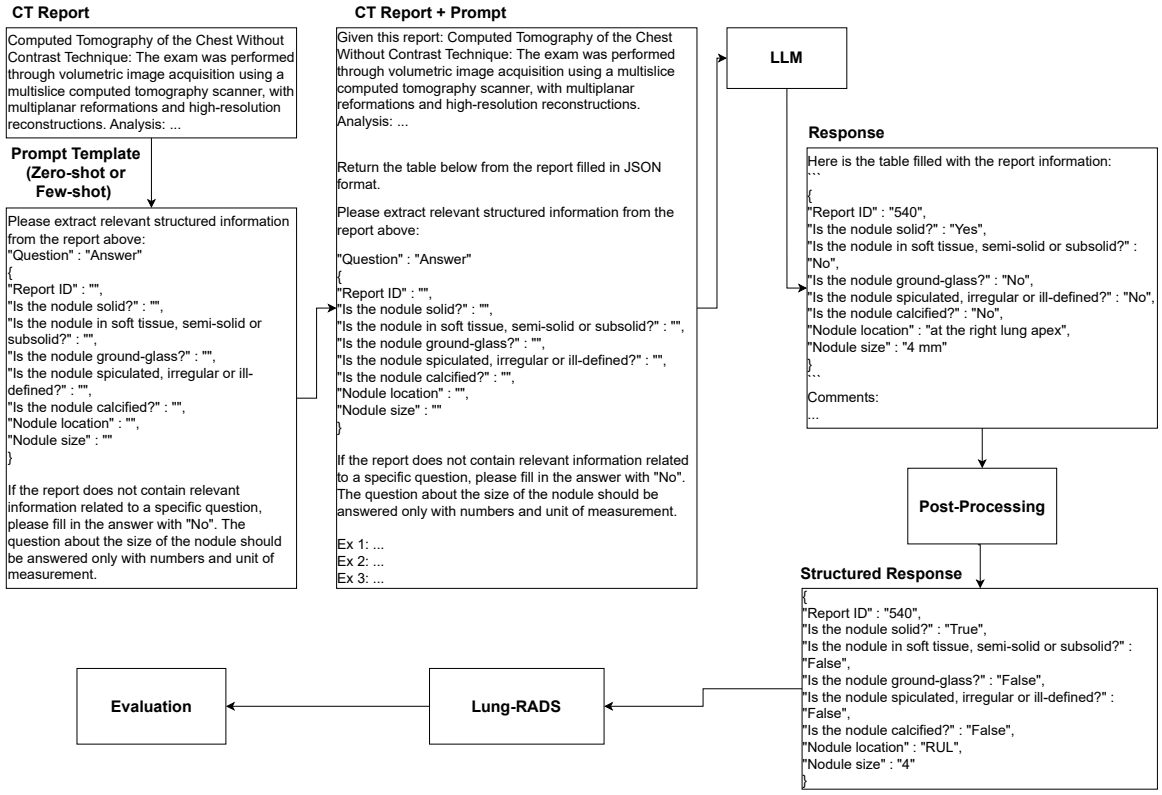


Figure 4.3: Methodology scheme applied in QA task for Lung-RADS classification.

4.7 Post-processing for Question Answering

The LLMs were instructed to extract only the answers from the provided table in the Prompts. However, the responses were only sometimes structured. Post-processing was applied to convert unstructured responses into a structured format to address this issue. This post-processing involved removing all text except the Questions and answers table.

Besides, when the LLM did not provide an answer in the table, leaving the space blank, or responded with 'Not informed,' the answer was considered 'No.' For Questions 2 to 6 (Table 4.2), the answers 'Yes' and 'No' were converted into Boolean values.

Regular expressions were used to identify the number and unit of measurement to extract the nodule size, which was then converted to millimeters. For the location Question, keywords such as 'right,' 'left,' 'middle,' 'upper,' and 'lower' were employed to categorize the answer into six formats: 'right upper lobe,' 'right middle lobe,' 'left lower lobe,' 'left upper lobe,' 'left lower lobe,' and 'others.' The answer 'others' was used when the specific location of the pulmonary nodule was unclear.

4.8 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. During this review, the radiologist was asked to:

- Assess the Lung-RADS index assigned by the ruled-based algorithm and indicate whether he agreed or disagreed with the generated classification;
- Provide a detailed justification for his evaluation;

This methodological approach allows for a comprehensive assessment of the AI tool's performance by comparing machine-generated classifications against 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. Analyzing discrepancies between AI classifications and radiologist interpretations reveals specific areas where the algorithm's decision-making can be improved.

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

4.9 Evaluation

To evaluate the effectiveness of the BiLSTM-CRF and BioBERTpt models for NER, and to assess the effectiveness of LLMs in the QA task using zero-shot and few-shot learning,

precision (P), recall (R), and the F1-score (F1) were used. These same metrics were also employed to evaluate the effectiveness of combining the result of these models with a rule-based algorithm for Lung-RADS classification.

- Precision for each Lung-RADS category evaluates the model's ability to avoid misclassifying exams into a particular category, indicating how well the model prevents erroneous positive classifications.

- Recall, also known as the true positive rate or sensitivity, measures the model's ability to correctly identify exams that genuinely belong to each Lung-RADS category. It captures how effectively the model retrieves relevant cases.

- F1-score balances these two metrics by combining precision and recall into a single score, representing the model's overall effectiveness for each Lung-RADS category. Calculated as the harmonic mean of precision and recall, the F1-score provides a robust measure of the model's performance in correctly classifying Lung-RADS categories while minimizing false positives and false negatives.

Chapter 5

Experimental Results and Discussion

Table 5.13 shows F1-scores of 27 BiLSTM-CRF models for the named entity task trained for 3 epochs in 10 data splits. The analysis of the table reveals a discrepancy in the models' F1-scores across different data splits. Specifically, models 19 (batch size = 16, word embedding size = 50, LSTM units = 50) and 24 (batch size = 16, word embedding size = 100, LSTM units = 200) achieved the lowest F1-scores, both scoring 0.74 in data split 6. In contrast, model 15 (batch size = 8, word embedding size = 100, LSTM units = 200) achieved the highest F1-score, recorded in data split 2.

The Friedman test performed shows that there is a difference between some models. The Nemenyi test was used to determine which specific groups are different from each other. Figure 5.1 shows a heat map that contains significant differences ($P\ value < 0.05$) between the models. Figure 5.2 shows a box plot with the F1-scores of these 27 models. Analysis of Figures 5.1 5.2 reveals that model 1 has statistically higher results on F1-score compared to models 19 ($P\ value = 0.000617$) and 20 ($P\ value = 0.028268$). In particular, model 1 exhibits low dispersion in its F1-score values. Given its simplicity (Word Embedding Size = 50, LSTM units = 50), its high average F1-score (0.86, within the top 5), its consistent F1-score across the 10 data splits, and the absence of statistically significant differences from other models, model 1 was selected for the Lung-RADS classification task.

Table 5.1 presents the results of the BiLSTM-CRF model (Model 1) trained for the named entity recognition task over 10 epochs across 10 data splits. This model achieved a macro F1-score of 0.86. This result is consistent with the performance observed after only 3 training epochs, which yielded the same F1-score. Table 5.2 displays the results of BioBERTpt fine-tuned for the same named entity recognition task. This model, fine-tuned for 10 epochs, achieved an impressive macro F1-score of 0.99 across the 10 data splits.

The proportion of consistent responses concerning lung nodule questions for Gemini 1.5 Flash, GPT-4o, and Llama 3 70B is detailed in Tables 5.3, 5.4, 5.5, 5.7, 5.6, and 5.8. The extraction of nodule attenuation, calcification, edges, location, and size demonstrated

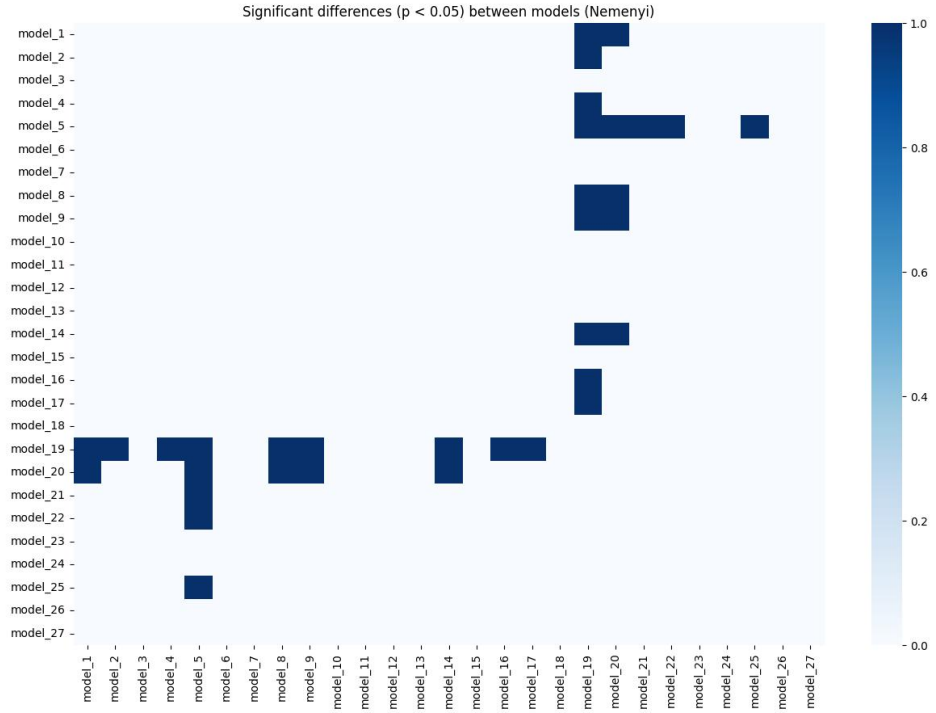


Figure 5.1: Nemenyi post-test 27 BiLSTM-CRF models

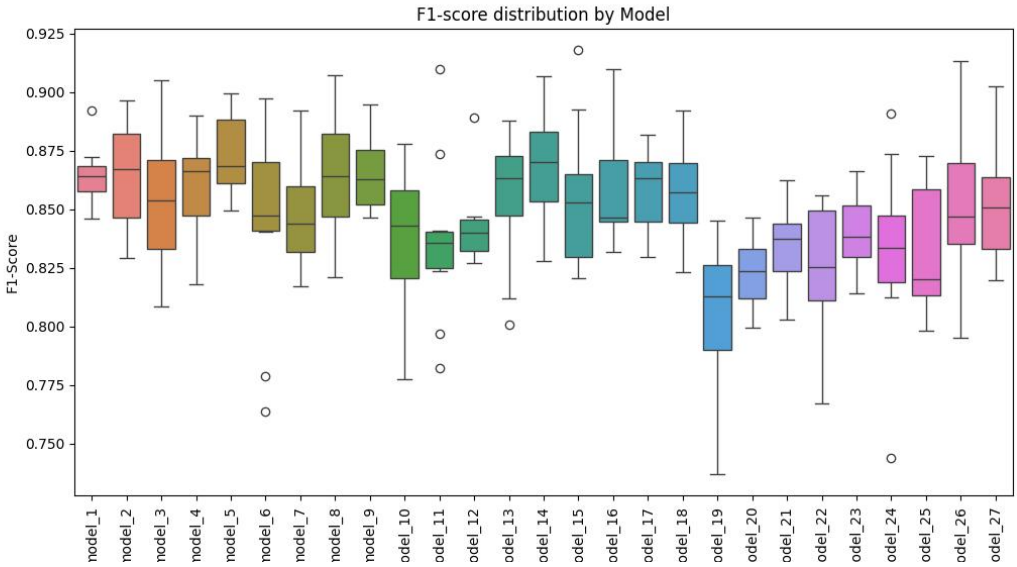


Figure 5.2: F1-scores of all entities evaluation of the 27 BiLSTM-CRF models

greater consistency across the six prompt templates for Gemini 1.5 Flash and Llama 3 70B. In contrast, GPT-4o showed reduced consistency for Question 1 (lung nodule attenuation) with zero-shot prompt 1 (0.87), Question 6 (lung nodule location) with zero-shot prompt 1 (0.79), and Question 6 with zero-shot prompt 2 (0.86). The possible reason for this happening is that the GPT-4o returns the size and attenuation of other findings described

Table 5.1: Evaluation BiLSTM-CRF Model in 10 Data Splits

Model	Batch Size	Embedding Dim	LSTM Units	Epochs	Dropout	Macro F1-Score in 10 Data Splits
BiLSTM-CRF	4	50	50	10	0.1	0.86

Table 5.2: Evaluation of the BioBERTpt Model in 10 Data Splits

Model	Batch Size	Maximum Length	Epochs	Learning Rate	Macro F1-Score in 10 Data Splits
BioBERTpt	4	512	10	3e-5	0.99

in the chest CT report.

Table 5.3: Evaluation Consistency of Response LLMs Zero-shot Prompt Template 1

LLM	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Gemini 1.5 Flash	0.98	0.99	0.99	1.00	1.00	0.98	0.99
GPT-4o	0.87	1.00	0.98	0.96	0.98	0.79	0.92
Llama 3 70B	0.99	0.98	1.00	0.99	1.00	0.97	1.00

Q1 - Is the nodule solid or soft tissue?
Q2 - Is the nodule semisolid or subsolid?
Q3 - Is the nodule ground-glass?
Q4 - Is the nodule spiculated, irregular, or poorly defined?
Q5 - Is the nodule calcified?
Q6 - Nodule Location
Q7 - Nodule Size

Table 5.4: Evaluation Consistency of Response LLMs Zero-shot Prompt Template 2

LLM	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Gemini 1.5 Flash	0.95	1.00	0.99	1.00	1.00	0.97	1.00
GPT-4o	0.97	0.97	0.99	0.98	0.91	0.86	0.94
Llama 3 70B	1.00	1.00	1.00	1.00	1.00	1.00	0.99

Q1 - Is the nodule solid or soft tissue?
Q2 - Is the nodule semisolid or subsolid?
Q3 - Is the nodule ground-glass?
Q4 - Is the nodule spiculated, irregular, or poorly defined?
Q5 - Is the nodule calcified?
Q6 - Nodule Location
Q7 - Nodule Size

Table 5.9 shows the precision, recal and F1-score metrics for Gemini 1.5 Flash, GPT-4o and Llama 3 70B in QA task using zero-shot prompt templates 1 and 2. The F1-score

Table 5.5: Evaluation Consistency of Response LLMs Few-shot with 5 Examples Prompt Template 1

LLM	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Gemini 1.5 Flash	0.99	1.00	1.00	1.00	0.98	0.99	1.00
GPT-4o	0.98	0.99	0.98	1.00	0.96	0.97	0.97
Llama 3 70B	1.00	1.00	1.00	1.00	1.00	0.98	0.98

Q1 - Is the nodule solid or soft tissue?
Q2 - Is the nodule semisolid or subsolid?
Q3 - Is the nodule ground-glass?
Q4 - Is the nodule spiculated, irregular, or poorly defined?
Q5 - Is the nodule calcified?
Q6 - Nodule Location
Q7 - Nodule Size

Table 5.6: Evaluation Consistency of Response LLMs Few-shot with 10 Examples Prompt Template 1

LLM	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Gemini 1.5 Flash	0.99	1.00	0.99	1.00	0.99	0.95	0.99
GPT-4o	0.97	1.00	0.98	0.99	0.95	0.96	0.96
Llama 3 70B	0.99	1.00	1.00	1.00	1.00	0.97	0.98

Q1 - Is the nodule solid or soft tissue?
Q2 - Is the nodule semisolid or subsolid?
Q3 - Is the nodule ground-glass?
Q4 - Is the nodule spiculated, irregular, or poorly defined?
Q5 - Is the nodule calcified?
Q6 - Nodule Location
Q7 - Nodule Size

Table 5.7: Evaluation Consistency of Response LLMs Few-shot with 5 Examples Prompt Template 2

LLM	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Gemini 1.5 Flash	1.00	1.00	1.00	1.00	1.00	0.98	0.99
GPT-4o	1.00	1.00	0.98	1.00	0.98	0.99	0.99
Llama 3 70B	1.00	1.00	1.00	1.00	1.00	0.99	0.98

Q1 - Is the nodule solid or soft tissue?
Q2 - Is the nodule semisolid or subsolid?
Q3 - Is the nodule ground-glass?
Q4 - Is the nodule spiculated, irregular, or poorly defined?
Q5 - Is the nodule calcified?
Q6 - Nodule Location
Q7 - Nodule Size

for question 2 was zero for all models and for the 6 prompts used because examples with true answers in the 100 test reports were not provided.

Analyzing Table 5.9 with zero-shot prompt template 1 reveals a low F1-score for

Table 5.8: Evaluation Consistency of Response LLMs Few-shot with 10 Examples Prompt Template 2

LLM	Q1	Q2	Q3	Q4	Q5	Q6	Q7
Gemini 1.5 Flash	0.99	1.00	1.00	0.99	1.00	0.99	1.00
GPT-4o	0.99	1.00	1.00	0.99	0.96	0.98	0.99
Llama 3 70B	1.00	1.00	1.00	0.99	1.00	0.98	0.98

Q1 - Is the nodule solid or soft tissue?
Q2 - Is the nodule semisolid or subsolid?
Q3 - Is the nodule ground-glass?
Q4 - Is the nodule spiculated, irregular, or poorly defined?
Q5 - Is the nodule calcified?
Q6 - Nodule Location
Q7 - Nodule Size

Gemini 1.5 Flash on questions 3 and 5. GPT-4o has a low F1-score on questions 1, 3, 4, and 5. Llama 3 70B also presents a low F1-score on Questions 1, 3, 4, and 5. The same analysis in Table with zero-shot prompt template 2 indicates a low F1-score for Gemini 1.5 Flash on Questions 1 and 3. GPT-4o demonstrates a low F1-score on questions 1, 3, and 4. Llama 3 70B shows a low F1-score on questions 3 and 4.

Examination of the responses from Gemini 1.5 Flash and GPT-4o to Question 1 shows that the presence of a "calcified nodule" in the report results in a "solid soft tissue attenuation nodule" being marked as true, regardless of clearly described nodule attenuation.

Reviewing the responses from Gemini 1.5 Flash, GPT-4o, and Llama 3 70B to question 3 reveals that the occurrence of terms like "ground-glass opacities," "centrilobular ground-glass micronodules," "opacities with ground-glass attenuation," or "ground-glass lesions" leads to "is the nodule ground-glass" being marked as true, regardless of unambiguous nodule attenuation.

Inspection of the responses from Gemini 1.5 Flash, GPT-4o, and Llama 3 70B to Question 4 demonstrates that the presence of terms such as "microlobulated contours," "expansive formation with soft tissue density," "ill-defined hypoattenuating parenchymal images," "areas of focal and irregular pleural thickening," "lobulated contours," or "roughly triangular morphology" results in "is the nodule spiculated or irregular" being marked as true, irrespective of clear nodule border information.

Analysis of the responses from Gemini 1.5 Flash, GPT-4o, and Llama 3 70B to question 5 indicates that the occurrence of a "hyperdense oval image" does not result in "is the nodule calcified" being marked as true, thus failing to capture nodule calcification information.

Analyzing Table 5.10 with few-shot prompt template 1 with 5 examples reveals a low F1-score for Gemini 1.5 Flash on Questions 3 and 5. GPT-4o exhibits a low F1-score on Questions 1, 3, 4. Llama 3 70B also presents a low F1-score on questions 3, 4. The

Table 5.9: LLMs Effectiveness using Zero-shot Prompt 1 and Prompt 2

No.	Question	Gemini 1.5 Flash - P1			GPT-4o - P1			Llama 3 70B - P1		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid or soft tissue	0.68	1.00	**0.81**	0.23	1.00	0.37	0.57	1.00	0.72
2	Is the nodule semisolid or subsolid	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	Is the nodule ground glass	0.11	1,00	**0.21**	0.10	1.00	0.19	0.05	1.00	0.10
4	Is the nodule spiculated or irregular	0.80	0.80	**0.80**	0.42	1.00	0.59	0.42	1.00	0.59
5	Is the nodule calcified	0.94	0.57	0.71	0.94	0.58	0.72	0.93	0.63	**0.75**
6	Nodule location	0.89	1.00	0.94	0.92	1.00	**0.96**	0.89	1.00	0.94
7	Nodule size	0.88	1.00	0.93	0.91	1.00	**0.95**	0.89	1.00	0.94
No.	Question	Gemini 1.5 Flash - P2			GPT-4o - P2			Llama 3 70B - P2		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid or soft tissue	0.50	1.00	0.66	0.68	0.76	0.72	0.85	1.00	**0.92**
2	Is the nodule semisolid or subsolid	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	Is the nodule ground glass	0.25	1.00	0.40	0.40	1.00	**0.57**	0.11	1.00	0.20
4	Is the nodule spiculated or irregular	0.83	1.00	**0.91**	0.57	0.80	0.67	0.42	1.00	0.59
5	Is the nodule calcified	0.96	0.97	**0.97**	0.76	0.98	0.85	0.95	1.00	**0.97**
6	Nodule location	0.90	1.00	0.95	0.93	1.00	**0.96**	0.92	1.00	**0.96**
7	Nodule size	0.90	1.00	0.95	0.90	1.00	0.95	0.93	1.00	**0.96**

same analysis in Table with few-shot prompt template 2 with 5 examples indicates a low F1-score for Gemini 1.5 Flash on Questions 1 and 3. GPT-4o demonstrates a low F1-score on Questions 3 and 4. Llama 3 70B shows a low F1-score on Question 3.

A potential explanation for these outcomes aligns with the explanations provided for zero-shot learning prompt results. Specifically, the occurrence of certain terms seems to hinder the LLMs' accurate retrieval of pulmonary nodule information.

Table 5.10: LLMs Effectiveness using Few-shot with 5 Examples Prompt 1 and Prompt 2

No.	Question	Gemini 1.5 Flash - P1			GPT-4o - P1			Llama 3 70B - P1		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid or soft tissue	0.43	1.00	0.61	0.00	0.00	0.00	0.81	1.00	**0.89**
2	Is the nodule semisolid or subsolid	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	Is the nodule ground glass	0.12	1,00	0.22	0.15	1.00	0.27	0.20	1.00	**0.33**
4	Is the nodule spiculated or irregular	0.83	1.00	**0.91**	0.45	1.00	0.62	0.62	1.00	0.77
5	Is the nodule calcified	0.98	0.75	0.85	0.95	0.68	0.80	0.96	0.90	**0.93**
6	Nodule location	0.90	1.00	0.95	0.91	1.00	0.95	0.94	1.00	**0.97**
7	Nodule size	0.92	1.00	**0.96**	0.89	1.00	0.94	0.92	1.00	**0.96**
No.	Question	Gemini 1.5 Flash - P2			GPT-4o - P2			Llama 3 70B - P2		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid or soft tissue	0.57	1.00	0.72	0.80	0.94	0.86	0.77	1.00	**0.87**
2	Is the nodule semisolid or subsolid	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	Is the nodule ground glass	0.28	1.00	0.44	0.40	1.00	0.57	0.66	1.00	**0.80**
4	Is the nodule spiculated or irregular	1.00	1.00	1.00	0.67	0.80	**0.73**	0.57	0.80	0.67
5	Is the nodule calcified	0.98	0.97	**0.97**	0.96	0.82	0.88	0.97	0.98	**0.97**
6	Nodule location	0.92	1.00	0.96	0.98	1.00	**0.99**	0.96	1.00	0.98
7	Nodule size	0.92	1.00	0.96	0.93	1.00	0.96	0.94	1.00	**0.97**

Analyzing the Table 5.11 with few-shot prompt template 1 with 10 examples reveals a low F1-score for Gemini 1.5 Flash on Questions 1 and 3. GPT-4o exhibits a low F1-score on questions 3 and 4. Llama 3 70B also presents a low F1-score on Question 3. The same analysis in Table with few-shot prompt template 2 with 10 examples indicates a low F1-score for Gemini 1.5 Flash on Questions 1 and 3. GPT-4o demonstrates a low

F1-score on Question 3. Llama 3 70B shows a low F1-score on Question 3 and 4.

Table 5.11: LLMs Effectiveness using Few-shot with 10 Examples Prompt 1 and Prompt 2

No.	Question	Gemini 1.5 Flash - P1			GPT-4o - P1			Llama 3 70B - P1		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid or soft tissue	0.48	1.00	0.65	0.72	0.94	0.82	0.77	1.00	**0.87**
2	Is the nodule semisolid or subsolid	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	Is the nodule ground glass	0.13	1,00	0.23	0.18	1.00	0.31	0.33	1.00	**0.50**
4	Is the nodule spiculated or irregular	1.00	1.00	**1.00**	0.33	0.80	0.47	0.71	1.00	0.83
5	Is the nodule calcified	0.98	0.82	0.89	0.95	0.70	0.81	0.97	0.95	**0.96**
6	Nodule location	0.90	1.00	0.95	0.92	1.00	0.96	0.96	1.00	**0.98**
7	Nodule size	0.92	1.00	**0.96**	0.90	1.00	0.95	0.92	1.00	**0.96**
No.	Question	Gemini 1.5 Flash - P2			GPT-4o - P2			Llama 3 70B - P2		
		P	R	F1	P	R	F1	P	R	F1
1	Is the nodule solid or soft tissue	0.53	1.00	0.69	0.79	0.88	0.83	0.81	1.00	**0.89**
2	Is the nodule semisolid or subsolid	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	Is the nodule ground glass	0.28	1.00	0.44	0.67	1.00	0.80	0.62	1.00	**0.77**
4	Is the nodule spiculated or irregular	1.00	1.00	**1.00**	0.57	0.80	0.67	0.62	1.00	0.77
5	Is the nodule calcified	0.98	0.95	0.97	0.96	0.80	0.87	0.97	0.98	**0.97**
6	Nodule location	0.90	1.00	0.95	0.96	1.00	0.98	0.98	1.00	**0.99**
7	Nodule size	0.92	1.00	0.96	0.93	1.00	0.96	0.95	1.00	**0.97**

Using few-shot learning with prompt 2 and 10 examples, Gemini 1.5 Flash, GPT-4o, and Llama 3 70B achieved a high consistency of 0.99 in their answers to the table questions. These models also demonstrated the highest F1-scores, with Gemini 1.5 Flash scoring 0.83, GPT-4o scoring 0.85, and Llama 3 70B scoring 0.89. The answers of these models, generated using the specified few-shot learning approach, were then used to assign the Lung-RADS index.

Table 5.12 shows precision, recall, and F1-score for five evaluated models BiLSTM-CRF, BioBERTpt, Gemini 1.5 Flash, Llama 3 70B, and GPT-4o for Lung-RADS classification.

Table 5.12: Lung-RADS Classification Result

Lung-RADS	BiLSTM-CRF			BioBERTpt			Gemini 1.5 Flash			GPT-4o			Llama 3 70B			Nº Ex.
Category	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
1	1.00	0.93	0.97	1.00	0.95	0.97	1.00	0.95	0.97	0.98	0.80	0.88	0.98	0.98	**0.98**	61
2	0.85	0.92	0.88	0.85	0.88	0.86	0.96	0.88	**0.92**	0.70	0.84	0.76	1.00	0.80	0.89	25
3	0.33	0.50	0.40	0.50	1.00	**0.67**	0.50	1.00	**0.67**	0.40	1.00	0.57	0.33	0.50	0.40	2
4A	0.78	1.00	**0.88**	0.78	1.00	**0.88**	0.78	1.00	**0.88**	0.83	0.71	0.77	0.86	0.86	0.86	7
4X	1.00	0.80	0.89	1.00	0.40	0.57	1.00	1.00	**1.00**	0.57	0.80	0.67	0.57	0.80	0.67	5
weighted avg	0.93	0.92	0.92	0.94	0.91	0.91	0.96	0.94	**0.95**	0.87	0.81	0.83	0.95	0.91	0.92	

All models demonstrated high effectiveness in classifying **Lung-RADS 1** nodules. This likely stems from the prevalence of this classification in the training data and the clear textual descriptions of these nodules in the reports. Additionally, the absence of similar characteristics, such as "calcification," which are associated with other findings, may have contributed to this effectiveness.

Of the twenty five **Lung-RADS 2** cases, the BiLSTM-CRF model incorrectly classified one case as Lung-RADS 3 because it failed to identify the pulmonary nodule as

subpleural. It also misclassified two other cases as Lung-RADS 2. In one case, the model missed the nodule's calcification, which should have resulted in a Lung-RADS 1 classification. In the other, the model misidentified the nodule's size, leading to an incorrect Lung-RADS 2 assignment.

The BioBERTpt model misclassified two cases: one as Lung-RADS 3 due to a missed subpleural nodule and another as Lung-RADS 4B due to incorrect size extraction. Additionally, this model misclassified three cases as Lung-RADS 2. Two of these should have been Lung-RADS 4X because the model failed to identify irregular nodule borders, and one should have been Lung-RADS 1 due to a missed calcification. These errors likely occurred because the models, during the named entity recognition task, were unable to correctly identify the characteristics of the pulmonary nodules, and the rule-based algorithm misconverted the extracted information into a structured format.

Analyzing Gemini 1.5 Flash, GPT-4o, and Llama 3 70B revealed the following errors:

- **Gemini 1.5 Flash** made mistakes in two cases. In the first, the model incorrectly extracted the nodule size information, leading to an incorrect Lung-RADS 4B classification. In the second, it failed to identify that the pulmonary nodule was subpleural, assigning a Lung-RADS 3 classification. Additionally, the model misclassified one case as Lung-RADS 2 because it failed to extract the "calcification" characteristic of the pulmonary nodule.

- **GPT-4o** made mistakes in three cases. In one, it failed to identify the pulmonary nodule as subpleural, resulting in a Lung-RADS 3 classification. In the second, it incorrectly extracted the irregular border characteristic, leading to a Lung-RADS 4X classification. In the third, it did not extract any relevant information, leading to a Lung-RADS 0 classification. Additionally, this model misclassified eight cases as Lung-RADS 2. In all eight cases, it failed to extract the calcification characteristic, which should have led to a Lung-RADS 1 classification.

- **Llama 3 70B** made mistakes in four cases. In the first, it incorrectly extracted the nodule size, leading to a Lung-RADS 4B classification. In the second, it did not extract any information from the nodule, leading to a Lung-RADS 0 classification. In the third, it failed to identify the nodule as subpleural, assigning a Lung-RADS 3 classification. In the fourth, it incorrectly extracted the irregular border characteristic, leading to a Lung-RADS 4X classification.

Regarding the **Lung-RADS 3 classification** , three models achieved an F1-score below 0.60, with BioBERTpt and Gemini 1.5 Flash achieving the highest F1-score of 0.67. This was the Lung-RADS category where the models performed the worst. Due to the small number of test examples (only two), the evaluation of the models in this category was limited.

Of the two Lung-RADS 3 cases:

- **The BiLSTM-CRF model** correctly classified only one. In the second case, it failed to extract the nodule's characteristics, resulting in a Lung-RADS 2 classification. Additionally, it incorrectly classified one case as Lung-RADS 3 because it failed to extract the calcification information, whereas the nodule should have been classified as Lung-RADS 1.
- **The BioBERTpt model** correctly classified both test cases but incorrectly classified one additional case as Lung-RADS 3. In this case, the nodule was calcified, but the model failed to extract this information, meaning the correct classification should have been Lung-RADS 1.
- **The Gemini 1.5 Flash model** also correctly classified both test cases as Lung-RADS 3 but misclassified one additional case as Lung-RADS 3 due to failure to extract calcification information, when the correct classification should have been Lung-RADS 1.
- **The GPT-4o model** correctly classified both test cases as Lung-RADS 3 but incorrectly classified two additional cases as Lung-RADS 3. In these instances, it failed to extract the calcification information, resulting in misclassifications when the nodules should have been categorized as Lung-RADS 1.
- **The Llama 3 70B model** correctly classified one of the two Lung-RADS 3 cases. In the second case, it incorrectly extracted irregular border characteristics, leading to a Lung-RADS 4X classification. Additionally, it misclassified one case as Lung-RADS 3 due to failure to extract calcification information, when the correct classification should have been Lung-RADS 1.

For **Lung-RADS 4A classification**, four models achieved an F1-score above 0.85, while GPT-4o had the lowest F1-score 0.77. Among the seven test cases:

- **The BiLSTM-CRF, BioBERTpt, and Gemini 1.5 Flash models** correctly classified all cases as Lung-RADS 4A.
- **The GPT-4o model** correctly classified five cases but misclassified two. In both cases, the model incorrectly extracted the irregular border characteristic, leading to a Lung-RADS 4X classification. Furthermore, it mistakenly assigned a Lung-RADS 4X classification to one case where the nodule was subpleural, failing to correctly identify this characteristic and leading to an incorrect Lung-RADS 4A classification.
- **The Llama 3 70B model** correctly classified six out of seven test cases as Lung-RADS 4A. The misclassification occurred due to incorrect extraction of the irregular

border characteristic, leading to a Lung-RADS 4X classification. Furthermore, it incorrectly assigned a Lung-RADS 4A classification to a case where the nodule was subpleural, failing to identify this characteristic.

For the **Lung-RADS 4X classification** Gemini 1.5 Flash model achieved an excellent F1-score of 1.00, while the BiLSTM-CRF model obtained an F1-score of 0.89. The GPT-4o and Llama 3 70B models achieved an F1-score of 0.67, and the BioBERTpt model had the lowest F1-score at 0.57.

- Of the five cases classified as Lung-RADS 4X, the BioBERTpt model correctly identified only two. In the other three cases, the model failed to identify the irregular border characteristic and thus classified the nodule as Lung-RADS 4A. Additionally, in two cases, the rule-based algorithm failed to correctly convert the model's results into a structured format, leading to information loss and the incorrect classification of the nodules as Lung-RADS 2.

- The BiLSTM-CRF model correctly classified four out of the five Lung-RADS 4X cases. In the only misclassified case, the model failed to extract the irregular border characteristic, leading to a Lung-RADS 4A classification.

- The GPT-4o model correctly classified four out of five Lung-RADS 4X cases. The single misclassification occurred because the model failed to extract the calcification characteristic, resulting in a Lung-RADS 1 classification. However, this model also incorrectly assigned a Lung-RADS 4X classification in three cases. In the first two cases, the model incorrectly extracted the irregular border characteristic, and these nodules should have been classified as Lung-RADS 4A. In the last case, the model also incorrectly extracted the irregular border characteristic, and the nodule should have been classified as Lung-RADS 2.

- The Llama 3 70B model correctly classified four out of five Lung-RADS 4X cases. The single misclassification occurred because the model failed to extract the calcification characteristic, leading to a Lung-RADS 1 classification. However, this model also incorrectly assigned a Lung-RADS 4X classification in three cases. In the first case, the model incorrectly extracted the irregular border characteristic, and this nodule should have been classified as Lung-RADS 4A. In the second case, the model incorrectly extracted the irregular border characteristic, and this nodule should have been classified as Lung-RADS 3. In the third case, the model also incorrectly extracted the irregular border characteristic, and the nodule should have been classified as Lung-RADS 2.

Therefore, the results demonstrated that all 5 models were effective in classifying lung nodules in Portuguese chest CT reports to assist radiologists with Lung-RADS indexing. The findings highlight an pathway for developing more adaptable NLP system by

leveraging deep learning and LLMs, which has significant implications for Lung-RADS classification tasks.

Typically, developing a machine learning-based NER system tailored to specific clinical entities requires creating a highly annotated corpus, a resource-intensive process demanding time and expert input from the medical domain. In our cases only 100 reports were annoted with the Lung-RADS category.

Our study wasn't the first that utilizes NLP to assign the Lung-RADS index from CT reports. Beyer et al. [Beyer et al., 2017] previously developed a rule-based algorithm to assign the Lung-RADS classification of lung nodules identified within structured, clinical reports of consecutive CT lung screening examinations in English idiom. Their NLP algorithm achieved an overall sensitivity of 0.75 and an overall specificity of 0.98 in identifying nodules with the Lung-RADS 3 and Lung-RADS 4 index. Our approach differs slightly from that of Beyer et al. [Beyer et al., 2017] in that they utilized a database with 1501 structured reports, whereas we employed a database with 963 unstructured reports. Their method relied on a rule-based algorithm to extract nodule characteristics, while our approach leverages deep learning models and LLMs for characteristics extraction of lung nodules using NER and QA tasks.

Gandomi et al. [Gandomi et al., 2024] further explored NLP for Lung-RADS index extraction from CT reports. They developed a rule-based algorithm for extract the Lung-RADS categories described in the report and compared its performance to that of both radiologists and LCS specialists. Across four ground truth sets of CT reports, their approach demonstrated high precision (0.99) and recall (0.99). Our approach differed from that of Gandomi et al. in several key aspects. Firstly, they utilized a significantly larger database comprising 24,060 reports, whereas our database consisted of 963 reports, and the findings in our study may not generalize well to a broader population, as the dataset might not capture the full variability present in real-world clinical reports. Secondly, Gandomi's database already included the Lung-RADS index within the reports themselves. In contrast, our study assigned the Lung-RADS index based on the characteristics of pulmonary nodules as described within the reports. They compared the algorithm's assigned Lung-RADS with those assigned by both radiologists and LCS members.

To our knowledge, our study is the first to use LLMs to assign Lung-RADS scores from chest CT reports in Portuguese. The Portuguese language presents challenges due to its regional variations, intricate medical terminology, nuanced sentence structures, and comparatively limited data resources. Additionally, most LLMs have been pretrained on English data, necessitating careful consideration to ensure accurate effectiveness in this medical context. Previous studies in Portuguese have focused on extracting named entities from electronic health records [Schneider et al., 2020] [da Rocha et al., 2023] [Oliveira et al., 2022]. Our work contribute with the state-of-the-art by demonstrating the successful assignment of Lung-RADS classifications from chest CT reports in Por-

tuguese, utilizing deep learning models and LLMs to extract the lung nodule characteristics necessary for calculating the Lung-RADS index.

Despite the encouraging results this work have some limitation:

- A significant limitation of this work stems from the author's lack of medical expertise. The data annotations, crucial for the analysis, were validated by a single radiologist. While this review provides a degree of medical oversight, it is acknowledged that a more robust validation process involving multiple radiologists would have been preferable. The single-reviewer approach may not capture the full spectrum of radiological interpretations, potentially impacting the generalizability of the findings;

- It is also essential to acknowledge that the data utilized in this study was collected solely from one hospital. It restricts the demographic and clinical variability of the dataset. Consequently, the results may not be representative of diverse patient populations or healthcare practices found in other institutions;

- The composition of the test dataset for Lung-RADS classification presents a further limitation. With only 100 reports, the distribution of Lung-RADS categories is notably skewed. The absence of Lung-RADS 0 and 4B cases, coupled with the dominance of Lung-RADS 1 and the limited representation of Lung-RADS 3, 4A, and 4X, creates an imbalanced dataset. This imbalance may disproportionately influence the model's effectiveness, potentially leading to biased evaluations and hindering its ability to accurately classify less frequent, but clinically significant, categories. Consequently, the model's observed performance may not accurately reflect its real-world effectiveness across a more balanced and diverse patient population;

- Additionally, considerations around API costs, LLM usage, and data sensitivity are critical when deploying such models in clinical settings. A more comprehensive analysis of the resources and expenses associated with traditional NLP, word embedding models, and LLM-based systems will be valuable for future research, helping clarify the feasibility and practical implications of LLMs in clinical NER tasks, particularly in Lung-RADS classification for Portuguese-language radiology reports.

Table 5.13: Evaluation 27 BiLSTM-CRF Models in 10 Data Splits

	Hyperparameters			Macro F1-Score									
Model	Batch Size	Word Embedding Size	LSTM Units	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8	Split 9	Split 10
1	4	50	50	0.86	0.87	0.87	0.85	0.89	0.86	0.85	0.86	0.86	0.87
2	4	50	100	0.86	0.89	0.83	0.84	0.86	0.83	0.87	0.87	0.90	0.89
3	4	50	200	0.85	0.91	0.87	0.86	0.88	0.83	0.81	0.84	0.86	0.83
4	4	100	50	0.82	0.86	0.86	0.87	0.89	0.82	0.87	0.87	0.89	0.84
5	4	100	100	0.87	0.89	0.87	0.86	0.88	0.85	0.86	0.85	0.90	0.89
6	4	100	200	0.84	0.86	0.85	0.78	0.90	0.84	0.89	0.76	0.87	0.85
7	4	200	50	0.83	0.89	0.85	0.84	0.89	0.83	0.82	0.83	0.85	0.86
8	4	200	100	0.82	0.91	0.85	0.85	0.90	0.82	0.87	0.85	0.88	0.88
9	4	200	200	0.85	0.89	0.86	0.86	0.88	0.85	0.85	0.87	0.85	0.88
10	8	50	50	0.82	0.88	0.84	0.80	0.84	0.78	0.82	0.86	0.87	0.86
11	8	50	100	0.83	0.91	0.84	0.78	0.84	0.83	0.82	0.84	0.87	0.80
12	8	50	200	0.84	0.89	0.84	0.85	0.84	0.83	0.83	0.85	0.83	0.83
13	8	100	50	0.80	0.89	0.87	0.86	0.87	0.81	0.84	0.85	0.87	0.87
14	8	100	100	0.83	0.91	0.88	0.87	0.89	0.87	0.86	0.84	0.88	0.85
15	8	100	200	0.85	0.92	0.83	0.87	0.89	0.82	0.86	0.83	0.82	0.86
16	8	200	50	0.84	0.91	0.85	0.83	0.88	0.85	0.84	0.85	0.87	0.87
17	8	200	100	0.85	0.88	0.87	0.87	0.88	0.84	0.86	0.83	0.86	0.84
18	8	200	200	0.82	0.89	0.84	0.86	0.86	0.87	0.87	0.84	0.85	0.87
19	16	50	50	0.82	0.83	0.81	0.80	0.83	0.74	0.81	0.79	0.84	0.77
20	16	50	100	0.82	0.85	0.83	0.81	0.85	0.80	0.82	0.83	0.82	0.80
21	16	50	200	0.83	0.84	0.84	0.82	0.86	0.80	0.84	0.85	0.81	0.83
22	16	100	50	0.81	0.84	0.81	0.86	0.77	0.78	0.81	0.85	0.85	0.84
23	16	100	100	0.85	0.87	0.84	0.83	0.86	0.83	0.84	0.83	0.85	0.81
24	16	100	200	0.84	0.84	0.87	0.83	0.89	0.74	0.81	0.85	0.83	0.82
25	16	200	50	0.81	0.85	0.81	0.82	0.82	0.80	0.86	0.81	0.87	0.87
26	16	200	100	0.83	0.91	0.85	0.85	0.88	0.83	0.79	0.85	0.86	0.87
27	16	200	200	0.82	0.90	0.83	0.83	0.89	0.86	0.87	0.84	0.85	0.85

Chapter 6

Conclusion

In this study, we used deep learning models and LLMs to extract lung nodule characteristics from chest CT reports written in Portuguese to automate the Lung-RADS classification. We found that all five evaluated models(BiLSTM-CRF, BioBERTpt, Gemini 1.5 Flash, GPT-4o and Llama 3 70B were effective for classifying lung nodules in Portuguese chest CT reports to assist radiologists with Lung-RADS indexing. Our findings highlight apathway for developing a more adaptable NLP system by leveraging deep learning and LLMs models, which has significant implications for Lung-RADS classification tasks. To our knowledge, our study is the first to use LLMs to assign Lung-RADS scores from chest CT reports in the Portuguese idiom.

Our findings underscore the potential of Deep Learning and LLMs models to support radiologists in accurately categorizing lung nodules according to Lung-RADS criteria, thereby simplifying the diagnostic process. By automating and improving the information extraction, these models are anticipated to reduce radiologists' workload and enhance the consistency of follow-up recommendations. Ultimately, we hope this will benefit patients by facilitating more timely and accurate detection and management of lung cancer.

6.1 Future Work

In order to improve the Lung-RADS classification system of pulmonary nodules described in chest CT reports, future plans are as follows:

- We aim to incorporate datasets from multiple institutions representing diverse clinical practices and patient populations;
- We aim to expand this scope to include newer models, such as Llama 3.3, Gemini 2.0 Pro, Claude 3.5, and Deep Seek to gain deeper insights into their effectiveness;
- Use different prompt techniques such as chain-of-thought and retrieval augmented generation to improve information extraction from pulmonary nodules described in

chest CT reports.

6.2 Scientific Contributions

The scientific works published with Qualis CAPES were:

- Results accepted in: 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE), 2023, Dayton. 2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE), Qualis A3, with the title: "Lung-RADS + AI: A Tool for Quantifying the Risk of Lung Cancer in Computed Tomography Reports".

- Results accepted in: 24º SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE, 2024, Goiânia. 2024: Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde, 2024. Qualis A4, with the title: "LungRads+AI: Automatização do Índice Lung-RADS em Laudos de TC de Tórax".

- Results accepted in: 2024 IEEE 24th International Conference on Bioinformatics and Bioengineering (BIBE), 2024, Kragujevac. 2024 IEEE 24th International Conference on Bioinformatics and Bioengineering (BIBE), 2024. Qualis A3, with the title: "Comparative Study of Large Language Models for Lung-RADS Classification in Portuguese CT Reports".

Appendices

Zero-shot Prompt Templates

Zero-Shot Prompt Template 1

Please extract relevant structured information from the above report:
"Question" : "Answer"
{
"Is the nodule solid or soft tissue?" : "",
"Is the nodule semisolid or subsolid?" : "",
"Is the nodule ground-glass?" : "",
"Is the nodule spiculated, irregular or poorly defined?" : "",
"Is the nodule calcified?" : "",
"Nodule location" : "",
"Nodule size" : ""
}

If the report does not contain relevant information related to a specific question, please fill in the answer to that question with "No". The nodule size question should be answered with numbers and unit of measure only.

Figure 6.1: Zero-shot Prompt 1 Template.

Zero-Shot Prompt Template 2

Please extract relevant structured information from the report above:
"Question" : "Answer"
{
"Is the nodule solid or soft tissue?" : "",
"Is the nodule semisolid or subsolid?" : "",
"Is the nodule ground-glass?" : "",
"Is the nodule spiculated, irregular or poorly defined?" : "",
"Is the nodule calcified?" : "",
"Nodule location" : "",
"Nodule size" : ""
}

The following are some requirements for extraction:

1. Please extract structured information for the pulmonary nodule mentioned in the report to fill in the table. In this process, you must disregard all findings described in the report except for: nodules, hyperdense oval image or hyperattenuating oval image.
2. If the report does not contain relevant information related to a specific question, please fill in the answer to that question with "No". The nodule size question should be answered with numbers and unit of measure only.

Here are some points of prior medical knowledge for your reference

1. Hyperdense oval image should be considered as a calcified pulmonary nodule.
2. Solid, soft parts and ground glass are mutually exclusive. Only one of the three questions can be "Yes", and the mixed ground glass opacity means that the tumor has components of solid and ground glass opacity.
3. Micronodule is a nodule in the lung less than 3 millimeters (mm) in diameter. In this context due to its small size we are not interested in extracting its characteristics. Therefore, its characteristics should not be extracted.

Figure 6.2: Zero-shot Prompt 2 Template.

Few-shot Prompt Templates

Few-Shot Prompt Template 1

Please extract relevant structured information from the above report:
"Question" : "Answer"
{
"Is the nodule solid or soft tissue?" : "",
"Is the nodule semisolid or subsolid?" : "",
"Is the nodule ground-glass?" : "",
"Is the nodule spiculated, irregular or poorly defined?" : "",
"Is the nodule calcified?" : "",
"Nodule location" : "",
"Nodule size" : ""
}

If the report does not contain relevant information related to a specific question, please fill in the answer to that question with "No". The nodule size question should be answered with numbers and unit of measure only.

The example report: ...

The example report with the table filled in: ...

Figure 6.3: Few-shot Prompt 1 Template.

Few-Shot Prompt Template 2

Please extract relevant structured information from the report above:
"Question" : "Answer"
{
"Is the nodule solid or soft tissue?" : "",
"Is the nodule semisolid or subsolid?" : "",
"Is the nodule ground-glass?" : "",
"Is the nodule spiculated, irregular or poorly defined?" : "",
"Is the nodule calcified?" : "",
"Nodule location" : "",
"Nodule size" : ""
}

The following are some requirements for extraction:

1. Please extract structured information for the pulmonary nodule mentioned in the report to fill in the table. In this process, you must disregard all findings described in the report except for: nodules, hyperdense oval image or hyperattenuating oval image.
2. If the report does not contain relevant information related to a specific question, please fill in the answer to that question with "No". The nodule size question should be answered with numbers and unit of measure only.

Here are some points of prior medical knowledge for your reference

1. Hyperdense oval image should be considered as a calcified pulmonary nodule.
2. Solid, soft parts and ground glass are mutually exclusive. Only one of the three questions can be "Yes", and the mixed ground glass opacity means that the tumor has components of solid and ground glass opacity.
3. Micronodule is a nodule in the lung less than 3 millimeters (mm) in diameter. In this context due to its small size we are not interested in extracting its characteristics. Therefore, its characteristics should not be extracted.

The example report: ...

The example report with the table filled in: ...

Figure 6.4: Few-shot Prompt 2 Template.

All the code used in this work is available in the Github repository: `https://github.com/tarcisiolf/Lung_RADS_Automation.git`

References

[ACR, 2022] ACR (2022). Lung-rads® v2022. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads. Accessed: 2023-05-01.

[Ahmad et al., 2025] Ahmad, S. Q., Pesola, F., Crosbie, P. A., Gabe, R., Hancock, N., Kennedy, M. P., Marshall, C., Quaife, S. L., Rogerson, S., Simmonds, I., and Callister, M. E. (2025). Adherence to community-based lung cancer screening in the yorkshire lung screening trial. *Lung Cancer*, 200:108086.

[Ainslie et al., 2023] Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. (2023). Gqa: Training generalized multi-query transformer models from multi-head checkpoints.

[Alsentzer et al., 2019] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. B. A. (2019). Publicly available clinical bert embeddings.

[Arlene et al., 2021] Arlene, C., Emma, D., Michael, P., Hang, D., Daniel, D., Andreas, G., Claire, G., Víctor, S.-P., Richard, T., William, W., Honghan, W., and Beatrice, A. (2021). A systematic review of natural language processing applied to radiology reports. *BMC Medical Informatics and Decision Making*, 21(1):179.

[Bankier et al., 2024] Bankier, A. A., MacMahon, H., Colby, T., Gevenois, P. A., Goo, J. M., Leung, A. N., Lynch, D. A., Schaefer-Prokop, C. M., Tomiyama, N., Travis, W. D., Verschakelen, J. A., White, C. S., and Naidich, D. P. (2024). Fleischner society: Glossary of terms for thoracic imaging. *Radiology*, 310(2):e232558. PMID: 38411514.

[Bedi et al., 2024] Bedi, S., Jain, S. S., and Shah, N. H. (2024). Evaluating the clinical benefits of llms. *Nature Medicine*.

[Beyer et al., 2017] Beyer, S. E., McKee, B. J., Regis, S. M., McKee, A. B., Flacke, S., El Saadawi, G., and Wald, C. (2017). Automatic Lung-RADS™ classification with a natural language processing system. *J Thorac Dis*, 9(9):3114–3122.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

[Bushberg et al., 2012] Bushberg, J. T., Seibert, A. J., Leidholdt, E. M., and Boone, J. M. (2012). *The essential physics of medical imaging; 3rd ed.* Lippincott Williams & Wilkins, Philadelphia, PA.

[Carrodeguas et al., 2019] Carrodeguas, E., Lacson, R., Swanson, W., and Khorasani, R. (2019). Use of machine learning to identify follow-up recommendations in radiology reports. *Journal of the American College of Radiology*, 16(3):336–343.

[Chen et al., 2024] Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2024). Unleashing the potential of prompt engineering in large language models: a comprehensive review.

[Chowdhery et al., 2022] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., and Gehrmann, S. (2022). Palm: Scaling language modeling with pathways.

[Christensen et al., 2024] Christensen, J., Prosper, A. E., Wu, C. C., Chung, J., Lee, E., Elicker, B., Hunsaker, A. R., Petranovic, M., Sandler, K. L., Stiles, B., Mazzone, P., Yankelevitz, D., Aberle, D., Chiles, C., and Kazerooni, E. (2024). Acr lung-rads v2022: Assessment categories and management recommendations. *Journal of the American College of Radiology*, 21(3):473–488.

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

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

[Demšar, 2006] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Ewoud et al., 2016] Ewoud, P., MM, B. L., Myriam, H. M., and A, K. J. (2016). Processamento de linguagem natural em radiologia: uma revisão sistemática. *Radiology*, 279(2):329–343.

[Fedus et al., 2022] Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

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

[Gandomi et al., 2024] Gandomi, A., Hasan, E., Chusid, J., Paul, S., Inra, M., Makhnevich, A., Raoof, S., Silvestri, G., Bade, B. C., and Cohen, S. L. (2024). Evaluating the accuracy of lung-rads score extraction from radiology reports: Manual entry versus natural language processing. *International Journal of Medical Informatics*, 191:105580.

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

[Goyal et al., 2018] Goyal, A., Gupta, V., and Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, 29:21–43.

[Grattafiori et al., 2024] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., and Vaughan, A. (2024). The llama 3 herd of models.

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

[Gu et al., 2021] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

[Hansell et al., 2008] Hansell, D. M., Bankier, A. A., MacMahon, H., McLoud, T. C., Müller, N. L., and Remy, J. (2008). Fleischner society: glossary of terms for thoracic imaging. *Radiology*, 246(3):697–722.

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

[Hu et al., 2024a] Hu, D., Liu, B., Zhu, X., Lu, X., and Wu, N. (2024a). Zero-shot information extraction from radiological reports using chatgpt. *International Journal of Medical Informatics*, 183:105321.

[Hu et al., 2024b] Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., Roberts, K., and Xu, H. (2024b). Improving large language models for clinical named entity recognition via prompt engineering.

[Huang et al., 2015] Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging.

[Jang et al., 2020] Jang, B., Kim, M., Harerimana, G., Kang, S.-u., and Kim, J. W. (2020). Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism. *Applied Sciences*, 10(17).

[Khurana et al., 2022] Khurana, D., Koli, A., Khatter, K., and Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744.

[Lee et al., 2019] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

[Linna and Kahn, 2022] Linna, N. and Kahn, C. E. (2022). Applications of natural language processing in radiology: A systematic review. *International Journal of Medical Informatics*, 163:104779.

[Lopes et al., 2019] Lopes, F., Teixeira, C., and Gonçalo Oliveira, H. (2019). Contributions to clinical named entity recognition in Portuguese. In Demner-Fushman, D., Cohen, K. B., Ananiadou, S., and Tsujii, J., editors, *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233, Florence, Italy. Association for Computational Linguistics.

[Luo and Chong, 2020] Luo, J. W. and Chong, J. J. (2020). Review of natural language processing in radiology. *Neuroimaging Clinics of North America*, 30(4):447–458. Machine Learning and Other Artificial Intelligence Applications.

[MacMahon et al., 2017] MacMahon, H., Naidich, D. P., Goo, J. M., Lee, K. S., Leung, A. N. C., Mayo, J. R., Mehta, A. C., Ohno, Y., Powell, C. A., Prokop, M., Rubin, G. D., Schaefer-Prokop, C. M., Travis, W. D., Van Schil, P. E., and Bankier, A. A. (2017). Guidelines for management of incidental pulmonary nodules detected on ct images: From the fleischner society 2017. *Radiology*, 284(1):228–243. PMID: 28240562.

[Mathias et al., 2020] Mathias, C., Prado, G. F., Mascarenhas, E., Ugalde, P. A., Zimmer Gelatti, A. C., Carvalho, E. S., Faroni, L. D., Oliveira, R., Cordeiro de Lima, V. C., and de Castro Jr., G. (2020). Lung cancer in brazil. *Journal of Thoracic Oncology*, 15(2):170–175.

[Minaee et al., 2024] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey.

[Nakayama et al., 2018] Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

[National Lung Screening Trial Research Team, 2019] 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] Nobel et al., J. M. (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.

[Nori et al., 2023] Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., and Horvitz, E. (2023). Can generalist foundation models outcompete special-purpose tuning? case study in medicine.

[of Biomedical Imaging and (NIBIB), 2024] of Biomedical Imaging, N. I. and (NIBIB), B. (2024). Computed tomography (ct). https://www.nibib.nih.gov/science-education/science-topics/computed-tomography-ct. Accessed: 2024-10-01.

[Oliveira et al., 2022] Oliveira, L. E. S. e., Peters, A. C., da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., and Moro, C. M. C. (2022). Semclinbr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1).

[OpenAI et al., 2024] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., and Altenschmidt, J. (2024). Gpt-4 technical report.

[Pandey et al., 2020] Pandey, M., Xu, Z., Sholle, E., Maliakal, G., Singh, G., Fatima, Z., Larine, D., Lee, B. C., Wang, J., van Rosendael, A. R., Baskaran, L., Shaw, L. J., Min, J. K., and Al'Aref, S. J. (2020). Extraction of radiographic findings from unstructured thoracoabdominal computed tomography reports using convolutional neural network based natural language processing. *PLoS One*, 15(7):e0236827.

[Polanco et al., 2024] Polanco, D., González, J., Gracia-Lavedan, E., Pinilla, L., Plana, R., Molina, M., Pardina, M., and Barbé, F. (2024). Multidisciplinary virtual management of pulmonary nodules. *Pulmonology*, 30(3):239–246.

[Quinta de Castro et al., 2018] Quinta de Castro, P. V., Félix Felipe da Silva, N., and da Silva Soares, A. (2018). Portuguese named entity recognition using lstm-crf. In Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., and Paetzold, G. H., editors, *Computational Processing of the Portuguese Language*, pages 83–92, Cham. Springer International Publishing.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[Rainio et al., 2024] Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

[Rajpurkar et al., 2016] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text.

[Sahoo et al., 2024] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications.

[Saxena et al., 2024] Saxena, Y., Chopra, S., and Tripathi, A. M. (2024). Evaluating consistency and reasoning capabilities of large language models.

[Schneider et al., 2020] Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. M. (2020). BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T., editors, *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

[Siegel et al., 2023] Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. *CA Cancer J Clin*, 73(1):17–48.

[Singhal et al., 2022] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., y Arcas, B. A., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. (2022). Large language models encode clinical knowledge.

[Singhal et al., 2023] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., y Arcas, B. A., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S. S., Barral, J., Webster, D., Corrado, G. S., Matias, Y., Azizi, S., Karthikesalingam, A., and Natarajan, V. (2023). Towards expert-level medical question answering with large language models.

[Sugimoto et al., 2021] Sugimoto, K., Takeda, T., Oh, J.-H., Wada, S., Konishi, S., Yamahata, A., Manabe, S., Tomiyama, N., Matsunaga, T., Nakanishi, K., and Matsumura, Y. (2021). Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116:103729.

[Sung et al., 2021] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

[Sánchez et al., 2018] Sánchez, M., Benegas, M., and Vollmer, I. (2018). Management of incidental lung nodules. *Journal of Thoracic Disease*, 10(Suppl 22).

[Team et al., 2024a] Team, G., Anil, R., Borgeaud, S., and Alayrac, J.-B. (2024a). Gemini: A family of highly capable multimodal models.

[Team et al., 2024b] Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., and Wang, S. (2024b). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

[Tho, 2022] Tho, Q. (2022). N/a modern approaches in natural language processing. *VNU Journal of Science: Computer Science and Communication Engineering*, 39(1).

[Touvron et al., 2023a] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). Llama: Open and efficient foundation language models.

[Touvron et al., 2023b] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E.,

Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023b). Llama 2: Open foundation and fine-tuned chat models.

[Vaswani et al., 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

[Wang et al., 2018] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., and Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49.

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

[Wu et al., 2023] Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. (2023). Pmc-llama: Towards building open-source language models for medicine.

[Xiong et al., 2023] Xiong, W., Liu, J., Molybog, I., Zhang, H., Bhargava, P., Hou, R., Martin, L., Rungta, R., Sankararaman, K. A., Oguz, B., Khabsa, M., Fang, H., Mehdad, Y., Narang, S., Malik, K., Fan, A., Bhosale, S., Edunov, S., Lewis, M., Wang, S., and Ma, H. (2023). Effective long-context scaling of foundation models.

[Zech et al., 2018] Zech, J., Pain, M., Titano, J., Badgeley, M., Schefflein, J., Su, A., Costa, A., Bederson, J., Lehar, J., and Oermann, E. K. (2018). Natural language-based machine learning models for the annotation of clinical radiology reports. *Radiology*, 287(2):570–580.

[Zhang et al., 2018] Zhang, K., Ren, W., and Zhang, Y. (2018). Attention-based bi-lstm for chinese named entity recognition. In Hong, J.-F., Su, Q., and Wu, J.-S., editors, *Chinese Lexical Semantics*, pages 643–652, Cham. Springer International Publishing.

[Zheng et al., 2021] Zheng, C., Huang, B. Z., Agazaryan, A. A., Creekmur, B., Osuj, T. A., and Gould, M. K. (2021). Natural language processing to identify pulmonary nodules and extract nodule characteristics from radiology reports. *Chest*, 160(5):1902–1914.

[Zheng et al., 2015] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. S. (2015). Conditional random fields as recurrent neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE.