Caracterização de imagens com COVID-19 usando Haralick, Transformada Wavelet e K-means

Pedro Moisés de Sousa
Faculdade de Engenharia Elétrica
Universidade Federal de Uberlândia
Uberlândia, Brasil
ORCID: 0000-0003-4563-0033

Ana *Claúdia Patrocinio*
Faculdade de Engenharia Elétrica
Universidade Federal de Uberlândia
Uberlândia, Brasil
ORCID: 0000-0002-1645-1112

***Abstract*— With the appearance of COVID-19, several researches were conducted to better understand and fight the disease. In this work, 1000 images were randomly selected from the BIMCV database, 500 from the COVID-19 disease and 500 from other lung diseases such as pneumonia, pneumothorax, etc. From there, Haralick and Wavelet attribute extraction techniques were used. Then, the extracted attributes and attributes (individual or combined) were analyzed using the KMEANS grouping technique, which checks the best characteristics (individual or combined) for separating COVID-19 from other lung diseases. Among 14 attributes of the Haralick technique, the attributes contrast, variance, sum mean, variance sum, sum entropy, entropy, difference variance, difference entropy, correlation information measure II and correlation coefficients achieved an accuracy in around 95.5% acting individually. Among the 8 attributes extracted from Wavelet, the vertical and diagonal standard deviation achieved accuracy of 95.3 and 95.5%, respectively. The results of the article show that the combination of attributes of the Haralick and Wavelet techniques did not outperform the results of the individual attributes.**

Keywords — Haralick, Covid-19, Wavelet, Clusters, K-means

I. INTRODUÇÃO

Em dezembro de 2019, um grupo de pacientes com pneumonia de causa desconhecida foi vinculado a um atacado de comidas marinhas em Wuhan, Província de Hubei, China e a epidemia espalhou rapidamente para outras partes do país e além se tornando uma pandemia, no presente momento [1,2]. Além disso, já matou mais de 500 mil pessoas no Brasil (26/07/2021) de acordo com os dados do ministério da saúde [3].

Através do sequenciamento imparcial de amostras dos pacientes, foi possível descobrir um novo tipo de beta coronavírus. Esse novo coronavírus, nomeado de 2019-nCoV, foi comparado a Síndrome Respiratória Aguda Severa (SARS) e a Síndrome Respiratória do Oriente Médio (MERS) apresentando uma maior capacidade de transmissão [1,2] .

Os coronavírus são vírus de RNA envelopados capazes de causar doenças respiratórias, entéricas, hepáticas e neurológicas em animais domésticos e selvagens bem como nos humanos [1,3]. Muitos vírus existem na natureza há muito tempo e devido a alta prevalência, ampla distribuição, grande diversidade genética, frequente recombinação genética e aumento das atividades de interface homem animal, é provável que aumente a disseminação de vírus dos hospedeiros naturais para os seres humanos. Tanto o SARS quanto o MERS foram ligados a uma origem zoonótica, sendo transmitidos de civetas de mercados e camelos dromedários, respectivamente. É provável que estes eventos, assim como o novo coronavírus serão eventos frequentes caso não sejam estabelecidas barreiras entre a sociedade humana e a natureza [1,4].

Atualmente o diagnóstico da doença é realizado através detecção do ácido nucleico viral usado a reação em cadeia da polimerase com transcrição reversa em tempo real (RT – PCR), e este é o método aceito como padrão ouro, mesmo possuindo uma taxa de acurácia entre 30 e 50% [6,7,8,9]. O problema deste método é a sua indisponibilidade em regiões e países afetados e a incapacidade de providenciar testes suficientes para os milhares de pacientes suspeitos [6]. Além disso, a demora para o recebimento dos resultados e a quantidade de resultados falsos- negativos fez com que os pesquisadores começassem a atuar em diferentes campos para driblar este problema [7]. Desta forma, vários estudos vêm sendo realizados para entendimento, auxílio ao prognóstico e combate da doença através de exames de radiografias de tórax ou de Tomografia Computadorizada (CT) [7]. Como os pacientes acometidos pela doença desenvolvem uma infecção nos pulmões, os exames de Tomografia Computadorizada se mostraram uma técnica de imagem eficiente para detecção da doença e classificação da sua progressão [6,8]. Entretanto, a imagem obtida pode apresentar similaridades com imagens obtidas de pacientes acometidos por outras doenças que causam pneumonia viral especificamente os vírus da mesma família (SARS e MERS) e essa diferenciação se torna um desafio para encontrar um exame de imagem que tenha uma boa acurácia e que auxilie os profissionais de saúde no prognostico da doença [7,9].

Assim sendo, o objetivo de vários estudos é desenvolver técnicas de imagens médicas capazes de ajudar no prognóstico do novo coronavírus, seja para os exames de CT, seja para os exames de Raio X [7]. O problema dos exames de CT é a sua menor disponibilidade, o seu maior custo e maior tempo de aquisição de imagem em relação ao exame de radiografia de tórax, nas regiões afetadas [6].

Assim, este trabalho propõe uma análise de uma base pública de imagens com pacientes com a doença COVID-19 e imagens com pacientes com outras doenças pulmonares. Esta análise será feita utilizando técnicas de extração de atributos (descritores de *Haralick* e coeficientes *Wavelets*), tais atributos serão submetidos a técnica de seleção de atributos por clusterização *k-means*, afim de avaliar quais os melhores atributos a serem utilizados na caracterização de imagens de pacientes com COVID-19 de outras doenças pulmonares.

II. METODOLOGIA

A. Banco de dados BIMCV

O banco de dados é composto por imagens do repositório Valencian Region Medical ImageBank (BIMCV) [1], dividido em :

- BIMCV-COVID19 + é um grande conjunto de dados com imagens de radiografias de tórax CR e imagens de tomografia computadorizada (CT) de **pacientes**

COVID-19, juntamente com seus achados radiográficos, patologias, reação em cadeia da polimerase (PCR), imunoglobulina G (Testes de diagnóstico de anticorpos IgG) e imunoglobulina M (IgM) e relatórios radiográficos do Banco de Dados de Imagens Médicas do Banco de Imagens Médicas da Região de Valência (BIMCV).

- O conjunto de dados BIMCV-COVID19- é um grande conjunto de dados com imagens de radiografia de tórax CR e imagens de tomografia computadorizada (CT) onde as doenças pulmonares não são COVID-19, juntamente com seus achados radiográficos, patologias, reação em cadeia da polimerase (PCR), imunoglobulina Testes de diagnóstico de anticorpos G (IgG) e imunoglobulina M (IgM) e laudos radiográficos do Medical Imaging Databank do Banco de Imagens Médicas da Região Valenciana (BIMCV).

B. Critérios de seleção da base de dados

Nesta etapa, selecionamentos 1000 imagens (radiografias de tórax- CR) do banco BIMCV, sendo 500 imagens BIMCV-COVID19 + e 500 imagens BIMCV-COVID19 -.

Para a escolha dessas imagens foram criados alguns critérios de inclusão e exclusão:

- Inclusão→ Imagens com o posicionamento RX tórax – antero-posterior (AP) (Fig.1); imagens de adultos; imagem de 16 bits/pixel.

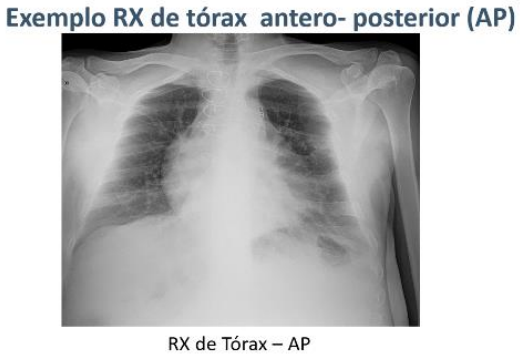

Fig.1. Posicionamento AP (Elaborado pelo autor).

- Exclusão→Imagens menores que 1000x1000 (Fig. 2a); imagens com posicionamento RX tórax-perfil(Fig. 2b); imagens de crianças(Fig. 2c);, imagens visualmente distorcidas ou que não apresentem a área de interesse que é o pulmão (Fig. 2d).

Fig. 2. Critérios de exclusão da base BIMCV (Elaborado pelo autor)

C. Extração de atributos

A partir das imagens de radiografias de tórax foram utilizadas duas técnicas para extração de atributos:

Extração de atributos *Haralick* [2], o qual é utilizado para o cálculo da textura de uma imagem digital, o qual é constituído por 14 medidas estatísticas, apresentada na TABELA 1. Este utiliza a matriz de co-ocorrência de níveis de cinza. A matriz de co-ocorrência é uma matriz quadrada que tem como tamanho a quantidade de níveis de cinza da imagem a ser analisada. As combinações de ocorrência entre os níveis de cinza são calculadas nos ângulos 0, 45, 90 e 135, os demais ângulos são calculados via simetria. Após o cálculo da matriz uma outra é calculada, a nova matriz é a de probabilidade de ocorrência das combinações entre os níveis de cinza. Com essa matriz são feitos cálculos dos 14 atributos de textura apresentados na TABELA 1. [2]

Extração de *Wavelet* são funções geradas a partir de uma única função (função base) chamada *wavelet* mãe por escalas e translações no domínio do tempo (frequência). A Transformada *Wavelet* Discreta (DWT) pode ser considerada como uma sequência de números que mostra uma certa função contínua. Quando as imagens digitais são tratadas em várias resoluções, o DWT é uma ferramenta matemática viável. Além de sua estrutura eficiente e altamente intuitiva para representação e armazenamento de imagens multirresolução, o DWT fornece uma visão poderosa das características espaciais e de frequência de uma imagem [3,4,5]. Há uma ampla gama de escolha para *wavelet* mãe, dentre elas: *daubechies, symlets, coiflet*, entre outras. Assim, de forma resumida a transformada *wavelet* executa primeiro uma etapa da transformada em todas as linhas, produzindo uma matriz em que o lado esquerdo contém os coeficientes *low pass* (*L*) de cada linha e o direito contém os coeficientes *high pass* (*H*). Em seguida, é aplicada a todas as colunas, resultando em quatro tipos de coeficientes, conforme ilustrado na Fig. 3.

- As características diagonais da imagem são geradas a partir de uma convolução com o filtro *high pass* em ambas as direções (HH).
- As características horizontais da imagem são geradas a partir de uma convolução do filtro *low pass* nas colunas, seguido pelo filtro *high pass* nas linhas (*LH*).
- As características verticais da imagem são geradas a partir de uma convolução do filtro high pass nas linhas, seguido pelo *filtro low* pass nas colunas (*LH*).
- O Coeficiente de aproximação da imagem é gerado a partir de uma convolução com o filtro *low pass* (LL) em ambas as direções.

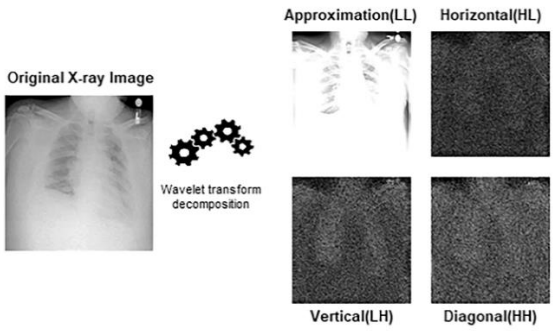

Fig. 3 Diagrama geral da transformada Wavelet (Elaborado pelo autor).

D. *Técnica de Clusterização K-means*

Para seleção dos melhores atributos extraídos, foi utilizada a técnica de clusterização *K-means*, que consiste em dividir os dados ou população em grupos distintos conforme a semelhança dos dados apresentados, ou seja, agrupar padrões conforme a semelhança apresentada. O agrupamento *K-means* [4,6,7] é um método comumente utilizado para subdividir automaticamente um conjunto de dados em **n** grupos. O algoritmo é que determinou o número de **2** de centroides e posteriormente seguiu refinando-os de forma iterativa. Primeiramente a cada dado é atribuído a um centroide com a menor distância **d**, posteriormente os centroides são ajustados para que seu valor seja a média dos dados próximos à ele [4,6,7].

A partir da extração de atributos *Haralick*, *Wavelet* e o agrupamento *k-means*, temos as seguintes etapas:

- Primeira etapa: calcular as 14 características da extração de atributos *Haralick*, calcular a média dos 4 ângulos para cada características do *Haralick* (TABELA *1*).

TABELA 1 média dos atributos *Haralick*

x	**Média dos ângulos 0, 45, 90 e 135**
1H	Segundo momento angular (Energia)
2H	Contraste
3H	Correlação
4H	Variância
5H	Momento de diferença inverso (Homogeneidade)
6H	Média da soma
7H	Soma da variância
8H	Entropia da soma
9H	Entropia
10H	Variância da Diferença
11H	Entropia da Diferença
12H	Medidas de informação da correlação I
13H	Medidas de informação da correlação II
14H	Coeficiente de Correlação Máximo

- Segunda etapa: aplicar a média dos atributos Haralick (um a um) no agrupamento *k-means*.
- Terceira etapa: aplicar a combinação das médias [exemplo: 1H:3H, se refere a combinação de 3 atributos: energia-1H, contraste-2H e correlação-3H, seguindo a ordem da Tabela1] dos atributos *Haralick* no agrupamento *k-means.*
- Quarta etapa: calcular a média e o desvio padrão da wavelet mãe coif5 dos coeficientes aproximação, horizontal, vertical e diagonal, gerando 8 características mostradas na TABELA *2*.

TABELA 2. atributos wavelet

y	***Atributos***
1W	Média do coeficiente aproximadada
2W	Média do coeficiente horizontal
3W	Média do coeficente vertical
4W	Média do coeficiente diagonal
5W	Desvio Padrão do coeficiente aproximadada
6W	Desvio Padrão do coeficiente horizontal
7W	Desvio Padrão do coeficente vertical
8W	Desvio Padrão do coeficiente diagonal

- Quinta etapa: aplicar a média e desvio padrão dos coeficientes da *wavelet* (um a um) no agrupamento *k-means*.
- Sexta etapa: aplicar a combinação das médias e desvios padrões [exemplo: 2W:4W, se refere a combinação de 3 atributos: média horizontal-2W, média vertical-3W e média diagonal-4W, seguindo a ordem da Tabela 2] dos coeficientes da *wavelet* no agrupamento *k-means*.
- Sexta etapa: Comparar os resultados.

III. RESULTADOS

O trabalho gerou resultados para cada etapa da metodologia, para um total de 1000 imagens de radiografias de tórax -CR, sendo 500 com a doença COVID-19 e 500 com outras patologias pulmonares, apresentados a seguir.

A. *Primeira e segunda etapas*

Para a média dos atributos de *Haralick*, os atributos variância, média da soma, soma da variância, entropia da soma, entropia da diferença, medidas de informação da correlação II e coeficientes de correlação máximo apresentaram uma maior acurácia de 95,50%. Sendo a correlação e medidas de informação da correlação II uma menor acurácia de 50%. Conforme mostra a Fig. 4.

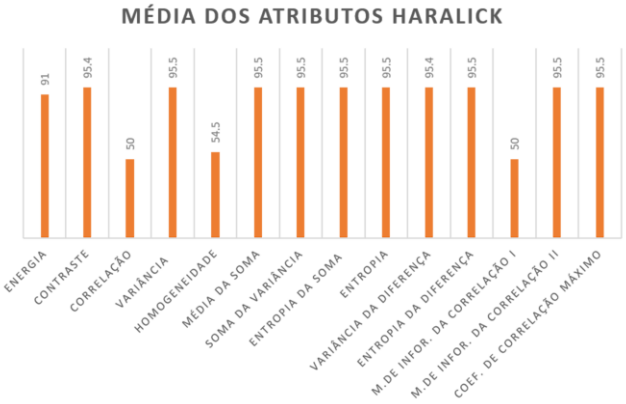

Fig. 4.Média dos atributos *Haralick* (Elaborado pelo autor).

B. *Terceira etapa*

Foram feitas combinações dos atributos *Haralick*, onde 1H:2H representa a combinação dos atributos Energia e

contraste, 4H:7H representa a combinação dos atributos variância, homogeneidade, média da soma, soma da variância, assim sucessivamente. A combinação 1H:2H(energia e contraste) apresentou uma acurácia de 91%, as combinações 4H:5-11H, 5H:6-11H, 6H:7-11H, 7H:8-11H, 8H:9-11, 9H:10-11H, 10H:11H, 13H:14H apresentaram acurácia de 95,5%, sendo o restante das combinações valores próximos de 50%. Mostrados na Fig. 5.

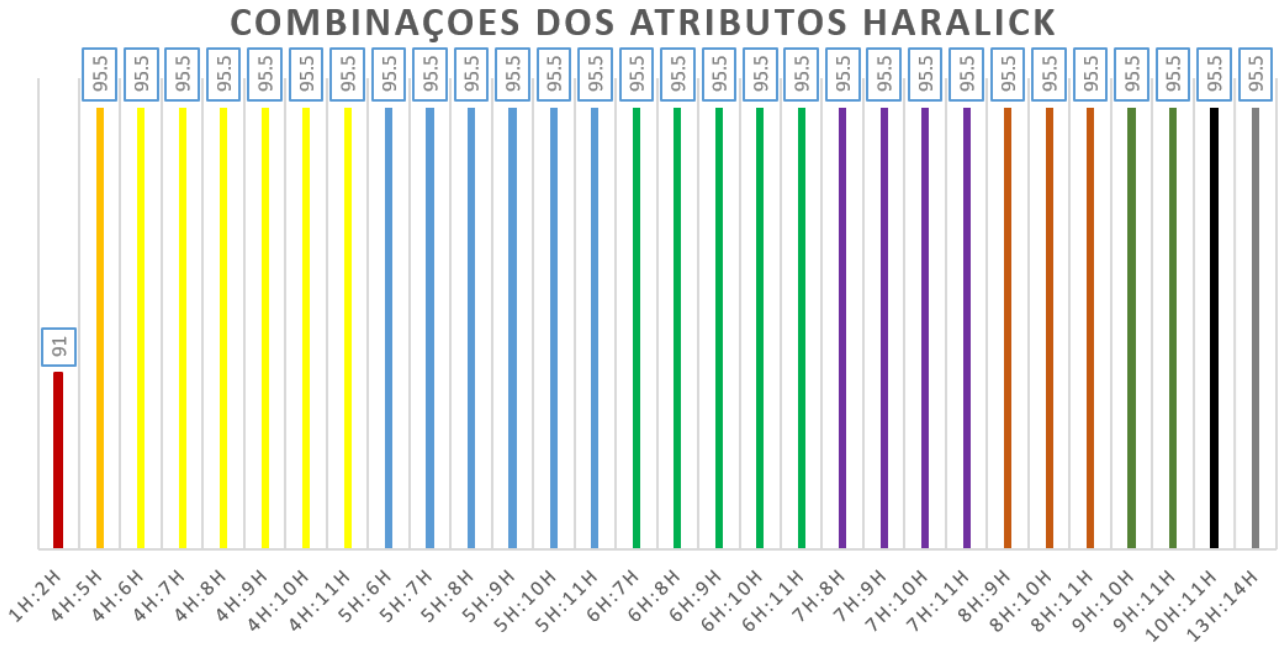

Fig. 5. Combinações dos atributos *Haralick* (Elaborado pelo autor).

C. *Quarta e quinta etapas*

Em relação a *wavelet*, foi utilizado a família *coiflet-5*, obtendo as médias e o desvios padrões dos coeficientes aproximada, horizontal, vertical e diagonal, sendo que os atributos desvio padrão vertical e diagonal apresentaram uma maior acurácia acima de 95% e o atributo desvio padrão aproximada apresentou uma menor acurácia com o valor de 54,5%. Conforme a Fig. 6.

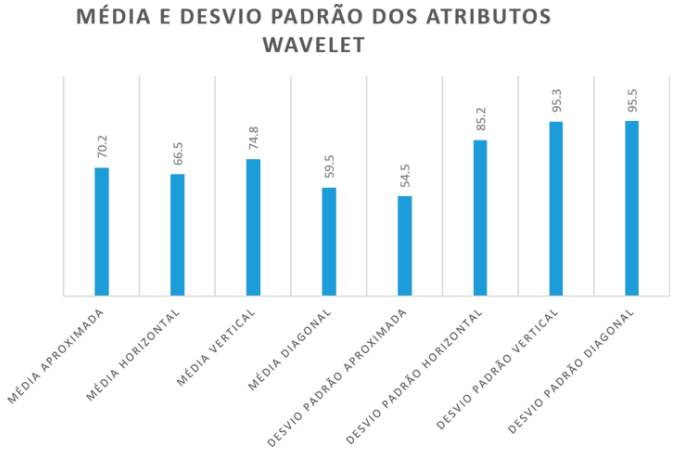

Fig. 6. Atributos da Transformada *Wavelet* (Elaborado pelo autor).

D. *Sexta etapas*

Foram feitas combinações dos atributos Wavelet, onde 1W:2W representa a combinação dos atributos média aproximada e horizontal, 3W:5W representa a combinação dos atributos média vertical, diagonal e desvio padrão aproximada, assim sucessivamente.

A combinação 7W:8W(desvio padrão vertical e diagonal) apresentaram maior acurácia com o valor de 95,3%, a combinação 4W:5W(média diagonal e desvio padrão aproximada) obteve a menor acurácia com o valor de 59,5%. O restante das combinações ficaram entre 66,5% e 87,6% conforme a Fig. 7.

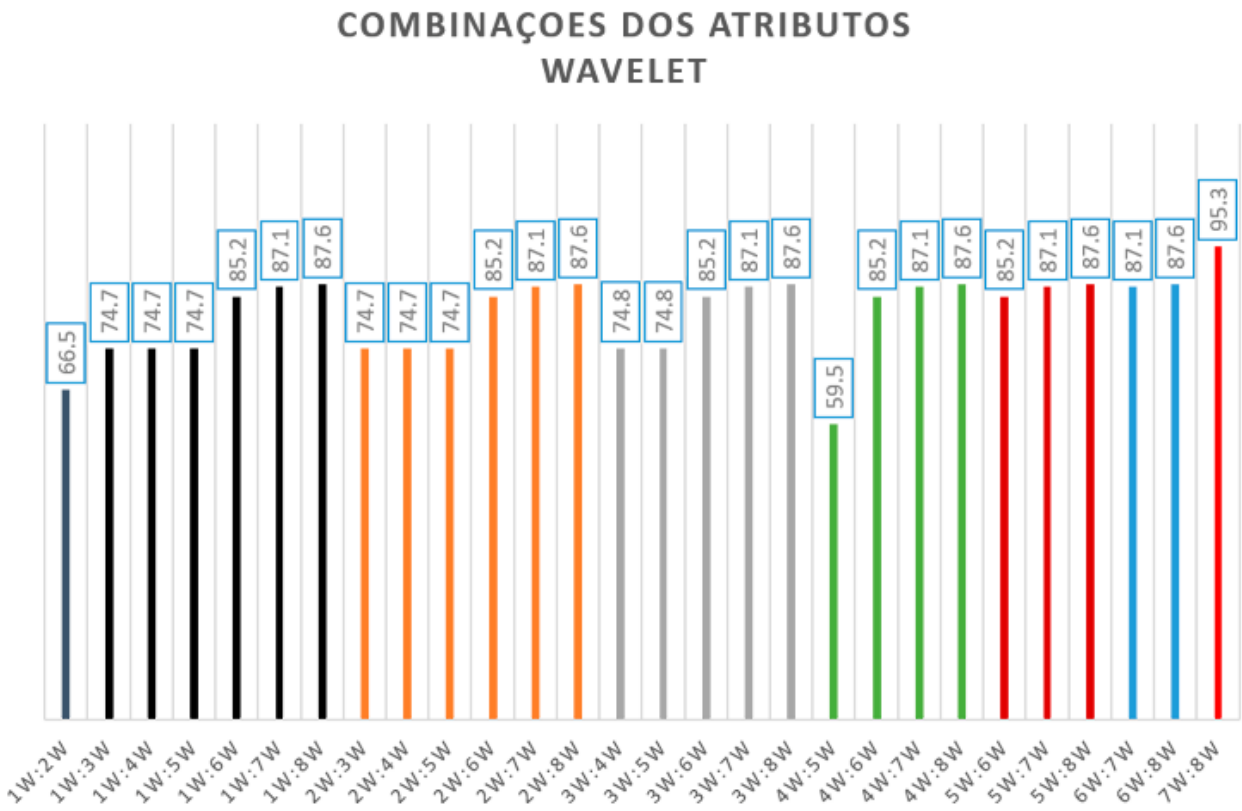

Fig. 7. Combinações dos atributos Wavelet (Elaborado pelo autor).

IV. Discussão

Sobre a base de dados, foi primordial criar critérios de inclusão e exclusão, para evitar erros nos cálculos, como por exemplo fazer combinações de atributos de adultos e crianças, que são muito diferentes por razões proporcionais. A primeira e a quarta etapa foram gerados os atributos individuais das técnicas *Haralick* e *Wavelet*, para serem utilizadas no agrupamento *k-means*. Entre 14 atributos da técnica de *Haralick*, os atributos contraste, variância, média da soma, soma da variância, entropia da soma, entropia, variância da diferença, entropia da diferença, medida de informação da correlação II e coeficientes de correlação obtiveram uma acurácia em torno de 95,5% atuando de forma individual no agrupamento *k-means*. As combinações dos atributos *Haralick*, não aumentaram a acurácia, mantendo os maiores valores do experimento individual de acordo com o atributo de cada grupo, por exemplo H5 (homogeneidade) na Fig. 4 tem acurácia de 54,5% e na Fig. 5 a combinação H5:H6 (homogeneidade, média da soma) tem acurácia de 95,5% por causa do atributo H6 (média da soma) que tem essa acurácia no experimento individual. Assim, os atributos individuais da técnica de *Haralick* podem ser considerados um ponto de partida para diferenciar os grupos COVID-19 e não COVID-19. Entre 8 atributos da técnica de *Wavelet*, as médias dos coeficientes de aproximação, horizontal, vertical e diagonal obtiveram uma acurácia entre 59,5% e 74,8%. O desvio padrão vertical e diagonal obteve acurácia 95,3% e 95,5%, respectivamente. Atuando de forma individual no agrupamento *k-means*. As combinações dos atributos *Wavelet* também não aumentaram a acurácia, chegando a uma acurácia de 95,3% conforme visto nas Fig. 6 e 7.

V. Conclusão

Com a técnica de extração de atributos *Haralick*, *os melhores atributos foram* contraste, variância, média da soma, soma da variância, entropia da soma, entropia, variância da diferença, entropia da diferença, medida de informação da correlação II e coeficientes de correlação alcançando uma acurácia de 95,5% inseridos de individual (um a um) no agrupamento *k-means*. Em relação a técnica de extração de atributos *wavelet*, os atributos desvio padrão vertical e diagonal obtiveram melhores resultados alcançado uma acuraria de 95,3% e 95,5%, respectivamente. Por fim, a combinação dos atributos das técnicas *Haralick* e *wavelet* não melhoraram a taxa de acerto dos agrupamentos individuais. O próximo passo seria e utilizar os melhores atributos de cada técnica de extração em outras técnicas de agrupamento como redes neurais, redes neurais profundas, etc.

Acknowledgment

Este estudo foi parcialmente financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001. E com recursos próprios da autora Ana Claudia Patrocinio.

Referências

[1] N. Zhu, D. Zhang, W. Wang, and others, "China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in China, 2019 [published January 24, 2020]," *N Engl J Med*.

[2] Wenjie Yang et al., "Clinical characteristics and imaging manifestations of the 2019 novel coronavirus

disease (COVID-19): A multi-center study in Wenzhou city, Zhejiang, China," *Journal of Infection*, 2020.

[3] ministério da Saúde. (2021, July) CORONAVÍRUS. [Online]. https://covid.saude.gov.br/

[4] Susan R. Weiss and Julian L. Leibowitz, "Coronavirus pathogenesis," in *Advances in virus research*.: Elsevier, 2011, vol. 81, pp. 85–164.

[5] Jie Cui, Fang Li, and Zheng-Li Shi, "Origin and evolution of pathogenic coronaviruses," *Nature Reviews Microbiology*, vol. 17, pp. 181–192, 2019.

[6] T. Ai et al., "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases [published online ahead of print February 26, 2020]," *Radiology*, vol. 10.

[7] Jianpeng Zhang, Yutong Xie, Yi Li, Chunhua Shen, and Yong Xia, "Covid-19 screening on chest x-ray images using learning based anomaly detection," *arXiv preprint arXiv:2003.12338*, 2020.

[8] Saban Ozturk, Umut Ozkaya, and Mucahid Barstugan, "Classification of coronavirus images using shrunken features," *medRxiv*, 2020.

[9] Ahmed T. Sahlol et al., "COVID-19 image classification using deep features and fractional-order marine predators algorithm," *Scientific reports*, vol. 10, pp. 1–15, 2020.

[10] Wei-cai Dai et al., "CT imaging and differential diagnosis of COVID-19," *Canadian Association of Radiologists Journal*, vol. 71, pp. 195–200, 2020.

[11] Michael Chung et al., "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology*, vol. 295, pp. 202–207, 2020.

[12] Maria de la Iglesia Vayá et al., "BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients," *arXiv preprint arXiv:2006.01174*, 2020.

[13] Robert M. Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, pp. 610–621, 1973.

[14] Maarten Jansen, *Noise reduction by wavelet thresholding*.: Springer Science & Business Media, 2012, vol. 161.

[15] C. Rafael, "Gonzalez, and Richard E. Woods," *Digital image processing*, 2006.

[16] Travis Williams and Robert Li, "Advanced image classification using wavelets and convolutional neural networks," in *2016 15th IEEE international conference on machine learning and applications (ICMLA)*, 2016, pp. 233–239.

[17] James MacQueen and others, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1967, pp. 281–297.

[18] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, and others, "Constrained k-means clustering with background knowledge," in *Icml*, vol. 1, 2001, pp. 577–584.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, pp. 436–444, 2015.

[20] David R. Martin et al., "A learning convolutional neural network can recognize common patterns of injury in gastric pathology," *Archives of pathology & laboratory medicine*, vol. 144, pp. 370–378, 2020.