LungRads+AI: Automatização do Índice Lung-RADS em Laudos de TC de Tórax

Tarcísio Lima Ferreira[1], **Marcelo Costa Oliveira**[1], **Thales Miranda de Almeida Vieira**[1]

[1]Instituto de Computação – Universidade Federal de Alagoas (UFAL)
Caixa Postal – 57072-900 – Maceió – AL – Brasil

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

Abstract. *Lung cancer is the second most commonly diagnosed cancer globally. It represents the deadliest form of malignant neoplasm, resulting in around 1.8 million fatalities in 2020. The Lung-RADS is a guideline used for screening and follow-up of suspected Lung lesions. In this context, the main objective of this work is to assess the effectiveness of three Named Entity Recognition techniques, CNN, BiSTM, and BERT, to extract characteristics of pulmonary nodules in chest CT reports and calculate the probability of malignancy index using the Lung-RADS guideline. Our top-performing model was BiLSTM-CRF, and this model achieved a precision of 96%, a recall of 88%, and a F1-score of 90%.*

Resumo. *O câncer do pulmão é o segundo câncer mais frequentemente diagnosticado. Representa a forma mais mortal de neoplasia maligna, resultando em cerca de 1,8 milhão de mortes em 2020. O Lung-RADS é uma diretriz utilizada para o rastreio e o acompanhamento de lesões pulmonares suspeitas. Neste contexto, o principal objetivo deste trabalho é avaliar a eficácia de três técnicas de Reconhecimento de Entidades Nomeadas, CNN, BiLSTM e BERT, para extrair características de nódulos pulmonares em relatórios de TC de tórax e calcular o índice de probabilidade de malignidade usando a diretriz Lung-RADS. O nosso modelo com melhor desempenho foi o BiLSTM-CRF, que obteve uma precisão de 96%, uma revocação de 88% e um F1-score de 90%.*

1. Introdução

O câncer é um problema crítico de saúde pública com grandes consequências em todo mundo. Segundo projeções da Sociedade Americana de Câncer, o câncer foi responsável por milhares de óbitos somente nos Estados Unidos em 2023 [Siegel and Miller 2023]. O câncer do pulmão é o segundo tipo câncer mais frequentemente diagnosticado em todo o mundo e, infelizmente, é a forma mais letal de neoplasia maligna, responsável por cerca de 1,8 milhão de mortes ao ano [Sung and Ferlay 2021]. A atual taxa de sobrevivência de 5 anos na fase inicial pode atingir 59,8%, mas diminui para 5% quando a doença atinge a sua fase avançada. Portanto, a detecção e o diagnóstico precoce do câncer do pulmão são essenciais para garantir um melhor tratamento e aumentar as chances de sobrevida [Vykoukal et al. 2022].

A utilização de diretrizes em programas de rastreio tem uma importância significativa, pois, visa minimizar a necessidade de exames de acompanhamento excessivos, oferecendo uma melhor orientação aos radiologistas, clínicos e pacientes

[MacMahon and Naidich 2017]. Várias sociedades profissionais, incluindo *American College of Chest Physicians* e a *Fleischner Society*, publicaram diretrizes para padronizar a interpretação e o relato de achados em exames de imagem pulmonar. Essas diretrizes oferecem aos profissionais de saúde orientações fundamentadas em evidências científicas, delineando abordagens de diagnóstico e tratamento apropriadas para pacientes que apresentam nódulos suspeitos ou já confirmados [MacMahon and Naidich 2017] [Gould et al. 2013] [Pinsky et al. 2015].

O *Lung-RADS* foi desenvolvido pelo *American College of Radiology* e serve como instrumento de garantia de qualidade para normalizar a comunicação dos resultados tomografia computadorizada (TC) no rastreio do câncer do pulmão e fornecer recomendações de gestão ao especialista. O seu objetivo é minimizar a ambiguidade na interpretação dos exames de TC de rastreio do câncer do pulmão e simplificar o monitoramento dos resultados dos pacientes [Lun 2023].

No entanto, apesar dos avanços na padronização da interpretação de exames de imagem, ainda existem lacunas significativas no processo de análise de laudos de TC. A natureza não estruturada ou semi-estruturada desses documentos dificulta a extração eficiente de informações relevantes, representando um desafio para a implementação de práticas automatizadas de diagnóstico e monitoramento [Fei et al. 2022] [Kang et al. 2019]. A extração manual dessas informações de forma estruturada é uma tarefa demorada, trabalhosa, sujeita a erros e cara [Wang et al. 2018].

Nesse sentido, a aplicação do Processamento de Linguagem Natural (PLN) nos laudos médicos surge como uma oportunidade valiosa para automatizar a classificação do *Lung-RADS* em textos não estruturados, fornecendo suporte em tempo real aos radiologistas ao sugerir a categoria apropriada do *Lung-RADS* e identificando laudos com informações insuficientes para calcular o índice de malignidade do *Lung-RADS* [Beyer et al. 2017]. Dessa forma, o PLN pode ter um papel essencial na redução de erros cometidos pelo especialista, minimizando a ocorrência de falsos positivos e falsos negativos [Mendoza et al. 2022].

Pesquisas anteriores mostraram que o uso de dados estruturados e não estruturados combinado com IA levaram à identificação de diversas informações importantes em textos clínicos, como a identificação de condições clínicas, sintomas, diagnósticos, medicamentos, exames e tratamentos [da Rocha et al. 2023] [Lopes et al. 2019].

Diante deste contexto, o objetivo principal deste trabalho é avaliar a eficácia de três técnicas de *Named Entity Recognition* (NER), incluindo as variantes CNN, BiLSTM e BERT, para extrair características de nódulos pulmonares em laudos médicos na língua portuguesa. Como objetivo secundário, desenvolvemos a ferramenta de inteligência artificial LungRads+AI capaz de calcular o índice *Lung-RADS* de maneira automática a partir do laudo da TC de tórax.

O restante deste artigo está organizado da seguinte forma. Na Seção 2 apresentamos a revisão da literatura, explorando trabalhos relacionados que contextualizam e fundamentam a pesquisa. Na Seção 3 apresentamos a metodologia do trabalho, começando pela descrição das categorias do Lung-RADS (seção 3.1), apresentação do conjunto de dados (Seção 3.2), seguido pelas etapas de pré-processamento dos dados (Seção 3.3). O reconhecimento de entidade nomeada é discutido na Seção 3.4, enquanto a classificação

de entidade nomeada é abordada na Seção 3.5. Além disso, as métricas selecionadas para avaliar o desempenho do trabalho são apresentadas na Seção 3.6. A Seção 4 dedica-se à apresentação dos experimentos realizados e os resultados obtidos. Finalmente, a Seção 5 abrange as conclusões derivadas do estudo, destacando também as limitações identificadas durante a pesquisa.

2. Trabalhos Relacionados

Utilizando PLN baseado em dicionário e correspondência de padrões, os autores [Gershanik et al. 2011] desenvolveram o *iSCOUT*. Esse aplicativo foi capaz de recuperar documentos e avaliar a consistência das detecções de nódulos pulmonares. Em um estudo conduzido por [Nobel et al. 2020], os autores utilizaram um modelo de PLN baseado em regras com etapas de pré-processamento de aprendizado de máquina para a classificação automática do estágio T de nódulos pulmonares em laudos radiológicos no idioma holandês. Utilizando um banco de dados com 147 laudos, o algoritmo alcançou uma precisão de 83% no conjunto de treinamento e 87% no conjunto de validação. [Zheng and Z. 2021] apresentaram um algoritmo de PLN para identificar a descrição de nódulos pulmonares em laudos médicos. O algoritmo identificou várias características associadas aos nódulos, incluindo lateralidade, lóbulo, atenuação, calcificação e borda. O trabalho utilizou 354.773 laudos de TC, e obteve sensibilidade de 98,6% e especificidade de 100% na identificação de nódulos pulmonares. Além disso, na identificação das características dos nódulos, o algoritmo PLN alcançou uma sensibilidade média de 95,1% e uma especificidade média de 98,38%.

Em trabalhos recentes, os sistemas de PLN baseados em aprendizado profundo ganharam destaque. Um método inteligente de gerenciamento de qualidade de relatórios radiológicos no idioma chinês foi desenvolvido por [Fei et al. 2022]. Os autores empregaram um modelo baseado na arquitetura BiLSTM + CRF para extrair atributos como: localização, forma, tipo do nódulo, solidez, quantidade, nível de risco, tamanho e recomendação de acompanhamento. Um conjunto de dados de 48.091 relatórios do hospital Terciário de Pequim foi utilizado no estudo. O modelo BiLSTM + CRF obteve precisão de 94,56%, revocação de 93,96% e *F1-score* de 93,07%. Nesta pesquisa foram identificadas e analisadas determinadas entidades, como *qualidade de gestão* e *nível de risco*.

No trabalho conduzido por [da Rocha et al. 2023], os autores propuseram um modelo de Rede Neural Convolucional (CNN) que foi treinado com dados não estruturados de registos médicos no idioma português para identificar sete entidades: sintomas, diagnósticos, medicamentos, condições, exames e tratamento. Dos 30.000 prontuários disponíveis para o estudo, 1.200 foram utilizados na construção do corpus. O modelo CNN foi utilizado para extração de entidades nomeadas e obteve precisão de 72,72%, revocação de 56,93% e *F1-score* de 63,87%.

Trabalhos recentes de PLN têm usado modelos baseados na arquitetura *transformers* [Patwardhan et al. 2023], como o BERT [Devlin et al. 2018]. Alguns trabalhos compararam o desempenho de modelos BERT e outros modelos de aprendizado profundo, como BiLSTM, para extrair termos clínicos de relatórios radiológicos [Sugimoto et al. 2021] [Liu et al. 2021].

Em [Sugimoto et al. 2021], foi avaliado o desempenho de três modelos (BiLSTM-

CRF, BERT e BERT-CRF) para a extração de sete termos clínicos de laudos de TC de tórax. Os pesquisadores utilizaram 118.078 relatórios de TC de tórax do Hospital Universitário de Osaka. Além dos relatórios, os modelos foram pré-treinados usando a Wikipedia. Dentre os modelos avaliados, o BiLSTM-CRF obteve o melhor desempenho utilizando *word embeddings* pré-treinadas com laudos de TC de tórax. O modelo obteve precisão média de 94,00%, revocação média de 94,42% e *F1-score* média de 94,19%.

3. Materiais e Métodos

Neste trabalho, usamos as bibliotecas Keras (v.2.12.0), Tensorflow (v.2.12.0) e Tensorflow Addons (v.0.20.0) para implementar o modelo Bi-LSTM e o campo aleatório condicional de cadeia linear (camada CRF). Usamos a biblioteca Pytorch (v.2.0.1) para realizar o *fine-tuning* do BERT Multilingual e aplicamos SpaCy (v.3.5.3) para implementar a o modelo baseado na arquitetura CNN. A Figura 1 apresenta a visão geral dos passos metodológicos utilizados neste trabalho que serão descritos em detalhes nas seções subsequentes.

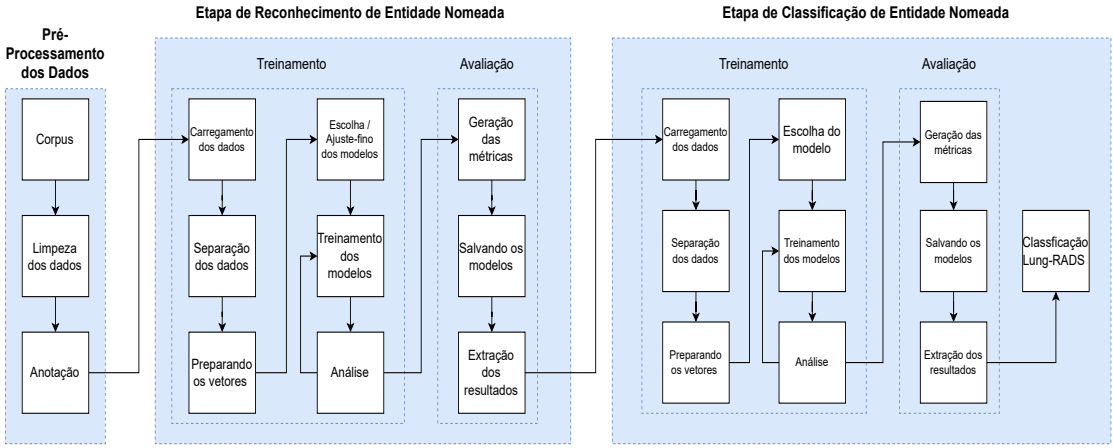

Figura 1. Esquema geral da metodologia.

3.1. *Lung-RADS*

Na classificação *Lung-RADS*, a Categoria 0 refere-se a avaliações incompletas, em que os pulmões não podem ser totalmente avaliados, exigindo exames de TC do tórax adicionais. A Categoria 1 indica um resultado negativo de rastreamento, sem nódulos pulmonares identificados, e requer um acompanhamento de 12 meses. Nódulos benignos, determinados por características específicas encontradas no exame de imagem, são categorizados como Categoria 2, com subcategorias que são especificadas de acordo com tamanho e características dos nódulos pulmonares. A Categoria 3 representa lesões provavelmente benignas, exigindo um acompanhamento de 6 meses para avaliação de estabilidade. As Categorias 4A e 4B estão associadas a nódulos suspeitos e muito suspeitos, respectivamente, cada uma com protocolos específicos de acompanhamento, incluindo intervalos mais curtos (3 meses) para o exame de TC e considerando modalidades de exames de imagem adicionais, como PET/CT. A Categoria 4X, abrange nódulos com características adicionais que aumentam a suspeita de câncer de pulmão. O *Lung-RADS* também incorpora um modificador ”S”para acomodar descobertas clinicamente significativas não relacionadas ao câncer de pulmão.

3.2. Conjunto de Dados

O conjunto de dados utilizado contém 963 laudos de TC de tórax no idioma português coletados entre 01/02/2022 a 03/04/2023 no Hospital Universitário da Universidade Federal de Alagoas. Diante da assinatura do Termo de Consentimento, os laudos foram obtidos de pacientes submetidos a TC de tórax para qualquer indicação. É importante ressaltar que todos os dados do paciente foram anonimizados.

Os 963 textos dos laudos foram divididos em proporções específicas: 90% foram utilizados para o treinamento e os 10% restantes para os testes. Ao analisar os 96 laudos utilizados na etapa de teste, observamos a presença de nódulos pulmonares em 79 laudos distribuídos nas categorias *Lung-RADS* da seguinte forma: **Categoria 0** (22 casos), **Categoria 1** (42 casos), **Categoria 2** (8 casos), **Categoria 3** (2 casos), **Categoria 4A** (1 caso), **Categoria 4B** (3 casos), e **Categoria 4X** (1 caso).

3.3. Pré-processamento dos Dados

A limpeza de dados foi realizada em todos os 963 laudos radiológicos com os seguintes passos: primeiro, os caracteres especiais foram removidos dos relatórios, então, espaços em branco foram inseridos entre palavras e sinais de pontuação, e cada relato foi segmentado em uma única frase.

Para o esquema de anotação adotamos o formato *Inside-Outside-Beginning* (IOB). Este esquema permite identificar e delimitar corretamente as entidades de um texto, facilitando o processamento e análise de informações específicas.

- O prefixo I indica que um rótulo pertence à uma entidade;
- O prefixo O indica que o *token* não pertence à nenhuma entidade;
- O prefixo B indica que o rótulo está no início de uma entidade;

Cada *token* é anotado individualmente, resultando em uma sequência de rótulos.

3.4. Anotação de Dados de Treinamento

Cada laudo foi carregado na ferramenta de anotação *Doccano* [Nakayama et al. 2018] na qual foi feita a rotulação do texto usando seis entidades nomeadas (ENs), que correspondem às características dos nódulos pulmonares. As ENs utilizadas foram: Atenuação, Calcificação, Bordas, Achado, Localização e Tamanho. Essas características foram escolhidas com base nas diretrizes do *Lung-RADS* [Lun 2023].

Como resultado, foi gerado um arquivo contendo as informações de rotulação de todos os laudos. Este arquivo contém o texto do laudo, as ENs rotuladas no texto e a posição inicial e final das ENs na sequência de texto do laudo.

3.5. Etapa de Treinamento para Reconhecimento de Entidade Nomeada

Cada laudo foi *tokenizado* e convertido em uma sequência de inteiros conforme o dicionário onde cada *tokens* do corpus corresponde a um número inteiro. Finalmente, cada sequência de números inteiros que representa um relatório foi preenchida até um tamanho fixo, uma vez que modelos como o BERT requerem um comprimento de sequência de entrada específico. O tamanho máximo das listas contendo os laudos convertidos foi de 497 *tokens*. Porém, em nossa abordagem, definimos o valor do comprimento máximo dos *tokens* como 512. Utilizamos essa quantidade específica de *tokens* porque ela está alinhada

com o limite de *tokens* de entrada do modelo BERT Multilíngue. Ainda, designamos *tokens* de preenchimento com um rótulo distinto: *–PADDING–*.

Neste trabalho, avaliamos três modelos baseados em arquiteturas de redes neurais distintas: BiLSTM-CRF, BERT e CNN, para realizar o reconhecimento de entidades nomeadas nas sequências resultantes. Cada modelo foi treinado para prever um rótulo de entidade no formato IOB para cada *token* da sequência de entrada. Assim, cada modelo treinado pode ser considerado um modelo de transformação de sequência (Seq2Seq).

Em todos os casos, a etapa de treinamento incluiu a busca pela combinação ótima de hiperparâmetros das redes neurais. Isso inclui parâmetros de arquitetura, funções de ativação e a função de perda. Para cada combinação testada, os pesos da rede são aprendidos (ou ajustados) e subsequentemente avaliados. Isso também se aplica aos pesos de camadas que já passaram por um pré-treinamento, como os pesos do modelo BERT. A seguir descrevemos cada uma das arquiteturas utilizadas neste trabalho.

3.5.1. BiLSTM-CRF

Recurrent Neural Networks (RNNs) foram projetadas com o objetivo principal de capturar e modelar dependências ou padrões de longo prazo em dados sequenciais, como o texto. Contudo, essas redes enfrentaram desafios como os problemas de explosão e desaparecimento do gradiente, que prejudicam sua eficácia na captura das dependências. Para lidar com essas limitações foi usada a *Long Short-Term Memory* (LSTM), uma arquitetura de rede recorrente. A principal característica de uma LSTM é sua célula de memória, que armazena e propaga informações ao longo do tempo. As LSTMs utilizam um mecanismo de portas que consiste em três portas principais: entrada, esquecimento e saída. Essas portas controlam o fluxo de informações para dentro, para fora e dentro da célula de memória, permitindo que a LSTM retenha ou descarte informações seletivamente em diferentes intervalos de tempo [Zhang et al. 2018].

RNNs bidirecionais podem processar a sequência de entrada em duas passagens: uma na direção direta e outra na direção reversa. Essa arquitetura faz isso com duas camadas ocultas separadas que capturam informações de suas respectivas direções e posteriormente as encaminham para a mesma camada de saída. Isto permite que a rede capture informações de contextos passados e futuros, permitindo uma compreensão mais rica dos dados sequenciais. Uma LSTM bidirecional é uma rede neural que compreende unidades LSTM que funcionam nas direções direta e reversa.

A saída da camada BiLSTM (do inglês, *Long Short Term Memory Bidirecional*) é alimentada na camada *Conditional Random Field* (CRF), aumentando a capacidade da rede de considerar os relacionamentos entre rótulos vizinhos. A camada CRF permite que a rede estabeleça dependências de rótulos significativos. Os hiperparâmetros utilizados para treinar o modelo BiLSTM+CRF foram: *word embeddings*, *batch size* e unidades LSTM, realizamos um *grid search* com os seguintes valores: Tamanho do *Word embedding* = $\{25, 100, 300\}$, Unidades LSTM = $\{25, 100, 300\}$, e *Batch size* = $\{4, 8, 16\}$.

A função de otimização *Adam* foi adotada com taxa de aprendizado de 0,01. Definimos uma taxa de *dropout* de 0,1 para mitigar o *overfitting*, e todos os modelos baseados na arquitetura BiLSTM+CRF foram treinados utilizando 10 épocas.

3.5.2. BERT

O BERT (do inglês, *Bidirectional Encoder Representations from Transformers*) é um modelo de linguagem baseado na arquitetura *transformer* que se concentra exclusivamente na codificação. Ele aprende representações vetoriais contextuais de palavras e frases, capturando nuances semânticas a partir de ambas as direções (esquerda para a direita e direita para a esquerda) no texto de entrada. O modelo BERT pode ser ajustado para tarefas específicas de PLN adicionando camadas específicas de tarefas ao modelo principal, como resposta às perguntas ou reconhecimento de entidade nomeada.

Assim, existem duas etapas principais para treinar um modelo BERT para tarefas específicas: pré-treinamento e *fine-tuning*. Neste trabalho, realizamos o *fine-tuning* de um modelo de BERT base multilíngue (BERT-Multilingual) com distinção entre maiúsculas e minúsculas e realizamos um *grid search* com os seguintes valores de hiperparâmetros: *Batch size* {4, 8, 16}, Épocas = {5, 10, 15}, e Taxa de Aprendizagem = {1e-5, 1e-4, 1e-3}. Utilizamos o *Autotokenizer* da biblioteca *Transformers* [Wolf and Debuth 2020] para *tokenizar* nossos relatórios. O Gradiente descendente estocástico foi o otimizador adotado para esse modelo.

3.5.3. Redes Neurais Convolucionais (CNN)

Nesta arquitetura, o processo começa por *tokenizar* a sentença de entrada, transformando-a em uma matriz de sentença onde cada linha corresponde a uma representação vetorial de um *token*. Essas representações vetoriais podem se originar de modelos pré-treinados *word2vec* ou *Glove*. Posteriormente, a matriz de sentença passa por operações convolucionais utilizando filtros lineares, resultando na criação de mapas de características com comprimentos variados.

Neste trabalho, para obter uma representação consistente e de comprimento fixo para cada mapa de características, aplicamos uma operação de agrupamento máximo 1D. Essas representações agrupadas de vários mapas de recursos foram combinadas em um vetor de estrutura de recursos de comprimento fixo de *nível superior*. Então esse vetor estruturado foi passado para uma função *softmax* para classificar a sentença. A aplicação de filtros convolucionais 1D ao texto de entrada do modelo baseado na arquitetura CNN implementado na biblioteca *spaCy* permite realizar a predição considerando o contexto das palavras vizinhas e de suas entidades, levando em conta que o reconhecimento de entidades nomeadas depende destas análises de vizinhança.

3.6. Etapa de Classificação da Entidade Nomeada

Os resultados dos modelos de reconhecimento de entidade nomeada fornecem a localização de entidades relevantes nos laudos de TC do tórax. Contudo, ainda é necessário reconhecer as características de cada entidade em cada um dos segmentos de texto do laudo. Por exemplo, um segmento de texto descrevendo o tamanho de um nódulo pode caracterizá-lo como pequeno, médio ou grande.

Dessa forma, pré-processamos os dados textuais para classificar o texto. O passo inicial foi converter todo o texto para minúsculas, para auxiliar na padronização do texto e no tratamento de palavras maiúsculas e minúsculas como idênticas. Posteriormente,

as pontuações foram removidas do texto, pois geralmente têm significado limitado no contexto de tarefas de classificação e podem introduzir ruído. Para simplificar ainda mais o texto e eliminar termos irrelevantes, palavras irrelevantes e palavras comumente usadas como "e", "o" e "é" também foram removidas.

Seguindo as etapas de pré-processamento, utilizamos a técnica TF-IDF (do inglês, *Term Frequency-Inverse Document Frequency*) para transformar os dados de texto em representações numéricas adequadas sendo utilizadas como entrada em um modelo de máquina de vetores de suporte (SVM). O TF-IDF capta a importância das palavras em cada documento considerando sua frequência e raridade em todo o corpus. Esta transformação não apenas mantém o contexto do texto, mas também atribui pesos apropriados às palavras, permitindo que o modelo SVM reconheça padrões e relacionamentos dentro dos dados de forma eficaz.

Treinamos diferentes modelos SVM para categorizar as seguintes entidades distintas: Achado, Atenuação, Bordas e Localização. A entidade **Achado** nesse contexto tem duas classificações principais: **nódulo** e **enfisema**. Na entidade **Atenuação**, existem três classes distintas: **sólido**, **partes moles** e **vidro fosco**. Por último, a entidade **Bordas** pode ser classificada em uma de duas categorias: **irregular/espiculada** e **regular**. Utilizamos expressões regulares (*regex*) para extrair as dimensões dos nódulos pulmonares identificados.

Diante da identificação das características do nódulo pulmonar, a ferramenta desenvolvida neste trabalho, denominada LungRads+AI, automatiza o cálculo do índice de probabilidade de malignidade de um nódulo pulmonar. Isso é realizado com base nos resultados das classificações das características do nódulo, conforme estabelecido pelas diretrizes do *Lung-RADS*.

3.7. Métricas e Avaliação

O desempenho dos modelos foi avaliado utilizando as métricas Precisão, Revocação e *F1-Score*

$$Precisão = \frac{VP}{VP + FP} \tag{1}$$

$$Revocação = \frac{VP}{VP + FN} \tag{2}$$

$$F1 - score = \frac{2 \cdot Precisão \cdot Revocação}{Precisão + Revocação} = \frac{2 \cdot VP}{2 \cdot VP + FP + FN} \tag{3}$$

onde VP, VN, FP e FN são verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, respectivamente. A Precisão, revocação e *F1-score* de cada entidade de cada modelo foram usados como métricas de avaliação dos 3 modelos apresentados nesse trabalho.

Tabela 1. Comparação dos modelos CNN, BiLSTM-CRF, e BERT para NER

Modelos	Precisão	Revocação	*F1-Score*
CNN	0,83	0,82	0,82
BiLSTM-CRF	**0,86**	**0,85**	**0,86**
BERT-Multilíngue	0,78	0,86	0,82

4. Resultados e Discussão

A Tabela 1 apresenta os resultados obtidos a partir da avaliação da eficácia das técnicas de NER utilizadas neste trabalho. Na avaliação consideramos a capacidade dos modelos em extrair as características de nódulos pulmonares presentes em laudos médicos não estruturados redigidos no idioma português do Brasil. O BiLSTM-CRF alcançou os melhores resultados, seguido pelo modelo CNN proposto por [da Rocha et al. 2023], enquanto o BERT-Multilíngue apresentou o desempenho menos satisfatório dentre os modelos analisados.

O BiLSTM-CRF obteve os melhores resultados ao utilizar os seguintes hiperparâmetros: *word embedding* de 300, 50 unidades LSTM e *Batch size* de 8. No entanto, mesmo após aumentarmos os valores dos hiperparâmetros, o desempenho do modelo BiLSTM-CRF permaneceu estável. Contudo, o desempenho do modelo decaiu com o aumento do *Batch size*.

Quanto ao modelo BERT-Multilíngue, os melhores resultados foram obtidos utilizando os seguintes hiperparâmetros: *Batch size* de 8, número de épocas 10 e taxa de aprendizagem: 1e-3. Observamos que um aumento superior a dez épocas não resultou em um impacto significativo no desempenho do modelo. Contudo, o uso de taxas de aprendizagem mais conservadoras (inferiores a 1e-3) comprometeram a capacidade de

Tabela 2. Desempenho BiLSTM-CRF em cada entidade

Entidade	Precisão	Revocação	*F1-Score*
Achado	0,90	0,92	0,91
Atenuação	0,76	0,76	0,76
Bordas	0,60	0,50	0,55
Calcificação	0,93	0,95	0,94
Localização	0,79	0,74	0,76
Tamanho	0,89	0,91	0,90

Tabela 3. Resultado da Classificação Lung-RADS

Categoria *Lung-RADS*	Precisão	Revocação	*F1-Score*
0	0,88	1,00	0,94
1	1,00	1,00	1,00
2	0,83	0,62	0,71
3	1,00	0,50	0,67
4A	1,00	1,00	1,00
4B	1,00	1,00	1,00
4X	1,00	1,00	1,00

generalização do modelo.

Diante da eficácia superior do modelo BiLSTM-CRF, optamos por utilizá-lo no LungRads+AI para realizar a classificação automática dos nódulos segundo o *Lung-RADS*.

O desempenho detalhado do modelo BiLSTM-CRF é apresentado na Tabela 2. As entidades "Achado" e "Calcificação" obtiveram as melhores métricas de avaliação, enquanto as "Bordas" tiveram a *F1-score* inferior às demais entidades. Este resultado provavelmente se deve ao fato da entidade "Bordas" conter um número inferior de *tokens* anotados em relação as demais entidades, pois continha 336 anotações de treinamento e 28 de teste, evidenciando que o modelo tende a não generalizar com poucas anotações.

Dentre os 79 casos de nódulos pulmonares, 22 (27,85%) apresentaram informações insuficientes sobre os achados. Logo, o LungRads+AI atribuiu a categoria *Lung-RADS* LR-0. Esta categoria foi atribuída para que uma nova TC de tórax pudesse ser realizada dentro do período de 1 a 3 meses com o objetivo de receber informações mais detalhadas sobre esses achados. O LungRads+AI classificou incorretamente quatro dos 22 casos, classificados como LR-0. Isso ocorreu porque o modelo BiLSTM-CRF falhou ao tentar extrair apropriadamente a entidade referente ao tamanho do nódulo, descrita no laudo. Esses quatro nódulos deveriam ter recebido a classificação LR-2. Os 18 casos restantes foram classificados corretamente como LR-0 devido à falta de informações sobre a atenuação nos nódulos.

O LungRads+AI classificou 42 (53,16%) achados como LR-1, oito (10,13%) achados foram classificados como (LR-2), sete (8,86%) achados foram classificados como categoria 3 ou 4, destes, dois (2,53%) da categoria LR-3, um (1,26%) da categoria LR-4A, três (3,80%) da categoria 4B e um achado (1,26%) da categoria 4X. Além disso, a ferramenta foi capaz de classificar nódulos benignos (LR-2) com precisão de 83%, revocação de 62% e *F1-score* de 71%. Em 1 de 8 casos, a classificação foi feita incorretamente porque o BiLSTM-CRF usou o tamanho do componente sólido do nódulo ao invés do tamanho do nódulo. Este nódulo deveria ter recebido classificação (LR-3). O LungRads+AI identificou nódulos benignos (LR-3) com precisão de 100%, revocação de 50% e *F1-score* de 67%; e identificou nódulos suspeitos (LR-4) com precisão e revocação de 100%.

5. Conclusão

Este trabalho avaliou a eficácia de modelos de redes neurais treinados para realizar o *Named Entity Recognition* (NER) a fim de automatizar a classificação de nódulos pulmonares em laudos da TC do tórax. Em nossos resultados, o BiLSTM-CRF obteve o melhor desempenho dentre os modelos avaliados (CNN e BERT), alcançando precisão de 86%, revocação de 85% e *F1-score* de 86%. Estes resultados podem ser atribuídos à especialização do modelo para tarefas de NER, pois, o BiLSTM-CRF utiliza uma camada BiLSTM para capturar os relacionamentos entre as palavras de uma frase, característica importante para identificar entidades compostas por múltiplas palavras. Ainda, o modelo usa a camada CRF para modelar as dependências entre os diferentes tipos de entidades.

O *Lung-RADS* é uma diretriz fundamental no rastreio do câncer de pulmão. Contudo, é reconhecido que o preenchimento manual do *Lung-RADS* é uma tarefa laboriosa, demorada e propensa a erros, dada a rotina clínica intensa do radiologista. Nesse contexto, a ferramenta LungRads+AI foi desenvolvida para automatizar o cálculo do Índice

Lung-RADS, atuando como um sistema de auxílio ao radiologista. A ferramenta obteve desempenho acima de 70% nas métricas de avaliação para as classes *Lung-RADS* 0, 1, 4A, 4B e 4X na classificação dos índices *Lung-RADS*. Dando continuidade ao trabalho, estamos em processo de coleta de mais relatórios de TC em três instituições hospitalares distintas, visando ampliar a diversidade dos laudos *Lung-RADS*.

Este trabalho foi apoiado pelas seguintes agências de pesquisa: Ministério da Saúde, CNPq, SESAU-AL e Fundação de Amparo à Pesquisa de Alagoas (FAPEAL).

Referências

(2023). Lung ct screening reporting data system. `https://www.acr.org/-/media/ACR/Files/RADS/Lung-RADS/Lung-RADS-2022.pdf`. Accessed: 2023-05-01.

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.

da Rocha, N. C., Barbosa, A. M. P., Schnr, Y. O., Machado-Rugolo, J., de Andrade, L. G. M., Corrente, J. E., and de Arruda Silveira, L. V. (2023). Natural language processing to extract information from portuguese-language medical records. *Data*, 8(1).

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fei, X., Chen, P., Wei, L., Huang, Y., Xin, Y., and Li, J. (2022). Quality management of pulmonary nodule radiology reports based on natural language processing. *Bioengineering (Basel)*, 9(6).

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.

Gould, M. K., Donington, J., Lynch, W. R., Mazzone, P. J., Midthun, D. E., Naidich, D. P., and Wiener, R. S. (2013). Evaluation of individuals with pulmonary nodules: When is it lung cancer?: Diagnosis and management of lung cancer, 3rd ed: American college of chest physicians evidence-based clinical practice guidelines. *Chest*, 143(5, Supplement):e93S–e120S.

Kang, S. K., Garry, K., Chung, R., Moore, W. H., Iturrate, E., Swartz, J. L., Kim, D. C., Horwitz, L. I., and Blecker, S. (2019). Natural language processing for identification of incidental pulmonary nodules in radiology reports. *J Am Coll Radiol*, 16(11):1587–1594.

Liu, H., Zhang, Z., Xu, Y., Wang, N., Huang, Y., Yang, Z., Jiang, R., and Chen, H. (2021). Use of bert (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in chinese radiology reports: Development of a computer-aided liver cancer diagnosis framework. *J Med Internet Res*, 23(1):e19689.

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.

MacMahon, H. and Naidich, D. P. (2017). Guidelines for management of incidental pulmonary nodules detected on ct images: From the fleischner society 2017. *Radiology*, 284(1):228–243. PMID: 28240562.

Mendoza, D. P., Petranovic, M., Som, A., Wu, M. Y., and Digumarthy, S. R. (2022). Lung-rads category 3 and 4 nodules on lung cancer screening in clinical practice. *American Journal of Roentgenology*, 219(1):55–65. PMID: 35080453.

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.

Nobel, J. M., Puts, S., Bakers, F. C. H., Robben, S. G. F., and Dekker, A. L. A. J. (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.

Patwardhan, N., Marrone, S., and Sansone, C. (2023). Transformers in the real world: A survey on nlp applications. *Information*, 14(4).

Pinsky, P. F., Gierada, D. S., Black, W., Munden, R., Nath, H., Aberle, D., and Kazerooni, E. (2015). Performance of Lung-RADS in the national lung screening trial: a retrospective assessment. *Ann Intern Med*, 162(7):485–491.

Siegel, R. L. and Miller (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1):17–48.

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, H. and Ferlay (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.

Vykoukal, J., Fahrmann, J. F., and Patel (2022). Contributions of circulating micrornas for early detection of lung cancer. *Cancers*, 14(17).

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, T. and Debuth, L. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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