Diagnóstico de Tuberculose em Imagens de Radiografia utilizando CvT

Carlos M. Santos Neto[1], **Anderson L. Silva**[1], **Alexandre C. P. Pessoa**[1], **Darlan B. P. Quintanilha**[1], **João D. S. de Almeida**[1], **Geraldo Braz Junior**[1], **João O. B. Diniz**[2]

[1]Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA)
São Luís – MA – Brasil

[2]Instituto Federal de Educação Tecnológica do Maranhão (IFMA)
Grajaú – MA – Brasil

{carlos.msn, anderson.silva, alexandre.pessoa}@nca.ufma.br,

{dquintanilha,jdallyson, geraldo}@nca.ufma.br, joao.bandeira@ifma.edu.br

Abstract. *Tuberculosis (TB) remains one of the leading causes of death from infectious diseases. In 2022, an estimated 10.6 million people worldwide contracted TB. Chest X-rays, a non-invasive medical examination used to detect pathologies in various areas of the chest, are a crucial tool in TB diagnosis. Recent advancements in the field of computer vision, particularly with the application of deep learning techniques, have led to significant progress in the automated detection of abnormalities in chest X-rays. This has opened the door to machine-aided diagnosis. In this work, we propose a method for diagnosing tuberculosis in radiographic images using the Convolutional Vision Transformers neural network. The results show relevant metrics, with an accuracy of 93.13%, an F1-score of 92.68%, and an AUC-ROC of 97.16%, using the public image databases Shezen and Montgomery County. These results are superior to the state of the art.*

Resumo. *A tuberculose (TB) é uma das maiores causadoras de morte por doenças infecciosas. Em 2022, estimou-se que no mundo 10,6 milhões de pessoas ficaram doentes com TB. A radiografia de tórax é um exame médico não invasivo que é utilizado para detectar patologias em diversas áreas do tórax, sendo uma ferramenta crucial no diagnóstico de TB. O desenvolvimento na área de visão computacional, com a utilização de técnicas de aprendizado profundo, levou a avanços significativos na detecção automática de anormalidades em imagens radiográficas, possibilitando a existência de diagnósticos auxiliados por máquina. Neste trabalho é proposto um método para o diagnóstico de tuberculose em imagens de radiografia utilizando a rede neural Convolutional Vision Transformers. Os resultados mostram métricas relevantes, com uma acurácia de 93,13%, um F1-score de 92,68% e uma AUC-ROC de 97,16%, utilizando as bases de imagens públicas Shezen e Montgomery County. Esses resultados são superiores ao estado da arte.*

1. Introdução

A tuberculose é uma das principais causas de morte por doenças infecciosas. Em 2022, estima-se que 10,6 milhões de pessoas em todo o mundo tenham

contraído tuberculose, mas apenas 7,5 milhões foram diagnosticadas e relatadas [World Health Organization 2023].

A radiografia de tórax (RXT) é uma ferramenta recomendada e amplamente utilizada para o rastreamento da tuberculose [Pande et al. 2015]. Caracteriza-se como um exame médico não invasivo utilizado para detectar patologias em diversas áreas do tórax, incluindo os pulmões, coração, vasos sanguíneos, ossos, entre outros. Essa técnica de imagem é frequentemente utilizada como uma ferramenta de triagem para a detecção precoce de doenças, permitindo que os médicos tomem decisões rápidas e precisas sobre o tratamento necessário.

Alguns pesquisadores apoiam a ideia de que o exame de imagem possa ser usado como uma ferramenta primária para detecção de anormalidade para triagem em áreas epidêmica [Mahbub et al. 2022, Diniz et al. 2023, Munzlinger et al. 2023]. A detecção precoce dessas doenças pode salvar vidas e melhorar significativamente a qualidade de vida dos pacientes [da Silva et al. 2021, Costa et al. 2021].

No entanto, a interpretação das radiografias de tórax pode ser desafiadora, especialmente em áreas com recursos limitados onde a disponibilidade de radiologistas é escassa [World Health Organization 2016]. Nesse contexto, avanços na inteligência artificial e visão computacional têm proporcionado novas oportunidades para o desenvolvimento de sistemas de *Computer Aided Diagnosis* (CAD) [Murphy et al. 2020]. Tais sistemas podem auxiliar os médicos na interpretação das imagens radiográficas, agilizando o processo de diagnóstico e reduzindo erros de interpretação.

Os *Vision Transformers* (ViT) [Dosovitskiy et al. 2020] são uma técnica de aprendizado profundo relativamente nova que demonstrou excelente desempenho em tarefas de classificação de imagem. Essa técnica foi originalmente proposta para processar imagens em sequências de *tokens*, semelhante ao processamento de linguagem natural com *transformers*. Em comparação às técnicas de convolução tradicionais usadas em redes neurais convolucionais, os *transformers* visuais demonstram melhor capacidade de capturar relações de longo alcance entre diferentes partes da imagem.

Diante desse cenário, o presente estudo propõe uma abordagem para o diagnóstico de tuberculose em imagens de radiografia utilizando a rede neural *Convolutional Vision Transformers*. O objetivo é desenvolver um sistema robusto e preciso que possa auxiliar os médicos na detecção precoce e precisa da tuberculose pulmonar, contribuindo para o controle eficaz da doença.

2. Trabalhos Relacionados

Nos últimos anos, houve um aumento significativo no interesse e na pesquisa sobre o uso da visão computacional no diagnóstico de tuberculose em imagens radiográficas, principalmente devido as suas vantagens como um custo operacional menos elevado. Uma revisão abrangente da literatura revela uma série de estudos relevantes que exploraram diversas abordagens e técnicas para essa finalidade.

[Islam et al. 2017] usam vários conjuntos de imagens radiográficas, tais como *Indiana RXT*, *JSRT* e *Shenzhen* [Jaeger et al. 2014], a fim de avaliar as Redes Neurais Convolucionais (CNNs). Os autores concluem que recursos superficiais fornecem consistentemente maior precisão de detecção em comparação com recurso profundos. Ao combinar

essas descobertas (*ensemble*), obtiveram maior precisão na detecção de anormalidades em radiografias de tórax, acurácia de 90% e AUC-ROC de 94%.

Com o propósito de implantar um modelo em dispositivos móveis, [Pasa et al. 2019] propuseram uma CNN mais simples e eficiente do que os modelos tradicionais, mantendo sua precisão, obtendo acurácia de 84,4% e AUC de 90%. O estudo obteve uma acurácia de 84,4% e uma AUC de 90%, utilizando conjuntos de imagens provenientes de duas bases públicas diferentes: a *NIH Tuberculosis Chest X-ray*, que é subdividida em duas bases, a *Montgomery County em Maryland* e a e a *Shenzhen*, e a *Belarus Tuberculosis Portal*. Além disso, os autores exploraram o uso de mapas de saliência e *grad-CAMs* como métodos de visualização da tuberculose nas radiografias. A rede CNN foi otimizada para eficiência e velocidade, não para maximizar a acurácia. A rede não foi pré-treinada com grandes bases de dados como a *ImageNet*, que pode ter limitado a exposição da rede a diversos padrões e características diferentes que poderiam ter melhorado a performance.

A *LightTBNet*, uma rede convolucional profunda baseada na *ResNet*, foi projetada para detectar tuberculose em imagens de radiografia de tórax em [Capellán-Martín et al. 2023]. O modelo alcançou uma acurácia de 90,6%, um F1-score de 90,7% e uma área sob a curva ROC (AUC) de 96,1%, utilizando um conjunto de dados híbrido ao combinar imagens das bases *Shenzen* e *Montgomery* durante o treinamento da rede neural.

O método proposto em [Evangelista and Guedes 2019] avalia um conjunto de redes neurais convolucionais tradicionais e define um *ensemble* de três arquiteturas profundas diferentes dessas redes, utilizando os conjuntos de imagens radiográficas, *Montgomery*, *JSRT* e *Shenzhen* no treinamento, alcançando uma precisão de 93,2 % e uma AUC ROC de 95,6% na tarefa proposta.

[Munzlinger et al. 2023] adaptam o modelo *ResNet-50* utilizando *transfer learning* para predizer doenças pulmonares em imagens de RXT, incluindo Covid-19, pneumonia e tuberculose. O método alcançou uma acurácia geral de aproximadamente 89% na predição dessas doenças.

Até o momento, os estudos relacionados ao diagnóstico de tuberculose em imagens de radiografia têm se concentrado principalmente no uso de CNNs e, mais recentemente, no uso de *Vision Transformers*. Entretanto, não foram identificados estudos recentes que abordem o problema em questão combinando as características das CNNs, utilizando camadas convolucionais iniciais para extrair características visuais locais, com os benefícios dos *transformers*, para modelar relações de longo alcance. Diante desse cenário, este estudo busca avaliar se as redes *transformers* têm a capacidade de superar as CNNs no diagnóstico de tuberculose em radiografias torácicas.

3. Materiais e Método

O método proposto está dividido em quatro etapas, conforme ilustrado na Figura 1. Em resumo, a primeira etapa descreve a base de imagens de radiografia de tórax. A segunda etapa consiste na etapa de pré-processamento para padronização das imagens. As seguintes, fazem parte da classificação realizada usando a rede neural. Por fim, os resultados são avaliados.

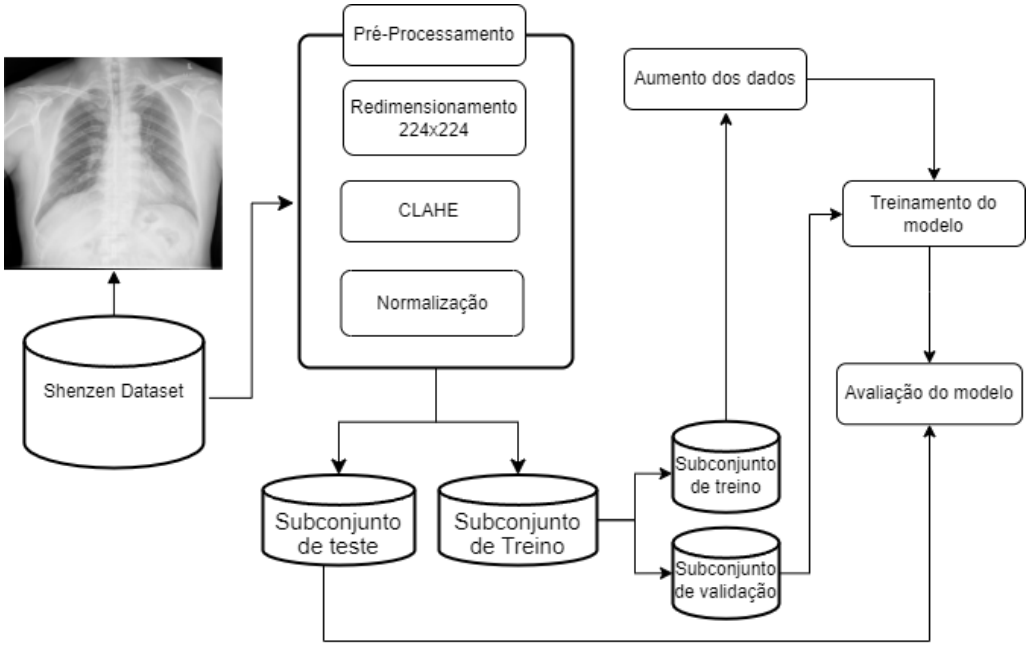

Figura 1. Fluxo de trabalho.

3.1. Aquisição de Imagens

O conjunto de dados utilizado para validar o método proposto para o diagnóstico de tuberculose foi o *Shenzen Dataset* [Jaeger et al. 2014]. Essa base contém 662 exames de RXT com uma resolução aproximada de 3000 x 3000 *pixels*.

O conjunto de dados utilizado para validação do método proposto para diagnóstico de tuberculose foi a *Shenzen Dataset* [Jaeger et al. 2014]. Esta base contém 662 exames de RXT com uma resolução aproximada de 3000×3000 *pixels*. Dentro deste conjunto, estão incluídos 326 casos normais e 336 casos com manifestações de tuberculose, exibindo uma variedade de 19 tipos de anormalidades. Todas as imagens são arquivos *png* nomeados seguindo a convenção de nomenclatura "CHNCXR_DDDD_L.png", onde "DDDD" representa o número da imagem e "L" pode assumir o valor 0 para imagens sem a presença de tuberculose e 1 para imagens com a presença de tuberculose, conforme exemplificado na Figura 2.

As anotações das anormalidades nas imagens positivas foram realizadas por dois radiologistas da *Chinese University of Hong Kong*, sendo inicialmente conduzidas por um radiologista júnior e posteriormente verificadas por um radiologista sênior, com consenso em todos os casos [Yang et al. 2022]. A base é projetada para destacar características relevantes para identificação de tuberculose, o que a caracteriza como uma abordagem *multi-label*, permitindo que cada imagem contenha múltiplas anotações de anormalidades.

Além das imagens, um arquivo CSV contendo 19 tipos de anormalidades anotadas está disponível, como derrame pleural, espessamento apical, nódulo único (não calcificado), espessamento pleural (não apical), nódulo calcificado, pequeno infiltrado (não linear), cavidade, densidade linear, infiltrado severo (consolidação), espessamento da fissura interlobar, nódulos agrupados (2mm-5mm de distância), infiltrado moderado (não linear), adenopatia, calcificação (além de nódulo e linfonodo), linfonodo calcificado,

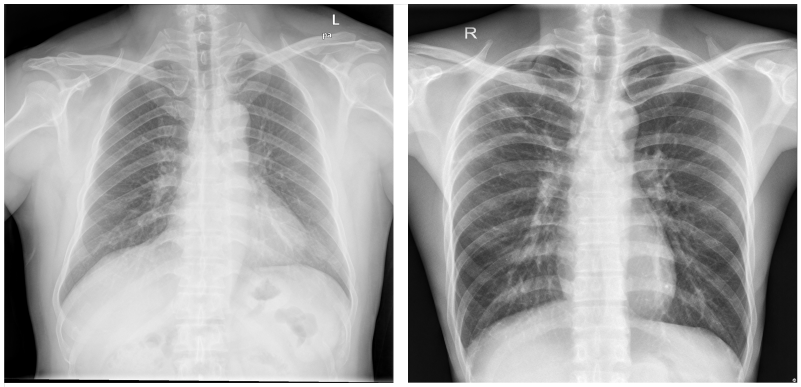

Figura 2. Exemplo de imagens de RXT da base *Shenzen*: (a) classe Normal e (b) classe com Tuberculose.

miliar, retração, outro e desconhecido, para as 336 imagens com tuberculose. A Figura 3 ilustra a frequência dessas anormalidades. Na abordagem de classificação binária, a presença de qualquer uma dessas anormalidades será considerada como "anormal".

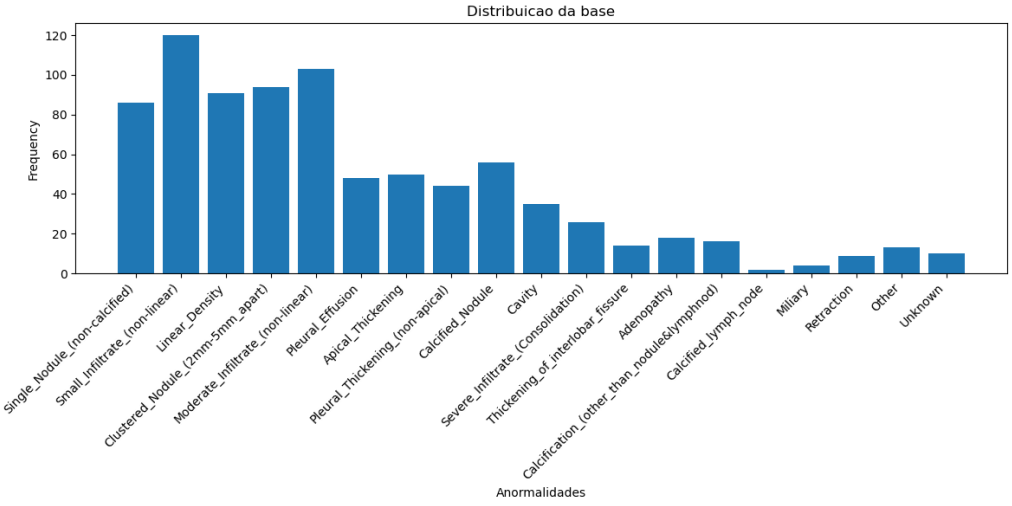

Figura 3. Frequência dos tipos de anormalidades presentes na base de imagens *Shenzen*.

3.2. Pré-processamento

Durante a análise das anotações da base de imagens, foram identificadas quatro imagens com anotações inconsistentes. Embora o arquivo CSV indicasse a ausência de tuberculose, a nomenclatura dos arquivos sugeria o contrário, indicando positividade para tuberculose. Desta forma, essas imagens foram excluídas da avaliação do método.

Em seguida, as imagens que possuíam pelo menos um dos tipos de tuberculose foram rotuladas como "anormais", enquanto aquelas que não apresentavam essas características foram rotuladas como "normais". Em seguida, o algoritmo de Equalização Adaptativa de Histograma com Limite de Contraste (CLAHE) [Siracusano et al. 2020] foi aplicado às imagens para melhorar o contraste local. Para facilitar a execução dos

experimentos com a rede neural, todas as imagens foram redimensionadas para 224×224 *pixels* e seus valores foram normalizados para o intervalo de [0, 1].

3.3. Classificação das imagens de radiografia

Para classificar as imagens de radiografia do tórax entre normais e anormais, utilizou-se o modelo preditivo da rede *Convolutional Vision Transformers* (CvT)[Wu et al. 2021]. Na Figura 4 (a) é apresentado o pipeline geral da rede.

A CvT representa uma melhoria na *Vision Transformer* (ViT) —em termos de desempenho e eficiência—ao introduzir convoluções na ViT para combinar o melhor de ambas as abordagens. Isso foi alcançado por meio de duas modificações principais: uma hierarquia *Transformers* que inclui um novo *convolutional token embedding* visando modelar contextos espaciais locais, desde bordas de baixo nível até primitivas semânticas de ordem superior, em uma abordagem hierárquica de vários estágios, similar às *convolutional neural networks* (CNNs); e um bloco *convolutional Transformer* caracterizado pelo uso de uma *convolutional projection* com o objetivo de alcançar a modelagem adicional do contexto espacial local e fornecer benefícios de eficiência ao permitir que aumente a amostragem das matrizes de chave (K) e valor (V) [Wu et al. 2021], como mostrado na Figura 4 (b). Essas modificações incorporaram as propriedades de uma CNN em uma arquitetura ViT.

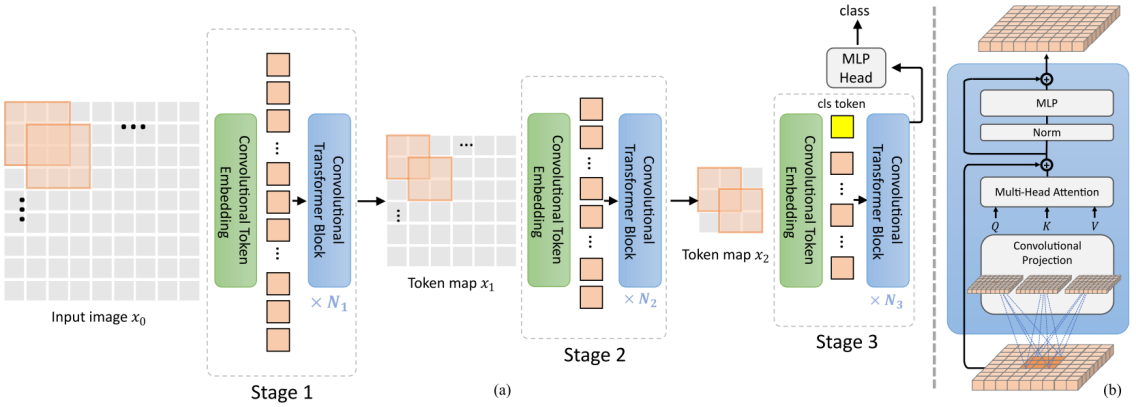

Figura 4. Fluxo de funcionamento CvT (a) *Pipeline* da CvT (b) Bloco convolucional *transformer* [Wu et al. 2021].

A rede possui algumas variantes que se diferenciam nos números de *Transformer blocks* em cada estágio e nas dimensões do recurso oculto utilizadas. Três estágios são adaptados: CvT-13 e CvT-21 foram definidos como modelos básicos, com 19,98 milhões e 31,54 milhões de parâmetros, respectivamente. Adicionalmente, foi experimentado um modelo mais amplo com uma dimensão de *tokens* maior para cada estágio, denominado CvT-W24 (W significa *Wide*), resultando em 298,3 milhões de parâmetros, para validar a capacidade de escalonamento da arquitetura proposta [Wu et al. 2021]. Esses modelos são detalhados na Tabela 1.

A variante escolhida para o experimento foi a CvT-13 devido à sua leveza e desempenho superior em relação a outras CNNs, como a ResNet-152, na base de dados *ImageNet*, alcançando uma acurácia *top-1* maior.

	Output Size	Layer Name	CvT-13	CvT-21	CvT-W24
Stage1	56×56	Conv. Embed.	7×7, 64, stride 4		7×7, 192, stride 4
	56×56	Conv. Proj. MHSA MLP	$\begin{bmatrix} 3 \times 3, 64 \\ H_1 = 1, D_1 = 64 \\ R_1 = 4 \end{bmatrix} \times 1$	$\begin{bmatrix} 3 \times 3, 64 \\ H_1 = 1, D_1 = 64 \\ R_1 = 4 \end{bmatrix} \times 1$	$\begin{bmatrix} 3 \times 3, 192 \\ H_1 = 3, D_1 = 192 \\ R_1 = 4 \end{bmatrix} \times 2$
Stage2	28×28	Conv. Embed.	3×3, 192, stride 2		3×3, 768, stride 2
	28×28	Conv. Proj. MHSA MLP	$\begin{bmatrix} 3 \times 3, 192 \\ H_2 = 3, D_2 = 192 \\ R_2 = 4 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 192 \\ H_2 = 3, D_2 = 192 \\ R_2 = 4 \end{bmatrix} \times 4$	$\begin{bmatrix} 3 \times 3, 768 \\ H_2 = 12, D_2 = 768 \\ R_2 = 4 \end{bmatrix} \times 2$
Stage3	14×14	Conv. Embed.	3×3, 384, stride 2		3×3, 1024, stride 2
	14×14	Conv. Proj. MHSA MLP	$\begin{bmatrix} 3 \times 3, 384 \\ H_3 = 6, D_3 = 384 \\ R_3 = 4 \end{bmatrix} \times 10$	$\begin{bmatrix} 3 \times 3, 384 \\ H_3 = 6, D_3 = 384 \\ R_3 = 4 \end{bmatrix} \times 16$	$\begin{bmatrix} 3 \times 3, 1024 \\ H_3 = 16, D_3 = 1024 \\ R_3 = 4 \end{bmatrix} \times 20$
Head	1×1	Linear	1000		
Params			19.98 M	31.54 M	276.7 M
FLOPs			4.53 G	7.13 G	60.86 G

Tabela 1. Arquitetura do CvT [Wu et al. 2021].

Além disso, para lidar com o desequilíbrio de classes, foi utilizada a função de perda *Focal Loss* [Lin et al. 2017]. Essa função foi especialmente projetada para tarefas de classificação binária, oferecendo uma solução eficaz para situações em que as classes não estão igualmente distribuídas, ou seja, uma classe ocorre com mais frequência em comparação com a outra. Ela ajusta o foco do treinamento para os exemplos mais desafiadores, considerando a probabilidade correta da classe (p), um fator de ajuste (α) para equilibrar as classes e um parâmetro de modulação (γ). Esses elementos juntos permitem que o modelo priorize exemplos que são mais difíceis de classificar corretamente, melhorando sua capacidade de lidar com o desequilíbrio de classes e, consequentemente, aumentando a precisão da classificação final. A equação da *Focal Loss* é dada por:

$$FL(p, y) = \begin{cases} -\alpha(1-p)^{\gamma} \log(p), & \text{Se } y = 1, \\ -(1-\alpha)^{\gamma} \log(1-p)) \log(1-p), & \text{Senão.} \end{cases} \tag{1}$$

3.4. Métricas de avaliação

Para avaliação do método proposto, foram utilizadas as métricas Área Sob a Curva ROC (AUC-ROC), acurácia, sensibilidade, precisão e F1-score, apresentadas nas Equações 2 a 6.

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(FPR)\, d(FPR) \tag{2}$$

$$\text{Acurácia} = (TP + TN)/(TP + TN + FP + FN) \tag{3}$$

$$\text{Precisão} = TP/(TP + FP) \tag{4}$$

$$\text{Sensibilidade} = TP/(TP + FN) \tag{5}$$

$$\text{F1-Score} = 2 * TP/(2 * TP + FN + FP) \tag{6}$$

Na definição das métricas de avaliação, TP refere-se aos verdadeiros positivos, representando os casos de tuberculose corretamente classificados; FP indica os falsos positivos, que são casos sem tuberculose erroneamente classificados como positivos; TN representa os verdadeiros negativos, abarcando os casos de não tuberculose corretamente classificados; e FN designa os falsos negativos, que são os casos de tuberculose incorretamente classificados como negativos.

4. Resultados

Nesta seção, descreve-se o resultado obtido com a aplicação do método proposto. Primeiramente é apresentado, e em seguida, é apresentado o desempenho do método e comparação com a literatura.

O experimento foi conduzido utilizando a linguagem de programação *Python*. Para normalizar as imagens, empregou-se o gerador de imagens do *Keras preprocessing*. A arquitetura da rede foi implementada utilizando a construção fornecida pelo *Keras* [Chollet et al. 2015] na biblioteca *cvt-tensorflow*. Durante a execução da rede, foi utilizada a GPU GTX 1660 TI com 6GB de VRAM.

A técnica de validação do modelo utilizada é a *Holdout*. A base de imagens foi dividida em 60% para treino, 20% para validação e 20% para teste, com estratificação pelos tipos de tuberculoses e a classe normal.

Além disso, foi aplicada a técnica de aumento de dados para aumentar o número de imagens de treinamento de 393 para 10.218. Essa técnica envolveu transformações geométricas de rotação (-20° a 20°), espelhamento horizontal e escala (0 a 0,1), resultando em uma maior variabilidade na base de treinamento.

Optou-se por utilizar o *Optuna* [Akiba et al. 2019] como ferramenta de otimização dos hiperparâmetros do modelo. Foram realizados trinta estudos, empregando uma abordagem logarítmica, para encontrar o valor máximo da curva ROC na validação. Os hiperparâmetros otimizados foram a taxa de aprendizado do otimizador ADAM e os parâmetros de α e γ da função de perda *Focal Loss*.

Esses estudos foram repetidos para refinar os limites da busca, possibilitando a identificação do melhor intervalo de valores com um treinamento mais estável. Após o refinamento e definição dos hiperparâmetros, realizou-se o experimento de classificação com a rede CvT-13. Os valores otimizados para o otimizador ADAM foram uma taxa de aprendizado de $1{,}4 \times 10^{-4}$, enquanto a focal loss empregou um valor de α de 0,49 e γ de 1,33. Os melhores resultados obtidos nos testes podem ser visualizados na Tabela 2, juntamente com a matriz de confusão na Tabela 3.

Tabela 2. Avaliação do método proposto no subconjunto de teste: resultados das métricas de Acurácia (ACC), Área sob a Curva ROC (AUC), Sensibilidade (SEN), Especificidade (ESP), Precisão (PREC) e F1-Score (F1)

ACC	SEN	ESP	PREC	AUC	F1
93,13%	89,06%	97,01%	96,61%	0,9716	92,68%

Tabela 3. Matriz de confusão do método proposto no subconjunto teste.

		Predito	
		Anormal	Normal
Real	Anormal	57	2
	Normal	7	65

Esses resultados foram obtidos aplicando a Equalização Adaptativa de Histograma com Limite de Contraste (CLAHE) [Siracusano et al. 2020] como técnica de melhora-

mento das imagens, como em outros estudos. Observa-se que a rede atinge desempenho comparável à outras CNNs bem estabelecidas, mesmo ao lidar com um conjunto de imagens maior. Por exemplo, o estudo [Evangelista and Guedes 2019] empregou uma combinação dos conjuntos de dados de *Shenzhen* e *Montgomery*. Essa comparação é apresentada na Tabela 4.

Tabela 4. Comparação dos resultados com outros modelos de outras arquiteturas CNNs

CNN	Otimizador	ACC	F1	SEN	ESP	PREC
Inception	ADAM	89,32%	87,25%	87,84%	90,38%	86,67%
ResNet	ADAM	87,64%	85,14%	85,14%	89,42%	85,13%
VGG16	ADAM	85,39%	81,69%	78,38%	90,38%	85,29%
AlexNet	ADAM	83,71%	77,86%	68,92%	94,23%	89,47%
Método proposto	ADAM	93,13%	92,68%	89,06%	97,01%	96,61%

A comparação entre os estudos que utilizaram os conjuntos de dados *Shenzhen* podem ser observada na Tabela 5. Os resultados referentes a base *Shenzhen* são consistentes e o resultado do método proposto é superior ao estado da arte. O método proposto apresenta uma abordagem inovadora ao combinar elementos das CNNs e dos *Transformers*. Seu potencial reside na capacidade de processar informações de forma abrangente e contextualizada, capturando relações complexas e aprendendo representações mais abstratas. De acordo com os resultados do diagnóstico de tuberculose em imagens de radiografia, a CvT oferece uma base sólida para pesquisas futuras. Sua flexibilidade e capacidade de generalização sugerem que ajustes adicionais na arquitetura e técnicas de treinamento podem levar a melhorias significativas no desempenho.

Tabela 5. Comparação dos resultados do modelo proposto no subconjunto de teste da base de dados Shenzhen com os resultados reportados em outros estudos.

Modelo/Implementação	ACC	AUC	SEN	ESP	PREC	F1
[Islam et al. 2017]	90,00%	0,9400	88,00%	92,00%	-	-
[Pasa et al. 2019]	84,40%	0,9000	-	-	-	-
[Capellán-Martín et al. 2023]	90,60%	0,9630	-	-	-	-
Método proposto	93,13%	0,9716	89,06%	97,01%	96,61%	92,68%

A base de dados *Montgomery County* [Jaeger et al. 2014] consiste em imagens de RXT que mostram anormalidades relacionadas à tuberculose, contendo uma quantidade de 80 casos normais e 58 anormais. Com o intuito de avaliar a capacidade de generalização do modelo proposto diante de uma maior variedade de características e padrões das imagens, como diferentes origens geográficas dos pacientes ou variações nos equipamentos de aquisição das imagens, a base de dados *Montgomery County* foi utilizada como um conjunto de validação externa. Isso significa que não foi realizado um novo treinamento específico para essa base de dados. Os resultados podem ser visualizados na Tabela 6.

Conforme exibido na Tabela 6, a validação externa demonstrou um bom desempenho, superando os resultados obtidos em [Pasa et al. 2019]. No entanto, o método proposto não conseguiu superar o desempenho relatado em [Capellán-Martín et al. 2023]. É

importante ressaltar que em [Pasa et al. 2019] a base Montgomery não foi utilizada apenas como validação externa, e que um subconjunto da mesma foi utilizado no treinamento. Em [Capellán-Martín et al. 2023] utilizaram um conjunto de dados híbrido ao combinar imagens das bases *Shenzen* e *Montgomery* durante o treinamento do modelo. Em seguida, utilizaram o modelo treinado para avaliar separadamente as bases de imagens *Shenzen* e *Montgomery*.

Tabela 6. Comparação dos resultados do modelo proposto no conjunto de validação externa a base de dados Montgomery com os resultados reportados em outros estudos.

Modelo/Implementação	ACC	AUC	SEN	ESP	PREC	F1
[Pasa et al. 2019]	79,00%	0,8110	-	-	-	-
[Capellán-Martín et al. 2023]	88,90%	0,9430	-	-	-	-
Método proposto	80,14%	0,8326	60,71%	93,75%	87,18%	71,57%

Uma análise completa do método proposto foi apresentada, revelando seu desempenho em comparação com estudos anteriores. A comparação com outras arquiteturas de CNNs estabelecidas evidenciou que o método proposto supera seus desempenhos, destacando-se como uma alternativa viável para o diagnóstico de tuberculose em imagens de radiografia. Ao expandir a análise para diferentes conjuntos de dados, como o *Montgomery County*, foi possível avaliar a capacidade de generalização do modelo.

É importante ressaltar que estratégias como o aumento de dados com operações geométricas contribuíram para melhorar a generalização do modelo, permitindo lidar com conjuntos de dados maiores e mais diversos. Esses resultados sugerem que a abordagem inovadora da rede CvT, combinando elementos de CNNs e Transformers, apresenta um potencial significativo para avanços futuros no diagnóstico de tuberculose por meio de imagens de radiografia.

5. Conclusão

Neste trabalho foi proposto um método para diagnóstico de tuberculose em imagens de radiografia utilizando a rede CvT, uma abordagem inovadora que combina elementos das CNNs e dos *Transformers*. Foram aplicadas técnicas de pré-processamento, como o CLAHE, e de aumento de dados, juntamente com a função de perda *focal loss* para lidar com o desequilíbrio de classes. Os resultados obtidos na classificação binária das imagens foram promissores, com uma acurácia de 93,13%, AUC-ROC de 97,16% e F1-score de 92,68%.

Embora tenha superado os resultados alcançados pelo estado da arte, ainda há oportunidades para aprimorar o desempenho do método. Uma estratégia em consideração é ajustar o número de blocos em cada estágio da rede CvT para determinar a configuração ideal que maximize o desempenho. Além disso, pretende-se explorar a técnica *Ensemble*, combinando múltiplos modelos de arquiteturas diversas, como CNNs e *Transformers*, cada um especializado na detecção de um tipo específico de tuberculose. Por fim, a utilização do método de validação cruzada nas redes é essencial para obter resultados mais consistentes e robustos, proporcionando uma avaliação mais confiável do desempenho em diversos cenários.

É importante ressaltar que as limitações computacionais impediram a utilização das versões com mais blocos da rede CvT, que poderiam fornecer insights adicionais sobre a necessidade de recursos computacionais. Por fim, embora o método proposto tenha produzido resultados aceitáveis, é crucial enfatizar que ele não substitui o diagnóstico de um radiologista profissional.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 , Fundação de Amparo a Pesquisa do Maranhão (FAPEMA), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Empresa Brasileira de Serviços Hospitalares (Ebserh) Brazil (Proc. 409593/2021-4).

Referências

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Capellán-Martín, D., Gómez-Valverde, J. J., Bermejo-Peláez, D., and Ledesma-Carbayo, M. J. (2023). A lightweight, rapid and efficient deep convolutional network for chest x-ray tuberculosis detection. *IEEE International Symposium on Biomedical Imaging*.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Costa, G. S. S., Paiva, A. C., Junior, G. B., and Ferreira, M. M. (2021). Covid-19 automatic diagnosis with ct images using the novel transformer architecture. In *Anais do XXI simpósio brasileiro de computação aplicada à saúde*, pages 293–301. SBC.

da Silva, G. L., de Oliveira, F. Y., Diniz, J. O., Diniz, P. S., Quintanilha, D. B., Silva, A. C., de Paiva, A. C., and de Cavalcanti, E. A. (2021). An automatic method for prostate segmentation on 3d mri scans using local phylogenetic indexes and xgboost. In *Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 165–176. SBC.

Diniz, J. O., Quintanilha, D. B., de Carvalho Filho, A. O., Gomes Jr, D. L., Silva, A. C., Braz Jr, G., de Paiva, A. C., and Luz, D. d. S. (2023). Detecção de covid-19 em imagens de raio-x de tórax através de seleção automática de pré-processamento e de rede neural convolucional. In *Anais do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 162–173. SBC.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Evangelista, L. G. C. and Guedes, E. B. (2019). Ensembles of convolutional neural networks on computer-aided pulmonary tuberculosis detection. *IEEE Latin America Transactions*, 17(12):1954–1963.

Islam, M. T., Aowal, M. A., Minhaz, A. T., and Ashraf, K. (2017). Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*.

Jaeger, S., Candemir, S., Antani, S., Wang, Y.-X. J., Lu, P.-X., and Thoma, G. R. (2014). Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Mahbub, M. K., Biswas, M., Gaur, L., Alenezi, F., Alenezi, F., and Santosh, K. C. (2022). Deep features to detect pulmonary abnormalities in chest x-rays due to infectious diseasex: Covid-19, pneumonia, and tuberculosis. *Information Sciences*.

Munzlinger, C., Yepes, I., and Rieder, R. (2023). Uso de uma rede neural convolucional para análise de exames de radiografia de pulmão com detecção de covid-19, pneumonia e tuberculose. In *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 25–30, Porto Alegre, RS, Brasil. SBC.

Murphy, K., Habib, S. S., Zaidi, S. M. A., Khowaja, S., Khan, A., Melendez, J., Scholten, E. T., Amad, F., Schalekamp, S., Verhagen, M., et al. (2020). Computer aided detection of tuberculosis on chest radiographs: An evaluation of the cad4tb v6 system. *Scientific reports*, 10(1):5492.

Pande, T., Pai, M., Khan, F. A., and Denkinger, C. M. (2015). Use of chest radiography in the 22 highest tuberculosis burden countries. *European Respiratory Journal*.

Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., and Pfeiffer, D. (2019). Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific reports*, 9(1):6268.

Siracusano, G., La Corte, A., Gaeta, M., Cicero, G., Chiappini, M., and Finocchio, G. (2020). Pipeline for advanced contrast enhancement (pace) of chest x-ray in evaluating covid-19 patients by combining bidimensional empirical mode decomposition and contrast limited adaptive histogram equalization (clahe). *Sustainability*, 12(20):8573.

World Health Organization (2016). *Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches*. World Health Organization.

World Health Organization (2023). Global tuberculosis report.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Liu, Y., and Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers.

Yang, F., Lu, P. X., Deng, M., Wáng, Y. X. J., Rajaraman, S., Xue, Z., Folio, L. R., Antani, S. K., and Jaeger, S. (2022). Annotations of lung abnormalities in the shenzhen chest x-ray dataset for computer-aided screening of pulmonary diseases. *Data*, 7(7).