Seleção de atributos e classificação de imagens radiográficas em paciente com COVID-19

Mariane Modesto Oliveira
Faculdade de Engenharia Elétrica
Universidade Federal de Uberlândia
Uberlândia, Brazil
ORCID: 0000-0003-4368-7087

Guilherme Brilhante Guimarães
Faculdade de Engenharia Elétrica
Universidade Federal de Uberlândia
Uberlândia, Brazil
ORCID: 0000-0001-8558-8445

Ana Claudia Patrocinio
Faculdade de Engenharia Elétrica
Universidade Federal de Uberlândia
Uberlândia, Brazil
ORCID: 0000-0001-9376-7689

Abstract - In December 2019 a disease appeared caused by a new type of coronavirus, Covid-19. The disease has become a pandemic, to date. One of the ways to avoid contamination is through social isolation and especially isolation and rapid diagnosis of the patient. For the diagnosis, it is necessary to perform the RT-PCR exam, through a blood sample, but as one of the characteristics of the disease is the damage to the lungs, it is possible to detect it through CT scans and X-rays. the similarity of the images of the results of patients diagnosed with Covid-19 and the results of patients who have other diseases. Thus, we used the K-means technique to differentiate radiographic images of patients with Covid-19 and those without the disease. Analyzing the Haralick texture descriptors without individual parameters, we observed that the highest hit rate occurred for the Entropy Difference descriptor with 95.83% hits, followed by the Inverse Moment Difference descriptor with 94.44% hits.

Keywords - covid-19, X-rays, clustering, K-mean

I. INTRODUÇÃO

A. Coronavirus

No ano de 2019 surgiu, em dezembro, uma série de casos de pneumonia causada por um novo tipo de coronavírus (SARS-CoV-2), iniciando na cidade de Wuhan, província de Hubei, China [1]. Estudos clínicos mostraram que a maioria dos pacientes infectados com o vírus apresentou infecção pulmonar [2]. Uma das formas iniciais de detectar a doença seria através do exame indicador de reação de transcrição da polimerase reversa (RT-PCR), entretanto um dos problemas apresentados por esse método é a sua baixa sensibilidade e com isso pacientes podem obter um resultado falso-negativo o que ocasiona o não recebimento do tratamento adequado além de retardar o isolamento do doente, que é um ponto crucial para evitar a contaminação de pessoas saudáveis [3]. Em 2019, a preocupação era que a doença não tivesse apenas um potencial endêmico, mas sim pandêmico e esse cenário se tornou realidade até o presente momento [4]. Uma das medidas para limitar a contaminação e o espalhamento do vírus é através do isolamento social, desta forma se torna imprescindível haver uma rápida detecção da doença.

Uma das formas de auxiliar no diagnóstico é através das imagens de tomografia computadorizada (TC) de tórax com o auxílio de ferramentas de inteligência artificial (AI), pois através delas é possível detectar anormalidades nos pulmões e acompanhar a evolução da doença [3,5]. Entretanto, se comparado aos exames de raios X, a tomografia computadorizada apresenta um custo maior e uma menor disponibilidade geográfica [2]. O problema é a similaridade existente entre radiografias de pacientes que possuem Covid-19 e pneumonia. Assim sendo, é necessário recorrer ao auxílio de ferramentas de AI, como por exemplo o deep learning, para que seja possível comparar as imagens e diferenciá-las com uma melhor eficiência [2].

B. Algoritmo K-means

Diariamente, é gerada uma grande quantidade de dados diversificados, contendo informações importantes e de interesses para diferentes ramos do conhecimento.

Como os dados, geralmente, são produzidos em grande quantidade, é necessário o desenvolvimento de ferramentas e métodos, cada vez mais sofisticados, que possam facilitar e permitir a extração, processamento e tratamentos dos dados; um dos métodos conhecidos, é o algoritmo de clusterização k-means [6].

O método de k-means é um poderoso algoritmo de aprendizagem não supervisionada e clusterização, capaz de tratar uma grande quantidade de dados, particionando e classificando em diferentes k grupos [6].

A capacidade de clusterização do k-means permite o cálculo de uma grande quantidade de interações, a cada interação no sistema, os k grupos são recalculados e reorganizados, quando o sistema atinge o limite de modificações, o algoritmo é interrompido. Os dados separados em k grupos, após o término da compilação do algoritmo, compartilham informações de interesses e/ou semelhanças [7].

As interações no algoritmo de k-means são calculadas medindo a distância entre os objetos (dados) de um determinado k grupo, existem diversas opções para calcular a distância, chebyshev, procrustes, euclidiana, mahalanobis, etc; a cada interação dos k grupos é recalculado a distância, classificando os objetos em diferentes k grupos [6][7], conforme Fig. 1.

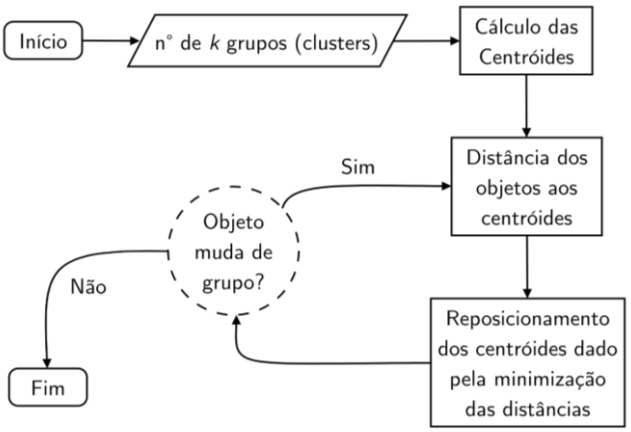

Fig. 1. Fluxograma de execução do algoritmo de k-means [2].

II. METODOLOGIA

A. *Grupo de Estudo*

Neste estudo utilizamos um banco de imagens com 72 radiografias de radiografias de tórax para a extração de atributos, composto por 36 imagens com laudos positivos (COVID) e 36 imagens com laudos negativos (N-COVID) para a doença Covid-19 (Sars-CoV-2) do banco de dados bimcv [8].

B. *Softwares e Técnicas*

Para analisarmos a acurácia do algoritmo k-means na classificação das imagens médicas do estudo, é necessário extrairmos alguns atributos, como as Medidas Globais de Intensidade, as características de textura de Haralick e os valores para a técnica de Wavelet.

Utilizando os softwares Image J e Octave, extraímos as Medidas Globais de Intensidade, sendo elas: média, desvio padrão, moda, mediana, mínimo, máximo, diferença da média mínima e máxima, e a porcentagem do maior valor; posteriormente, comparamos os valores dos atributos de intensidade.

As características de textura de Haralick foram extraídas utilizando o software Octave, para essa técnica de textura foram utilizados quatro ângulos relacionados com a matriz de co-ocorrência, sendo realizado o cálculo para os 14 descritores de textura de Haralick: uniformidade e energia (segundo momento angular), contraste, correlação, variância, momento da diferença inversa, média da soma, variância da soma, entropia da soma, entropia, variância da diferença, entropia da diferença, medida de informação de correlação I e II, e máximo coeficiente de correlação.

A partir das imagens originais aplicamos a técnica de Wavelet, obtendo as medidas da média e desvio nas seguintes configurações: originais, aproximadas, horizontais, verticais e diagonais. A partir das características extraídas de Haralick e Wavelet, aplicamos o algoritmo de k-means para determinar o resultado individual e das combinações nas técnicas de estudo.

III. RESULTADOS E DISCUSSÃO

A. *Medidas Globais de Intensidade*

Os valores extraídos para as Medidas Globais de Intensidade (Atributos de Intensidade) são apresentados nas Fig. 2 e Fig. 3:

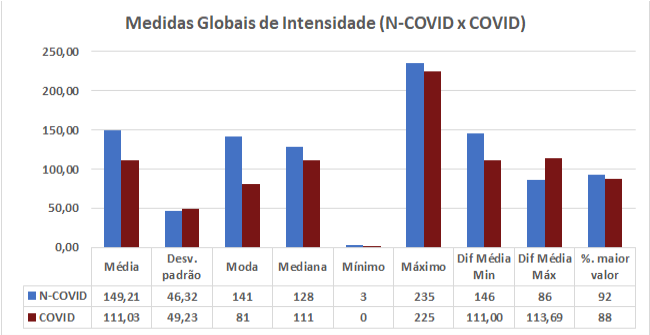

Fig. 2. Comparativo das Medidas Globais de Intensidade para os grupos de estudos N-COVID x COVID.

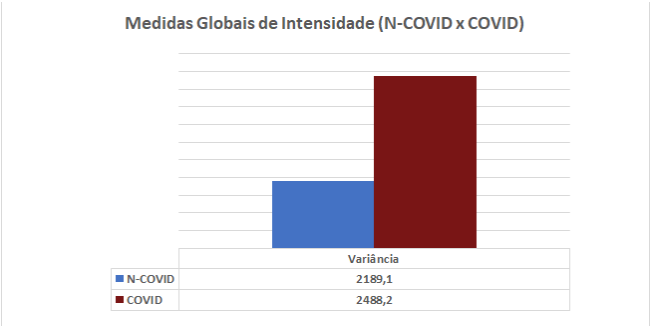

Fig. 3. Comparativo da Medidas Globais de Intensidade (variância) para os grupos de estudos N-COVID x COVID.

Analisando os parâmetros que foram extraídos, de maneira quantitativa, é possível observar que os únicos que apresentaram valores superiores para imagens com laudo positivo para a doença Covid-19, foram o desvio padrão, a variância e a diferença média máxima. Os demais valores foram superiores para imagens cujo resultado foi negativo para a Covid-19.

B. *Descritores de Textura de Haralick*

Para os atributos de Haralick, foram construídos 14 gráficos (Figs. 4-17), descrevendo cada um dos descritores de textura de Haralick. Os 4 primeiros elementos dos gráficos representam a média de cada um dos 4 diferentes ângulos da matriz de co-ocorrência, o quinto elemento corresponde à média dos valores dos ângulos.

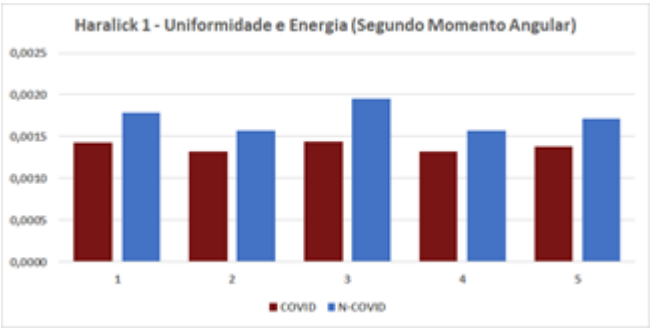

Fig. 4. Comparativo do descritor Uniformidade e Energia de Haralick para os grupos de estudos N-COVID x COVID.

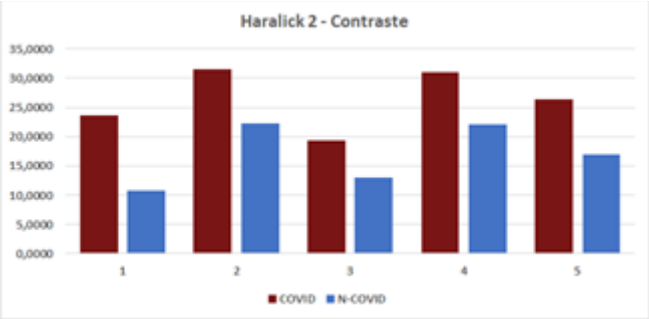

Fig. 5. Comparativo do descritor Contraste de Haralick para os grupos de estudos N-COVID x COVID.

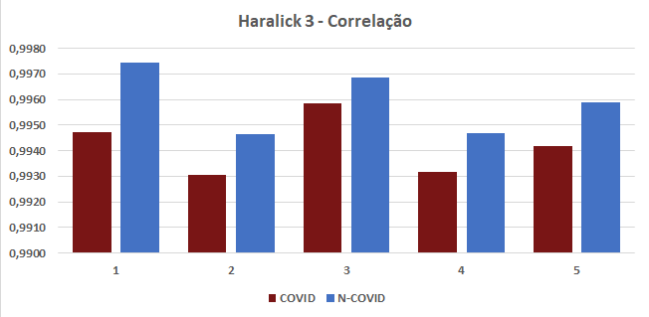

Fig. 6. Comparativo do descritor Correlação de Haralick para os grupos de estudos N-COVID x COVID.

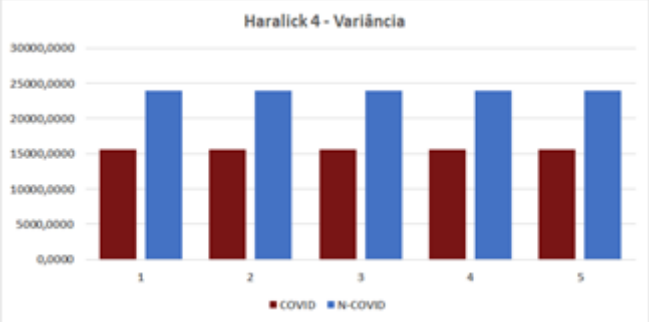

Fig. 7. Comparativo do descritor Variância de Haralick para os grupos de estudos N-COVID x COVID.

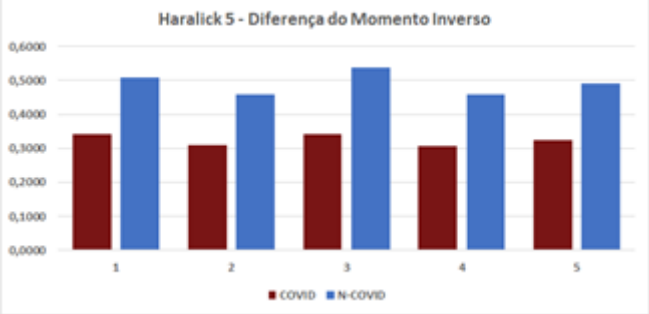

Fig. 8. Comparativo do descritor Diferença do Momento Inverso de Haralick para os grupos de estudos N-COVID x COVID.

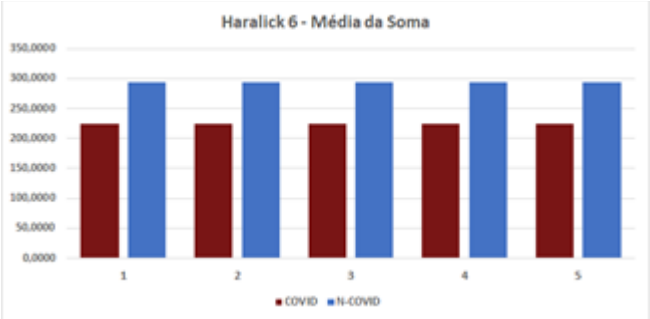

Fig. 9. Comparativo do descritor Média da Soma de Haralick para os grupos de estudos N-COVID x COVID.

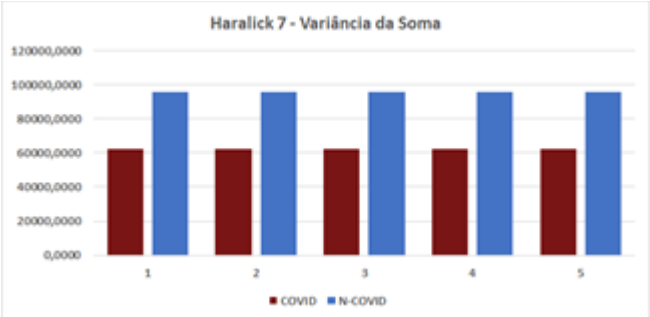

Fig. 10. Comparativo do descritor Variância da Soma de Haralick para os grupos de estudos N-COVID x COVID.

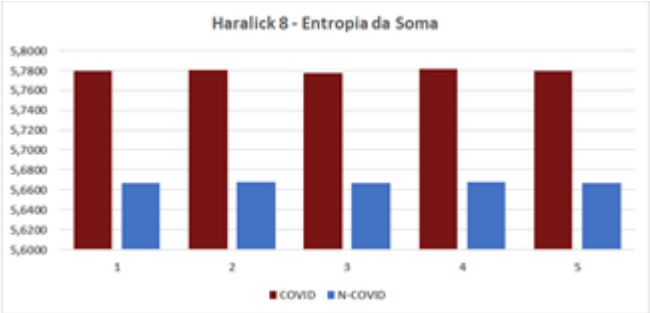

Fig. 11. Comparativo do descritor Entropia da Soma de Haralick para os grupos de estudos N-COVID x COVID.

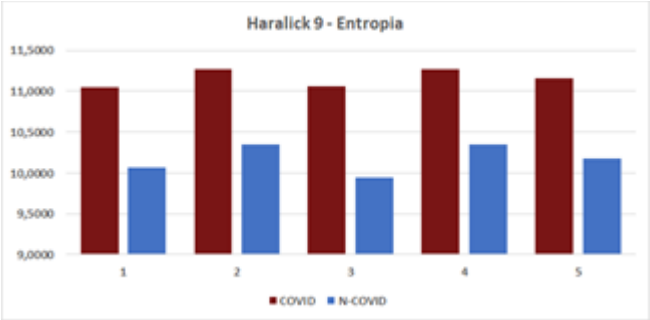

Fig. 12. Comparativo do descritor Entropia de Haralick para os grupos de estudos N-COVID x COVID.

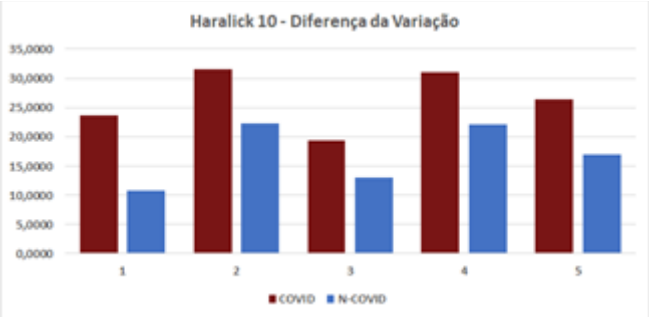

Fig. 13. Comparativo do descritor Diferença de Variação de Haralick para os grupos de estudos N-COVID x COVID.

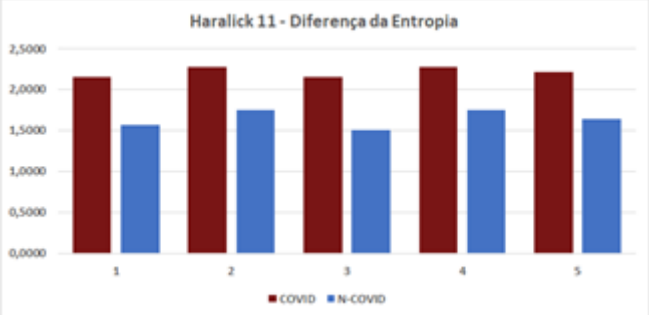

Fig. 14. Comparativo do descritor Diferença da Entropia de Haralick para os grupos de estudos N-COVID x COVID.

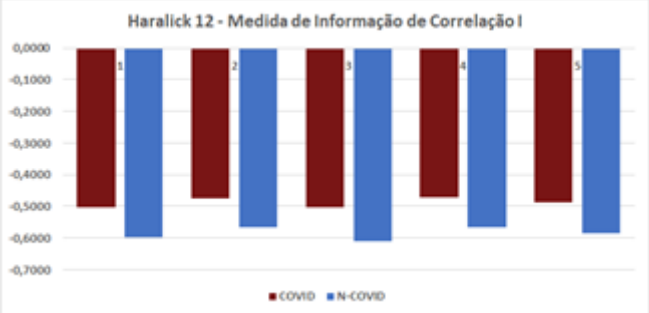

Fig. 15. Comparativo do descritor da Medida de Informação de Correlação I de Haralick para os grupos de estudos N-COVID x COVID.

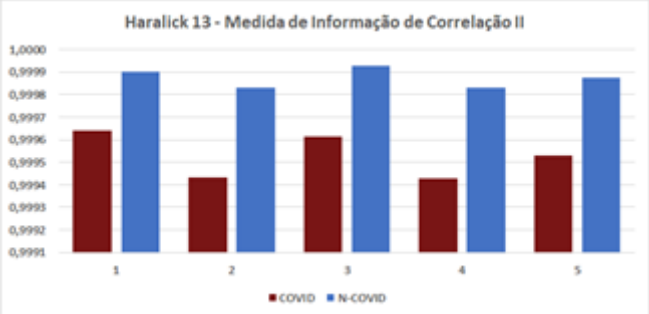

Fig. 16. Comparativo do descritor da Medida de Informação de Correlação II de Haralick para os grupos de estudos N-COVID x COVID.

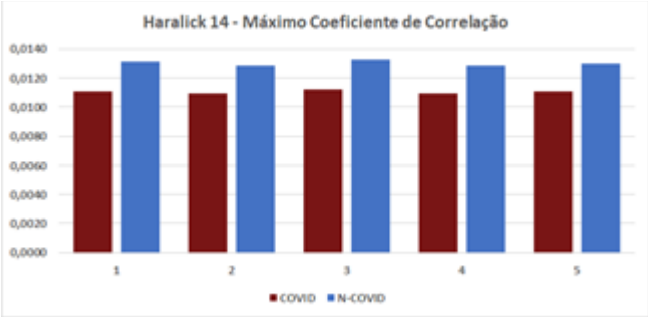

Fig. 17. Comparativo do descritor do Máximo Coeficiente de Correlação de Haralick para os grupos de estudos N-COVID x COVID.

Dentro dos 14 descritores de textura de Haralick, 6 deles apresentaram valores maiores para imagens com laudo positivo para Covid-19, os gráficos do Contraste, Entropia da Soma, Entropia, Diferença da Variância, Diferença da Entropia e Medida de Informação de Correlação I.

C. Técnica de Wavelet

Após aplicação da técnica de Wavelet, obtemos os gráficos (Fig. 18 e Fig. 19) para os atributos de Wavelet:

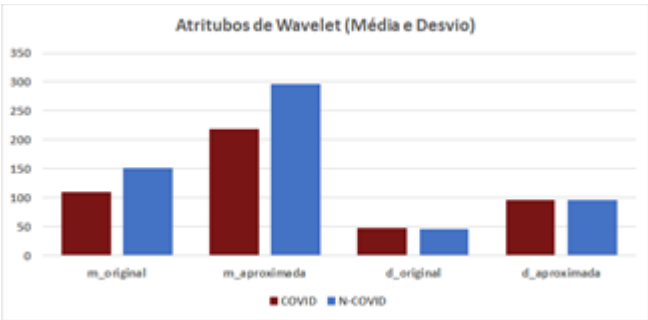

Fig. 18. Comparativo dos atributos de Wavelet para os grupos de estudos N-COVID x COVID.

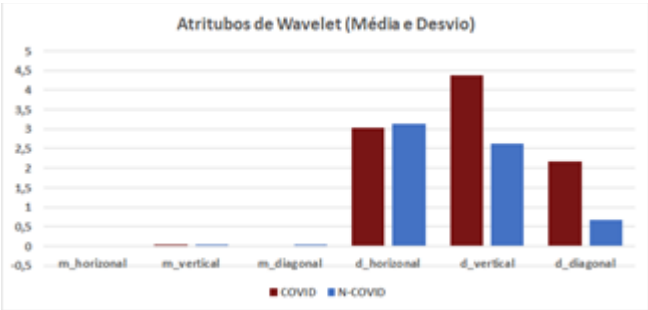

Fig. 19. Comparativo dos demais atributos de Wavelet para os grupos de estudos N-COVID x COVID.

D. Aplicação do Algoritmo de k-means

Após aplicação do algoritmo de k-means obtemos os dados da acurácia para o agrupamento das classes, utilizando os parâmetros de forma individual e posteriormente combinados dois a dois.

Analisando os descritores de textura de Haralick no parâmetro individual, observamos que a maior taxa de acertos ocorreu para o descritor Diferença de Entropia com 95,83% de acertos, seguido do descritor Diferença do Momento Inverso com 94,44% de acertos. Aplicando a combinação dois a dois, a taxa de 95,83% foi obtida com as seguintes

combinações: 10 e 11; 10 e 14; 11 e 12; 11 e 13; 11 e 14; 12 e 14; seguido da taxa de 94,44% para as combinações 1 e 3; 10 e 12; 10 e 13.

Utilizando o algoritmo de k-means para os atributos de Wavelet, obtemos do parâmetro analisado na forma individual, 98,61% para o desvio diagonal, seguido de 90,28% do desvio vertical. Aplicando a combinação dois a dois, os atributos desvio vertical e desvio diagonal apresentaram a maior porcentagem de acertos, 93,06%, seguido da combinação dos atributos desvio horizontal e desvio diagonal com 90,28% de acertos. No entanto, não foi capaz de superar o resultado obtido pelo parâmetro de forma individual.

Comparando a extração dos descritores de textura de Haralick com os resultados obtidos pelo algoritmo k-means, é possível observar que o descritor 11 (Diferença de Entropia) obteve o maior resultado para imagens com Covid-19, também apresentou o melhor resultado para o agrupamento dos dois tipos de imagens (N-COVID e COVID).

Para os atributos de Wavelet, o desvio diagonal apresentou a melhor porcentagem de acertos, obtendo resultados próximos de 100% para a forma individual e valores altos para as combinações dois a dois (Haralick e Wavelet).

IV. CONCLUSÃO

O algoritmo de k-means é uma técnica que pode ser utilizada como técnica de seleção de atributos para classificação de problemas complexos, como diagnóstico por imagem. Para os atributos de Wavelet as porcentagens de acertos foram bem altas, com 98,61% na forma individual e 93,06% na combinação dois a dois, para o Haralick obtemos 95,83% em ambas.

Para trabalhos futuros, é recomendado realizar a combinação das duas técnicas, Haralick e Wavelet, a fim de verificar se é possível aumentar a probabilidade de acertos do algoritmo de k-means.

Até o presente momento, não foi possível comparar este trabalho com outros trabalhos presentes na literatura.

REFERÊNCIAS

[1] XU, Xi et al. Imaging and clinical features of patients with 2019 novel coronavirus SARS-CoV-2. European journal of nuclear medicine and molecular imaging, v. 47, n. 5, p. 1275-1280, 2020.

[2] ZHANG, Jianpeng et al. Covid-19 screening on chest x-ray images using deep learning based anomaly detection. arXiv preprint arXiv:2003.12338, v. 27, 2020.

[3] NARIN, Ali; KAYA, Ceren; PAMUK, Ziynet. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. Pattern Analysis and Applications, p. 1-14, 2021.

[4] YANG, Wenjie 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, v. 80, n. 4, p. 388-393, 2020.

[5] DAI, Wei-cai et al. CT imaging and differential diagnosis of COVID-19. Canadian Association of Radiologists Journal, v. 71, n. 2, p. 195-200, 2020.

[6] SOUSA, Maria Cristina Cordeiro Sousa. Uma análise do algoritmo K-means como introdução ao aprendizado de máquinas. 2020.

[7] OLIVEIRA, Arthur Scardini et al. Comparação entre os algoritmos K-Means e Dynamic Cluster em imagens digitais. Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574), n. 2, 2018.

[8] https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/#1590858128006-9e640421-6711.