Classificação de Câncer de Pulmão por Meio de Imagens Histopatológicas Usando Redes Neurais Convolucionais

1st Luiz Carlos Cosmi Filho
Departamento de Engenharia Elétrica
Universidade Federal do Espírito Santo
Vitória, Brasil
luiz.cosmi@edu.ufes.br

2nd Mateus Sobrinho Menines
Departamento de Engenharia Elétrica
Universidade Federal do Espírito Santo
Vitória, Brasil
mateus.menines@edu.ufes.br

***Resumo*—Lung cancer is the leading cause of cancer-related deaths worldwide, with smoking accounting for about 85% of cases, according to the World Health Organization. Typically diagnosed in advanced stages, early detection can significantly increase survival rates. Diagnostic methods include physical exams, imaging tests, and histopathological examinations. Using artificial intelligence, the histopathological diagnosis of lung cancer can provide significant help. In that regard, the objective of this study is to propose a artificial intelligence algorithm capable of classifying histopathological images to help specialists in the diagnosis. To accomplish that, we used a set of tissue images (LC25000) that includes benign cells, lung adenocarcinoma, and squamous cell carcinoma from the James A. Haley Veterans' Hospital in the United States. Then, we applied deep learning algorithms using CNNs architectures and transfer learning techniques in the classification task and evaluate their performance metrics.**

***Index Terms*—deep learning, computational intelligence, lung cancer, convolutional neural network, transfer learning.**

I. Introdução

De acordo com a *World Health Organization* [1], o câncer de pulmão é a principal causa de mortes relacionadas ao câncer em todo o mundo, sendo que o hábito de fumar é responsável por aproximadamente 85% dos casos.

O câncer de pulmão geralmente é diagnosticado em estágios avançados, quando as opções de tratamento são limitadas. A detecção precoce pode aumentar significativamente as taxas de sobrevivência.

A *World Health Organization* menciona vários métodos de diagnóstico para o câncer de pulmão, incluindo exame físico, exames de imagem como radiografias, tomografias e ressonância magnética, e a coleta de amostras de tecido para exame histopatológico. Este último exame analisa os tecidos para detectar possíveis lesões nas células.

Por meio da coleta dos tecidos, o Hospital de Veteranos James A. Haley, localizado Flórida nos Estados Unidos, obteve um conjunto de imagens dos tecidos, conhecido como LC25000 [2]. Esse conjunto inclui separações entre células benignas, adenocarcinoma pulmonar (o tipo mais comum de câncer de pulmão) e carcinoma escamoso (o segundo tipo mais comum) [3].

A análise das imagens obtidas pelo hospital permite a aplicação de um algoritmo de classificação que possa diferenciar entre as três classes do banco de dados para obter resultados das análises de tecidos dos indivíduos. Algoritmos avançados de aprendizado de máquina, tais como redes neurais profundas, ganharam popularidade nos últimos anos na análise de imagens médicas ao permitir melhorar a eficiência e a confiabilidade no diagnóstico [4], [5].

Neste contexto, este artigo explora o desempenho algumas arquiteturas de redes neurais convolucionais (CNN - *Convolutional Neural Network*), um algoritmo de aprendizado de máquina que permite a classificação de imagens sendo eficiente devido a capacidade obter características da imagem de maneira diferenciável [6]. Nesse trabalho, esse algoritmo tem como objetivo classificar as imagens histopatológicas entre as classes presentes no banco de dados.

No entanto, um desafio recorrente nessa área é a escassez de grandes conjuntos de dados anotados. Um grandes conjuntos de dados anotados geralmente é um requisito para o treinamento robusto e de alto desempenho de modelos de aprendizado de máquina. Nesse sentido, nesse artigo explora-se o uso de técnicas de transferência de aprendizagem (do inglês, *transfer learning*) [7]. Essa técnica permite utilizar modelos treinados em conjuntos de dados extensos e diversos e aplicá-los a contextos médicos específicos. Ao utilizar modelos pré-treinados, aproveita-se características genéricas previamente aprendidas das imagens pelas rede neurais, possibilitando uma extração de características de imagens médicas mais eficiente.

Portanto, pode-se listar como as principais contribuições deste trabalho:

- A aplicação de técnicas de aprendizado de máquina para a classificação de imagens histopatológicas de câncer de pulmão;
- A avaliação do desempenho de CNNs utilizando técnicas de transferência de aprendizado nessa tarefa.

Para um melhor entendimento deste trabalho, as próximas seções irão abordar, respectivamente: trabalhos relacionados da área (Seção II), a metodologia da proposta desenvolvida nesse trabalho (Seção III), os experimentos e resultados obtidos

(Seção IV) e, por fim, as conclusões e trabalhos futuros (Seção V).

II. Trabalhos relacionados

Alguns trabalhos podem ser relacionados ao utilizar imagens de células cancerígenas para ser aplicadas em algoritmos de inteligência computacional. O trabalho de [8], os autores propuseram uma metodologia de classificação de imagens histopatológicas utilizando técnicas de extração de características e algoritmos de aprendizado de máquina (XGBoost, SVM, RF, LDA, MLP and LightGBM). Tal abordagem foi escolhida visando obter uma melhor interpretabilidade da tarefa classificação, buscando entender quais características são mais importantes. Os resultados obtidos utilizando o algoritmo XGBoost apresentou o melhor desempenho, 99% de acurácia e 98.8% de média harmônica F1.

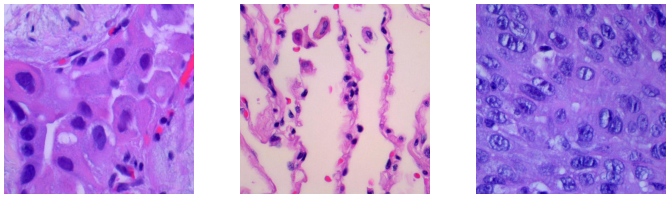
Por outro lado, diversos trabalhos propõem arquiteturas de CNN para classificação de imagens histopatológicas, tal como em [9]. Em [10], uma arquitetura de CNN inicialmente projetada para classificação de cores de objetos em imagens foi adaptada para classificação de imagens histopatológicas, tanto de pulmão quanto de cólon.

Já em [11], apresenta uma combinação de CNNs com o algoritmo LightGBM [12] para classificação de imagens histopatológicas de câncer de pulmão. Nesse sentido, a CNN é utilizada para extração de características e o algoritmo LightGBM para classificação do resultado fornecido pela CNN. Os autores obtiveram resultados de acurácias e sensibilidade acima de 99%.

III. Metodologia

O conjunto de imagens histopatológicas de câncer de pulmão utilizada nesse trabalho é o LC25000 [2]. O banco de dados contém 25.000 imagens (RGB) com resolução de 728x728 de tecidos do pulmão e cólon, distribuídas em 5 classes, com cada classe tendo 5.000 imagens: (i) Adenocarcinoma de Cólon; (ii) Tecido Benigno do Cólon; (iii) Adenocarcinoma Pulmonar; (iv) Tecido Pulmonar Benigno; (v) Carcinoma de Células Escamosas Pulmonar.

Esses tecidos foram preparados com a técnica de coloração de hematoxilina e eosina. Neste trabalho, apenas as imagens de tecidos pulmonares serão utilizadas. Exemplos de imagens de cada classe podem ser vistos na Figura 1.



(a) (b) (c)

Figura 1: (a) Adenocarcinoma Pulmonar; (b) Tecido Pulmonar Benigno; (c) Carcinoma de Células Escamosas Pulmonar.

Para o contexto deste trabalho, os modelos foram avaliados utilizando o procedimento de validação cruzada *k-fold* com 5 divisões. Sendo que, em cada divisão, 20% da base dados é utilizada para teste, 20% para validação e 60% para treino. É importante mencionar que essas divisões são estratificadas. Ou seja, a separação dos dados ocorre de forma proporcional às diferentes classes presentes nos dados.

A validação cruzada é um técnica que avalia o desempenho de modelos especialmente em situações onde os dados não estão bem disponíveis ou distribuídos em conjuntos de treinamento, validação e teste. Nesse sentido, oferecendo uma medida mais confiável sobre a capacidade dos modelos avaliados.

A. Algoritmo de treinamento

Duas arquiteturas diferentes de redes neurais foram consideradas nesse trabalho, são elas:

- Resnet50 [13]: CNN de 50 camadas apresenta conexões de atalho que ignoram algumas camadas, dessa forma evitando o problema do desaparecimento do gradiente (do inglês, *vanishing gradient problem*);
- VGG-19 [14]: CNN composto por 19 camadas, usa filtros de convolução pequenos (3x3) com um número crescente de filtros em cada camada subsequente.

Dessa forma, os pesos dessas redes neurais pré-treinadas no banco de dados ImageNet [15] foram utilizados para transferência de aprendizado. Na Figura 2, pode-se observar um esquemático representando o procedimento de transferência de aprendizado utilizado nesse trabalho durante o treinamento. Observe que apenas as últimas camadas foram treinadas, todas as camadas convolucionais foram congeladas. No caso da Resnet50, última e única camada linear foi treinada. Já no caso da VGG19, as três últimas camadas lineares foram treinadas.

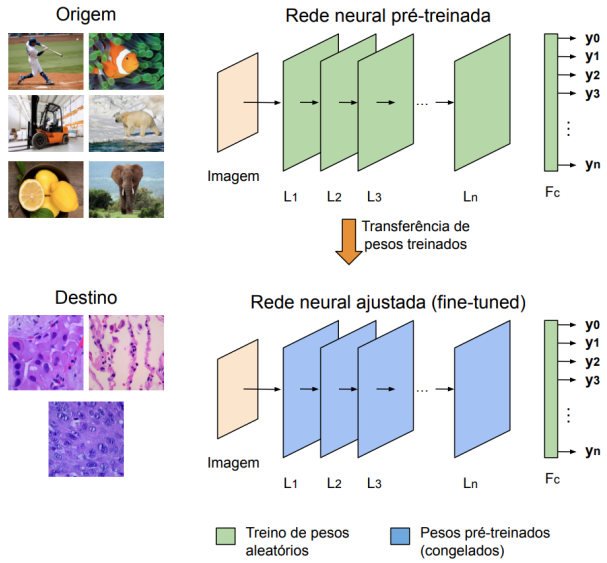


Figura 2: Esquemático representando o procedimento de transferência de aprendizado.

É importante mencionar que todas imagens foram redimensionadas para resolução de 256x256 e normalizadas usando os valores de média e desvio padrão da base de dados ImageNet [15]. Além disso, durante o treinamento técnicas de *Data Augmentation* foram utilizadas, são elas: rotações de 90° e espelhamentos aleatórios.

Ambos os modelos foram treinados sobre os conjuntos de treinamento durante 30 épocas e com lotes de 30 imagens, utilizando como a função de perda de entropia cruzada e o algoritmo de otimização Adam [16]. De cada treinamento, os pesos escolhidos para inferência sobre o conjunto de teste foram aqueles que minimizavam a função de perda no conjunto de validação. Além disso, a taxa de aprendizado para todos os treinamentos foi fixada em 0.0001.

B. Métricas para avaliação dos algoritmos

É importante ressaltar que o problema de classificação do base de dados utilizada neste trabalho possui 3 classes. Logo, o cálculo de métricas devem refletir esse aspecto, podendo ser computadas por meio de médias macro, micro ou com pesos. Como a base de dados escolhida é balanceada, possuindo o mesmo número de amostras para cada classe, escolheu-se a média macro para apresentar os resultados. Para avaliar os modelos treinados, as seguintes métricas macro foram selecionadas:

- Acurácia: Mede a porcentagem de amostras que foram corretamente classificadas, sendo calculada pela Eq. (1), onde N_C é o número de amostras corretamente classificadas e N é o número total de amostras.

$$A = \frac{N_C}{N} \tag{1}$$

- Precisão macro: Mede a proporção de amostras que foram classificadas como positivas e que realmente são positivas. A Eq. (2) calcula a média macro da precisão a partir da matriz de confusão, sendo C, o número de classes e VP_i e FP_i são, respectivamente, a quantidade de verdadeiros positivos e falsos positivos para cada classe.

$$P = \frac{1}{C}\sum_{i=1}^{C}\frac{VP_i}{VP_i + FP_i} \tag{2}$$

- *Recall* macro: Calcula a proporção de amostras positivas que foram classificadas como positivas, medindo a capacidade do algoritmo de recuperar todas as amostras positivas. O cálculo da média macro do *Recall* é realizada pela Eq. (3), sendo, C o número de classes e VP_i e FN_i são, respectivamente, a quantidade de verdadeiros positivos e falsos negativos para cada classe.

$$R = \frac{1}{C}\sum_{i=1}^{C}\frac{VP_i}{VP_i + FN_i} \tag{3}$$

- *F1* macro: Calcula a média harmônica entre a precisão e *recall*. A Eq. (4) calcula a média macro F1 a partir da precisão e do *recall* de cada classe, Onde F_1 é a média harmônica entre a precisão e *recall*, P_i e R_i são, respectivamente, a precisão e *recall* para cada classe.

$$F_1 = \frac{1}{C}\sum_{i=1}^{C}\frac{2P_iR_i}{P_i + R_i} \tag{4}$$

C. Tabelas e gráficos para avaliação dos algoritmos

Neste trabalho, é apresentada o gráfico da curva do perda (do ingles, *loss curve*) e tabela com as métricas para a avaliação do modelo.

A análise da *loss curve* é utilizada para verificar a variação da função de perda no modelo durante o treinamento através das épocas. A curva exibe os valores de perda do conjunto de validação e no conjunto de treinamento ao longo de cada iteração permitindo avaliar o modelo se está convergindo para uma solução e identificar um possível *Overfitting* ou *Underfitting* [17].

IV. Resultados e Experimentos

Os experimentos foram realizados utilizando um computador com sistema operacional Ubuntu Server 22.04, processador Intel® Xeon® E5-2660 v4 @ 56x 3.2 GHz, com 64 Gb de memória RAM, 1 placa gráfica NVIDIA GeForce RTX 3090 e 3 NVIDIA GeForce RTX 3060. O código desenvolvido está disponível no *Github*.

A. 1ª parte do experimento

Para validar o treinamento utilizando *transfer learning* com as duas redes neurais convolucionais (VGG-19 e ResNet50), o experimento envolveu a obtenção das métricas por meio da validação cruzada dos algoritmos. Optou-se por esse método devido à ausência de uma divisão clara no banco de dados entre conjuntos de teste, validação e treinamento.

A Tabela I apresenta quatro métricas avaliadas em cada modelo durante o procedimento de validação cruzada: Acurácia, F1 Macro, Precisão Macro e *Recall* Macro. Para cada métrica, são fornecidos os valores médios e os desvios padrões resultantes das cinco subdivisões do banco de dados.

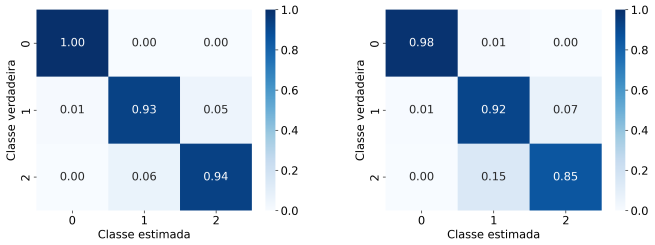
	Acurácia		F1 Macro		Precisão Macro		Recall Macro	
	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão
VGG19	0.9550	0.0059	0.9549	0.0058	0.9551	0.0058	0.9550	0.0059
Resnet50	0.9186	0.0035	0.9188	0.0034	0.9208	0.0034	0.9186	0.0035

Tabela I: Médias e desvios padrões de diversas métricas sobre os conjuntos de testes fornecidos durante o procedimento de validação cruzada com 5 divisões.

A partir da Tabela I dos resultados obtidos, foi possível observar que a VGG-19 com pesos das camadas convolucionais congelados e 3 camadas lineares treinadas teve um maior desempenho quando comparada à Resnet50 com pesos das camadas convolucionais congelados e 1 camada linear treinada. Acredita-se que a maior quantidade de camadas lineares seja a razão por trás do maior desempenho observado pela VGG-19, que permitiu-a obter para esses treinamentos, melhores valores de métricas.

B. 2ª parte do experimento

Para a segunda parte do experimento, é gerado a média entre todas as 5 divisões do banco de dados na validação cruzada para gerar uma matriz de confusão de cada rede. A Figura 3 apresenta as duas matrizes de confusão.



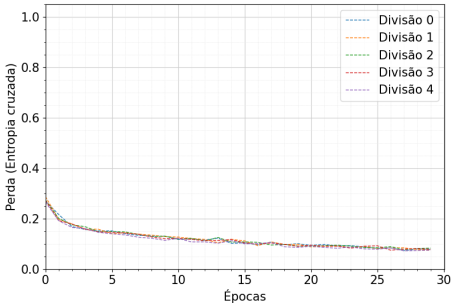
(a) (b)

Figura 3: Médias das matrizes de confusão sobre os conjuntos de testes fornecidos durante o procedimento de validação cruzada com 5 divisões. (a) VGG-19 e (b) Resnet50.

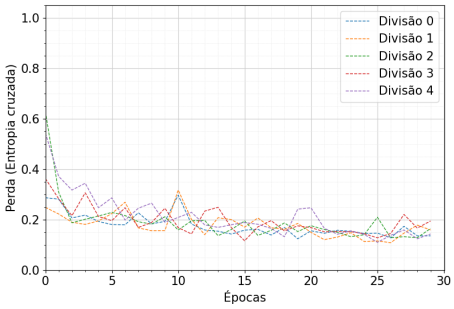
A partir da análise das matrizes de confusão, é possível discutir diversos resultados. Observa-se que a matriz de confusão da VGG-19 apresenta valores superiores, com uma maior quantidade de verdadeiros positivos em cada classe. Embora a ResNet50 apresente valores aceitáveis para a classificação das três classes, a taxa de verdadeiros positivos para a classe 2 é inferior em comparação com as outras taxas de acerto entre as duas matrizes. Esses dados indicam que a VGG-19 obteve um desempenho médio superior na validação cruzada em relação a Resnet50.

C. 3ª parte do experimento

A terceira e última parte do experimento consiste na avaliação da curva de perda para cada treinamento da CNN. Este gráfico é de suma importância para determinar se o modelo apresentou *underfitting* ou *overfitting*. Em geral, a perda observada na curva de treinamento é inferior à da curva de validação. Se a perda de treinamento continuar diminuindo enquanto a perda de validação se estabiliza, indica-se a ocorrência de *overfitting* no modelo. Sendo assim, a Figura 4 e a Figura 5 exibem as curvas de perda de cada divisão do bando de dados pelo processo de validação cruzada entre a rede VGG-19 e a Resnet50 respectivamente, sendo possível avaliar se o modelo está bem generalista ou sofreu um *overfitting* no processo de treinamento.



(a)



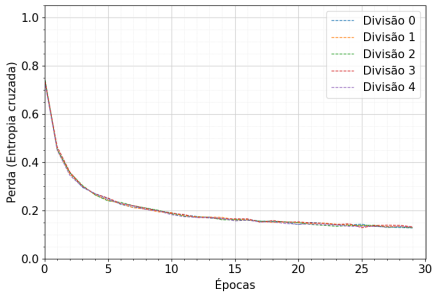
(b)

Figura 4: Função de perda ao longo do treinamento da rede neural VGG-19 para os conjuntos de treino e validação fornecidos durante o procedimento de validação cruzada com 5 divisões. (a) Treino e (b) Validação.

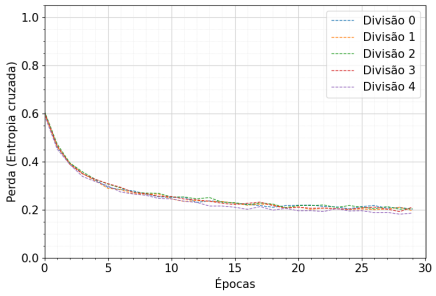
A Figura 4 (a) apresenta a curva de perda do treinamento da rede VGG-19 para as cinco divisões de treino. Observa-se que as curvas praticamente se sobrepõem, indicando uma perda (entropia cruzada) aproximada de 0,1 em todas as divisões. Esta similaridade sugere que o banco de dados é homogêneo, contendo características semelhantes entre os dados, o que resulta em um modelo robusto durante o treinamento. Para avaliar se o modelo sofreu *overfitting* ao longo das 30 épocas, é necessário analisar a Figura 4 (b), que apresenta as curvas de perda durante o processo de validação da VGG-19 para cada divisão, em comparação com a curva de perda mostrada na Figura 4 (a).

Pela Figura 4 (b), é observado uma perda variando entre 0,1 e 0,2, com variações entre esses valores nas cinco curvas. Observa-se também uma semelhança que é, de fato, pior do que as curvas no processo de treinamento, mas que converge para valores próximos entre si. Essa análise dos valores de perda nas Figuras 4 (a) e (b) indica que o modelo não sofreu *overfitting* durante o processo de validação cruzada, uma vez que as curvas convergem para pontos semelhantes ao longo das épocas.

Após a análise das curvas de perda da validação cruzada da rede VGG-19, a Figura 5 apresenta as curvas de perda da rede Resnet50, permitindo uma comparação entre as duas redes no final.



(a)



(b)

Figura 5: Função de perda ao longo do treinamento da rede neural Resnet50 para os conjuntos de treino e validação fornecidos durante o procedimento de validação cruzada com 5 divisões. (a) Treino e (b) Validação.

A Figura 5 (a) apresenta a curva de perda do treinamento da rede Resnet50 para cada divisão na validação cruzada. A análise é similar àquela feita para a Figura 4 (a) da rede VGG-19, mostrando curvas semelhantes e indicando uma perda ligeiramente superior a 0,1. Na Figura 5 (b), são exibidas as curvas de validação para todas as divisões realizadas na validação cruzada, que também convergem para um valor comum de aproximadamente 0,2. Isso indica que o modelo gerado em cada divisão não sofreu *overfitting* para esse banco de dados.

Embora o *overfitting* seja analisado pela curva de perda, é importante notar que vários fatores podem contribuir para esse problema, como a arquitetura da rede, a natureza dos dados no banco de dados e outros aspectos. O *overfitting* é uma análise mais conceitual, dependendo de como o modelo será utilizado para a classificação neste experimento. Nesse contexto, ambos os modelos (VGG-19 e ResNet50) apresentaram um desempenho aceitável na validação cruzada, conforme observado nas curvas de perda, nas matrizes de confusão da Figura 3 e na Tabela I.

V. Conclusões

Ao lidar com banco de dados que imagens, é possível a utilização de CNN pré treinadas para realizar o *transfer-learning* e treinar na ultima camada, uma rede classificatória para diferenciar entre as classes do banco de dados utilizado neste trabalho.

Apesar de simples, pode-se verificar que é possível a utilização das redes VGG-19 e Resnet50 pré-treinadas sobre o conjunto ImageNet como extratores de características eficientes para classificação de imagens histopatológicas, chegando à acurácias relativamente altas. Nesse sentido, verificou-se na prática que a utilização de modelos de aprendizado de máquina na análise de imagens médicas podem permitir melhorar a eficiência e a confiabilidade nos diagnósticos de pacientes com câncer de pulmão.

Como trabalhos futuros, propõe-se explorar o uso de outras técnicas de transferência de aprendizado. Como por exemplo, não congelar os pesos e treinar toda a rede neural a partir dos pesos treinados sobre o conjunto ImageNet. Ou realizar o procedimento de descongelamento dos pesos de forma gradual, a partir de um número de épocas. Nesse sentido, pode-se analisar se o uso dessas técnicas leva a uma melhora no aprendizado ou não e comparar os resultados.

Referências

[1] World Health Organization. Lung cancer fact sheet, 2023. Accessed: 2024-07-02.

[2] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

[3] OncoCenter Médicos. Câncer de pulmão: Quais os principais tipos?, 2023. Accessed: 2024-07-02.

[4] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

[5] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[7] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[8] Aya Hage Chehade, Nassib Abdallah, Jean-Marie Marion, Mohamad Oueidat, and Pierre Chauvet. Lung and colon cancer classification using medical imaging: A feature engineering approach. *Physical and Engineering Sciences in Medicine*, 45(3):729–746, 2022.

[9] Bijaya Kumar Hatuwal and Himal Chand Thapa. Lung cancer detection using convolutional neural network on histopathological images. *Int. J. Comput. Trends Technol*, 68(10):21–24, 2020.

[10] Sanidhya Mangal, Aanchal Chaurasia, and Ayush Khajanchi. Convolution neural networks for diagnosing colon and lung cancer histopathological images. *arXiv preprint arXiv:2009.03878*, 2020.

[11] Esraa A.-R. Hamed, Mohammed A.-M. Salem, Nagwa L. Badr, and Mohamed F. Tolba. An efficient combination of convolutional neural network and lightgbm algorithm for lung cancer histopathology classification. *Diagnostics*, 13(15), 2023.

[12] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.