Ensembles of Convolutional Neural Networks on Computer-Aided Pulmonary Tuberculosis Detection

L. Coimbra, and E. Guedes

***Abstract*—Pulmonary tuberculosis is one of the top 10 causes of death in the world in which Latin America represents 3% of the worldwide incidence with a mortality rate of 7.3%. According to goals of World Health Organization to end the global tuberculosis epidemy by 2030, the development of novel diagnosis strategies is crucial. Taking this context into account, the present work shows results on the use of ensembles of Convolutional Neural Networks to aid Computer Vision diagnosis of tuberculosis from patients' chest X-ray. Our solution comprises an ensemble of three different deep architectures of such networks having accuracy higher than 93% on the proposed task. The solution was obtained and evaluated in real-world data from public datasets, favouring reproducibility, surpasses human experts performance reported by literature in these very same datasets, uses canonical architectures of convolutional neural networks, with and without Transfer Learning, and requires minimal effort on data preparation and no previous feature extraction.**

***Index Terms*—Pulmonary tuberculosis, Convolutional Neural Networks, Deep Learning.**

I. Introdução

A *Tuberculose Pulmonar* (TB) é uma doença infectocontagiosa que acomete 1/3 da população mundial em sua forma latente, sendo uma das 10 maiores causas de mortes no mundo, ultrapassando índices relativos à HIV/AIDS e Malária, segundo a Organização Mundial de Saúde (OMS). Em 2015, a América Latina representou cerca de 3% da carga mundial de TB, com mais de 268 mil casos novos estimados e mais de 18.500 óbitos, cujo principais países afetados foram Brasil, Peru, México e Haiti [1].

A OMS tem como meta principal acabar com a epidemia global de TB até 2030, diminuindo 90% das mortes ocasionadas por esta enfermidade e promovendo uma redução de 80% em novos casos. Para alcançar esta meta, tem-se como um dos pilares a pesquisa e a inovação com vistas a desenvolver métodos de diagnóstico mais eficientes, a criação de novos medicamentos e a elaboração de estratégias de tratamento e acompanhamento, colocando pacientes no centro dos serviços de diagnóstico e prevenção e consolidando políticas e sistemas de apoio com a partipação do governo, da sociedade e de entidades privadas [2].

Atualmente, o diagnóstico de TB é feito por meio do estudo da história de adoecimento do paciente e por exames clínicos, devendo ser confirmado por exames específicos, a exemplo da baciloscopia [3]. O exame de raio-X da região anteroposterior torácica costuma ser de grande importância na triagem dos pacientes, pois possui baixo custo, é não invasivo e possui alta sensibilidade para anomalias na região pulmonar, mas é sujeito à baixa especificidade, pois não há anomalias definitivas para caracterizar TB, e sua interpretação possui significativa variação interobservador, fatores que representam barreiras na adoção deste exame em maior escala [4].

O recente progresso das técnicas de *Deep Learning* para tarefas de Visão Computacional também têm alcançado a análise de imagens médicas, com diversas contribuições e muitas perspectivas interessantes [5], como, por exemplo, no diagnóstico de lesões pigmentadas de pele [6]. No caso da Radiologia, em especial, tais técnicas têm colaborado em tarefas de detecção de lesão ou doença, classificação e diagnóstico, segmentação e quantização, especialmente nos âmbitos da imagiologia cardiotorácica e mamária [7], [8]. As *Redes Neurais Convolucionais* (CNNs, do inglês *Convolutional Neural Networks*) sobressaem-se como boas extratoras de características, altamente eficientes no reconhecimento e localização em imagens, o que tem feito diversos grupos de análises de imagens médicas considerar o seu uso junto à outras metodologias de *Deep Learning* para aplicação em diversos cenários de diagnóstico médico apoiado por computador [9].

Diante do exposto, este trabalho tem por objetivo apresentar os resultados da proposição e avaliação de comitês de CNNs aplicados à tarefa de Visão Computacional de classificar imagens radiográficas da região anteroposterior do tórax no tocante à presença ou ausência de TB. O treino e o teste de diversas arquiteturas canônicas bem como a agregação das mesmas mediante votação majoritária compõem a estratégia adotada neste trabalho para o problema em questão. Os resultados obtidos revelam uma acurácia de 93,82%, os quais foram corroborados por outras métricas de desempenho aferidas.

Para apresentar o que se propõe, este artigo está organizado como segue: um panorama da incidência de TB e desafios para a América Latina são expostos na Seção II. Os trabalhos relacionados são listados e discutidos na Seção III. A solução proposta e sua metodologia de obtenção é detalhada na Seção IV, na qual são apresentados os dados experimentais, a tarefa de Aprendizado de Máquina considerada e ainda os modelos propostos a serem treinados e testados. Os resultados e a discussão são mostrados na Seção V. Por fim, as considerações finais e perspectivas futuras encontram-se na Seção VI.

II. Tuberculose Pulmonar na América Latina: Panorama e Desafios

Apesar de um esforço global no combate à TB e da crescente melhoria da taxa de sucesso no tratamento de pacientes acometidos por esta enfermidade, a incidência de TB ainda caracteriza uma realidade alarmante em vários países. Sozinha, a América Latina concentra 3% de toda a ocorrência de TB

Lucas Gabriel Coimbra Evangelista, Elloá B. Guedes – Research Group on Intelligent Systems, Amazonas State University, Manaus, Amazonas, Brazil. E-mails: lgce.eng@uea.edu.br, ebgcosta@uea.edu.br.

no mundo, com uma taxa de mortalide de 7,3% [10]. São mais de 268 mil novos casos estimados por ano, nos quais Brasil (33%), Peru (14%), México (9%) e Haiti (8%) possuem maior incidência e juntos concentram mais da metade de toda ocorrência de TB na América Latina [11].

De acordo com dados da OMS, o Peru tem uma taxa de incidência de 116 indivíduos acometidos por TB por 100.000 habitantes, enquanto o Brasil tem uma taxa de 44, México de 22 e Haiti de 181 [12]. De modo geral, somente em 2017, a incidência de TB na América do Sul foi de 46,2 indivíduos por 100.000 habitantes, enquanto nos países caribenhos e demais países da América Central, incluindo México, esta taxa foi de 61,2 e 25,9 indivíduos por 100.000 habitantes, respectivamente [10]. Não há uma uniformidade na incidência de TB na América Latina, pois países como Jamaica, Costa Rica, República Dominica, Antígua e Barbuda registram baixa incidência, menor que 10 por 100.000 habitantes. Dos 15 países caribenhos, outros 12 possuem taxas similares [10], [12].

No geral, percebe-se que a incidência de TB na América Latina vêm reduzindo a uma taxa de 1,7% ao ano desde os anos 2000, mas sabe-se que esta redução encontra-se abaixo da meta anual necessária de 5,3% proposta pela OMS para erradicação da doença a partir de 2030 [13]. Em relação à taxa de mortalidade, a OMS almeja uma redução da mesma em 95% até 2035 e, embora a redução na taxa de mortalidade tenha melhorado em 2,5% entre os anos de 2000 a 2017, os esforços para obter uma taxa de melhoria anual de 12% até 2020 mostram-se difíceis de serem alcançados [14], o que demanda mais esforços no combate à TB. No âmbito latino-americano das pesquisas voltadas ao combate à TB, ainda citam-se dificuldades em identificar prioridades, a necessidade de fomentar trabalhos com dados epidemiológicos locais e a importância em construir uma rede científica latino-americana de pesquisa em TB [13].

III. Trabalhos Relacionados

Sistemas especialistas consistiram nas primeiras respostas adotadas para auxiliar no diagnóstico de TB. Nesta perspectiva, um sistema especialista foi previamente proposto com vistas a replicar as decisões de especialistas experientes no diagnóstico de TB mediante sintomas dos pacientes e sua intensidade, como febre, dor abdominal, lesões de pele, teste sanguíneo, entre outros [15]. Embora os resultados obtidos sejam relativamente recentes, essa abordagem é considerada superada pela visão atual da literatura, por ser fortemente dependente de intervenção humana, amostras clínicas e testes diagnósticos [16].

Os trabalhos de Elveren e Yumuşak [17], Er et al. [18] e de El-Solh e colaboradores [19] baseiam-se na utilização de redes neurais artificiais *multi-layer perceptron* para distinção de pacientes acometidos por TB de indivíduos saudáveis a partir da dados oriundos de testes laboratoriais e exames preliminares – presença de tosse, pressão alta, presença de líquido dos pulmões, quantidade de leucócitos no sangue, hemoglobina, nível do fosfatase alcalina, etc., diferenciando-se quanto ao modo de treinamento das redes (busca de parâmetros por algoritmos genéticos, por exemplo) e arquitetura utilizada. Os resultados apontam acurácia de 93,30% e 93,93%, respectivamente [18], [19]. No entanto, todos esses trabalhos apresentam limitações no tocante à necessidade de intervenção humana, custos envolvidos para a realização de exames, dentre outros.

Abordagens computacionais mais autônomas consideradas pela literatura levam em conta que utilizar imagens digitais para o diagnóstico de TB são uma opção mais fácil e mais barata, principalmente do ponto de vista da pouca intervenção humana no processo de preparação dos modelos e algoritmos. Desta forma, muitas abordagens de reconhecimento de padrões foram desenvolvidas, como: segmentação de pulmões, supressão óssea, detecção das vias aéras pulmonares, e extração de características relevantes [20]. A perspectiva do desenvolvimento de abordagens computacionais para apoio ao diagnóstico de TB a partir de imagens radiográficas é destacada positivamente pela OMS, especialmente levando em conta aspectos práticos, tais como a disponibilidade de equipamentos portáteis para raio-X, a baixa dosagem de radiação dispensada pelos equipamentos mais modernos e até mesmo as tecnologias de comunicação para transmissão eletrônica das imagens [4].

Em relação ao uso de técnicas de *Deep Learning*, Cao *et al.* propuseram a extração de regiões das imagens pulmonares utilizando uma busca seletiva seguida da classificação das mesmas com CNNs e da agregação final com uma Máquina de Vetores de Suporte de natureza linear [21]. Neste trabalho, a proposta é de que as imagens radiográficas sejam digitalizadas com o auxílio de dispositivos móveis, para facilitar o uso da solução em regiões com poucos recursos, e da posterior classificação multi-rótulo do tipo de manifestação sugestiva de TB. Uma avaliação da solução com dados oriundos de uma base de dados privada contendo 4.071 imagens foi conduzida pelos autores e os resultados mostraram uma precisão de 62,07% para a tarefa proposta.

Similarmente, Hwang et. al utilizaram 11 mil imagens de radiografias torácicas, coletadas a partir de três bases dados, para treinar e testar CNNs da arquitetura canônica AlexNet sujeitas à técnicas de Transferência de Aprendizado com pesos oriundos da base de dados ImageNet [22]. Este trabalho considerou o treino das CNNs a partir de cada base de dados individualmente, em que os melhores resultados alcançados foram aferidos para exemplos de teste provenientes do banco de imagens do Instituto Coreano de Tuberculose, com AUC igual a 0,967 e acurácia igual a 0,903.

Os autores Lakhani e Sundaram utilizaram quatro bases de dados supervisionadas, das quais uma não se encontra publicamente disponível, e adotaram arquiteturas canônicas de CNNs, como AlexNet e GooLeNet, e também testaram a melhoria das métricas de avaliação com e sem o uso de técnicas de aumento de dados e Transferência de Aprendizado [23]. Os melhores resultados foram alcançados com pesos pré-treinados e aumento de dados, em um modelo baseado em *ensemble* de duas CNNs, com AUC de 0,99 e acurácia igual a 98,9%. Esses resultados são importantes, mas os autores reforçam que não substituem a interpretação médico-radiológica humana.

Uma característica em comum verificada nestes trabalhos reside na preparação dos modelos com dados não amplamente disponíveis em bases de dados públicas e gratuitas. Este fator compromete a reprodutibilidade dos resultados produzidos por estes autores, dificulta a análise comparativa de resultados e obsta melhorias nas soluções propostas. Levando em conta tais aspectos, um dos objetivos deste trabalho consiste em contornar as dificuldades observadas baseando-se na preparação de CNNs com dados abertos e combinando-as com o uso de comitês.

IV. Solução Proposta

A solução proposta neste trabalho consiste em abordar a detecção de TB a partir de imagens de raio-X da região do tórax como uma tarefa de classificação mediante aprendizado supervisionado. Os dados experimentais utilizados, a preparação dos mesmos, métricas de desempenho, modelos e técnicas considerados são descritos detalhadamente nas subseções a seguir.

A. Dados Experimentais: Aquisição e Preparação

Inicialmente, foi feito um levantamento de bases de dados públicas e gratuitas de imagens de raio-X torácicas acompanhadas de rotulações indicativas da presença e ausência de TB oriundas de profissionais da área de Saúde. Deste levantamento, embora nenhum resultado tenha sido concernente à América Latina, as seguintes bases de dados satisfizeram os critérios elencados:

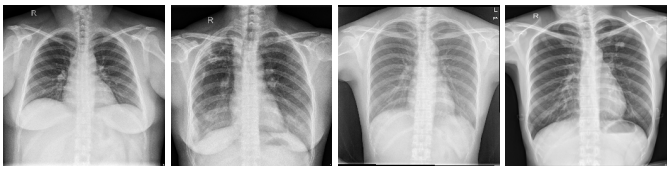
1) **Conjunto Montgomery County X-Ray**. Esta base de dados contém 138 imagens totais, das quais 58 são positivas para TB e as demais são oriundas de indivíduos saudáveis. Essas imagens foram adquiridas pelo programa de controle de TB do Departamento de Saúde e Serviços Humanos do Município de Montgomery, nos Estados Unidos. As imagens desta base de dados possuem dimensões de 4.020×4.892 pixels [24];
2) **Conjunto Shenzhen Hospital X-Ray**. Contém 662 imagens, algumas das quais são pediátricas, sendo 336 representativas de TB e as demais de indivíduos saudáveis. Estas imagens foram adquiridas durante um mês de rotina hospitalar em um estabelecimento médico na China e possuem dimensões variadas, mas em torno de 3.000×3.000 pixels [24]; e
3) **Base de dados JSRT**. Criada pela Sociedade Japonesa de Tecnologia Radiológica (JSRT), contém 154 exemplos sugestivos de nódulos pulmonares malignos e benignos e outros 93 exemplos oriundos de pacientes saudáveis, sem a indicação de nódulos. Estas imagens, com dimensões de 2048×2048 pixels, foram coletadas em 13 centros médicos japoneses ao longo de dois anos com vistas a fomentar pesquisas com a utilização de imagens médicas [25].

A base de dados utilizada neste trabalho foi um produto da reunião das bases de dados supracitadas, mas com algumas considerações específicas. Primeiramente, a partir de uma análise dos metadados disponíveis na base de dados JSRT, foi possível verificar que nem todas as radiografias nela contidas e com indicação de nódulos eram sugestivas de TB. Assim, desta base de dados foram selecionados apenas os 93 exemplos de imagens indicativas de indivíduous saudáveis. Após esta seleção, excetuando-se a classificação indicativa de presença ou ausência de TB, todos os demais metadados foram descartados. Por último, mediante as diferentes dimensões das imagens ao longo das bases de dados consideradas, foi realizada uma padronização mediante redimensionamento para o tamanho 256×256 pixels, seguindo sugestões disponíveis na literatura e considerando uma diminuição na sobrecarga dos custos de tempo e recursos computacionais tipicamente requeridos por CNNs. A Tabela I sintetiza a quantidade de exemplos na base de dados consolidada.

TABELA I
Base de dados consolidada: origem dos exemplos e distribuição por classe

	Exemplos Positivos	**Exemplos Negativos**
Montgomery County X-Ray	58	80
Shenzhen Hospital X-Ray	336	326
JSRT	0	93
Total	394	499
Percentual	44,12%	55,88%

É relevante salientar que a base de dados consolidada encontra-se desbalanceada, o que ensejará a posterior utilização de métricas de desempenho compatíveis. Ademais, é importante ressaltar que esta base de dados é realística para o problema em questão, pois possui exemplos de manifestações de TB incidentes em diferentes regiões geográficas, com múltiplos graus de intensidade, envolvendo indivíduos de ambos os sexos, com diferentes idades e características físicas. A Figura 1 ilustra alguns exemplos presentes na base de dados consolidada.



(a) F, Negativo (b) F, Positivo (c) M, Negativo (d) M, Positivo

Fig. 1. Amostras de exemplos presentes na base de dados com indicação de sexo (M denotando Masculino, e F denotando Feminino) e presença (positivo) ou ausência (negativo) de TB.

Considerando a tarefa de Aprendizado de Máquina abordada, foi efetuada uma partição aleatória dos dados com vistas à realização de validação cruzada do tipo *holdout*. Nesta etapa, definiu-se uma divisão de 70% dos dados para treinamento dos modelos, 10% para validação e 20% para testes, em que nesta última partição foram aferidas as métricas de desempenho dos modelos.

Com vistas a evitar *overfitting* nas CNNs a serem propostas e também visando uma melhor generalização dos modelos, considerou-se também a utilização de técnicas de *data augmentation*. Esta abordagem promove a geração de mais dados de treinamento a partir dos exemplos já disponíveis neste conjunto por meio do aumento artificial de amostras

segundo transformações aleatórias que resultam em imagens verossímeis [26]. Assim, considerou-se o treinamento dos modelos mediante três conjuntos de treinamento distintos:

1) **Abordagem Original**. Considera apenas os exemplos disponíveis após a partição *holdout* sem quaisquer alterações;
2) **Abordagem de Aumento Abrupto**. Considera operações de espelhamento vertical, rotação em $90°$ ou $270°$, e também rotações em ângulos escolhidos aleatoriamente entre $0°$ e $10°$ para a esquerda e para a direita. Tais transformações foram aplicadas até que o conjunto de treinamento passasse a ser composto por 10.000 imagens;
3) **Abordagem de Aumento Suave**. Considera apenas modificações mais sutis nas imagens originais com rotações em ângulos escolhidos aleatoriamente entre $0°$ e $10°$ para a esquerda e para a direita. O mesmo quantitativo de 10.000 imagens resultantes para treinamento foi adotado.

É importante salientar que nas abordagens que fizeram uso de *data augmentation* não foram consideradas operações que promovessem a ampliação ou recorte seletivos de partes das imagens, sob o risco de descartar alguma característica determinante para a detecção de TB, e nem tampouco adotou-se uma estratégia típica de espelhamento horizontal, pois esta viola a hipótese de simetria horizontal que pode resultar na não detecção das condições de *dextrocardia isolada* ou *situs inversus*, nas quais há transposição horizontal do coração ou de diversas vísceras torácicas e abdominais, respectivamente [27]. Ressalta-se ainda que o aumento de dados segundo as duas abordagens propostas restringiu-se apenas aos dados de treinamento, não tendo sido aplicados aos dados de validação e nem tampouco aos dados de teste.

B. Modelos, Parâmetros e Hiperparâmetros

As CNNs são inspiradas na organização do córtex visual de animais e possuem uma topologia hierárquica composta de camadas convolucionais, de amostragem e completamente conectadas. A convolução substitui a multiplicação de matrizes presente nas redes neurais artificiais *multilayer perceptron* e colabora para a captura de relações espaciais da entrada, o que, em especial, favorece o bom desempenho deste modelo em problemas de Visão Computacional. Para o ajuste de pesos em CNNs mediante tarefas de Aprendizado Supervisionado, aplica-se o tradicional algoritmo *backpropagation*, no qual três fatores caracterizam-se como cruciais para promover um bom aprendizado: a interação esparsa, o compartilhamento de parâmetros e a representação equivariante [28].

O sucesso do aprendizado de CNNs em um problema é sensível à uma arquitetura que se ajuste bem ao mesmo. No processo de encontrar esta arquitetura, porém, várias decisões de projeto são necessárias e envolvem desde a determinação do número de camadas e seus tipos, a ordenação das mesmas e até a escolha de hiperparâmetros, uma tarefa desafiadora, que demanda tempo e exige conhecimento altamente especializado, especialmente em virtude do grande número de opções possíveis dentre as escolhas a serem feitas. Há também os aspectos de custos computacionais para avaliar as arquiteturas propostas e ainda os *tradeoffs* de viés e variância a considerar mediante o domínio do problema.

Para contornar a dificuldade previamente mencionada em projetar arquiteturas de CNNs para a tarefa de classificação binária de detecção de TB a partir das imagens radiográficas, optou-se por treinar, testar e avaliar o uso de arquiteturas de CNNs canônicas na literatura, em especial algumas das que obtiveram bom desempenho no desafio *ImageNet Large-Scale Visual Recognition Challenge* (ILSVRC) [29]. Levando isto em consideração, as seguintes arquiteturas foram elencadas [26], [30]: (i) *LeNet*, uma das primeiras arquiteturas propostas de CNNs aplicada inicialmente no reconhecimento de dígitos manuscritos [31]; (ii) *AlexNet*, uma CNN com maior quantidade de camadas profundas em relação às suas antecessoras e que faz uso de operações de regularização e *dropout* [32]; (iii) *Inception-v3*, uma topologia que considera ramificações internas, demanda pouca memória, é mais eficiente e geralmente resulta em alta acurácia [33]; (iv) *VGG*, a qual eleva o número de camadas profundas, entre 16 e 19, em conjunto com um grande número de convoluções com filtros pequenos, de tamanho 3×3, e, assim como a AlexNet, também utiliza *dropout* com vistas a evitar *overfitting* [34]; e (v) *ResNet*, cuja arquitetura introduz conexões residuais as quais reinjetam representações prévias no fluxo posterior de dados, o que ajuda na prevenção de perda de informação no decorrer do fluxo de processamento, culminando em melhoria no desempenho [35].

No âmbito hiperparâmetros, considerou-se a variação na escolha do otimizador para o *backpropagation*, dentre os algoritmos *Stochastic Gradient Descent* (SGD) e *Adaptive Moment Estimation* (Adam), os quais exploram diferentes estratégias iterativas para minimização do erro no gradiente descendente [36]. Ressalta-se que a taxa de aprendizado inicial foi definida como sendo igual a 10^{-3}. Além do otimizador, considerou-se ainda o número de épocas e o *patience*. Tomou-se o número de épocas como sendo igual a 100 mediante a abordagem original, em virtude do menor quantitativo da base de dados de treinamento, e como sendo igual a 200 para as duas estratégias de aumento de dados. No caso do *patience*, os valores de 15 e 35 foram adotados mediante a ausência e a presença de aumento de dados, respectivamente, com vistas a monitorar variações no decréscimo no desempenho do modelo durante a etapa de treinamento perante o conjunto de validação, promovendo uma parada precoce no aprendizado com vistas a evitar *overfitting*.

Uma vez estabelecidos os dados de treinamento e teste, bem como os modelos a serem usados e seus respectivos parâmetros e hiperparâmetros, observou-se que há modelos substancialmente mais profundos que outros, em especial, as arquiteturas GoogLeNet, VGG e ResNet. Assim, para estes casos, também foi explorado um maior número de épocas, igual a 500, além de *patience* iguais a 50 e 35. Estes valores foram obtidos de maneira experimental, conforme sugestões empíricas comumentes utilizadas na literatura [26].

C. Avaliação de Desempenho

Considerando o objetivo desta tarefa, duas métricas de perfomance foram utilizadas para avaliar o desempenho das

CNNs propostas perante o conjunto de testes: a acurácia e o F-Score. A acurácia é obtida como segue:

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (1)$$

em que TP, TN, FP e FN denotam, respectivamente, valores verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Esta métrica sintetiza o desempenho do modelo perante a proporção de classificações corretas dentre o total de classificações, fornecendo uma noção intuitiva do grau de acertos nas previsões. Porém, considerando que há um desbalanceamento entre as classes, o F_β-Score considera a média harmônica ponderada entre precisão e revocação, tal como segue:

$$F_\beta\text{-Score} = (1 + \beta^2)\frac{\text{Precisão} \cdot \text{Revocação}}{(\beta^2 \cdot \text{Precisão}) + \text{Revocação}}, \quad (2)$$

em que a precisão e a revocação são dadas por:

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3) \qquad \text{Revocação} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

É importante ressaltar que o F_β-Score considera a revocação como sendo β vezes mais importante que a precisão. Na avaliação de desempenho dos modelos será considerado $\beta = 1$, resultando no chamado F-Score, em que precisão e revocação possuem a mesma ponderação.

D. Transferência de Aprendizado e Comitês de Modelos

Além do uso de CNNs canônicas individualmente treinadas e avaliadas com dados do domínio do problema, também considerou-se adotar duas estratégias que costumam ser eficazes para obtenção de modelos de *Deep Learning* eficientes para problemas de Visão Computacional, a citar: a transferência de aprendizado e o uso de comitês de modelos.

A transferência de aprendizado é uma abordagem para adaptar e aplicar os conhecimentos adquiridos por um modelo em uma tarefa de origem em uma tarefa alvo [37]. Com a possibilidade de promover melhores pontos de partida de desempenho em comparação aos modelos dedicados, inclusive demandando menos dados para tanto, faz uso dos pesos de arquiteturas de CNNs pré-treinadas em grandes bases de dados, partindo do princípio de que a hierarquia espacial de características aprendidas por estes modelos nestas bases de dados massivas são genéricas o suficiente para outros problemas de Visão Computacional [26]. Porém, a depender das discrepâncias entre as características dos domínios e dos tipos de tarefa, tem-se dificuldades em alcançar transferência positiva de conhecimento [38].

No escopo deste trabalho, considerou-se a transferência de aprendizado mediante a técnica de extração de características nas CNNs a partir da base de dados Imagenet [29], pois esta última tem sido amplamente adotada como um *benchmark* para tarefas de Visão Computacional, possui uma quantidade significativa de exemplos, da ordem de 15 milhões de imagens coloridas distribuídas segundo cerca de mil classes distintas, e há na literatura estudos que sugerem que arquiteturas pré-treinadas com esta base generalizam bem em outras bases de dados [39].

Dado o ônus computacional de treinar as arquiteturas de CNNs elencadas previamente perante a base ImageNet, optou-se pela busca em *frameworks* por CNNs já pré-treinadas. Como resultado desta busca, apenas as arquiteturas VGG e Inception satisfizeram tais critérios e serão utilizadas na tarefa alvo mediante transferência de aprendizado. Levando isto em consideração, a primeira camada destas redes foi substituída por uma equivalente, mas que compatibilizasse a dimensão da entrada para 256×256 pixels, e mesmo em escala de cinza, as imagens da tarefa foram fornecidas no formato RGB, visando maximizar o aproveitamento dos pesos pré-treinados.

No caso dos comitês, uma estratégia oriunda do *Ensemble Learning*, a sua adoção teve em vista agregar as diferentes CNNs previamente propostas e combiná-las com o intuito de obter um classificador final cujo desempenho pudesse superar o dos modelos individuais para o mesmo fim [40]. Os relatos na literatura que mostram a eficiência deste tipo de combinação em diferentes problemas foram o principal motivador para a adotar a referida estratégia [41].

Tendo em mente que a composição de bons comitês contempla aspectos de diversidade, com vistas a indução de viéses distintos a serem mitigados perante etapa de agregação [42], considerou-se a proposição de três tipos de comitês distintos, todos com decisão baseada em votação por maioria:

1) **Comitê 1**. Comitê compreendido pelas três CNNs com melhor desempenho individual;
2) **Comitê 2**. Comitê composto apenas das três CNNs com arquiteturas mais profundas e melhor desempenho; e, por fim,
3) **Comitê 3**. Composto por cinco CNNs com melhor desempenho, porém com arquiteturas distintas.

Uma vez que foram estabelecidos todos os critérios para preparação dos dados, implementação, avaliação, transferência de aprendizado e combinação das CNNs, partiu-se para a etapa de implementação. A plataforma Python teve papel central na codificação das metodologia proposta, com destaque para os *frameworks* Keras e Sci-Kit Learn. O *hardware* computacional utilizado constitiu de um computador *desktop* com processador Intel Core i7, 16 GB de memória principal e 1 TB de memória secundária, em que utilizou-se aceleração em *hardware* no treinamento com duas placas gráficas NVidia GTX 1080 com 11 GB cada. O tempo aproximado de treinamento e teste da solução proposta consistiu em 25 h de execução ininterruptas.

V. Resultados e Discussão

Considerando a metodologia para obtenção da solução proposta, as arquiteturas de CNNs selecionadas foram treinadas individualmente conforme especificado e os resultados de suas avaliações encontram-se dispostos na Tabela II, agrupados conforme a abordagem relativa à ausência ou presença de aumento de dados e seu respectivo tipo, neste último caso. As siglas TP, FP, FN e TN possuem mesma semântica da Eq. (1).

De maneira geral, observa-se que o comportamento das CNNs não foi homogêneo sequer perante uma mesma arquitetura nem tampouco no âmbito da abordagem utilizada,

TABELA II
RESULTADO DA AVALIAÇÃO DE CNNS INDIVIDUAIS

CNN	Otimizador	TP	FP	FN	TN	Acurácia	F-Score
LeNet	SGD	61	11	13	93	86,52%	0,8356
LeNet	Adam	74	104	0	0	41,57%	0,5873
AlexNet	SGD	59	9	15	95	86,52%	0,8310
AlexNet	Adam	43	9	31	95	77,53%	0,6825
Inception	SGD	38	4	36	100	77,52%	0,6552
Inception	Adam	72	63	2	41	63,48%	0,6890
VGG16	SGD	0	0	74	104	58,42%	0,0000
VGG16	Adam	55	12	16	92	84,00%	0,7971
ResNet	SGD	62	14	12	90	85,39%	0,8539
ResNet	Adam	61	12	13	92	85,08%	0,8212

(a) Resultados para Abordagem Original.

CNN	Otimizador	TP	FP	FN	TN	Acurácia	F-Score
LeNet	SGD	59	17	15	87	82,02%	0,7867
LeNet	Adam	74	104	0	0	41,57%	0,5873
AlexNet	SGD	64	19	10	85	83,71%	0,8153
AlexNet	Adam	54	14	20	90	80,90%	0,7606
Inception	SGD	61	21	13	83	80,90%	0,7821
Inception	Adam	63	7	11	97	89,89%	0,8750
VGG16	SGD	39	39	35	65	58,42%	0,5132
VGG16	Adam	49	18	25	86	75,84%	0,6950
ResNet	SGD	63	33	11	71	75,28%	0,7412
ResNet	Adam	52	9	22	95	82,58%	0,7704

(b) Resultados para Abordagem de Aumento Abrupto.

CNN	Otimizador	TP	FP	FN	TN	Acurácia	F-Score
LeNet	SGD	54	11	20	93	82,58%	0,7770
LeNet	Adam	0	0	74	104	58,42%	0,0000
AlexNet	SGD	55	11	19	93	83,15%	0,7857
AlexNet	Adam	51	6	23	98	83,71%	0,7786
Inception	SGD	60	8	14	96	87,64%	0,8451
Inception	Adam	65	10	9	94	89,32%	0,8725
VGG16	SGD	54	7	20	97	84,83%	0,8000
VGG16	Adam	58	10	16	94	85,39%	0,8169
ResNet	SGD	61	9	13	95	87,64%	0,8472
ResNet	Adam	63	11	11	93	87,64%	0,8514

(c) Resultados para Abordagem de Aumento Suave.

fatores que ressaltam a importância na proposição e exploração dos diversos cenários distintos considerados na concepção da solução proposta. O desempenho de algumas CNNs, em especial, foi contraproducente no aprendizado da tarefa de classificação de TB, resultando em performance risível. Um dos fatores que pode justificar tal fato é o desvanecimento ou a explosão do gradiente descendente durante a etapa de treinamento [43]. Nesta etapa, o melhor desempenho segundo *F-Score* e acurácia foi observado para a arquitetura Inception mediante a abordagem de aumento abrupto. Nota-se que nenhuma CNN individual superou o patamar de 90% na performance de ambas métricas examinadas.

Destes primeiros resultados obtidos, nota-se que a abordagem suave foi a que resultou na menor média de valores falsos positivos ($8{,}3 \pm 3{,}2$) quando comparada às abordagens original ($23{,}8 \pm 31{,}4$) e de aumento abrupto ($28{,}1 \pm 26{,}9$). Acredita-se que o adicional de imagens gerado artificialmente e fornecido durante o treino tenha permitido à rede um melhor aprendizado das características relativas à estrutura do tórax, propiciando o reconhecimento das estruturas típicas. Por outro lado, não observou-se diferença estatística desta abordagem quando comparada às demais no tocante aos falsos negativos, o que evidencia que as abordagens são equivalentes no tocante às falhas de diagnóstico dos casos positivos. Ao examinar a dispersão dos valores de F-Score agrupados por arquitetura das CNNs independentemente do tipo de abordagem de treino, conforme Fig. 2, percebe-se que, até então, nenhuma das arquiteturas se sobressai das demais no tocante ao desempenho aferido por tal métrica.

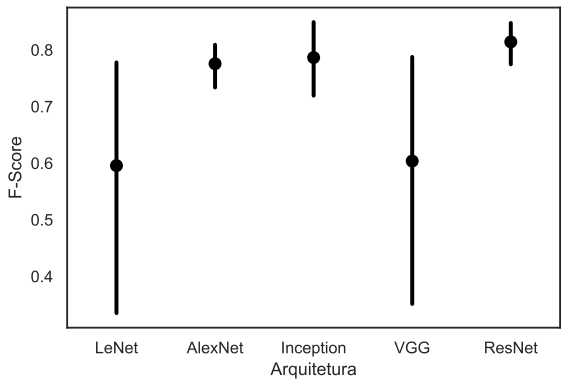


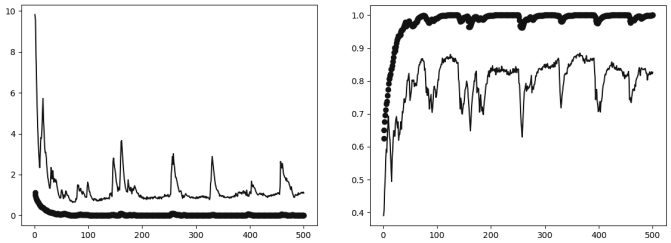
Fig. 2. Gráfico da dispersão do F-Score das CNNs agrupadas quanto à arquitetura, independentemente do tipo de abordagem adotada para treinamento.

Em seguida, tendo em mente avaliar o desempenho das CNNs com maior número de camadas profundas perante a tarefa de classificação de TB sujeitas à um treinamento com um número de épocas substancialmente maior, obteve-se os resultados ilustrados na Tabela III, na qual encontram-se sintetizados apenas os 5 modelos que obtiveram melhor desempenho. Excetuando-se pela CNN ResNet com acurácia de $87{,}08\%$ para a referida tarefa e que utilizou o otimizador SGD, todos os demais resultados foram obtidos utilizando o otimizador Adam. É possível observar que um maior número de épocas propiciou um melhor aprendizado por estas arquiteturas profundas, independentemente da abordagem adotada, o que é corroborado por uma menor média de resultados falsos positivos ($6{,}6 \pm 1{,}0$) e também de falsos negativos ($13{,}4 \pm 3{,}3$) quando comparado aos cenários anteriores.

TABELA III
CINCO MELHORES RESULTADOS DA AVALIAÇÃO DE CNNS INDIVIDUAIS PROFUNDAS TREINADAS COM 500 ÉPOCAS

CNN	Abordagem	TP	FP	FN	TN	Acurácia	F-Score
Inception	Original	67	8	7	96	91,57%	0,8993
Inception	Suave	60	5	14	99	89,33%	0,8633
ResNet	Suave	60	7	14	97	88,20%	0,8511
Inception	Abrupto	58	6	16	98	87,64%	0,8406
ResNet	Suave	58	7	16	97	87,08%	0,8345

Observando o melhor desempenho registrado dentre as CNNs individuais treinadas especificamente para a tarefa de classificação de TB, ressalta-se a CNN Inception sem aumento de dados e sujeita à um treinamento com 500 épocas. Ao monitorar o aprendizado da mesma, percebeu-se a importância de usar uma estratégia de validação do treinamento e parada antecipada, pois houve recorrente tendência de superajustamento aos dados de treino, conforme ilustrado na Figura 3.



(a) Perda (b) Acurácia

Fig. 3. Gráficos do histórico do treinamento ao longo das épocas da CNN Inception para monitoramento da perda e acurácia nos conjuntos de treinamento (denotado de forma pontilhada) e validação (denotado por traço contínuo).

No tocante à transferência de aprendizado, observou-se uma transferência positiva em todas as arquiteturas consideradas com pesos oriundos da ImageNet, conforme Tabela IV. Nota-se que o patamar de desempenho das CNNs de arquitetura VGG foi melhorado em decorrência da transferência de aprendizado quando comparado ao aprendizado especializado no problema da TB. É interessante notar que, na média, a transferência de aprendizado promoveu uma significativa diminuição das previsões do tipo falso negativo ($8{,}5 \pm 2{,}0$), porém, não corroborou para a diminuição média dos erros do tipo falso positivo, possuindo o maior valor médio dentre todas as estratégias avaliadas até então ($38{,}5 \pm 21{,}5$). Uma interpretação possível para tal resultado é que os pesos oriundos do ImageNet tenham sido úteis em capturar casos mais particulares e específicos da TB mas que, em compensação, resultou em perda no refinamento de reconhecer casos saudáveis.

TABELA IV
RESULTADOS PARA AS CNNS SUJEITAS A TRANSFERÊNCIA DE APRENDIZADO

CNN	Origem	Otimizador	TP	FP	FN	TN	Acurácia	F-Score
VGG16	ImageNet	Adam	63	13	9	93	87,64%	0,8514
VGG16	ImageNet	SGD	54	22	5	97	84,83%	0,8000
Inception	ImageNet	SGD	22	54	10	92	64,04%	0,4074
Inception	ImageNet	Adam	11	65	10	92	57,87%	0,2268

Por fim, houve então a construção dos três comitês propostos, cujo desempenho encontra-se sintetizado na Tabela V. Ressalta-se que, para tanto, não houve necessidade de novos treinamentos nas CNNs já existentes, o que é um aspecto positivo no tocante aos custos computacionais envolvidos. Para uma melhor entendimento dos comitês, as estatísticas de desempenho das CNNs que os integram foram repetidas como referência.

A obtenção de tais comitês sintetiza a solução proposta no âmbito do presente trabalho, em que os melhores resultados foram observados no Comitê 2 no tocante à acurácia e ao F-Score. É interessante notar que os Comitês 1 e 3 possuem desempenho resultante inferior até mesmo do que algumas das CNNs que os compõem, o que não resulta em mitigação dos resultados errôneos por meio da votação majoritária. Assim, o Comitê 2 é a solução de referência para a abordagem proposta de detecção de tuberculose a partir de imagens de raio-X no escopo deste trabalho.

TABELA V
RESULTADO PARA OS COMITÊS DE MODELOS

CNN	TP	FP	FN	TN	Acurácia	F-Score
Inception	67	8	7	96	91,57%	0,8993
Inception	63	7	11	97	89,89%	0,8570
Inception	60	5	14	99	89,33%	0,8633
Comitê 1	56	18	10	94	84,27%	0,8000

(a) Comitê com três melhores CNNs.

CNN	TP	FP	FN	TN	Acurácia	F-Score
Inception	67	8	7	96	91,57%	0,8993
ResNet	60	7	14	97	88,20%	0,8511
VGG	63	13	9	93	87,64%	0,8514
Comitê 2	68	6	5	99	93,82%	0,9252

(b) Comitê com três melhores CNNs de arquiteturas profundas.

CNN	TP	FP	FN	TN	Acurácia	F-Score
Inception	67	8	7	96	91,57%	0,8993
ResNet	60	7	14	97	88,20%	0,8511
VGG	63	13	9	93	87,64%	0,8514
AlexNet	59	9	15	95	86,52%	0,8310
LeNet	61	11	13	93	86,51%	0,8356
Comitê 3	56	18	3	101	88,20%	0,8421

(c) Comitê com cinco melhores CNNs de arquiteturas distintas.

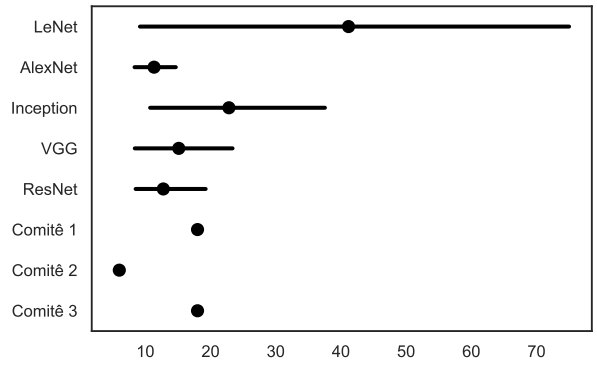
Com vistas a enfatizar a relevância da solução proposta no contexto da detecção de TB, partiu-se então para uma etapa posterior de validação da solução de referência. Nesta validação, os valores de FP e FN foram analisados e também outras métricas de desempenho foram coletadas tanto para o Comitê 2 quanto para as CNNs individuais que o integram.

No tocante aos resultados incorretos das classificações, correspondentes aos valores FPs e FNs, as Figuras 4a e 4b sintetizam a distribuição dos mesmos, respectivamente, segundo as arquiteturas avaliadas e também perante os comitês produzidos. A partir da observação de tais erros, percebe-se que a LeNet foi a menos homogênea em seus resultados. O Comitê 1 possui um desempenho equiparável à uma CNN dentre as avaliadas, pois os valores de FP e FN encontram-se no intervalo de erro obtido pelas arquiteturas individuais avaliadas. O Comitê 3, embora possua o menor valor FN observado, não foi efetivo em mitigar os erros do tipo FP.

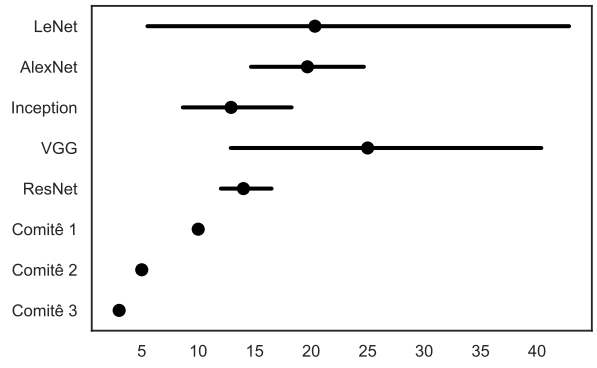
Ao analisar a relação entre FP e FN e o impacto destes erros na acurácia, uma das métricas de referência adotadas para obtenção da solução de referência, percebe-se que o Comitê 2 foi, de fato, mais efetivo na redução de ambas perante o conjunto de testes, como pode ser visto na Figura 5.

A Tabela VI sintetiza as demais métricas obtidas para o Comitê 2. A precisão, dada na Eq. (3), é uma medida de quão relevante um resultado positivo é. A revocação, obtida como mostrado na Eq. (4), diz respeito à proporção de resultados corretos positivos obtidos. Utilizando os resultados obtidos para os dados de teste da Tabela V(b), observa-se que ambas as métricas foram superiores a 90%.

A partir das probabilidades de classificação do Comitê 2 foi obtida a curva ROC ilustrada na Fig. 6, na qual ressalta-se



(a) Falso Positivo



(b) Falso Negativo

Fig. 4. Gráficos da dispersão dos valores FP e FN agrupados por arquitetura de CNN e também relativos aos comitês.

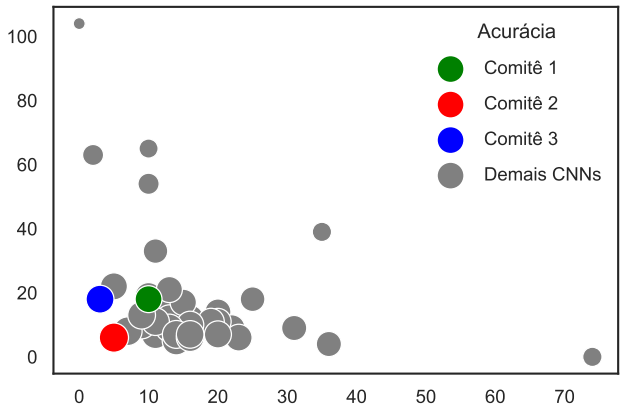


Fig. 5. Gráfico de dispersão dos valores FN (eixo horizontal) versus valores FP (eixo vertical). Cada ponto cinza corresponde à uma CNN, enquanto os pontos coloridos referem-se aos comitês. A área dos pontos é proporcional à acurácia na etapa de testes.

que a área sob a curva dos modelos individuais é menor que a área sob a curva do Comitê 2 obtido a partir da combinação destas.

Também foi calculada a área sob a curva da precisão-revocação, a qual ressalta o bom desempenho do comitê obtido perante uma base de dados desbalanceada. Menciona-se ainda a especificidade, a habilidade do comitê em lidar com observações negativas, a qual mostrou-se superior a 90%. Em relação aos erros do comitê, capturados pelas taxas de

TABELA VI
OUTRAS MÉTRICAS DE DESEMPENHO DO COMITÊ 2

Métrica	Inception	ResNet	VGG	Comitê 2
Precisão	0.89333	0.89552	0.84722	0.93150
Revocação	0.90540	0.81081	0.82432	0.91891
AUC ROC	0.95556	0.92587	0.92658	0.95601
Precisão-Revocação AUC	0.95576	0.88587	0.91258	0.92768
Especificidade	0.92307	0.93269	0.89423	0.95192
Taxa de Falso Positivo	0.07692	0.06730	0.10576	0.04807
Taxa de Falso Negativo	0.09459	0.18918	0.17567	0.08108
Taxa de Falsa Descoberta	0.10666	0.10447	0.15277	0.06849
F_2-Score	0.90296	0.82644	0.82880	0.92140
Coeficiente Kappa de Cohen	0.82687	0.75378	0.72136	0.87254
Coeficiente de Correlação de Matthews	0.82692	0.75633	0.72156	0.87260

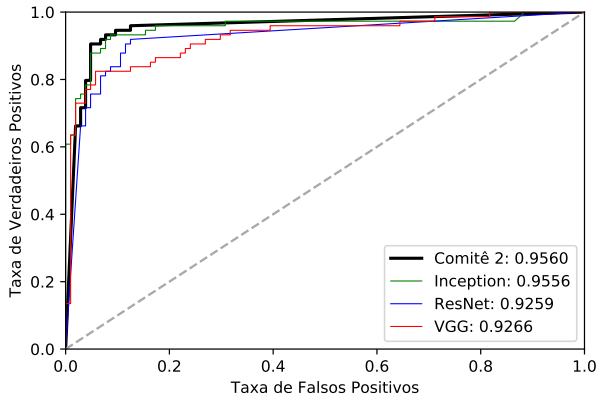


Fig. 6. Curva ROC para o Comitê 2 e CNNs integrantes.

falsos positivos, negativos e de descoberta, todas mostraram-se inferior a 10%.

As métricas remanescentes na Tabela VI provêem uma avaliação mais detalhada do comitê proposto. O F_2-Score, obtido conforme Eq. (2) com $\beta = 2$, afere o classificador com o dobro de penalidade ao classificar exemplos positivos como negativos em comparação quando $\beta = 1$. O Coeficiente Kappa de Cohen sumariza quão bem o comitê é quando comparado a um classificador aleatório que se baseia apenas na frequência de classes e, neste cenário, pode ser interpretado como sendo forte [44]. Por fim, o Coeficiente de Correlação de Matthews captura a correlação entre as classes previstas e as observadas.

Ainda no que diz respeito ao Comitê 2, partiu-se para o cálculo de métricas de diversidade do mesmo [45]. A primeira delas, a medida de dupla falha (do inglês, *double-fault measurement*), denotada por d, afere o quantitativo de erros de classificação cometidos por todos os integrantes do comitê para uma dada entrada. Neste caso, o valor observado de d foi igual a zero. A estatística Q sinaliza, quando positiva, a tendência dos classificadores integrantes de um comitê produzirem as mesmas saídas. A média da estatística Q sobre todos os pares de classificadores do Comitê 2 foi igual a -0.65366, mostrando uma boa diversidade na solução proposta, visto que esta métrica reside no intervalo $[-1,1]$. Por fim, o coeficiente Kappa de Fleiss, que afere a confiabilidade

da concordância entre os integrantes do comitê foi igual a 0.64129, que possui interpretação sugerida de concordância substancial. Tais métricas possuem coerência com o desempenho obtido pelo Comitê 2, pois, em especial, valores baixos da medida de dupla falha e da estatística Q são normalmente associados a uma alta acurácia [46].

Em relação aos trabalhos análogos disponíveis na literatura, Jaeger et al. propuseram uma solução para diagnosticar TB a partir de segmentação pulmonar e extração de características relevantes com modelos de Aprendizagem Máquina, no qual obteve acurácia de $78{,}3\%$ e 84% utilizando as bases de dados dos conjuntos Montgomery County e Shenzhen Hospital, respectivamente, ultrapassando a performance de especialistas humanos [24]. O Comitê 2 proposto supera a melhor acurácia deste trabalho relacionado em um percentual de $11{,}7\%$ o que, por conseguinte, também supera o desempenho de especialistas humanos sob as condições propostas por estes autores.

Como citado na Seção III, Lakhani e Sundaram obtiveram uma acurácia de $98{,}9\%$ nesta tarefa, mas reportaram uso de uma base de dados adicional e privada, oriunda da Universidade Thomas Jefferson [23]. Uma comparação equânime dos resultados aqui apresentados com este trabalho relacionado torna-se não trivial, pois os exemplos particulares podem ter fornecido novas características cruciais para a melhoria na classificação geral. Apesar disso, ressalta-se que o presente trabalho baseia-se inteiramente em bases de dados públicas e gratuitas e em arquiteturas canônicas de CNNs, o que favorece a sua reprodutibilidade.

VI. Considerações Finais

Com vistas a colaborar com o desenvolvimento de uma estratégia que auxilie na detecção de TB a partir de imagens radiográficas do tórax, alinhada com a demanda por métodos de diagnóstico mais eficientes que contribuam com as metas da OMS para erradicar a epidemia global desta enfermidade até 2030, especialmente no âmbito da América Latina, o presente trabalho propôs e avaliou a utilização de CNNs na referida tarefa. Embora não tenha se baseado em casos típicos de TB na América Latina em virtude da ausência de bases de dados públicas e gratuitas para este fim, considerou padrões disponíveis em bases de dados abertas contendo diversos graus desta enfermidade em pacientes de ambos os sexos e de diferentes idades. Para detecção dos padrões relativos à TB, considerou a utilização de diversas arquiteturas profundas de CNNs sujeitas a diferentes hiperparâmetros, estratégias de treinamento (individual e mediante transferência de aprendizado) e de combinações em comitês com agregação por votação majoritária. Os resultados obtidos ressaltam um comitê de três modelos compostos pelas arquiteturas Inception, ResNet e VGG como sendo mais adequado para esta tarefa, resultando em acurácia de $93{,}82\%$ e F-*Score* igual a $0{,}9252$.

É interessante ressaltar que este resultado reforça a não-trivialidade na proposição de CNNs para esta tarefa de Visão Computacional no âmbito do diagnóstico médico, em que dados oriundos de pacientes reais podem ser advindos de diferentes equipamentos de captura, sujeitos a variações fisiológicas dos indivíduos e do grau de acometimento da doença, aliados à ausência de características *sui generis* de TB. Embora o melhor desempenho aqui obtido seja comparável ao estado da arte, não se tem a pretensão de que os resultados aqui apresentados venham a substituir a interpretação médico-radiológica humana. Outros trabalhos que avaliem os riscos e benefícios do uso destes modelos precisam ser cuidadosamente elaborados e avaliados. Menciona-se também a pouca quantidade de exemplos disponíveis na base de dados consolidada, a qual mostra-se um limitador na aferição da solução proposta para fins práticos, especialmente em termos estatísticos.

Em trabalhos futuros, sugere-se conduzir avaliações mais robustas na solução proposta, a exemplo de validação cruzada, com vistas a melhor circunscrever vantagens e limitações. Ademais, seria de extrema importância avaliar o desempenho do Comitê 2 em casos de TB endêmicos da América Latina, visando aferir se as características aprendidas por este mostram-se relevantes para as manifestações locais desta enfermidade. Neste sentido, ressalta-se a importância de consolidar bases de dados de imagens médicas desta região para este tipo de tarefa. Sugere-se ainda a avaliação de outras bases de dados para transferência de aprendizado em CNNs com vistas a alavancar o desempenho na tarefa de detecção de TB, obtendo melhores resultados com menos esforço computacional.

Agradecimentos

Os autores agradecem o apoio financeiro e material provido pela Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) por meio do Edital PPP 04/2017. Os autores agradecem o apoio e as sugestões providas por Thaís Ferreira, Carlos Maurício Figueiredo e Marcos Filipe Salame.

Referências

[1] OMS, "Global Tuberculosis Report 2017," Geneva, Switzerland, 2017, available on http://www.who.int/tb/publications/global_report/en/. Acessed February 18, 2020.

[2] ——, "The end TB strategy – global strategy and targets for tuberculosis prevention, care and control after 2015," Geneva, Switzerland, 2014, available on https://www.who.int/tb/strategy/End_TB_Strategy.pdf?ua=1. Acessed February 18, 2020.

[3] D. Heemskerk, M. Caws, B. Marais, and J. Farrar, *Tuberculosis in Adults and Children*, 1st ed., ser. SpringerBriefs in Public Health 2. New York, NY, USA: Springer International Publishing, 2015.

[4] OMS, *Chest Radiography in Tuberculosis detection – Summary of current WHO recommendations and guidance on programmatic approaches*. Geneva, Switzerland: World Health Organization, 2016.

[5] J. G. Lee, S. Jun, Y. W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: General overview," *Korean J. Radiol.*, vol. 18, no. 4, pp. 570–584, 2017.

[6] J. Lima, L. Araújo, F. Silva, and C. Figueiredo, "Pigmented dermatological lesions classification using convolutional neural networks ensemble mediated by multilayer perceptron network," *IEEE Latin American Transactions*, vol. 17, no. 11, pp. 1902–1908, 2019.

[7] M. P. McBee, O. A. Awan, A. T. Colucci, C. W. Ghobadi, N. Kadom, A. P. Kansagra, S. Tridandapani, and W. F. Auffermann, "Deep Learning in radiology," *Academic radiology*, vol. 25, no. 11, pp. 1472–1480, 2018.

[8] L. Lu, Y. Zheng, G. Carneiro, and L. Yang, *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Switzerland: Springer, 2017.

[9] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial Deep Learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.
[10] M. Woodman, I. L. Haeusler, and L. Grandjean, "Tuberculosis genetic epidemiology: A latin american perspective," *Genes*, vol. 10, no. 1, p. 53, 2019.
[11] OMS, "Tuberculosis countries profiles," Geneva, Switzerland, 2018, available on https://extranet.who.int/sree/Reports?op=Replet&name=/WHO_HQ_Reports/G2/PROD/EXT/TBCountryProfile&ISO2=BR&outtype=pdf. Acessed February 18, 2020.
[12] Word Bank Group, "Incidence of tuberculosis (per 100,000 people)," Washington, D.C., USA, 2019, available on https://data.worldbank.org/indicator/SH.TBS.INCD?view=map. Acessed February 18, 2020.
[13] D. R. Silva, G. B. Migliori, and F. C. de Queiroz Mello, "Série Tuberculose 2019," *Jornal Brasileiro de Pneumologia*, vol. 45, no. 2, pp. e20 190 064–e20 190 064, 2019.
[14] OMS, "Tuberculosis in the Americas," Geneva, Switzerland, 2018, available on http://iris.paho.org/xmlui/bitstream/handle/123456789/49510/PAHOCDE18036_eng?sequence=1&isAllowed=y. Acessed February 18, 2020.
[15] A. A. Imianvan and J. Obi, "Fuzzy cluster means expert system for the diagnosis of tuberculosis," *Global Journal of Computer Science & Technology*, vol. 11, no. 6, pp. 41–48, 2011.
[16] S. Russell and P. Norvig, *Artificial Intelligence – A Modern Approach*, 3rd ed. Nova Jersey: Prentice Hall, 2010.
[17] E. Elveren and N. Yumuşak, "Tuberculosis disease diagnosis using artificial neural network trained with genetic algorithm," *Journal of medical systems*, vol. 35, no. 3, pp. 329–332, 2011.
[18] O. Er, F. Temurtas, and A. Ç. Tanrıkulu, "Tuberculosis disease diagnosis using artificial neural networks," *Journal of medical systems*, vol. 34, no. 3, pp. 299–302, 2010.
[19] A. A. El-Solh, C.-B. Hsiao, S. Goodnough, J. Serghani, and B. J. B. Grant, "Predicting active pulmonary tuberculosis using an artificial neural network," *Chest*, vol. 116, no. 4, pp. 968–973, 1999.
[20] S. Jaeger, A. Karargyris, S. Candemir, J. Siegelman, L. Folio, S. Antani, and G. Thoma, "Automatic screening for tuberculosis in chest radiographs: a survey," *Quantitative imaging in medicine and surgery*, vol. 3, no. 2, p. 89, 2013.
[21] Y. Cao, C. Liu, B. Liu, M. J. Brunette, N. Zhang, T. Sun, P. Zhang, J. Peinado, E. S. Garavito, L. L. Garcia, and W. H. Curioso, "Improving tuberculosis diagnostics using Deep Learning and mobile health technologies among resource-poor and marginalized communities," in *Proceedings of the IEEE First Conference on Connected Health: Applications, Systems and Engineering Technologies*, Washington, USA, 2016, pp. 274–281.
[22] S. Hwang, H.-E. Kim, J. Jeong, and H.-J. Kim, "A novel approach for tuberculosis screening based on deep convolutional neural networks," in *Medical Imaging 2016: Computer-Aided Diagnosis*, vol. 9785. Bellingham, Washington, USA: International Society for Optics and Photonics, 2016, p. 97852W.
[23] P. Lakhani and B. Sundaram, "Deep Learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.
[24] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wang, P.-X. Lu, , and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 6, no. 4, pp. 475–477, 2014.
[25] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. ichi Komatsu, M. Matsui, H. Fujita, Y. Kodera, , and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.
[26] F. Chollet, *Deep Learning with Python*, 1st ed. Shelter Island, New York: Manning Publications, 2017.
[27] K. L. Moore, A. F. Dalley, and A. M. R. Agur, *Anatomia – Orientada para a Clínica*, 8th ed. Editora Guanabara Koogan, 2018.
[28] "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.
[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
[30] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, *A Guide to Convolutional Neural Networks for Computer Vision*, 1st ed., ser. Synthesis Lectures on Computer Vision. San Rafael, California, USA: Morgan & Claypool, 2018.
[31] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," 2012.
[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1. Boston, Massachusetts, USA: IEEE, 2015, pp. 1–9.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[36] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv:1609.04747, 2016.
[37] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: The MIT Press, 2016, vol. 1.
[38] D. Sarkar, R. Bali, and T. Ghosh, *Hands-On Transfer Learning with Python Implement Advanced Deep Learning and Neural Network Models Using TensorFlow and Keras*, 1st ed. Livery Place 35, Livery Street, Birmingham: Packt Publishing, 201.
[39] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. Long Beach, CA: IEEE, 2019, pp. 2661–2671.
[40] L. Rokach, "Ensemble-based classifiers," *Artif Intell Rev*, vol. 33, pp. 1–39, 2010.
[41] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. USA: Springer Science+Business Media, 2012.
[42] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.
[43] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.
[44] M. L. McHugh, "Interrater reliability: the Kappa statistic." *Biochemia medica*, no. 3, pp. 276–82, 2012.
[45] M. P. Ponti Jr., "Combining classifiers: from the creation of ensembles to the decision fusion," in *2011 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials*, vol. 1. Brasil: IEEE, 2011, pp. 1–10.
[46] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

Lucas Gabriel Coimbra Evangelista Undergraduate in Computer Engineering at the State University of Amazonas, Lucas is part of the university's Intelligent Systems Laboratory and works in the ICTS Group.

Elloá B. Guedes PhD in Computer Science, Elloa holds an assistant professor position at Amazonas State University. Co-founder of the Intelligent Systems Laboratory, she currently leads the institution's Intelligent Systems Research Group, working on research and development of solutions based on Machine and Deep Learning.