

Patrick Ryan Sales dos Santos
Orientador: Antonio Oseas De Carvalho Filho

Índices COVID: Uma Abordagem Baseada em Textura para Classificar Lesões Pulmonares

Picos - PI
12 de janeiro de 2021

Patrick Ryan Sales dos Santos
Orientador: Antonio Oseas De Carvalho Filho

Índices COVID: Uma Abordagem Baseada em Textura para Classificar Lesões Pulmonares

Monografia submetida ao Curso de Bacharelado em Sistemas de Informação como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação. Orientador: Prof. Dr. Antonio Oseas de Carvalho Filho

Universidade Federal do Piauí
Campus Senador Helvídio Nunes de Barros
Bacharelado em Sistemas de Informação

Picos - PI
12 de janeiro de 2021

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Campus Senador Helvídio Nunes de Barros
Biblioteca Setorial José Albano de Macêdo
Serviço de Processamento Técnico

S237i Santos, Patrick Ryan Sales dos
Índices COVID: uma abordagem baseada em textura para classificar lesões pulmonares / Patrick Ryan Sales dos Santos – 2021.
47 f.; CD-ROM 4 ¾ pol.
Monografia (Graduação em Sistemas de Informação) – Universidade Federal do Piauí, Picos-PI, 2021.
“Orientador: Prof. Antônio Oseas de Carvalho Filho”
1. COVID 19. 2. Tomografia computadorizada. 3. Análise de textura 3D. 4. Diversidade filogenética. I. Título.

CDD 006.6

Elaborada por Maria José Rodrigues de Castro CRB 3: CE-001510/O

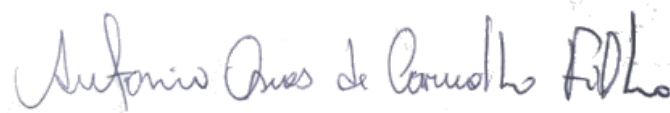
ÍNDICES COVID: UMA ABORDAGEM BASEADA EM TEXTURA PARA
CLASSIFICAR LESÕES PULMONARES

PATRICK RYAN SALES DOS SANTOS

Monografia APROVADA como exigência parcial para obtenção do grau de Bacharel em
Sistemas de Informação.

Data de Aprovação

Picos – PI, 19 de janeiro de 2021.



Prof(a). Antonio Oseas de Carvalho Filho



José Anatiel Gonçalves Santos Landim (UFPI/PPGEE)



Edson Damasceno Carvalho (UFPI/PPGEE)

Agradecimentos

Primeiramente agradeço a Deus, meu Pai por me dar o fôlego da vida, e por sempre abençoar-me de todas as formas possíveis, toda honra e toda glória é para ti meu Deus.

Agradecer a minha família, por me apoiar e acreditar que era possível, agradecer a minha companheira Rosana Mota, que aguentou de forma esplêndida meu mau humor por noites perdidas, amo-te de zero a dez. 'Bem'

Agradecer aos meus colegas e amigos, pelo companheirismo e por todos os dias bons que passamos juntos, buscando um lugar para estudar ou para jogar UNO. 'Panelinha'

Agradecer a minha querida amiga, Vitória Brito, que por anos aguentou-me reclamar de metodologia científica, que me ajudou sempre que eu necessitei, e sim, me apoiou quando tive medo de seguir. 'Chefa'

Agradecer aos meus professores, que fizeram parte desta caminhada, levo em meu coração todos os conselhos que me deram, agradecer de forma especial, ao meu orientador Antonio Oseas, que teve paciência para me ensinar, digo, construir-me como mente pensante, sou eternamente grato pelos puxões de orelha. 'Cuida'

E por fim, agradecer a mim, por nunca desistir dos meus sonhos, e buscar mesmo quando o mundo diz que não é possível.

Salmos 56:3

Mas eu, quando estiver com medo, confiarei em ti.

Resumo

A COVID-19 é uma doença infecciosa causada por um tipo recém descoberto de coronavírus, denominado SARS-CoV-2. Desde a descoberta da doença, no final de 2019, a COVID-19 tornou-se uma preocupação mundial, principalmente pelo seu elevado grau de contágio. Em outubro de 2020, o número de casos confirmados de COVID-19 reportados à *World Health Organization* já ultrapassava 36 milhões no mundo, enquanto o número de mortes ultrapassa 1 milhão. Em virtude dos impactos causados pela doença, a literatura tem intensificado seus esforços no estudo de abordagens voltadas à detecção de COVID-19, visando auxiliar e facilitar o processo de diagnóstico da doença. Este trabalho propõe a aplicação de descritores de textura baseados nas relações filogenéticas entre as espécies para a caracterização de volumes segmentados de exames de *Computed Tomography (CT)* e posterior classificação das regiões em COVID-19, sólidas ou não nódulos. Para avaliar o método proposto, foram utilizadas imagens provindas de três conjuntos de dados distintos. Os resultados mostraram-se promissores, com uma acurácia de 99,25%, *recall* de 99,25%, *precision* de 99,24%, F1 de 99,24% e *AUC* de 0,96. O trabalho apresenta um método robusto, simples e eficiente, que pode ser facilmente aplicado em imagens 2D e/ou 3D, sem limitações em relação à dimensionalidade das imagens.

Palavras-chaves: COVID-19; Tomografia Computadorizada; Análise de textura 3D; Diversidade Filogenética.

Abstract

COVID-19 is an infectious disease caused by a newly discovered type of coronavirus, called SARS-CoV-2. Thenceforth the discovery of the disease in late 2019, COVID-19, has become a worldwide concern, mainly due to its high degree of contagion. As of October 2020, the number of confirmed cases of COVID-19 reported to the World Health Organization has already exceeded 36 million worldwide, while the number of deaths exceeds 1 million. Due to the impacts of the disease, the literature has intensified its efforts to study approaches aimed at detecting COVID-19, aiming to help and facilitate the disease diagnosis process. This work proposes application texture descriptors based on phylogenetic relationships between species to characterize segmented volumes of CT and subsequent classification of regions in COVID-19, solid or non-nodule. To evaluate our method, we used images from three different datasets. The results were promising, with an accuracy of 99.25%, recall of 99.25%, a precision of 99.24%, f1-score of 99.24%, and AUC of 0.96. Our work presents a robust, simple, and efficient method, which can be easily applied to 2D and/or 3D images without limitations regarding the image's dimensionality.

Lista de ilustrações

Figura 1 – Exemplo de árvore filogenética para alguns primatas adaptados de (BA-XEVANIS; OUELLETTE, 2004).	19
Figura 2 – (a) Exemplo de uma região analisada, (b) Representação da árvore filogenética extraída do exemplo. Autoria própria.	21
Figura 3 – Método proposto. Autoria própria.	30
Figura 4 – Representação do PCA com três componentes para as características extraídas no caso de teste desta seção. Autoria própria.	36
Figura 5 – Resultado da matriz de confusão. Autoria própria.	36
Figura 6 – Representação de uma matriz de confusão com casos de acerto e erro do classificador. Autoria própria.	37
Figura 7 – Histograma das imagens selecionadas da matriz de confusão. Autoria própria.	38
Figura 8 – Representação dos índices extraídos das imagens selecionadas da matriz de confusão. Autoria própria.	39

Lista de tabelas

Tabela 1 – Correspondência entre os termos da biologia e o método proposto. . . .	19
Tabela 2 – Distâncias calculadas para a árvore filogenética mostrada na Figura 2.	22
Tabela 3 – Cálculo dos termos L_i e A_i do índice PD.	23
Tabela 4 – Resumo do trabalho relacionado.	28
Tabela 5 – Distribuição de imagens entre conjuntos de dados.	31
Tabela 6 – Parâmetros principais usados nos classificadores.	32
Tabela 7 – Experimentos realizados no trabalho.	34
Tabela 8 – Resultados usando k-fold, com $k = 10$	34
Tabela 9 – Resultados usando k-fold, com $k = 5$	34
Tabela 10 – Comparação com trabalhos relacionados.	40

Lista de abreviaturas e siglas

Acc	Acurácia
CAD	<i>Computer-Aided Detection</i>
CAD-X	<i>Computer-Aided Diagnosis</i>
CNN	<i>Convolutional Neural Network</i>
COVID-19	Coronavirus
CT	<i>Computed Tomography</i>
DWT	<i>Discrete Wavelet Transform</i>
F1	Pontuação F1
FN	Falso Negativo
FP	Falso Positivo
GGO	<i>Ground-Glass Opacities</i>
GLCM	<i>Grey Level Co-occurrence Matrix</i>
GLRLM	<i>Grey Level Run Length Matrix</i>
GLSZM	<i>Grey-Level Size Zone Matrix</i>
LDP	<i>Local Directional Pattern</i>
LIDC-IDRI	<i>Lung Image Database Consortium Image Collection</i>
MNND	<i>Mean Nearest Neighbor Distance</i>
MPD	<i>Mean Phylogenetic Distance</i>
PC	<i>Pulmonary Consolidation</i>
PD	<i>Phylogenetic Diversity</i>
PE	<i>Pleural Effusion</i>
Prec	Precisão
PSR	<i>Phylogenetic Species Richness</i>
PSV	<i>Phylogenetic Species Variability</i>

<i>Rec</i>	<i>Recall</i>
<i>RF</i>	<i>Random Forest</i>
<i>RT-PCR</i>	<i>Reverse Transcription Polymerase Chain Reaction</i>
<i>SPD</i>	<i>Sum of Phylogenetic Distances</i>
<i>SVM</i>	<i>Support-Vector Machine</i>
TN	Verdadeiro Negativo
TP	Verdadeiro Positivo
<i>VOI</i>	<i>Volume of Interest</i>
<i>WHO</i>	<i>World Health Organization</i>

Lista de símbolos

- Δ Índice de diversidade taxonômica
- Δ^* Índice de distinção taxonômica

Sumário

1	Introdução	14
1.1	Objetivos	16
1.2	Organização do Trabalho	16
2	Referencial Teórico	17
2.1	COVID-19	17
2.2	Extração de Características por Textura	18
2.2.1	Índices de diversidade filogenética	18
2.2.2	Exemplo de cálculo de índice para uma imagem	21
2.3	Classificação	25
2.4	Métricas de Validação	26
3	Trabalhos Relacionados	27
4	Metodologia	30
4.1	Aquisição de imagens	30
4.2	Extração de características	31
4.3	Classificação	31
4.4	Validação	32
5	Resultados	33
5.1	Resultados obtidos pelo método proposto	33
5.1.1	Estudo de caso	35
5.1.2	Comparação com trabalhos relacionados	37
6	Discussões	41
7	Conclusão	43
8	Publicações	44
	Referências	45

1 Introdução

Desde a descoberta do novo coronavírus (COVID-19) na China, no final de 2019, a doença tornou-se uma preocupação mundial, principalmente pela sua rápida disseminação. Em março de 2020, 80,565 casos do novo vírus foram confirmados na China e 95,333 no mundo (SHAN et al., 2020). Em janeiro de 2021, o número de casos confirmados reportados à *World Health Organization* (WHO) já ultrapassava 86 milhões, enquanto o número de mortes ultrapassa 1.8 milhão (WHO, 2021). A COVID-19 é uma doença infecciosa causada por um tipo de vírus descoberto recentemente, denominado SARS-CoV-2. A maioria das pessoas infectadas com esse vírus tem doença respiratória leve a moderada e se recupera sem tratamento especial. No entanto, idosos e pessoas com doenças pré-existentes, como doenças cardiovasculares, diabetes, doenças respiratórias crônicas e câncer têm maior probabilidade de desenvolver formas graves de COVID-19 (WHO, 2021; GORBALENYA et al., 2020; REMUZZI; REMUZZI, 2020). Devido à indisponibilidade de medicamentos específicos para COVID-19, o diagnóstico precoce é fundamental para a cura e controle da doença (SINGH et al., 2020).

Os testes de laboratório aplicados na detecção do COVID-19 atualmente baseiam-se, principalmente, na *reverse transcription polymerase chain reaction* (RT-PCR) (JAISWAL et al., 2020). Contudo, diante do cenário de pandemia, é comum a escassez desses testes em alguns lugares, o que pode atrasar o diagnóstico e tratamento clínico e, conseqüentemente, favorecer a propagação da doença. Além disso, esse tipo de exame pode produzir uma alta taxa de falsos negativos (SHAN et al., 2020). Como alternativa aos testes laboratoriais, outros tipos de exames têm sido utilizados pelos hospitais para o diagnóstico da doença, como por exemplo a *Computed Tomography* (CT) (SHAN et al., 2020).

A CT proporciona insumos radiológicos claros do paciente acometido pela COVID-19, portanto é considerada eficiente para a avaliação dos casos suspeitos da doença (BERNHEIM et al., 2020). Contudo, o diagnóstico através da CT depende da análise das imagens pelo especialista, sendo este um processo demorado, repetitivo e propenso a erros, uma vez que analisar muitas imagens pode ser exaustivo, principalmente no cenário pandêmico, onde há vários exames a serem diagnosticados (FILHO et al., 2018). Normalmente, os casos de COVID-19 compartilham algumas características nas imagens tomográficas, como a presença de *ground-glass opacities* (GGO) em estágios iniciais e *pulmonary consolidation* (PC) em estágios avançados (WANG et al., 2020). O *pleural effusion* (PE) também pode ocorrer em casos de COVID-19, mas é menos comum do que as demais lesões. No entanto, embora essas características sejam frequentes na maior parte dos casos, as imagens de algumas pneumonias virais também apresentam tais características, o que pode acabar confundindo o diagnóstico, ocultando outras doenças mais graves (WANG et al., 2020).

Devido à importância do diagnóstico precoce da COVID-19, várias técnicas computacionais vêm sendo desenvolvidas para serem integradas aos sistemas de detecção auxiliado por computador (CAD, do inglês *Computer-Aided Detection*) e diagnóstico auxiliado por computador (CAD-X do inglês *Computer-Aided Diagnosis*). O objetivo desses sistemas é melhorar a precisão do diagnóstico e auxiliar os especialistas em tomada de decisão, dando aos pacientes maiores chances de receber tratamento antes que a doença progrida.

Neste trabalho, é proposto a aplicação de um método eficiente para extrair características importantes dos volumes de CT, sendo essa etapa denominada em um sistema de visão computacional de, extração de características. Como as lesões provocadas pela COVID-19 não apresentam uma forma definida, a textura pode ser a propriedade mais relevante para fornecer informações da doença que sejam eficientes para o processo de classificação. Um dos motivos da despadronização na forma das lesões é que na segmentação manual, os próprios especialistas costumam delimitar uma região maior que o volume de interesse (*VOI*), do inglês *Volume of Interest*, a fim de extrair mais detalhes que possam contribuir para o diagnóstico da doença. Por outro lado, muitos sistemas CAD aplicam métodos de segmentação que produzem regiões candidatas com formatos semelhantes (FILHO et al., 2014a; NETTO et al., 2012), podendo acarretar uma classificação errônea. Em virtude de tais fatores, foi empregado apenas a análise por textura na solução.

Para medir essa textura, foram utilizados conceitos da biologia referentes aos índices de diversidade filogenética, aplicados para descrever as relações filogenéticas entre as espécies, a fim de distinguir as *VOIs* entre COVID-19, sólidas e tecido saudável. Reunimos *VOIs* de lesões provindas de três conjuntos de dados diferentes, a fim de formar um conjunto heterogêneo que possa ser utilizado para avaliar a robustez e confiabilidade da nossa solução.

A seguir, destaca-se as principais contribuições deste trabalho:

- O presente método leva em consideração as *VOIs* das *CTs*, sem a necessidade de dividi-las em slices;
- O método trabalha com as *VOIs* em sua dimensão original, dispensando a aplicação de quaisquer técnicas de redimensionamento;
- Utiliza-se técnicas de outras áreas do conhecimento, adaptando-as ao contexto das imagens pulmonares. Especificamente, é proposto a caracterização da textura das imagens com base nas relações filogenéticas entre as espécies;
- Desenvolve-se um método escalável, uma vez que pode ser facilmente empregado em imagens 2D ou 3D, sem restrições em relação à quantização das imagens;
- Emprega-se uma avaliação do método em conjuntos de dados distintos, realizando um processo de classificação multiclasse para a distinção das lesões provocadas pela COVID-19 e por outros tipos de doenças.

1.1 Objetivos

Diante do contexto apresentado, o propósito deste trabalho é explorar e analisar a textura na sua viabilidade da utilização das característica filogenéticas, então assim classificar os volumes de tecidos pulmonares, entre solida, COVID-19 e não nódulo em imagens de *CT*, provindas de conjuntos de imagens distintos.

De maneira específica, pretende-se:

1. Avaliar a viabilidade do uso dos índices para a identificação da doença, utilizando bases públicas de imagens de *CT*, com e sem balanceamento das classes;
2. Construir uma metodologia para classificação automática do covid-19, que possa ser aplicada no auxílio ao diagnóstico da doença.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte forma: o Capítulo 2 explana sobre os conceitos fundamentais para o entendimento da metodologia proposta neste trabalho; o Capítulo 3 aborda os trabalhos da literatura relacionados à detecção de COVID-19; o Capítulo 4 discorre sobre o método proposto; o Capítulo 5 aponta os resultados alcançados com a execução da metodologia; o Capítulo 6 discute os resultados obtidos; e, finalmente, o Capítulo 7 mostra a conclusão deste trabalho.

2 Referencial Teórico

Este capítulo apresenta os principais conceitos para a compreensão da proposta deste trabalho. A Seção 2.1 relata sobre o COVID-19, abordando as formas atuais de diagnóstico da doença. A Seção 2.2 discorre sobre a extração de características por textura, onde é detalhando sobre como é possível extrair informações de uma imagens analisando a textura da mesma. A Subseção 2.2.1 Descreve os índices de diversidade filogenética, extraídos dos exames de *CTs*, a Subseção 2.2.2 exemplifica de forma simples a utilização e os cálculos realizados pelos índices de diversidade filogenética, a Seção 2.3 discorre como a classificação foi realizada para as classes solida, COVID-19 e não nódulo, e por fim a Seção 2.4 descreve como a validação dos resultados foi obtida.

2.1 COVID-19

Em dezembro de 2019, a COVID-19 foi descoberto na cidade de Wuhan, China. Esse surto se espalhou exponencialmente por todo o mundo, sendo declarada como uma pandemia. Os mais prevalentes dos sintomas clínicos de pacientes com COVID-19 são febre, seguida por tosse, fadiga e dispneia. Essa doença pode levar a problemas respiratórios agudos, síndrome de angústia, insuficiência renal aguda, choque e morte. (ZHAO et al., 2020).

Os critérios de diagnósticos da COVID-19 são avaliação laboratorial de secreções respiratórias adquiridas do aspirado endotraqueal, lavagem broncoalveolar ou swab nasofaríngeo / orofaríngeo. Atualmente, exames laboratoriais, como a RT-PCR tornou-se a avaliação padrão para o diagnóstico de infecção por COVID-19. No entanto, os resultados do teste RT-PCR podem ser falsamente negativos devido a insuficiência amostra ou erro de laboratório (OZKAYA; OZTURK; BARSTUGAN, 2020).

A COVID-19 é uma doença inflamatória causada por quadro agudo grave de síndrome respiratória, que pode se manifestar como um amplo espectro de sintomas que variam de poucos ou nenhum sintoma para pneumonia grave que pode evoluir para aguda síndrome de dificuldade respiratória e morte (MILLET et al., 2021). Embora os mecanismos moleculares que conduzem a gravidade da doença permaneçam obscuros, a associação clínica de mediadores com casos graves sugerem que a inflamação excessiva é fundamental para um resultado clínico ruim (TAN et al., 2021).

A indução de processos inflamatórios na célula hospedeira muitas vezes requer o envolvimento de inflamassomas, que são plataformas de proteínas que se agregam no citosol em resposta a diferentes estímulos (RODRIGUES et al., 2021).

2.2 Extração de Características por Textura

Uma das tarefas mais complexas na análise de imagens está na definição de um conjunto de características que possam descrever de maneira concreta cada região contida em uma imagem, de modo a serem utilizados em processos de mais alto nível (PEDRINI; SCHWARTZ, 2008). Em outras palavras, a etapa de caracterização e representação da imagem consiste em uma etapa de fundamental importância no modelo clássico do processamento de imagem digital proposto por (CAREY et al., 2006). Nesta etapa, as propriedades da imagem podem ser representadas através de modelos matemáticos para servirem de entrada para a análise de reconhecimento e classificação de padrões.

Em linhas gerais, a etapa de extração de características pode ser dividida em duas categorias de análises, sendo elas por: textura e forma. Em uma análise por textura, o objetivo geral é descrever aspectos da imagem no que diz respeito a suavidade, rugosidade e regularidade (CAREY et al., 2006). Já em uma análise baseada na forma da imagem, o intuito é extrair informações que mensuram sobre propriedades morfológicas da imagem.

Segundo (HARALICK; SHANMUGAM; DINSTEN, 1973), textura é definida como a informação de uma região relacionada a coeficientes de uniformidade, densidade, aspereza, regularidade, intensidade, dentre outras características da imagem. A análise de textura é relevante em imagens digitais, uma vez que possibilita distinguir regiões da imagem que apresentam características semelhantes (AZEVEDO; CONCI; VASCONCELOS, 2018).

2.2.1 Índices de diversidade filogenética

Nesta Seção, apresenta-se a justificativa para os índices propostos para caracterização de textura. Cada índice corresponde a uma característica, totalizando oito características que serão extraídas de cada imagem analisada.

A filogenética é um ramo da biologia que estuda as relações evolutivas entre as espécies, e é possível descrever a semelhança entre elas. Em árvores filogenéticas, as folhas representam espécies e os nós representam ancestrais comuns. A figura 1 ilustra um exemplo de árvore filogenética, representando a relação genética entre a espécie de macaco e a espécie humana, onde é possível observar que geneticamente, homem e chimpanzé estão mais próximos que os outros pares existentes na árvore (BAXEVANIS; OUELLETTE, 2004)

A combinação de árvores filogenéticas e índices de diversidade filogenética é usada para analisar as relações evolutivas entre as espécies e pode medir a variação das espécies presentes em uma comunidade (CIANCIARUSO; SILVA; BATALHA, 2009). Para aplicar esses conceitos à caracterização de imagens de tomografia computadorizada, é necessário compreender a correspondência entre as definições da biologia e as utilizadas neste trabalho. A tabela 1 descreve esta correspondência.

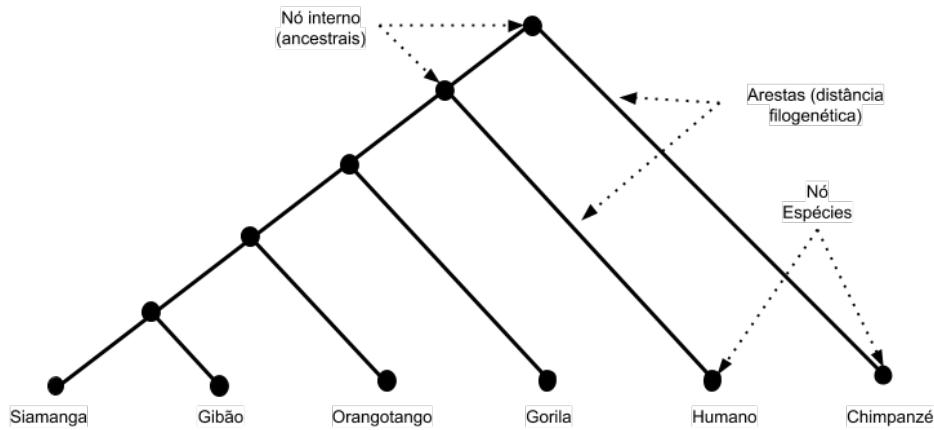


Figura 1 – Exemplo de árvore filogenética para alguns primatas adaptados de (BAXEVANIS; OUELLETTE, 2004).

Tabela 1 – Correspondência entre os termos da biologia e o método proposto.

Biologia	Método
Comunidade	Volume segmentado
Espécies	Valor de voxel
Indivíduos	Número de voxels em uma espécie
Características fenotípicas das espécies	Relação de textura entre espécies

Neste trabalho, é apresentado os índices de Diversidade Filogenética (PD), do inglês *Phylogenetic Diversity* (FAITH, 1992), Soma das Distâncias Filogenéticas (SPD), do inglês *Sum of Phylogenetic Distances* (HELMUS et al., 2007), Distância Média do Vizinho Mais Próximo (MNND), do inglês *Mean Nearest Neighbor Distance* (WEBB; LOSOS, 2000), Variabilidade Filogenética das Espécies (PSV), do inglês *Phylogenetic Species Variability* (HELMUS et al., 2007), Riqueza de Espécies Filogenéticas (PSR), do inglês *Phylogenetic Species Richness* (HELMUS et al., 2007), Distância Filogenética Média (MPD), do inglês *Mean Phylogenetic Distance* (WEBB; LOSOS, 2000), índice de diversidade taxonômica (Δ) (CLARKE; WARWICK, 1998) e índice de distinção taxonômica (Δ^*) (CLARKE; WARWICK, 1998). Cada índice corresponde a um descritor. Portanto, para cada imagem analisada, teremos oito feições compondo o vetor de feições.

O índice PD (FAITH, 1992) é uma medida responsável por fornecer a soma das distâncias dos ramos filogenéticos da árvore. Assim, quando o comprimento do ramo é longo, a espécie possivelmente corresponde a grupos mais distintos. A equação (2.1) apresenta a fórmula para o cálculo do PD, onde B identifica o número de ramos da árvore, L_i representa a extensão do ramo i (o número de bordas no ramo i), e A_i refere-se à abundância média de espécies que compartilham o ramo i .

$$PD = B \times \frac{\sum_i^B L_i A_i}{\sum_i^B A_i}. \quad (2.1)$$

O SPD é um índice filogenético responsável por calcular a soma das distâncias entre os pares de espécies presentes na árvore (HELMUS et al., 2007). A equação (2.2) pode calcular este índice, onde S representa o número de espécies na comunidade, d_{nm} representa a distância entre as espécies m e n , e a_m e a_n correspondem à abundância das espécies m e n , respectivamente.

$$SPD = \left(\frac{S(S-1)}{2} \right) * \frac{\sum \sum_{m<n} d_{mn} a_m a_n}{\sum \sum_{m<n} a_m a_n}. \quad (2.2)$$

O MNND calcula a média ponderada da distância filogenética do vizinho mais próximo de cada espécie (WEBB; LOSOS, 2000). Os pesos são a abundância de espécies. A equação (2.3) apresenta a fórmula para o cálculo deste índice, onde S representa o número de espécies na comunidade, d_{nm} representa a distância entre as espécies m e n , e a_m corresponde à abundância de espécies de m . No caso de d_{nm} , n se refere ao parente mais próximo da espécie m .

$$MNND = \sum_m^S \min(d_{mn}) a_m, \quad (2.3)$$

O índice PSV mede a variação entre duas espécies em uma comunidade a fim de quantificar a relação filogenética entre elas. O PSV pode ser calculado usando a Equação (2.4), onde C é uma matriz e trC é a soma dos valores diagonais desta matriz, $\sum c$ representa a soma de todos os valores da matriz, n é o número total de espécies e \bar{c} é a média dos valores fora da diagonal da matriz.

$$PSV = \frac{n(trC) - \sum c}{n(n-1)} = 1 - \bar{c}, \quad (2.4)$$

O índice PSR calcula a riqueza de espécies presentes em uma comunidade com base em sua variabilidade (HELMUS et al., 2007). Portanto, conforme mostrado pela Equação (2.5), o cálculo é feito multiplicando o número de espécies (n) pelo PSV.

$$PSR = nPSV. \quad (2.5)$$

O MPD calcula a média das distâncias filogenéticas analisando a combinação de todos os pares de espécies na comunidade (WEBB; LOSOS, 2000). Para isso, utiliza o número total de espécies, indicado por N , a distância filogenética entre cada par de espécies, denotada por d_{ij} , e também uma variável $p_i p_j$, que terá valor 1 se a espécie está presente e 0 se não estiver presente. A equação (2.6) apresenta a fórmula para o cálculo do MPD.

$$MPD = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^N d_{ij} p_i p_j}{\sum_{i=1}^{N-1} \sum_{j=i+1}^N p_i p_j}, \quad (2.6)$$

O valor de Δ fornece a distância filogenética média entre indivíduos de espécies (CLARKE; WARWICK, 1998). Para isso, esse índice leva em consideração o número de indivíduos da espécie e a relação taxonômica entre eles. A fórmula para calcular Δ é definida pela

Equação (2.7), onde x_i ($i = 1, \dots, S$) representa a abundância das espécies i , x_j ($j = 1, \dots, S$) representa a abundância das espécies j , S indica o número total de espécies, n denota o número total de indivíduos, e w_{ij} expressa a distância taxonômica entre as espécies i e j .

$$\Delta = \frac{\sum \sum_{i < j} w_{ij} x_i x_j}{[n(n-1)/2]}, \quad (2.7)$$

Finalmente, o índice Δ^* , definido pela Equação (2.8), expressa a distância taxonômica média entre dois indivíduos de espécies distintas (CLARKE; WARWICK, 1998). Neste cálculo, x_i ($i = 1, \dots, S$) é a abundância das espécies i , x_j ($j = 1, \dots, S$) é a abundância das espécies j , S indica o número total de espécies, n indica o número total de indivíduos e w_{ij} expressa a distância taxonômica entre as espécies i e j .

$$\Delta^* = \frac{\sum \sum_{i < j} w_{ij} x_i x_j}{\sum \sum_{i < j} x_i x_j}, \quad (2.8)$$

2.2.2 Exemplo de cálculo de índice para uma imagem

As equações descritas na Subseção 2.2.1 são apresentadas em relação à biologia. Para facilitar a compreensão do cálculo dos índices nas imagens, apresentamos nesta subseção um exemplo de imagem bidimensional, da qual foi extraída a árvore filogenética e calculada as distâncias e os oito índices. Foi usado uma pequena imagem para que os cálculos não ficassem extensos no papel. A figura 2 mostra a imagem que foi usada como exemplo, seguida da árvore filogenética extraída. É possível ver que a imagem possui diversidade de espécies em relação aos *pixels*. É importante ressaltar que a imagem ilustrada na Figura 2 é apenas um exemplo e, portanto, não consideramos os verdadeiros valores de pixel das cores utilizadas.

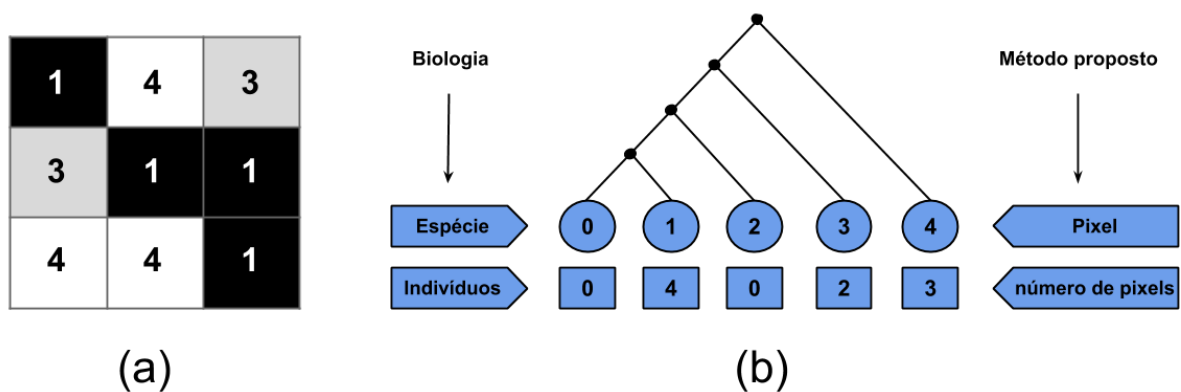


Figura 2 – (a) Exemplo de uma região analisada, (b) Representação da árvore filogenética extraída do exemplo. Autoria própria.

Para calcular as distâncias filogenéticas com base na árvore filogenética, consideramos as seguintes equações:

$$D_{ij} = j - i + 1, \text{ se } i = 0, \text{ e}$$

$$D_{ij} = j - i + 2, \text{ se } i \neq 0,$$

onde i e j são duas espécies diferentes.

Com base nas equações acima, observa-se na Tabela 2 as distâncias obtidas para a árvore filogenética da Figura 2.

Tabela 2 – Distâncias calculadas para a árvore filogenética mostrada na Figura 2.

i	j	D_{ij}
0	1	$D_{01} = 1 - 0 + 1 = 2$
0	2	$D_{02} = 2 - 0 + 1 = 3$
0	3	$D_{03} = 3 - 0 + 1 = 4$
0	4	$D_{04} = 4 - 0 + 1 = 5$
1	2	$D_{12} = 2 - 1 + 2 = 3$
1	3	$D_{13} = 3 - 1 + 2 = 4$
1	4	$D_{14} = 4 - 1 + 2 = 5$
2	3	$D_{23} = 3 - 2 + 2 = 3$
2	4	$D_{24} = 4 - 2 + 2 = 4$
3	4	$D_{34} = 4 - 3 + 2 = 3$

Nesta implementação dos índices, apresenta-se a árvore filogenética para uma estrutura de histograma. Assim, cada posição do histograma se refere às espécies, que são as intensidades da imagem, e cada valor se refere à abundância, que é o número de *pixels* de cada intensidade. Assim, é possível calcular as distâncias usando o histograma. Com base na imagem, na árvore filogenética e nas distâncias calculadas, calcula-se os índices descritos na Subseção 2.2.1.

O PD definido na Equação (2.1) considera o número de ramos da árvore (B), a extensão de cada ramo (L_i) e a média da abundância de espécies em cada ramo (A_i). Em relação às imagens, B equivale ao número de espécies menos um, L_i representa a distância entre as espécies do ramo i e A_i refere-se à abundância média das espécies conectadas ao ramo i pelo número de espécies também conectadas a esse ramo. Os valores de L_i são os mesmos da Tabela 2, enquanto a abundância das espécies pode ser vista na Figura 2. Como o cálculo do PD é mais extenso que os outros, foi descrito usando a Tabela 3.

Com base nos valores calculados na Tabela 3, o valor PD para o exemplo de imagem resultou em:

$$PD = B * \frac{\sum_i^B L_i A_i}{\sum_i^B A_i} = 4 * \frac{65,38}{18,04} = 14,49$$

Para calcular o SPD, representado pela Equação (2.2), precisamos dos valores de: distância entre espécies (d_{mn}); número de espécies (S), que nesse caso é o número de

Tabela 3 – Cálculo dos termos L_i e A_i do índice PD.

i	j	L_i	A_i	$L_i * A_i$
0	1	2	$(0 + 4) / 2 = 2$	$2 * 2 = 4$
0	2	3	$(0 + 4 + 0) / 3 = 1,33$	$3 * 1,33 = 3,99$
0	3	4	$(0 + 4 + 0 + 2) / 4 = 1,5$	$4 * 1,5 = 6$
0	4	5	$(0 + 4 + 0 + 2 + 3) / 5 = 1,8$	$5 * 1,8 = 9$
1	2	3	$(4 + 0) / 2 = 2$	$3 * 2 = 6$
1	3	4	$(4 + 0 + 2) / 3 = 2$	$4 * 2 = 8$
1	4	5	$(4 + 0 + 2 + 3) / 4 = 2,25$	$5 * 2,25 = 11,25$
2	3	3	$(0 + 2) / 2 = 1$	$3 * 1 = 3$
2	4	4	$(0 + 2 + 3) / 3 = 1,66$	$4 * 1,66 = 6,64$
3	4	3	$(2 + 3) / 2 = 2,5$	$3 * 2,5 = 7,5$
Sum			18,04	65,38

intensidades do histograma; e abundância de espécies (a), que representa o número de *pixels* da intensidade em análise. Substituindo os valores em cada equação, tem-se:

$$\begin{aligned} \Sigma \Sigma_{m < n} d_{mn} a_m a_n &= (2 * 0 * 4) + (3 * 0 * 0) + (4 * 0 * 2) + (5 * 0 * 3) + (3 * 4 * 0) + \\ &(4 * 4 * 2) + (5 * 4 * 3) + (3 * 0 * 2) + (4 * 0 * 3) + (3 * 2 * 3) = 110 \end{aligned}$$

$$\begin{aligned} \Sigma \Sigma_{m < n} a_m a_n &= (0 * 4) + (0 * 0) + (0 * 2) + (0 * 3) + (4 * 0) + (4 * 2) + (4 * 3) + \\ &(0 * 2) + (0 * 3) + (2 * 3) = 26 \end{aligned}$$

$$\left(\frac{S * (S - 1)}{2} \right) = \left(\frac{5 * (5 - 1)}{2} \right) = 10$$

$$SPD = 10 * \frac{110}{26} = 42,30$$

O MNND, representado pela Equação (2.3), requer apenas a distância de uma espécie ao parente mais próximo. Portanto, S indica o número de espécies (número de intensidades no histograma), m denota a espécie em questão (intensidade m), n representa o parente mais próximo de m (intensidade n , que se refere à intensidade após m), e a_m denota a abundância da espécie m (número de *pixels* com intensidade m). Como nossa árvore filogenética tem apenas um caminho de uma espécie para outra, o caminho mínimo é o único caminho entre as espécies. Assim, para a imagem, o MNND resulta no seguinte valor:

$$MNND = \sum_m^S \min(d_{mn}) a_m = (2 * 0) + (3 * 4) + (3 * 0) + (3 * 2) = 18,00$$

Para calcular o PSV, representado pela Equação (2.4), é considerado as operações aplicadas à imagem, que é a matriz C . Assim, trC se refere à soma da imagem diagonal, $\sum c$ representa a soma de todos os *pixels* da imagem e \bar{c} se refere à média dos *pixels* fora da diagonal. Além disso, n representa o número de espécies da comunidade que possuem indivíduos, ou seja, a quantidade de intensidades que estão presentes na imagem.

$$PSV = \frac{n(trC) - \sum c}{n(n-1)} = \frac{3 * 3 - 22}{3 * (3 - 1)} = \frac{-13}{6} = -2,16$$

$$1 - \bar{c} = 1 - 3,16 = -2,16$$

O PSR, representado pela Equação (2.5), é apenas o PSV multiplicado pelo número de espécies da comunidade que possuem indivíduos, ou seja, a quantidade de intensidades de imagem.

$$PSR = n * PSV = 3 * (-2,16) = -6,48$$

Em relação ao MPD, definido pela Equação (2.6), considera-se a soma das distâncias entre as espécies (d_{ij}) multiplicada pelas variáveis p_i e p_j , que recebem o valor 0 se a espécie não estiver presente ou 1 se a espécie estiver presente. Assim, quando existe uma intensidade do histograma na imagem, p recebe 1, caso contrário p recebe 0.

$$\sum_{i=1}^{N-1} \sum_{j=i+1}^N d_{ij}p_i p_j = (2 * 0 * 1) + (3 * 0 * 0) + (4 * 0 * 1) + (5 * 0 * 1) + (3 * 1 * 0) +$$

$$(4 * 1 * 1) + (5 * 1 * 1) + (3 * 0 * 1) + (4 * 0 * 1) + (3 * 1 * 1) = 12$$

$$\sum_{i=1}^{N-1} \sum_{j=i+1}^N p_i p_j = (0 * 1) + (0 * 0) + (0 * 1) + (0 * 1) + (1 * 0) + (1 * 1) + (1 * 1) +$$

$$(0 * 1) + (0 * 1) + (1 * 1) = 3$$

$$MPD = \frac{12}{3} = 4.00$$

No cálculo de Δ , definido pela Equação (2.7), w_{ij} representa a distância entre as espécies (intensidades) i e j , x refere-se ao abundância de espécies (número de *pixels* nas intensidades), e n é o número de indivíduos na comunidade (número de *pixels* no histograma). Dessa forma, podemos calcular Δ da seguinte maneira:

$$\sum \sum_{i < j} w_{ij} x_i x_j = (2 * 0 * 4) + (3 * 0 * 0) + (4 * 0 * 2) + (5 * 0 * 3) + (3 * 4 * 0) +$$

$$(4 * 4 * 2) + (5 * 4 * 3) + (3 * 0 * 2) + (4 * 0 * 3) + (3 * 2 * 3) = 110$$

$$[n * (n - 1) / 2] = [9 * (9 - 1) / 2] = 36$$

$$\Delta = \frac{110}{36} = 3,05$$

Finalmente, o cálculo de Δ^* , definido pela Equação (2.8), é semelhante ao de *triangulo*, com a diferença que no caso de Δ^* o denominador é a soma da multiplicação das abundâncias das espécies. Substituindo os valores, encontra-se:

$$\begin{aligned} \Sigma \Sigma_{i < j} w_{ij} x_i x_j &= (2 * 0 * 4) + (3 * 0 * 0) + (4 * 0 * 2) + (5 * 0 * 3) + (3 * 4 * 0) + \\ &(4 * 4 * 2) + (5 * 4 * 3) + (3 * 0 * 2) + (4 * 0 * 3) + (3 * 2 * 3) = 110 \end{aligned}$$

$$\begin{aligned} \Sigma \Sigma_{i < j} x_i x_j &= (0 * 4) + (0 * 0) + (0 * 2) + (0 * 3) + (4 * 0) + (4 * 2) + (4 * 3) + \\ &(0 * 2) + (0 * 3) + (2 * 3) = 26 \end{aligned}$$

$$\Delta^* = \frac{110}{26} = 4,23$$

2.3 Classificação

Para o reconhecimento do padrão das características extraídas, foram utilizados dois classificadores comuns na literatura: RF (BREIMAN, 2001) e reforço de gradiente extremo (XGBoost) (CHEN et al., 2015). Para validar os modelos, foi aplicado a técnica de validação cruzada. Essa técnica envolve a divisão aleatória do conjunto de imagens em k *folds*, de tamanho aproximadamente igual. A primeira *fold* é tratada como um conjunto de validação, e o método é treinado nas outras $k - 1$ *folds*. É importante ressaltar que cada imagem na amostra de dados é atribuída a um grupo individual e permanece nesse grupo durante o procedimento. Isso significa que cada amostra tem a oportunidade de ser usada no conjunto de validação 1 vez e usada para treinar o modelo $k - 1$ vezes. Cada *fold* tem uma média de 10% do número de imagens em cada classe. A seguir está uma breve descrição dos classificadores usados.

- RF é um algoritmo de regressão e classificação desenvolvido por Breiman (2001). Nesse método, as previsões são feitas a partir de árvores de decisão. Cada árvore no RF fornece uma previsão de classe, então a classe com mais votos torna-se a saída do modelo para a amostra em questão.
- XGBoost é uma técnica de aprendizado de máquina otimizada desenvolvida por (CHEN et al., 2015), que é baseada em árvores de decisão e usa uma estrutura gradiente crescente. Este algoritmo foi projetado para ser flexível e eficiente, por isso seus parâmetros podem ser facilmente alterados (CARVALHO et al., 2020). O XGBoost pode ser aplicado a problemas de regressão e classificação.

2.4 Métricas de Validação

Para avaliar os modelos de classificação obtidos na etapa anterior, foram utilizadas as seguintes métricas: acurácia, *recall*, precisão, pontuação F1 e *AUC*. Para calcular essas métricas, é necessário analisar a matriz de confusão, que é calculada com base em quatro valores: verdadeiro positivo (TP), falso positivo (FP), falso negativo (FN) e verdadeiro negativo (TN). Esses valores servem para indicar o número de amostras classificadas correta e incorretamente.

A acurácia (Acc) calcula a capacidade de um teste de diagnóstico de identificar tanto os verdadeiros positivos quanto os verdadeiros negativos em uma amostra, ou seja, a proporção de acertos nos dados classificados. A equação (2.9) mostra a fórmula desta métrica.

$$Acc = \frac{TN + TP}{TN + TP + FN + FP}. \quad (2.9)$$

Em um teste diagnóstico, a *recall* (rec), também conhecida como sensibilidade, mede a capacidade de identificar verdadeiros positivos, ou seja, indivíduos que têm a doença. É uma métrica fundamental na análise, pois indica a eficácia do método. A equação (2.10) apresenta a fórmula para o cálculo de *recall*.

$$Rec = \frac{TP}{TP + FN}. \quad (2.10)$$

A precisão (prec), expressa pela Equação (2.11), mede a proporção de positivos corretamente classificados entre todos aqueles classificados como positivos.

$$Prec = \frac{TP}{TP + FP}. \quad (2.11)$$

É possível combinar precisão e *recall* para medir o desempenho da classificação, e é isso que pontuação F1 (F1) faz. Conforme visto na Equação (2.12), esta métrica calcula uma média harmônica de precisão e *recall*, sendo uma medida comumente usada em problemas de classificação desequilibrada.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \quad (2.12)$$

Também foi usado a análise da área sob a curva ROC (*AUC*) para avaliar o método. A curva ROC (*receiver operating feature*) (ERKEL; PATTYNAMA, 1998) fornece um gráfico de desempenho de um método de classificação a partir de diferentes limiares, analisando duas métricas calculadas por meio da matriz de confusão: a taxa de verdadeiro positivo (TPR) e a taxa de falso positivo (FPR). A partir deste gráfico, a *AUC* pode medir quão bem o método de classificação foi capaz de distinguir os dados; portanto, quanto mais próximo de 1 for o resultado, melhor será o método de separação de classes.

3 Trabalhos Relacionados

Desde a declaração da pandemia do COVID-19, pela WHO, os esforços no campo científico têm sido intensificados em prol do desenvolvimento de metodologias para auxiliar o diagnóstico da doença, baseando-se em imagens médicas, como exames de raio-X e tomografias computadorizadas. É destacado, nesta Seção, alguns trabalhos relevantes sobre o tema.

Os trabalhos de [Abbas, Abdelsamea e Gaber \(2020\)](#) e [Narin, Kaya e Pamuk \(2020\)](#), apresentaram metodologias baseadas em imagens de raio-X e *Deep Learning* para a classificação de imagens em COVID-19 e non-COVID-19. Os resultados alcançados foram promissores, com acurácia acima de 90%. Devido à escassez na disponibilidade de dados de *CT* públicos da COVID-19, o trabalho de [Zhao et al. \(2020\)](#) e [He et al. \(2020\)](#) construíram seu próprio conjunto de dados, chamado COVID-CT, composto por 349 imagens de *CT* de COVID-19 e 463 de non-COVID-19. Já no estudo de [Ozkaya, Ozturk e Barstugan \(2020\)](#), dois subconjuntos foram extraídos de 150 imagens de *CT*, sendo que cada subconjunto contém 3000 *patches* de COVID-19 e 3000 *patches* de non-COVID-19. No estudo de [Wang et al. \(2020\)](#) também aplicou imagens de *CT* e abordagens de aprendizado profundo para a tarefa de classificação das imagens em casos positivos e negativos para a COVID-19.

No trabalho de [Barstugan, Ozkaya e Ozturk \(2020\)](#), foram utilizadas 150 imagens de *CT*, divididas em 4 subconjuntos, que abrangem *patches* de dimensões 16x16, 32x32, 48x48, 64x64, respectivamente. A abordagem empregada no trabalho de [Barstugan, Ozkaya e Ozturk \(2020\)](#), aplica técnicas como *Grey Level Co-occurrence Matrix* (GLCM), *Grey Level Run Length Matrix* (GLRLM), *Grey-Level Size Zone Matrix* (GLSZM), *Local Directional Pattern* (LDP), e *Discrete Wavelet Transform* (DWT) para extrair características das imagens e, em seguida, classificá-las em COVID-19 ou non-COVID-19 utilizando o *Support-Vector Machine* (SVM). Na abordagem proposta por [Jaiswal et al. \(2020\)](#), um modelo pré-treinado da arquitetura *DenseNet201* é aplicado para realizar a classificação de imagens de *CT* em COVID-19 ou non-COVID-19, aplicando os pesos da *ImageNet*.

Diferente das abordagens anteriores, o trabalho de [Zheng et al. \(2020\)](#) propõe uma metodologia para a detecção de COVID-19 utilizando volumes 3D de *CTs*. Nessa abordagem, cada volume é segmentado através de uma *UNet* pré-treinada e posteriormente classificado por meio de uma rede neural profunda 3D *weakly-supervised*, denominada *DeCoVNet*. As *CTs* utilizadas nos experimentos foram coletadas de 540 pacientes (313 com COVID-19 e 229 sem COVID-19).

A Tabela 4 aponta um resumo dos trabalhos descritos.

Como observado no decorrer desta seção, desde o surto da COVID-19, várias abordagens vêm sendo exploradas para a detecção da doença através da análise de imagens

Tabela 4 – Resumo do trabalho relacionado.

Trabalho	Método	Tipo de exame	Amostra	Resultados
Abbas, Abdelsamea e Gaber (2020)	<i>Decompose, Transfer, and Compose (DeTraC)</i>	X-ray	206	Acurácia: 95,12% Sensibilidade: 97,91% Especificidade: 91,87%
Narin, Kaya e Pamuk (2020)	<i>Pre-trained CNNs</i>	X-ray	100	Acurácia: 98,00% Recall: 96,00% Especificidade: 100% Precisão: 100% F1: 98,00%
Zhao et al. (2020)	<i>Multi-task learning and contrastive self-supervised learning</i>	CT	812	Acurácia: 89,00% F1: 90,00% AUC: 0,98
Ozkaya, Ozturk e Barstugan (2020)	<i>Pre-trained CNNs + SVM</i>	CT	150	Acurácia: 98,27% Sensibilidade: 98,93% Especificidade: 97,60% Precisão: 97,63% F1: 98,28%
Wang et al. (2020)	<i>Selection of ROIs + Inception</i>	CT	453	Acurácia: 82,90% Sensibilidade: 84,00% Especificidade: 80,50%
He et al. (2020)	<i>Self-supervised transfer learning</i>	CT	746	Acurácia: 86,00% F1: 85,00% AUC: 0,94
Barstugan, Ozkaya e Ozturk (2020)	<i>GLCM + GLRLM + GLSZM + LDP + DWT + SVM</i>	CT	150	Acurácia: 98,71% Sensibilidade: 97,56% Especificidade: 99,68% Precisão: 99,62% F1: 98,58%
Jaiswal et al. (2020)	<i>Pre-trained DenseNet201</i>	CT	2492	Acurácia: 96,25% Recall: 96,29% Especificidade: 96,21% Precisão: 96,29% F1: 96,29%
Zheng et al. (2020)	<i>3D UNet + 3D DeCoVNet</i>	CT	540	Sensibilidade: 90,70% Especificidade: 91,10% AUC: 0,959

médicas. Nos estudos correlatos, é possível observar a crescente aplicação de métodos de *deep learning* na detecção de lesões provocadas pela COVID-19, inclusive, as *Convolutional Neural Network* (CNN) são exploradas em diversos domínios de imagem, sendo o estado da arte no reconhecimento de imagens (LECUN; BENGIO; HINTON, 2015).

No entanto, alguns fatores tornam as CNNs um método complexo e muitas vezes limitado a um problema específico. Um desses fatores é a parametrização, visto que as CNNs requerem uma ampla quantidade de parâmetros, que precisam ser ajustados durante o treinamento até que o modelo se torne adequado ao problema a que se propõe (RAKHLIN et al., 2018; GOLATKAR; ANAND; SETHI, 2018; NAWAZ et al., 2018). Para contornar esse problema, alguns estudos aplicam a transferência de aprendizado, que permite a utilização de pesos pré-definidos; entretanto, essa abordagem pode não proporcionar um resultado válido em todos os casos, uma vez que esses pesos, sendo treinados para outra finalidade, podem não ser adequados para o problema em questão.

Outro fator que limita o uso de CNNs é o número de amostras necessárias no treinamento para melhorar o desempenho da classificação. Inclusive, por este motivo os estudos Jaiswal et al. (2020), Zheng et al. (2020), Abbas, Abdelsamea e Gaber (2020), Zhao et al. (2020), He et al. (2020) aplicam técnicas de aumento de dados. Além disso, dependendo da arquitetura, é necessário modificar o tamanho das imagens para adaptá-las à entrada da rede, como empregado nos estudos Narin, Kaya e Pamuk (2020), Zhao et al. (2020),

He et al. (2020). Contudo, o redimensionamento das imagens pode ocasionar a perda de informações valiosas para a classificação.

Dos trabalhos mencionados anteriormente, apenas Zheng et al. (2020) utilizou volumes de *CT* em seu método. Em contrapartida, apresentamos uma abordagem simples e eficiente para a detecção de COVID-19 em imagens de *CT*, utilizando volumes 3D das lesões. A extração de características com os índices de diversidade filogenética não requer uma parametrização complexa, não se limita à quantidade de bits das imagens, não necessita de padronização no tamanho das entradas e é escalável, já que é possível ser aplicada em imagens 2D ou 3D. Além disso, em união aos classificadores *Random Forest (RF)* e *XGBoost*, os índices mostraram-se promissores mesmo sem a aplicação de técnicas de aumento de dados.

4 Metodologia

Esta seção descreve a metodologia proposta para a classificação de volumes de *CT* em COVID-19, sólidas ou não nódulos. As imagens utilizadas neste estudo foram adquiridas utilizando a *Lung Image Database Consortium Image Collection* (LIDC-IDRI) (III et al., 2011) e do repositório *MedSeg* (MEDSEG, 2020), sendo que, deste último, utilizamos dois conjuntos com imagens de COVID-19. Na etapa de extração de características, foram aplicados os índices de diversidade filogenética para mensurar o comportamento da textura das imagens. Por fim, na etapa de classificação, as características extraídas serviram de entrada para dois classificadores, que foram testados e avaliados. Os algoritmos de extração e classificação desenvolvidos neste estudo estão disponíveis publicamente no repositório do [GitHub](#). A Figura 3 descreve o fluxo de trabalho da metodologia.

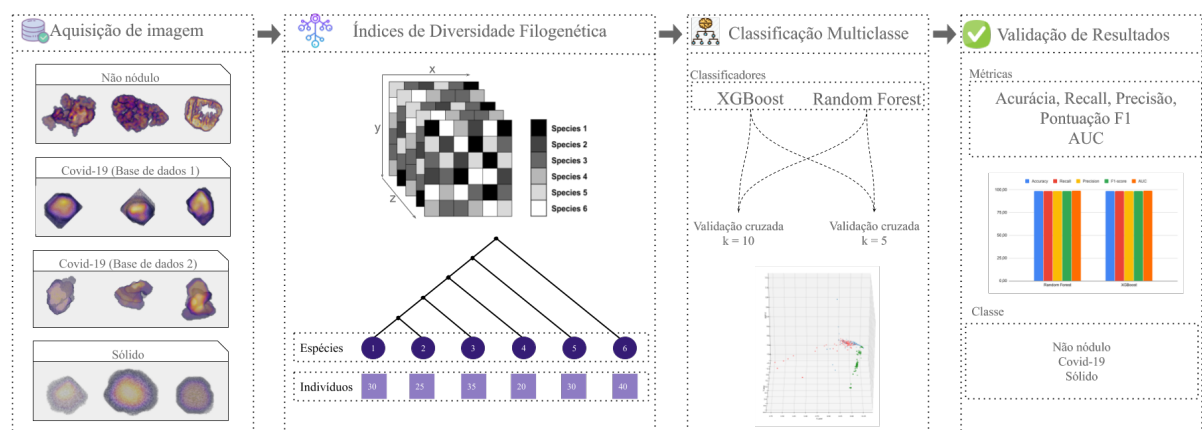


Figura 3 – Método proposto. Autoria própria.

4.1 Aquisição de imagens

Para avaliar o método, utilizou-se três conjuntos de *VOIs* extraídos da base *LIDC-IDRI* e do repositório *MedSeg*. i) no primeiro conjunto são extraídas as *VOIs* que apresentam lesões sólidas e regiões de não nodulares (regiões saudáveis). No caso das lesões sólidas, é utilizado a biblioteca *pyLIDC* (HANCOCK, 2016) para gerar as *VOIs* dos exames com base nas anotações dos especialistas; ii) no segundo conjunto de imagens são extraídas as *VOIs* de não nódulos, isso é, imagens saudáveis, as regiões foram geradas aplicando o método proposto por Filho et al. (2014b), que garante que as *VOIs* de não nódulos não possuem intersecção com as *VOIs* nodulares; iii) o terceiro conjunto de imagens foram adquiridos através do repositório *MedSeg* (MEDSEG, 2020), que fornece alguns conjuntos de dados externos de vários tipos de exames de CT, inclusive exames diagnosticados com

a COVID-19. Assim, utiliza-se dois conjuntos de dados distintos com lesões provocadas pela COVID-19, isso é, podem haver regiões com lesões *GGO*, *PC* e *PE*.

Indicamos na Tabela 5 a quantidade de imagens em cada conjunto utilizado.

Tabela 5 – Distribuição de imagens entre conjuntos de dados.

Conjunto de dados	Diagnóstico	Sample
LIDC	Sólidas	1679
	Não nódulos	17742
COVID-19 (Dataset 1)	<i>GGO</i> , <i>PC</i> e <i>PE</i>	215
COVID-19 (Dataset 2)	<i>GGO</i> , <i>PC</i> e <i>PE</i>	274

4.2 Extração de características

Como mencionado no Capítulo 2, foi aplicada uma análise por textura para extrair as características das imagens por meio dos índices PD, SPD, MNND, PSV, PSR, MPD, (Δ) e (Δ^*). Os índices foram implementados na linguagem de programação *python*. A extração de características ocorreu para todos os exemplos, gerando uma representação da árvore filogenéticas das espécies, que pode ser abstraído para o histograma da imagem e por fim, a extração dos indexes ocorre, embora a complexidade das imagens seja de forma 3D a construção do histograma mantêm-se independente das dimensões da imagem. É importante ressaltar que a representação de *bits* que a imagem é construída influencia diretamente no tamanho do histograma, como as imagens analisadas pelo presente trabalho pertencem a ordem de 16 *bits*, o processamento e extração percorrem por todos os valores, sem perda de informações. Um outro ponto relevante, está na transladação dos valores negativos presentes na imagem, onde foi realizado uma técnica que move os valores negativos para positivos, acarretando com que o menor valor negativo assuma o valor zero e assim sucessivamente.

4.3 Classificação

Como mencionado no Capítulo 2, foram aplicados os classificadores RF e XGBoost. A escolha desses classificadores, está na sua larga utilização pela literatura, sendo ferramentas robustas já validadas pela academia. Outro ponto pelo qual foi escolhido esses classificadores, está no fato de que o RF é um algoritmo baseado em árvores de decisão, e o XGBoost em *machine learning*, assim tendo uma análise e testificação mais abrangente das características extraídas por intermédio dos índices filogenéticos.

Os classificadores, foram utilizados na linguagem de programação *python*, com auxílio da biblioteca *sklearn*, que por sua vez facilita a utilização dessas ferramentas, envolvendo-os e abstraíndo sua complexidade.

A Tabela 6, apresenta os principais parâmetros utilizados nos modelos.

Tabela 6 – Parâmetros principais usados nos classificadores.

Classificador	Parâmetros Usados
RF	$number\ of\ estimators\ (number\ of\ trees) = 100,$ $min\ samples\ split = 2,$ $min\ samples\ leaf = 1,$ $max\ number\ of\ features = "auto" (\sqrt{number\ features})$ $bootstrap = True,$ $max\ depth = None\ (unlimited)$
XGBoost	$max\ depth = 6,$ $learning\ rate = 0.1,$ $number\ of\ estimators\ (number\ of\ trees) = 100,$ $booster = "gbtree",$ $objective = "binary:logistic",$ $gamma = 0,$ $max\ delta\ step = 0$

4.4 Validação

Como mencionado no Capítulo 2, foram utilizadas as métricas Acc, Rec, Prec, F1 e AUC para avaliar os resultados da classificação. As métricas escolhidas para realizar a validação dos resultados obtidos pelo presente trabalho, são comumente utilizadas pela literatura, por conseguirem demonstrar de forma clara os erros e acertos obtidos pelos classificadores.

Por se tratar de uma classificação multiclasse, aplica-se as métricas descritas no Capítulo 2 usando a estratégia *one-vs.-rest*, ou seja, considerando uma classe em relação às demais. Considere X como uma das classes no conjunto de imagens. Em outras palavras, para a validação de X , consideramos TP como amostras positivas de X e TN como amostras positivas de outras classes. Além disso, FP são todas as amostras classificadas como X , mas não são X , e FN são todas as amostras de X que foram previstas como outras classes. Esse processo é realizado na validação de cada amostra. Por fim, para solucionar o desequilíbrio do conjunto de imagens, é calculada uma média ponderada das métricas, sendo os pesos o número de amostras em cada classe.

5 Resultados

Esta seção, apresenta os resultados dos testes realizados nos conjuntos de dados descritas na Seção 4.1. As características das *VOIs* foram extraídas a partir de descritores de textura, com base na diversidade filogenética, conforme especificado na Seção 4.2. Para a classificação das regiões em saudáveis, solidas ou COVID-19, foram utilizados os classificadores apresentados na Seção 4.3. Para garantir que todas as amostras sejam treinadas e testadas pelo menos uma vez, foi utilizado o método *k-fold*, onde aplica-se $k = 5$ e $k = 10$, optou por variar o valor de k para garantir uma quantidade maior de amostras nos conjuntos de testes.

Esta seção divide-se em três partes: i) apresenta-se os resultados alcançados pelo método; ii) realiza-se uma exploração dos resultados obtidos através da análise de alguns casos de acerto e erro do método; e por fim, iii) compara-se os resultados do método proposto com os resultados dos estudos relacionados.

5.1 Resultados obtidos pelo método proposto

Foi dividido os testes em três experimentos, apresentados na Tabela 7. Cada experimento contempla combinações de classificação com as bases utilizadas, sendo que COVID-19 (1) e COVID-19 (2) referem-se às duas bases de *VOIs* de lesões provocadas pela COVID-19, fornecidas pela *MedSeg*. O intuito de executar o método em diferentes cenários de teste é avaliar a potencialidade das características diante de imagens providas de bases distintas. Os três cenários fornecem desafios importantes para avaliação do método proposto, no entanto, considera-se como melhor resultado os dispostos no cenário 1, pois, acredita-se que este se aproxima mais de um ambiente clínico real, uma vez que as lesões diagnosticadas pelos especialistas possuem origens distintas.

Outro ponto importante é que, como visto na Tabela 5, os conjuntos de dados utilizados neste trabalho são desbalanceados, o que pode induzir à interpretações errôneas dos resultados em virtude da discrepância entre a proporção de dados em cada classe. Por este motivo, cada experimento foi realizado duas vezes: uma com os dados desbalanceados (maior aproximação com cenário clínico real) e outra com os dados balanceados. Em relação ao balanceamento, é aplica apenas um *downsample* nos dados, baseando-se na quantidade de dados da base minoritária. Como o foco do nosso trabalho não está em técnicas de balanceamento, não implementamos abordagens robustas para balancear os dados, apenas uma técnica simples para que fosse possível analisar o desempenho do método nas duas situações.

A Tabela 8 aponta os resultados alcançados através dos experimentos realizados com 10 *folds* no *cross validation*, enquanto a Tabela 9 mostra os resultados com 5 *folds*. A

referência de cada experimento está descrito na Tabela 7.

Tabela 7 – Experimentos realizados no trabalho.

Experimentos	Descrição/Classes	Amostra
1	não nódulos × sólidas × COVID-19 (1) e COVID-19 (2)	19,868
2	não nódulos × sólidas × COVID-19 (1)	19,624
3	não nódulos × sólidas × COVID-19 (2)	19,653

Tabela 8 – Resultados usando k-fold, com k = 10.

Cenário de teste	Classificador	Experimento	Acc (%)	Rec (%)	Prec (%)	F1 (%)	AUC
Sem balanceamento	RF	1	99,09 ± 0,14	99,09 ± 0,14	99,09 ± 0,14	99,06 ± 0,16	0,95 ± 0,01
		2	99,32 ± 0,12	99,32 ± 0,12	99,29 ± 0,14	99,27 ± 0,15	0,92 ± 0,01
		3	99,68 ± 0,08	99,68 ± 0,08	99,68 ± 0,08	99,66 ± 0,09	0,96 ± 0,01
	XGBoost	1	99,25 ± 0,15	99,25 ± 0,15	99,24 ± 0,16	99,23 ± 0,16	0,96 ± 0,00
		2	99,33 ± 0,17	99,33 ± 0,17	99,30 ± 0,20	99,30 ± 0,19	0,93 ± 0,01
		3	99,81 ± 0,06	99,81 ± 0,06	99,81 ± 0,06	99,80 ± 0,06	0,98 ± 0,00
Com balanceamento	RF	1	92,23 ± 1,89	92,23 ± 1,89	92,30 ± 1,92	92,22 ± 1,88	0,94 ± 0,01
		2	91,31 ± 3,42	91,31 ± 3,42	91,66 ± 3,40	91,29 ± 3,45	0,93 ± 0,02
		3	94,94 ± 2,53	94,94 ± 2,53	95,05 ± 2,50	94,95 ± 2,52	0,96 ± 0,01
	XGBoost	1	92,81 ± 1,74	92,81 ± 1,74	92,87 ± 1,72	92,80 ± 1,74	0,95 ± 0,01
		2	92,25 ± 3,99	92,25 ± 3,99	92,59 ± 3,89	92,24 ± 4,00	0,94 ± 0,02
		3	96,31 ± 2,12	96,31 ± 2,12	96,42 ± 2,11	96,32 ± 2,11	0,97 ± 0,01

-Negrito é o melhor resultado considerando o cenário 1.

Tabela 9 – Resultados usando k-fold, com k = 5.

Cenário de teste	Classificador	Experimento	Acc (%)	Rec (%)	Prec (%)	F1 (%)	AUC
Sem balanceamento	RF	1	99,09 ± 0,11	99,09 ± 0,11	99,08 ± 0,11	99,05 ± 0,11	0,95 ± 0,00
		2	99,32 ± 0,07	99,32 ± 0,07	99,29 ± 0,09	99,26 ± 0,08	0,92 ± 0,00
		3	99,69 ± 0,04	99,69 ± 0,04	99,68 ± 0,04	99,68 ± 0,04	0,96 ± 0,00
	XGBoost	1	99,25 ± 0,15	99,25 ± 0,15	99,24 ± 0,16	99,23 ± 0,16	0,96 ± 0,00
		2	99,33 ± 0,17	99,33 ± 0,17	99,30 ± 0,20	99,30 ± 0,19	0,93 ± 0,01
		3	99,81 ± 0,06	99,81 ± 0,06	99,81 ± 0,06	99,80 ± 0,06	0,98 ± 0,00
Com balanceamento	RF	1	92,30 ± 0,73	92,30 ± 0,73	92,36 ± 0,68	92,31 ± 0,72	0,94 ± 0,00
		2	90,85 ± 1,33	90,85 ± 1,33	91,14 ± 1,32	90,84 ± 1,32	0,93 ± 0,00
		3	95,50 ± 1,25	95,50 ± 1,25	95,63 ± 1,20	95,50 ± 1,25	0,97 ± 0,00
	XGBoost	1	92,09 ± 1,77	92,09 ± 1,77	92,13 ± 1,73	92,08 ± 1,78	0,94 ± 0,01
		2	91,01 ± 2,62	91,01 ± 2,62	91,25 ± 2,74	90,97 ± 2,62	0,93 ± 0,01
		3	95,63 ± 1,32	95,63 ± 1,32	95,67 ± 1,32	95,63 ± 1,32	0,97 ± 0,00

-Negrito é o melhor resultado considerando o cenário 1.

Analisando os resultados das Tabelas 8 e 9, percebe-se que o XGBoost apresentou os melhores resultados em praticamente todos os experimentos realizados, principalmente nos cenários de teste onde a execução do método encontra-se sem o balanceamento. Possivelmente, os resultados de classificação da classe majoritária influenciaram nos resultados gerais, por este motivo optou-se por avaliar a potencialidade do método diante de dados balanceados. Nesse caso, é possível perceber que mesmo com o balanceamento os resultados continuaram promissores, atingindo acurácia de 96,31%, *recall* de 96,31%, *precision* de 96,42%, F1 de 96,32% e *AUC* de 0,97 no experimento 3 usando k=10. No caso do cenário ideal, que seria o experimento 1, o método apresentou os melhores resultados através do classificador XGBoost. Sem balanceamento, os testes desse cenário resultaram em acurácia de 99,25%, *recall* de 99,25%, *precision* de 99,24%, F1 de 99,24% e *AUC* de 0,96. Com balanceamento, o método alcançou acurácia de 92,81%, *recall* de 92,81%, *precision* de 92,87%, F1 de 92,80% e *AUC* de 0,95.

Em relação à variação nos experimentos realizados, nosso intuito foi avaliar a robustez do método diante de diferentes tipos de bases. Os testes mostraram que mesmo com heterogeneidade das bases, os resultados foram consistentes e promissores em ambos os classificadores. Um outro ponto importante a ser discutido é que o experimento 3 apresentou os melhores resultados em ambos os cenários de teste e classificadores. Acredita-se que isso deva-se ao fato de que a base 2 de COVID-19 apresenta uma similaridade maior entre as lesões, o que pode ter contribuído para uma melhor distinção das características de cada classe por parte dos classificadores. Pelos resultados, é possível perceber que a base 1 de COVID-19 não contempla este mesmo padrão de similaridade, o que pode confundir o classificador em algumas amostras.

5.1.1 Estudo de caso

Nesta Subseção, foi realizada uma exploração dos resultados obtidos através da análise de alguns casos de acerto e erro do método. Para essa análise, foi escolhida o melhor resultado do cenário 1 com balanceamento, onde alcançou-se acurácia de 92,81%, *recall* de 92,81%, *precision* de 92,87%, F1 de 92,80% e *AUC* de 0,95. Optou-se por mostrar apenas o resultado com o balanceamento para proporcionar uma visualização mais clara dos resultados, sem tantas amostras.

A primeira demonstração é o *plot* das características extraídas com os índices propostos, onde foi aplicado o algoritmo *Principal Component Analysis* (PCA) para reduzir a dimensionalidade dos dados preservando as informações mais importantes. O PCA resultante, representado pela Figura 4, possui três componentes. É possível observar que a classe não nodular, que teve maior assertividade em relação aos resultados obtidos, ficou espacialmente mais distante das demais classes, o que facilitou o trabalho dos classificadores. Em contrapartida, as classes solida e COVID-19 tiveram a maior parte das amostras espacialmente próximas, mas é notável a possibilidade de traçar uma fronteira de decisão entre elas.

Na próxima análise, foi selecionada aleatoriamente uma amostra de cada item da matriz de confusão (Figura 5) para o caso de teste abordado nesta seção. A Figura 6 apresenta as imagens selecionadas em forma de matriz de confusão. Os tracejados vermelhos representam as amostras onde o classificador errou, enquanto os verdes representam os acertos. É possível notar que, visivelmente, a lesão da classe não nodular apresenta a textura mais distinta entre as demais, comprovando assim o PCA mostrado na análise anterior. Da mesma forma, percebe-se que, visualmente, a Figura 6(b) possui uma similaridade de textura com a Figura 6(d), justificando assim a classificação errônea da classe solida com a de COVID-19. Já em relação à lesão da Figura 6(c), nota-se que ela apresenta uma textura mais próxima da classe solida do que as demais. Por fim, em relação à lesão Figura 6(e), acredita-se que a textura presente se distanciou do padrão encontrado

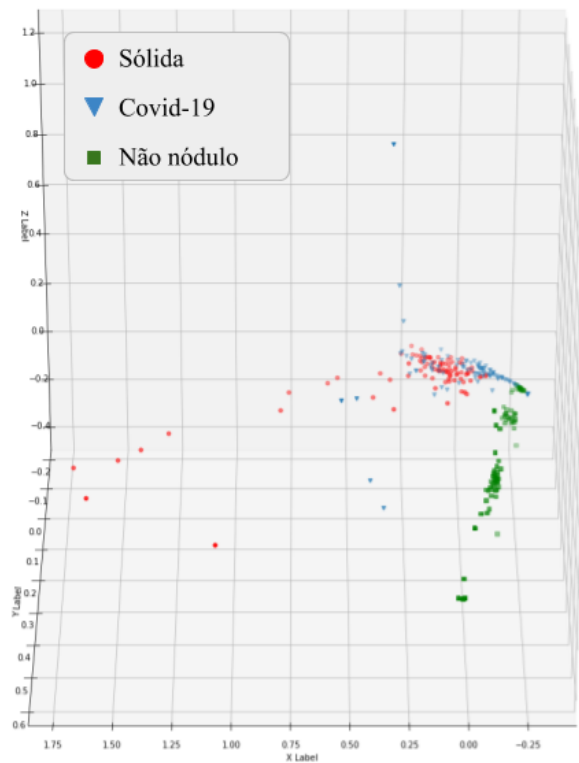


Figura 4 – Representação do PCA com três componentes para as características extraídas no caso de teste desta seção. Autoria própria.

nas lesões de COVID-19 e sólida, por esse motivo o classificador identificou-a como uma classe não nodular.

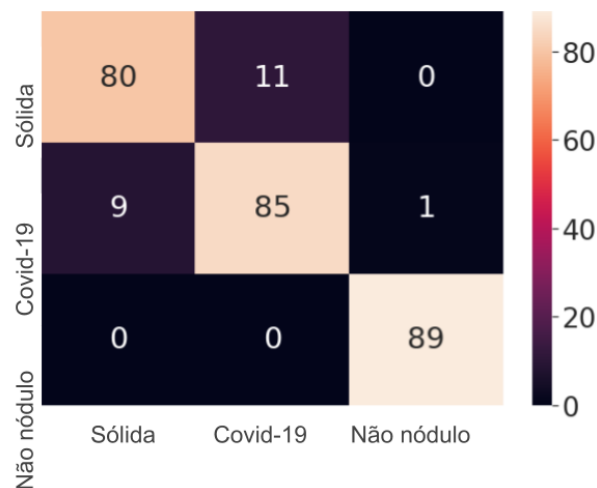


Figura 5 – Resultado da matriz de confusão. Autoria própria.

Também observa-se o histograma das imagens selecionadas na análise anterior para fornecer uma visualização da distribuição dos *voxels* das imagens, ilustrada na Figura 7. Além disso, a fim de apresentar uma visualização das características dessas imagens, foi criada a Figura 8. Sobre a Figura 8, é possível notar que o gráfico da Figura 8(e) foge do padrão visto nos outros gráficos da Figura 8, sendo ela a única instância classificada

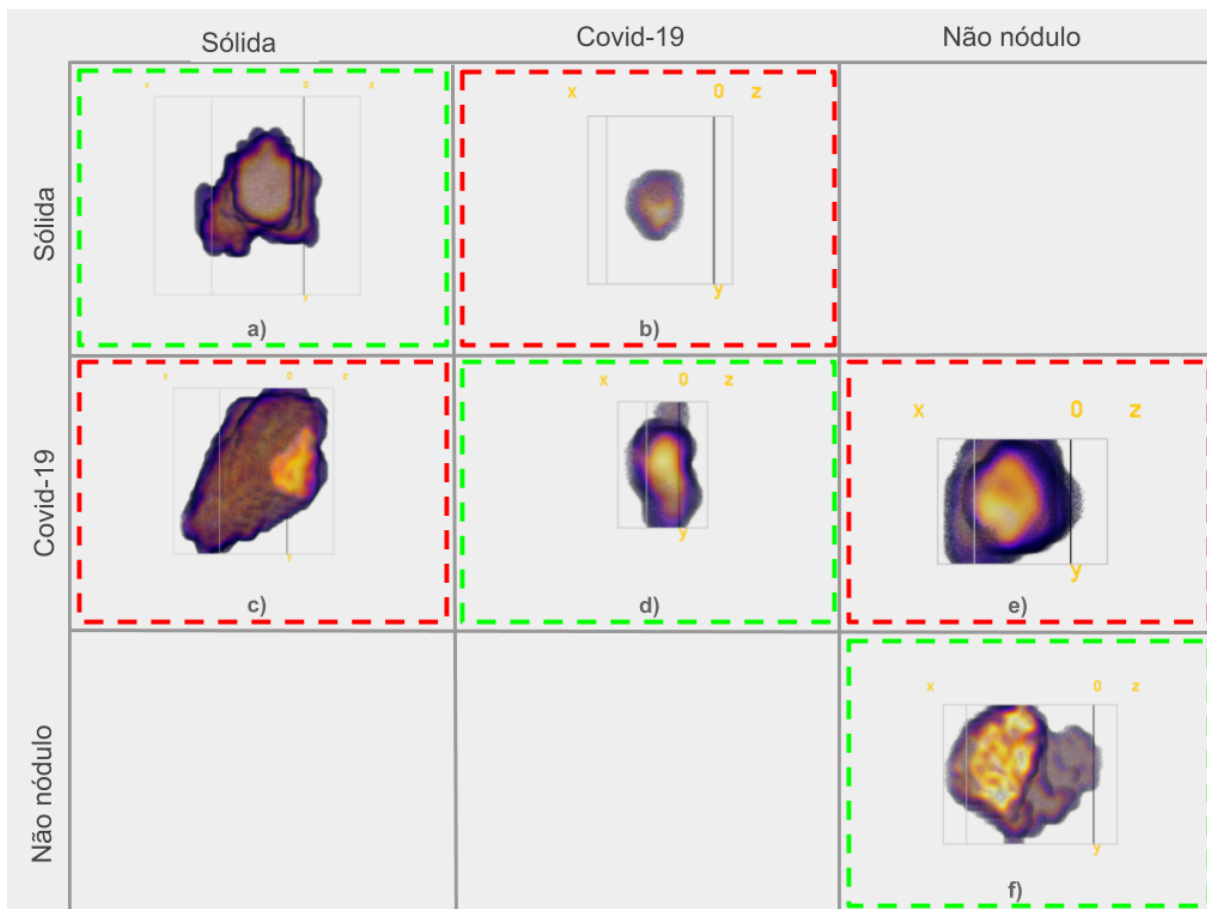


Figura 6 – Representação de uma matriz de confusão com casos de acerto e erro do classificador. Autoria própria.

erroneamente como não nodular. Em relação aos gráficos das Figuras 8(d) e 8(b), é possível notar que eles possuem similaridade em relação aos valores de índice, o que pode ter motivado a classificação errônea como COVID-19, sendo ela uma lesão solida. O mesmo acontece no caso dos gráficos das Figuras 8(a) e 8(c), onde há uma similaridade entre os valores, conseqüentemente, dificultando o processo de classificação

5.1.2 Comparação com trabalhos relacionados

Nesta Subseção, foi realizado uma comparação dos resultados obtidos neste trabalho com os estudos apresentados no Capítulo 3. O objetivo dessa comparação é apenas relacionar os resultados alcançados pelo método proposto com o estado da arte, sem prejudicar ou diminuir nenhum dos métodos mencionados. A Tabela 10 aponta esta comparação. Em relação aos resultados do método proposto, destaca-se na tabela apenas os resultados do classificador XGBoost no cenário 1, pelo fato de que esse é o cenário mais próximo de um ambiente real, já que contempla imagens de diferentes fontes, aumentando a confiabilidade da solução.

Observa-se, através da Tabela 10, que, de modo geral, o trabalho proposto apresenta

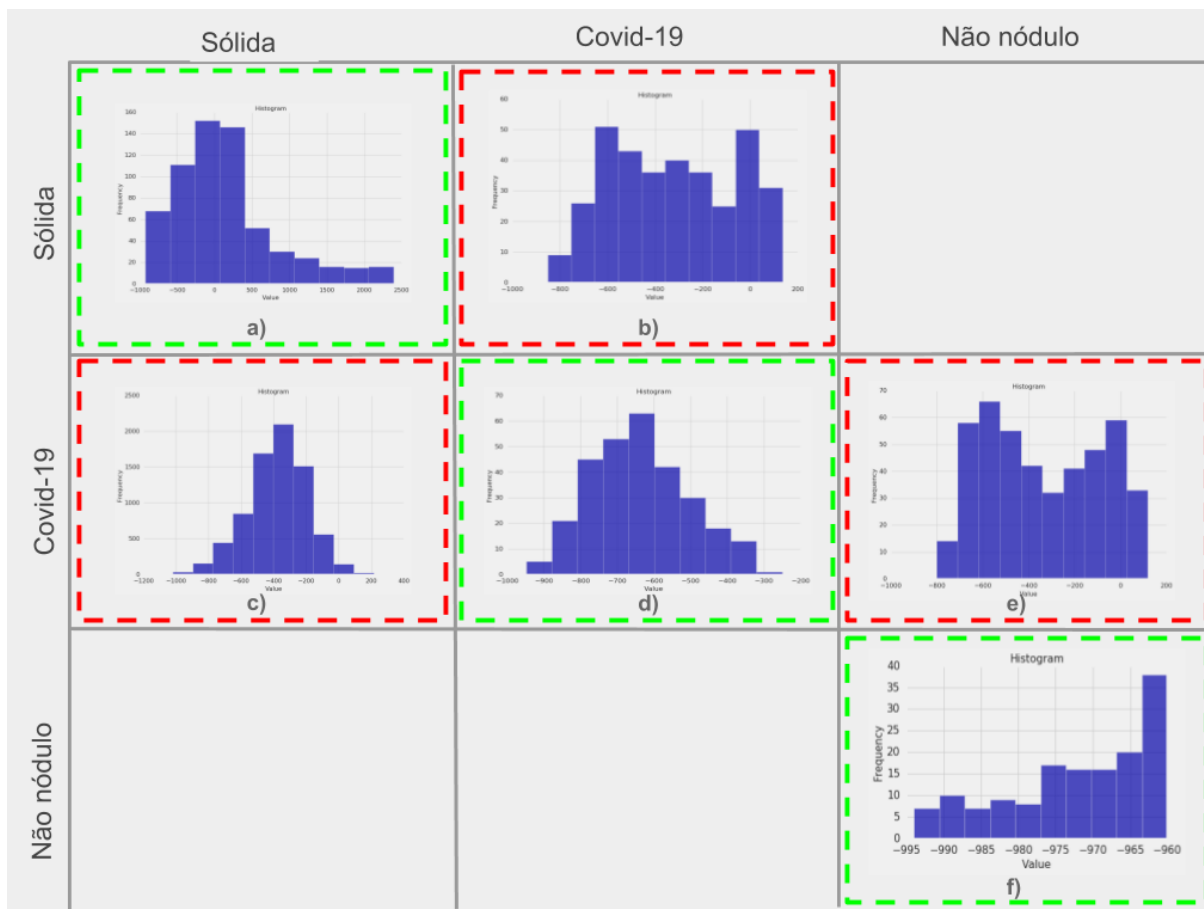


Figura 7 – Histograma das imagens selecionadas da matriz de confusão. Autoria própria.

resultados promissores tão quanto os demais estudos abordados. Inclusive, alcançamos acurácia, recall e F1 superiores aos demais. Vale ressaltar que apenas a nossa abordagem aplica uma classificação multiclasse, o que por si só, já torna a solução mais complexa. Além disso, foi utilizado o maior conjunto de imagens entre todos os estudos relacionados.

O ideal para um sistema CAD é atingir um equilíbrio entre as métricas de validação, pois deve diagnosticar corretamente indivíduos saudáveis e doentes. Diante dos resultados apresentados, observa-se que o x método se mostrou promissor na classificação das *VOIs* em todos os cenários de teste, enfatizando a relevância dos índices filogenéticos para a descrição das imagens de CT.



Figura 8 – Representação dos índices extraídos das imagens selecionadas da matriz de confusão. Autoria própria.

Tabela 10 – Comparação com trabalhos relacionados.

Trabalho	Tipo de exame	Objetivo	Acc (%)	Rec (%)	Prec (%)	F1 (%)	AUC
(ABBAS; ABDELSAMEA; GABER, 2020)	X-ray	COVID-19 e Não COVID-19	95,12	97,91	-	-	-
(NARIN; KAYA; PAMUK, 2020)	X-ray	COVID-19 e Não COVID-19	98,00	96,00	100	98,00	-
(ZHAO et al., 2020)	CT	COVID-19 e Não COVID-19	89,00	-	-	90,00	0,98,00
(OZKAYA; OZTURK; BARSTUGAN, 2020)	CT	COVID-19 e Não COVID-19	98,27	98,93	97,63	98,28	-
(WANG et al., 2020)	CT	COVID-19 e Não COVID-19	82,9	84,00	-	-	-
(HE et al., 2020)	CT	COVID-19 e Não COVID-19	86,00	-	-	85,00	0,94
(BARSTUGAN; OZKAYA; OZTURK, 2020)	CT	COVID-19 e Não COVID-19	98,71	97,56	99,62	98,58	-
(JAISWAL et al., 2020)	CT	COVID-19 e Não COVID-19	96,25	96,29	96,29	96,29	-
(ZHENG et al., 2020)	CT	COVID-19 e Não COVID-19	-	90,7	-	-	0,95
Método proposto	Sem Balanceamento	COVID-19, Solido e Não nódulo	99,25	99,25	99,24	99,24	0,96
	Com Balanceamento	COVID-19, Solido e Não nódulo	92,81	92,81	92,87	92,80	0,95

6 Discussões

Conforme apresentado no decorrer deste trabalho, a classificação de COVID-19 não é uma tarefa trivial, principalmente por se tratar de um problema recente, onde ainda não foi amplamente abordado pela literatura. Uma das principais dificuldades na exploração de abordagens voltadas à esse problema é a escassez de dados de exames.

Nos estudos relacionados, é possível notar os esforços na classificação da COVID-19 usando diferentes abordagens, sendo a maioria baseadas em *deep learning*. No entanto, embora os modelos de CNN tenham se mostrado abordagens cada vez mais capazes de representar imagens de forma eficiente, por meio de mapas de características e camadas profundas, essa eficácia depende de vários fatores que muitas vezes tornam o uso das CNNs mais complexo do que outros métodos aplicados ao mesmo problema, por exemplo, i) o problema do balanceamento de classes para obter um bom aprendizado nos modelos; ii) parametrização adequada ao problema; iii) exigir bom hardware dependendo do problema; iv) ampla quantidade de dados para identificar os padrões corretos; e, finalmente, v) cada arquitetura requer uma dimensão diferente das imagens de entrada.

Em relação aos trabalhos apresentados no Capítulo 3, apenas (ZHENG et al., 2020) aplicou volumes de CT em seu método. Já a abordagem proposta neste trabalho apresenta um método simples e eficiente para a detecção de COVID-19 em exames de CT, utilizando volumes 3D das lesões. Os índices de diversidade filogenética não necessitam de uma parametrização complexa, também não se limitam à quantidade de bits das imagens, não requerem padronização no tamanho das entradas e são escaláveis, sendo possível aplicá-los a imagens 2D e 3D. Além disso, em união aos classificadores utilizados nos experimentos, os índices mostraram-se promissores mesmo sem a aplicação de técnicas de aumento ou pré-processamento dos dados.

A seguir, alguns pontos relevantes sobre o trabalho:

- O método proposto foi testado em dois cenários de classificação: um deles usando o *cross validation* com $k=10$ e outro com $k=5$. O intuito em variar o parâmetro k era avaliar como o método se comporta quando é treinado com quantidades de dados diferentes;
- Foram realizados três experimentos combinando as bases utilizadas neste trabalho, visando avaliar o método com amostras de imagens provindas de diferentes fontes. Cada um dos experimentos foi executado com as amostras de dados balanceadas e não balanceadas, proporcionando uma visão mais crítica sobre como o método se comporta diante de conjuntos desbalanceados ou não;

- Os resultados alcançados mostraram-se promissores e consistentes em todos os cenários de teste e classificadores utilizados, motivando a aplicação dos descritores em ambientes reais;
 - Os índices caracterizam as imagens em suas dimensões originais, sem a necessidade de um redimensionamento, preservando a integridade das características das regiões;
 - Foram utilizados os oito índices para a análise de textura pelo fato de que essa combinação pode caracterizar melhor a lesão, uma vez que cada índice mede separadamente alguma propriedade que o outro não é capaz de medir. Dessa forma, os índices combinados podem fornecer aos classificadores uma descrição mais robusta das imagens.
-

7 Conclusão

Este trabalho propôs um método capaz de descrever e classificar regiões de *CT* em três classes: COVID-19, sólidas e não nódulos. Para tanto, foram utilizadas características de textura baseadas nos índices de diversidade filogenética. A abordagem alcançou acurácia de 99.25% e AUC de 0.96 nos dados desbalanceados, enquanto que nos dados balanceados obteve acurácia de 92.81% e AUC de 0.95. Acredita-se que os resultados promissores se devam ao fato de os índices filogenéticos podem potencializar as características de textura das *VOIs*, uma vez que cada um deles contribui para que um atributo específico da região seja adicionado aos demais para se chegar a uma distinção satisfatória das classes.

Assim, acredita-se que o método apresentado neste trabalho seja capaz de integrar uma ferramenta CAD, podendo ser aplicada em casos reais no auxílio ao diagnóstico de COVID-19. Como abordado durante o trabalho, a detecção de COVID-19 vem sendo um desafio para os profissionais da saúde, visto que é um problema recente que, apesar de não apresentar uma alta taxa de letalidade, precisa ser diagnosticado em seus estágios iniciais para evitar a progressão da doença e o contágio para outras pessoas. Sendo assim, a abordagem traz benefícios tanto para o especialista, que pode contar com uma segunda opinião durante o diagnóstico, quanto para o paciente, visto que a detecção precoce é fundamental para que o paciente alcance o tratamento em tempo hábil, aumentando as chances de cura.

8 Publicações

O proponente do presente trabalho alcançou uma publicação a nível internacional, a saber:

- P. R. Sales dos Santos, V. de Carvalho Brito, A. O. de Carvalho Filho, F. Henrique Duarte de Araújo, R. de Lira Rabêlo e M. Joseph Mathew, "A Capsule Network-based for identification of Glaucoma in retinal images", **2020 IEEE Symposium on Computers e Communications (ISCC)**, Rennes, France, 2020, pp. 1-6, doi: 10.1109/ISCC50000.2020.9219708.
- Patrick R. S. dos Santos, Lucas B. M. de Souza, Samuel P. B. D. Lélis, Hector B. Ribeiro, Fabbio A. S. Borges, Romuere R. V. Silva, Antonio Oseas Carvalho Filho, Flavio H. D. Araujo, Ricardo de Andrade Lira Rabêlo, e Joel J. P. C. Rodrigues, "Prediction of COVID-19 using Time-SlidingWindow: the case of Piauí State - Brazil", **IEEE International Conference on E-health Networking, Application and Services**, China, 2021.

Além desta, foram submetidas as seguintes publicações, sendo que uma delas se trata deste trabalho e as outras exploram a COVID-19 e os índices filogenéticos de forma distinta:

- Título: COVID-index: a texture-based approach to classification of lung lesions. Submetido em: **Pattern Recognition**.
- Título: Texture descriptors based on species relationships for classification of skin lesions. Submetido em: **Journal of Digital Imaging**.

Referências

- ABBAS, A.; ABDELSAMEA, M. M.; GABER, M. M. Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network. *arXiv preprint arXiv:2003.13815*, 2020. Citado 3 vezes nas páginas 27, 28 e 40.
- AZEVEDO, E.; CONCI, A.; VASCONCELOS, C. *Computação gráfica: Teoria e prática: geração de imagens*. [S.l.]: Elsevier Brasil, 2018. Citado na página 18.
- BARSTUGAN, M.; OZKAYA, U.; OZTURK, S. Coronavirus (covid-19) classification using ct images by machine learning methods. *arXiv preprint arXiv:2003.09424*, 2020. Citado 3 vezes nas páginas 27, 28 e 40.
- BAXEVANIS, A. D.; OUELLETTE, B. F. *Bioinformatics: a practical guide to the analysis of genes and proteins*. [S.l.]: John Wiley & Sons, 2004. Citado 3 vezes nas páginas 8, 18 e 19.
- BERNHEIM, A. et al. Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection. *Radiology*, Radiological Society of North America, p. 200463, 2020. Citado na página 14.
- BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 25.
- CAREY, C. L. et al. Additive deleterious effects of methamphetamine dependence and immunosuppression on neuropsychological functioning in hiv infection. *AIDS and Behavior*, Springer, v. 10, n. 2, p. 185, 2006. Citado na página 18.
- CARVALHO, E. D. et al. Breast cancer diagnosis from histopathological images using textural features and cbir. *Artificial Intelligence in Medicine*, Elsevier, v. 105, p. 101845, 2020. Citado na página 25.
- CHEN, T. et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, p. 1–4, 2015. Citado na página 25.
- CIANCIARUSO, M. V.; SILVA, I. A.; BATALHA, M. A. Phylogenetic and functional diversities: new approaches to community ecology. *Biota Neotropica*, SciELO Brasil, v. 9, n. 3, p. 93–103, 2009. Citado na página 18.
- CLARKE, K.; WARWICK, R. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, v. 35, p. 523 – 531, 08 1998. Citado 3 vezes nas páginas 19, 20 e 21.
- ERKEL, A. R. van; PATTYNAMA, P. M. Receiver operating characteristic (roc) analysis: Basic principles and applications in radiology. *European Journal of Radiology*, v. 27, n. 2, p. 88 – 94, 1998. ISSN 0720-048X. Citado na página 26.
- FAITH, D. P. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, v. 61, n. 1, p. 1 – 10, 1992. ISSN 0006-3207. Disponível em: <<http://www.sciencedirect.com/science/article/pii/0006320792912013>>. Citado na página 19.

FILHO, A. O. de C. et al. Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. *Artificial intelligence in medicine*, Elsevier, v. 60, n. 3, p. 165–177, 2014. Citado na página 15.

FILHO, A. O. de C. et al. Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. *Artificial intelligence in medicine*, Elsevier, v. 60, n. 3, p. 165–177, 2014. Citado na página 30.

FILHO, A. O. de C. et al. Classification of patterns of benignity and malignancy based on ct using topology-based phylogenetic diversity index and convolutional neural network. *Pattern Recognition*, Elsevier, v. 81, p. 200–212, 2018. Citado na página 14.

GOLATKAR, A.; ANAND, D.; SETHI, A. Classification of breast cancer histology using deep learning. In: SPRINGER. *International Conference Image Analysis and Recognition*. [S.l.], 2018. p. 837–844. Citado na página 28.

GORBALENYA, A. E. et al. Severe acute respiratory syndrome-related coronavirus: The species and its viruses – a statement of the coronavirus study group. *bioRxiv*, 2020. Disponível em: <<https://www.biorxiv.org/content/early/2020/02/11/2020.02.07.937862>>. Citado na página 14.

HANCOCK, M. C. *Pylidc - An object relational mapping for the LIDC dataset using SQLAlchemy*. 2016. <https://pylidc.github.io/>. Citado na página 30.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, Ieee, n. 6, p. 610–621, 1973. Citado na página 18.

HE, X. et al. Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medRxiv*, Cold Spring Harbor Laboratory Press, 2020. Citado 4 vezes nas páginas 27, 28, 29 e 40.

HELMUS, M. et al. Phylogenetic measures of biodiversity. *The American Naturalist*, [The University of Chicago Press, The American Society of Naturalists], v. 169, n. 3, p. E68–E83, 2007. ISSN 00030147, 15375323. Citado 2 vezes nas páginas 19 e 20.

III, S. G. A. et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, Wiley Online Library, v. 38, n. 2, p. 915–931, 2011. Citado na página 30.

JAISWAL, A. et al. Classification of the covid-19 infected patients using densenet201 based deep transfer learning. *Journal of Biomolecular Structure and Dynamics*, Taylor & Francis, p. 1–8, 2020. Citado 4 vezes nas páginas 14, 27, 28 e 40.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group, v. 521, n. 7553, p. 436, 2015. Citado na página 28.

MEDSEG. *COVID-19 CT segmentation dataset*. 2020. <http://medicalsegmentation.com/covid19/>. Citado na página 30.

MILLET, G. P. et al. Altitude and covid-19: Friend or foe? a narrative review. *Physiological Reports*, Wiley Online Library, v. 8, n. 24, p. e14615, 2021. Citado na página 17.

- NARIN, A.; KAYA, C.; PAMUK, Z. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*, 2020. Citado 4 vezes nas páginas 27, 28, 29 e 40.
- NAWAZ, W. et al. Classification of breast cancer histology images using alexnet. In: SPRINGER. *International Conference Image Analysis and Recognition*. [S.l.], 2018. p. 869–876. Citado na página 28.
- NETTO, S. M. B. et al. Automatic segmentation of lung nodules with growing neural gas and support vector machine. *Computers in biology and medicine*, Elsevier, v. 42, n. 11, p. 1110–1121, 2012. Citado na página 15.
- OZKAYA, U.; OZTURK, S.; BARSTUGAN, M. Coronavirus (covid-19) classification using deep features fusion and ranking technique. *arXiv preprint arXiv:2004.03698*, 2020. Citado 4 vezes nas páginas 17, 27, 28 e 40.
- PEDRINI, H.; SCHWARTZ, W. R. *Análise de imagens digitais: princípios, algoritmos e aplicações*. [S.l.]: Thomson Learning, 2008. Citado na página 18.
- RAKHLIN, A. et al. Deep convolutional neural networks for breast cancer histology image analysis. In: SPRINGER. *International Conference Image Analysis and Recognition*. [S.l.], 2018. p. 737–744. Citado na página 28.
- REMUZZI, A.; REMUZZI, G. Covid-19 and italy: what next? *The Lancet*, Elsevier, v. 395, n. 10231, p. 1225–1228, 2020/08/11 2020. Disponível em: <[https://doi.org/10.1016/S0140-6736\(20\)30627-9](https://doi.org/10.1016/S0140-6736(20)30627-9)>. Citado na página 14.
- RODRIGUES, T. S. et al. Inflammasomes are activated in response to sars-cov-2 infection and are associated with covid-19 severity in patients. *Journal of Experimental Medicine*, The Rockefeller University Press, v. 218, n. 3, 2021. Citado na página 17.
- SHAN, F. et al. Lung infection quantification of covid-19 in ct images with deep learning. *arXiv preprint arXiv:2003.04655*, 2020. Citado na página 14.
- SINGH, D. et al. Classification of covid-19 patients from chest ct images using multi-objective differential evolution-based convolutional neural networks. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*, April 2020. ISSN 0934-9723. Disponível em: <<https://europepmc.org/articles/PMC7183816>>. Citado na página 14.
- TAN, X. et al. Molecular stratification by bcl2a1 and aim2 provides additional prognostic value in penile squamous cell carcinoma. *The Lancet Oncology*, Elsevier, v. 22, n. 3, p. 1364, 2021. Citado na página 17.
- WANG, S. et al. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). *MedRxiv*, Cold Spring Harbor Laboratory Press, 2020. Citado 4 vezes nas páginas 14, 27, 28 e 40.
- WEBB, C. O.; LOSOS, A. E. J. B. Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. *The American Naturalist*, [The University of Chicago Press, The American Society of Naturalists], v. 156, n. 2, p. 145–155, 2000. ISSN 00030147, 15375323. Citado 2 vezes nas páginas 19 e 20.

WHO, W. H. O. *Coronavirus disease (COVID-19) outbreak situation*. 2021. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>. Accessed on 2021-01-01. Citado na página 14.

ZHAO, J. et al. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 2020. Citado 5 vezes nas páginas 17, 27, 28, 29 e 40.

ZHENG, C. et al. Deep learning-based detection for covid-19 from chest ct using weak label. *medRxiv*, Cold Spring Harbor Laboratory Press, 2020. Citado 5 vezes nas páginas 27, 28, 29, 40 e 41.



**TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA
“JOSÉ ALBANO DE MACEDO”**

Identificação do Tipo de Documento

- () Tese
- () Dissertação
- (X) Monografia
- () Artigo

Eu, **Patrick Ryan Sales dos Santos**, autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de 02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar, gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação **Índices COVID: Uma Abordagem Baseada em Textura para Classificar Lesões Pulmonares** de minha autoria, em formato PDF, para fins de leitura e/ou impressão, pela internet a título de divulgação da produção científica gerada pela Universidade.

Picos-PI, 22 de março de 2021.

Patrick Ryan Sales dos Santos

Assinatura

Patrick Ryan Sales dos Santos

Assinatura