Elaboração de uma base de imagens de tomografia computadorizada do tórax de pacientes COVID-19 e outras doenças pulmonares

P.M. Souza[1], M.S.D. Pinheiro[1], I.H. Silva[1], J.S. Paixão[1], G. M. Pereira[1], M.R.S. Lima[1], P.P.R Campisi[1], A.T.S.C. Saldanha[1], R.D.S. Pereira[1] , T.A.A.Macedo[1] e A.C. Patrocínio[1]

[1]Universidade Federal de Uberlândia/Engenharia Biomédica, Faculdade de Engenharia Elétrica, Uberlândia, Brasil

***Abstract*— Since the emergence of COVID-19, several institutions have started to share information (contagion information, radiographic images, computerized chest images, etc.) to help the academic community in their research for understanding, aiding the prognosis and combating the disease. Medical images are used to aid in the detection of different diseases and with COVID-19 it is no different. Computed tomography (CT) images were collected, totaling 558 patients (347,096 slices), 543 patients from COVID-19 positive individuals, constituting the COVID image base, and 15 patients with other infectious lung diseases or inflammatory, which gave rise to NON-COVID imaging. An image selection methodology was used, which included inclusion and exclusion criteria, which composed a CT image base about the COVID-19 disease for future research, totaling 556 patients (199,092 slice), being 541 patients from the COVID database and 15 patients from the NON-COVID base.**

***Keywords*— Covid-19, computed tomography, DICOM, images bases.**

I. INTRODUÇÃO

Uma radiografia serve como registro para investigar as alterações na saúde de pacientes sintomáticos e assintomáticos [1, 2, 3, 4, 5, 6]. Alguns exemplos de diagnóstico por imagem mais utilizados são a radiografia de tórax, por possuir um custo baixo e inúmeras informações quando bem avaliada; e a tomografia computadorizada que também ocupa um papel de destaque, sendo superior a radiografia, em termos de melhor detecção, nas alterações torácicas [7]. O uso dessas técnicas é essencial para ajudar radiologistas a diferenciar infecções virais, patogêneses semelhantes que podem ter características de imagem similares, alguns exemplos de doenças detectáveis são: tuberculose, doenças fúngicas, COVID-19, entre outras [8].

A tomografia computadorizada (TC) é baseada na diferença de absorção do feixe dos raios-X dos tecidos do corpo humano, possibilitando a visualização deles. Quanto maior for a absorção no tecido, mais claro ele aparecerá na imagem. A imagem resultante representa um corte anatômico, no entanto não são imagens de aquisição planar. As imagens são de aquisição tomográfica, onde sinais são reconstruídos em imagens bidimensionais. O processamento digital dessas imagens possibilita gerar imagens volumétricas(3D) e até subtrair estruturas, o que torna o uso dessa tecnologia ainda mais atrativo [9, 10, 11, 12].

Em dezembro de 2019 e janeiro de 2020 em Wuhan, província de Hubei, na China, ocorreram os primeiros casos de pneumonia infectada por coronavírus (2019-nCoV), COVID-19 [13, 14]. A qual apresenta um espectro clínico que varia de infecções assintomáticas a quadros graves [15].

O Brasil teve seu primeiro caso confirmado no dia 26 de fevereiro de 2020. Desse dia até 03 de agosto de 2021 foram registrados 19.938.358 casos e um total de 556.834 óbitos [16].

A principal confirmação do diagnóstico é dada pela RT-PCR, a transcrição reversa seguida de reação em cadeia polimerase que possui como amostra a secreção nasal e o seu tempo para o resultado é de 3-4h [17, 18]. Similarmente, exames de radiografia possuem grande importância para investigação diagnóstica e avaliação prognóstica. Como a radiografia de tórax possui baixa resolução de contraste e não mostra sinais da infecção nos estágios iniciais, nem sempre a radiografia é eficiente, e pode ser utilizado um exame complementar com a tomografia computadorizada [19]. Porém, em pacientes com uma infecção severa é observável consolidação multifocal bilateral, parcialmente fundida em consolidação maciça com pequenos derrames pleurais e até mesmo pode ser manifestado como “pulmão branco” [20, 21].

Nas imagens de tomografia computadorizada (TC) são encontradas opacidades em vidro fosco e consolidação pulmonar, também é possível observar uma morfologia arredondada e uma distribuição pulmonar periférica [22, 21]. Ou seja, para pacientes com testes patogênicos positivos é possível encontrar achados típicos nas imagens de TC ou radiografia de tórax. Tais características visualizadas, principalmente nas imagens tomográficas, auxiliam os radiologistas a diferenciar infecções com patogênese semelhante e que podem ter características de imagem similares [8].

Sabendo-se disso, este trabalho tem como objetivo coletar e organizar bases de dados de imagens radiológicas (radiografia de tórax e tomografia computadorizada) de indivíduos positivos para o COVID-19 e outras doenças pulmonares infecciosas ou inflamatórias.

II. METODOLOGIA

Para a elaboração da base, foi utilizada a infraestrutura computacional do Laboratório de Engenharia Biomédica-UFU (BIOLAB), assim como máquinas pessoais dos colaboradores.

A. Base de imagens

Para que o projeto fosse desenvolvido de forma propícia, inicialmente ocorreu a aquisição de uma base de dados fornecida integralmente por um hospital público. A mesma vem de uma colaboração com um médico radiologista e pertence a modalidade de aquisição radiológica nomeada tomografia computadorizada (TC).

A base coletada tem o formato de armazenamento DICOM, pois este sistema permite a transmissão de imagens médicas junto com as informações associadas a elas. Desta forma, é possível que haja a interoperabilidade de imagens entre equipamentos de fornecedores distintos e diferentes plataformas [23]. O intercâmbio de informações facilita a utilização de imagens digitais, pelo fato de simplificar a expansão e o desenvolvimento dos trabalhos realizados. Esse formato possui a característica de agrupar as informações (tags) em série. Ou seja, as informações contidas, como exemplo, os dados do paciente, não serão separadas da imagem [24]. Com isso, o sistema possui a vantagem da conservação da qualidade da imagem mesmo após o compartilhamento, devido à preservação da densidade de pixels, dimensões e números de bits por pixel [24, 25]. Um exemplo do formato DICOM é apresentado na Fig.1.

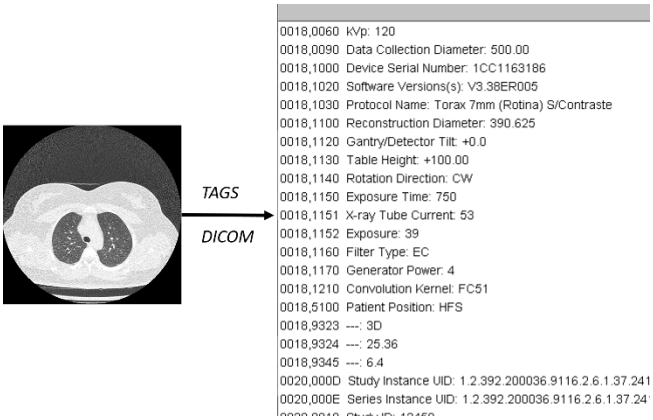

Fig.1. Exemplo do formato DICOM.

A TABELA 1 apresenta a base de dados no estágio inicial, com um total exames tomográficos de 558 pacientes (totalizando 347,096 fatias), sendo que 543 pacientes positivos para o COVID-19 chamado de COVID e 15 pacientes para outras doenças pulmonares infecciosas ou inflamatórias chamado NÃO-COVID.

TABELA 1. Base inicial

Base Inicial	Pacientes	Total de fatias
COVID	543	334,968
NÃO-COVID	15	12,128

Cada paciente tem várias fatias, mostrando vários cortes do pulmão, como mostra a Fig.2. Cada fatia tem o formato DICOM, com o tamanho 512x512 de resolução espacial.

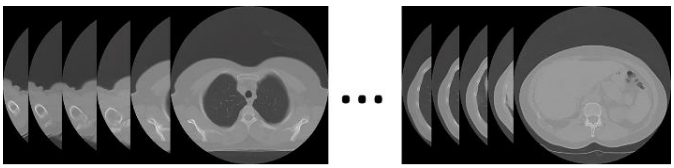
Fig.2. Exemplo de um exame de TC de um paciente.

Os equipamentos e espessuras da base COVID e NÃO-COVID são apresentados na TABELA

TABELA 2. Equipamento/espessura das fatias

Equipamentos	Espessura da Fatia
Siemens	1mm
Siemens	1,7mm
Siemens	5mm
Toshiba	1mm
Toshiba	1,7mm
Philips	5mm
Hitachi	1,3mm

B. Fase de anonimização das imagens

Todas as fatias dos pacientes das bases COVID E NÃO-COVID foram anonimizadas, sendo atribuído um número identificador correlacionado ao identificador único do prontuário do hospital para futuro acompanhamento do desfecho. Os autores e colaboradores do projeto assinaram um Termo de Confidencialidade de Uso de Dados (TCUD). Assim, os riscos da pesquisa foram mínimos, pois trata-se de um estudo retrospectivo, de casos de uma população restrita de indivíduos que já realizaram exames de radiografia de tórax e TC e foram diagnosticados, onde as informações dos mesmos se encontram protegidas.

C. Critérios de inclusão

- Foram incluídos na base COVID, pacientes que fizeram investigação devido a síndrome respiratória aguda, com sinais de agravamento e com teste para COVID-19 positivos.
- Foram incluídos na base NÃO-COVID, pacientes para outras doenças pulmonares infecciosas ou inflamatórias.
- Foram incluídas imagens com resolução de 512x512 pixels.
- Foram incluídas imagens de pacientes adultos.

D. Critérios de exclusão

- Segundo orientação do radiologista, foram descartados 40% do total de fatias de cada exame, sendo 20% das fatias iniciais e 20% das fatias finais. Esta seleção é importante para ter uma base mais consistente e focada na área pulmonar, pois as fatias iniciais e finais não são as ideais para salientar este órgão como mostra a Fig.3

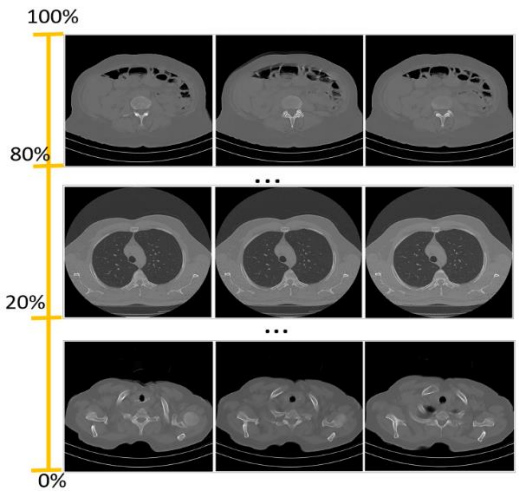

Fig.3. Esquema de seleção de fatias da TC.

- Foi feita uma inspeção visual para verificar se o exame correspondia a uma tomografia computadorizada de pulmão completa, tão quanto para excluir fatias que se encontravam em outra posição. (conforme mostra a Fig.4)

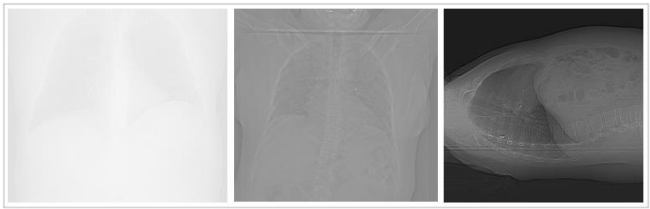

Fig.4. Esquema de inspeção de exclusão da base.

- Foram excluídas imagens que não estivessem inicialmente no formato de imagem DICOM e que não contassem com uma resolução de 512x512.
- Foram excluídas imagens de exames que não pertencessem a pacientes adultos (idade maior que 18 anos).

E. Estatística das tags DICOM

Após a seleção das imagens, foi construída uma planilha para cada paciente, onde ficaram disponíveis as informações das *tags DICOM* de todas as fatias. Logicamente, qualquer atributo de identificação de paciente já estava previamente ocultado. Apresentado na Fig.5

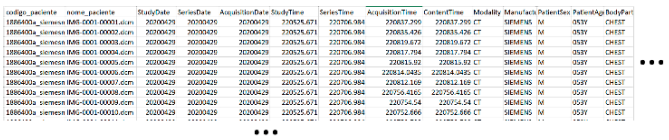

Fig.5. Planilha Excel das tags DICOM.

Posteriormente à criação destes documentos, foram selecionados, de maneira aleatória,15 pacientes da base COVID e 15 da base NÃO-COVID, e selecionadas quatro *tags* para uma visualização estatística destes dados. Esses atributos foram:

- *PatientAge* (idade do paciente);
- *XrayTubeCurrent* (corrente aplicada no tubo de raios X em mA);
- *ExposureTime* (Tempo de exposição em ms);
- *Peak Kilo* voltage (tensão de pico em kVp).

Foi calculada a média (conforme Equação 1) destes atributos, com auxílio da linguagem de programação Python. O processo foi automatizado, assim evitando possíveis falhas de cálculo, para os 15 pacientes acometidos de COVID-19 e

separadamente foi feito o mesmo processo para os 15 pacientes que possuem outros problemas pulmonares.

$$\bar{x} = \frac{\sum x}{n} \qquad (1)$$

Após aplicar a equação (1) foram obtidos os seguintes resultados de acordo com a TABELA 2:

TABELA 2. Média das *tags*: tempo de exposição, corrente, idade, tensão de pico.

Base	Tempo de exposição/ms	Corrente/mA	Idade	Tensão de pico/kvp
COVID	2253,145	85,394	52,8	121,989
NÃO-COVID	3496,729	109,011	58,1	122,667

III. RESULTADOS E DISCUSSÃO

O trabalho M. Roberts, et al. [26], relata armadilhas comuns e recomendações para usar o aprendizado de máquina para detectar e prognosticar para COVID-19 usando radiografias de tórax e tomografias computadorizadas, sendo as seguintes recomendações em relação a base de dados:

- Verificar se as imagens são de pacientes COVID-19 através de exames RT-PCR ou testes de anticorpos positivos. [26]

Resposta: Sobre a base de dados deste trabalho, todos os pacientes foram confirmados através de exames laboratoriais acerca da doença COVID-19.

- Verificar se houve a combinação de dados demográficos entre as coortes, por exemplo incluir imagens pediátricas no grupo de imagens de adultos. [26]

Resposta: Em relação a imagens pediátricas, foram encontrados dois pacientes com idade de 14 anos, os quais foram excluídos da base. Assim, para o restante dos pacientes coletados, ficaram no intervalo entre 23 e 90 anos. Conforme a TABELA 3.

TABELA 3. Idades mínimas e máximas da base.

Base	Idade Mínima	Idade Máxima
COVID	23	90
NÃO-COVID	31	88

- Verificar se a imagem é compactada ou de baixa resolução, contendo informações tais como fabricante, espessura de corte, etc.). [26]

Resposta: todas as imagens foram obtidas no formato DICOM, com resolução de 12 bits armazenada em 16 bits, além de conter as *tags DICOM*, as quais permitem a exploração dos parâmetros de aquisição das imagens.

Assim, a base deste trabalho é inicialmente apresentada na TABELA 1, com um total de 558 pacientes (347,096 fatias), sendo que 543 pacientes de indivíduos positivos para o COVID-19 e 15 pacientes para outras doenças pulmonares infecciosas ou inflamatórias. Após as etapas de verificação e seleção da base de dados, a base resultou em um total de 556 pacientes, mas com uma redução de fatias para 199,092, aproximadamente 40% de redução. Essa redução é proporcional ao corte inicial de 20% somado ao corte final de 20% das fatias de cada paciente, totalizando os 40% de redução da base, conforme mostra a TABELA 4.

TABELA 4. Base Final

Base Consolidada	Pacientes	Total de fatias
COVID	541	196,274
NÃO-COVID	15	2,818

A estrutura final da base é apresentada na Fig.6, a qual apresenta a hierarquia de diretórios, sendo um diretório principal chamado Base_HOSPITAL_UFU, contendo dois diretórios:

- COVID contendo 541 pacientes (pacientes positivos para o COVID-19).
- NÃO_COVID contendo 15 pacientes (pacientes para outras doenças pulmonares).

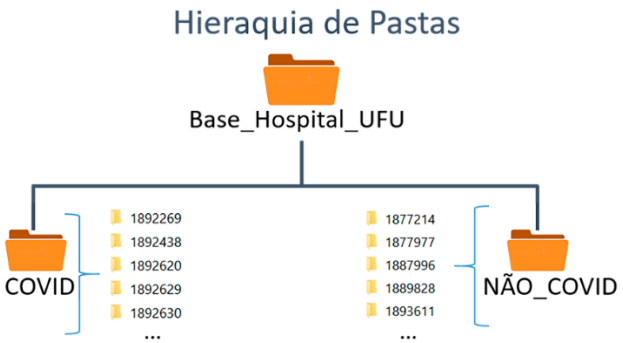

Fig.6. Esquema da base de dados do Hospital de Clínicas da UFU.

Em relação a estatística das tags DICOM da amostra de 15 pacientes de cada grupo, apresentada na TABELA 2, pode-se notar que o valor médio para tempo de exposição da base COVID é de 2253,145 enquanto para base NÃO-COVID é de 3496,729 essa característica de aquisição se mostra com diferença de média mais considerável entre as bases, já a corrente , tem-se o valor de média equivalente a 85,394 para exames de pacientes com CONVID e 109,011 pacientes NÃO-COVID. A idade dos pacientes, também tem

uma certa variação entre as duas bases, enquanto a base COVID tem pacientes com uma média de 52,8 anos a base NÃO-COVID consta com uma média de idade de 58,1 anos, por fim, a tensão de pico se apresenta como a medida com maior proximidade de valor médio sendo 121,989 para a base COVID e 122,667 para a base NÃO-COVID.

IV. CONCLUSÃO

A base de imagens deste trabalho seguiu as recomendações da literatura [26], assim foram criados os critérios de inclusão: todos os pacientes com a doença COVID-19 confirmados através de exames, todas as imagens recebidas têm o formato DICOM; exclusão: imagens de menores de 18 anos foram excluídas da base, imagens de outros tipos de exames, outro tipo de armazenamento diferente do DICOM, etc.

Assim, esta base de imagens de TC fornece informações mais confiáveis e valiosas para profissionais e pode, por exemplo:

- Extrair atributos através de técnicas de processamento digital de imagens, para uma análise estatística dos mesmos, a fim de escolher os atributos significativos para separação das classes COVID e NÃO-COVID;
- Treinar métodos de aprendizado profundo com as imagens e rótulos com o intuito de auxiliar os radiologistas no processo de tomada de decisão.

A fim gerar um melhoramento e balanceamento da base e de auxiliar a comunidade acadêmica em suas pesquisas, novos pacientes devem ser coletados, principalmente aqueles com outras doenças pulmonares.

ACKNOWLEDGMENT

Este estudo foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001.

REFERÊNCIAS

[1] W. D. Martins, "Wilhelm Conrad Roentgen e a Descoberta dos Raios-X," *Archives of Oral Research,* vol. 1, 2005.

[2] A. M. Xavier, A. G. d. Lima, C. R. M. Vigna, F. M. Verbi, G. G. Bortoleto, K. Goraieb, C. H. Collins e M. I. M. S. Bueno, "Marcos da história da radioatividade e tendências atuais," *Química Nova,* vol. 30, p. 83–91, 2007.

[3] F. C. Francisco, W. Maymone, A. C. P. Carvalho, V. F. M. Francisco e M. C. Francisco, "Radiologia: 110 anos de história," *Rev Imagem,* vol. 24, p. 281–6, 2005.

[4] J. M. L. CASTILHO, C. A. R. LOPRETO, P. R. J. R. Buzo e R. Basaglia, "A evolução dos aparelhos de Raios-X," *Recuperado em,* vol. 29, 2017.

[5] A. C. P. Carvalho, "História da tomografia computadorizada," *Revista Imagem,* vol. 29, p. 61–66, 2007.

[6] M. V. T. Navarro, H. J. D. Leite, J. d. C. Alexandrino e E. A. Costa, "Controle de riscos à saúde em radiodiagnóstico: uma perspectiva histórica," *História, Ciências, Saúde-Manguinhos,* vol. 15, p. 1039–1047, 2008.

[7] D. Capone, J. M. Jansen, A. J. Lopes, C. d. C. Sant'Anna, M. O. T. Soares, R. d. S. Pinto, H. R. d. Siqueira, E. Marchiori e R. B. Capone, "Diagnóstico por imagem da tuberculose pulmonar," *Pulmão RJ,* vol. 15, p. 166–74, 2006.

[8] H. J. Koo, S. Lim, J. Choe, S.-H. Choi, H. Sung e K.-H. Do, "Radiographic and CT features of viral pneumonia," *Radiographics,* vol. 38, p. 719–739, 2018.

[9] A. P. Mourão, Tomografia computadorizada: tecnologias e aplicações, Difusão Editora, 2018.

[10] A. C. M. Dovales, A. A. de Souza e L. H. S. Veiga, "Tomografia computadorizada no Brasil: frequência e padrão de uso em pacientes internados no Sistema Único de Saúde (SUS)," *Revista Brasileira de Física Médica,* vol. 9, p. 11–14, 2015.

[11] M. d. Saúde, "datasus," [Online]. Available: http://www2.datasus.gov.br/DATASUS/index.php?area=0201&id=1421686. [Acesso em 04 08 2021].

[12] D. L. U. Ferreira, "Análise da distribuição de aparelhos de tomografia computadorizada no Brasil 2008-2020," 2021.

[13] Y. Zheng, C. Xiong, Y. Liu, X. Qian, Y. Tang, L. Liu, E. L.-H. Leung e M. Wang, "Epidemiological and clinical characteristics analysis of COVID-19 in the surrounding areas of Wuhan, Hubei Province in 2020," *Pharmacological research,* vol. 157, p. 104821, 2020.

[14] T. P. Velavan e C. G. Meyer, "The COVID-19 epidemic," *Tropical medicine & international health,* vol. 25, p. 278, 2020.

[15] M. d. Saúde, "O que é a Covid-19?," 04 Abril 2021. [Online]. Available: https://www.gov.br/saude/pt-br/coronavirus/o-que-e-o-coronavirus. [Acesso em 4 Agosto 2021].

[16] M. d. Saúde, "Entenda qual é a situação do País na pandemia," 03 Agosto 2021. [Online]. Available: https://www.gov.br/saude/pt-br/vacinacao/#o-que-e-covid.. [Acesso em 04 Agosto 2021].

[17] "Carta Médica do National Health Office," 2020.

[18] N. Younes, D. W. Al-Sadeq, H. Al-Jighefee, S. Younes, O. Al-Jamal, H. I. Daas, H. Yassine e G. K. Nasrallah, "Challenges in laboratory diagnosis of the novel coronavirus SARS-CoV-2," *Viruses,* vol. 12, p. 582, 2020.

[19] M.-Y. Ng, E. Y. P. Lee, J. Yang, F. Yang, X. Li, H. Wang, M. M.-s. Lui, C. S.-Y. Lo, B. Leung e P.-L. Khong, "Imaging profile of the COVID-19 infection: radiologic findings and literature review," *Radiology: Cardiothoracic Imaging,* vol. 2, p. e200034, 2020.

[20] M. Young, Technical writer's handbook, University Science Books, 2002.

[21] S. Jamil, N. Mark, G. Carlos, C. S. D. Cruz, J. E. Gross e S. Pasnick, "Diagnosis and management of COVID-19 disease," *American journal of respiratory and critical care medicine,* vol. 201, pp. P19-P20, 2020.

[22] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu, Y. Yang e Z. A. Fayad, "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology,* vol. 295, p. 202–207, 2020.

[23] A. Moreira, A. R. Durão e A. Correia, "Aplicação da norma DICOM em Medicina Dentária," *Revista Portuguesa de Estomatologia, Medicina Dentária e Cirurgia Maxilofacial,* vol. 53, p. 117–122, 2012.

[24] D. M. Saez e others, “Avaliação da influência dos formatos DICOM e JPEG na reprodutibilidade de pontos cefalométricos em Telerradiografia digital em Norma Frontal,” 2009.

[25] M. D. Levin, “Digital technology in endodontic practice,” em *Cohen's Pathways of the Pulp*, Elsevier, 2011, p. 969–1006.

[26] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer e others, “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans,” *Nature Machine Intelligence*, vol. 3, p. 199–217, 2021.