COPPE
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia UFRJ

COC 891 – DEEP LEARNING

CLASSIFICAÇÃO DE IMAGENS DE RAIO X DE TÓRAX PARA IDENTIFICAR PACIENTES COM PNEUMONIA CAUSADA OU NÃO PELA COVID-19

Éric Kauati Saito
Paula Brandão Furlan

Rio de Janeiro
Janeiro de 2021

Sumário

1 Introdução

Em tempos em que a pandemia causada pelo COVID-19 continua sendo um tópico bastante preocupante, diversos estudos vêm sendo realizados acerca dessa doença, seus males e suas consequências. Seus impactos totais estão sendo e por muito tempo ainda serão alvos de estudos. Até o presente momento cerca de 95.612.831 casos foram confirmados de acordo com a Organização Mundial de Saúde (janeiro de 2021).

Alguns especialistas acreditam que a utilização de tecnologias com base em Inteligência Artificial possa ajudar no sistema de saúde, sendo que alguns hospitais já estão avaliando o seu uso. O artigo [1] que foi utilizado como orientação para este trabalho aponta três vantagens da utilização de radiografia de tórax:

- Rápida triagem
- Disponibilidade e Acessibilidade
- Portabilidade

Essas vantagens, para alguns especialistas, são discutíveis, principalmente a utilização de radiografia de tórax como diagnóstico. Sendo que o Ministério da Saúde (janeiro de 2021) aponta como um dos exames de diagnóstico laboratorial, o RT-PCR (swab), até o oitavo dia de início de sintomas.

Este trabalho não entra no mérito dessas discussões, e tenta abordar o problema de classificação de imagens de radiografia utilizando-se de banco de dados públicos e utilizando para isso as Redes Neurais Convolucionais (CNN – Convolutional Neural Network). O objetivo desse trabalho é classificar imagens de radiografia de tórax entre: Sem Pneumonia; Pneumonia não COVID-19; COVID-19.

1.1 Pesquisa Bibliográfica

No trabalho de [1] foi proposta uma arquitetura denominada CovidNet, composta de múltiplas camadas de convolução aliadas ao uso de *Depth Wise Convolution* para a classificação das imagens de radiografia de tórax em três classes. O código fonte utilizado para obter os dados necessários, dividi-los nos dois grupos (90 % para treino e 10 % para

teste) e realizar o treinamento e teste da rede proposta foram disponibilizados pela equipe do projeto em [2].

Com essa arquitetura, no grupo de teste, obtiveram uma acurácia de 93,3%, sendo que a arquitetura possuía 11,75 milhões de parâmetros. E obtiveram os seguintes resultados de sensibilidade: 95,0% para a classe sem pneumonia; 94,0% para a classe Pneumonia não COVID-19 e 93,3% para a classe com COVID-19. E os seguintes resultados de valor preditivo positivo: 90,5% para a classe sem pneumonia; 91,3% para a classe Pneumonia não COVID-19 e 98,9% para a classe com COVID-19.

2 Tecnologia

A implementação deste trabalho foi feita na linguagem Python utilizando como IDE o Jupyter Notebook, com a versão 3.8 do Python e a biblioteca do Tensorflow (versão 2.3.1). Os códigos fornecidos pela equipe do artigo [1,2] foram feitos em uma versão antiga da biblioteca, o que impossibilitou o reaproveitamento de boa parte do código, gerando considerável retrabalho para reescrever os códigos no novo formato. Além disso, o artigo não utilizou nenhum tipo de validação nos dados de treino, então uma validação cruzada teve que ser implementada.

Inicialmente tínhamos disponíveis dois computadores para o processamento das redes, porém devido a problemas técnicos e disponibilidade de uso, na metade do percurso do trabalho, apenas uma das máquinas foi utilizada. Essa máquina possui as seguintes configurações:

- Placa de Vídeo: NVIDIA GeForce MX150 4GB
- Processador: i7 - 8º geração
- Memória RAM: 16GB

2.1 Banco de Dados

O banco de dados utilizado para este trabalho foi construído com base no artigo [1,2]. Para isso foi utilizado cinco diferentes bancos de dados públicos de imagens de radiografia de tórax:

- COVID-19 Image Data Collection [3];
- COVID-19 Chest X-Ray Dataset Initiative [4];
- ActualMed COVID-19 Chest X-ray Dataset Initiative [5];
- RSNA Pneumonia Detection Challenge dataset [6];
- COVID-19 radiography database [7].

Esses bancos de dados continuam sendo atualizados e até o momento em que esse trabalho foi feito havia no total:

- 8.066 imagens de pacientes sem pneumonia;
- 5.575 imagens de pacientes com pneumonia não COVID-19;
- 617 imagens de pacientes com COVID-19;

(a) *(b)* *(c)*

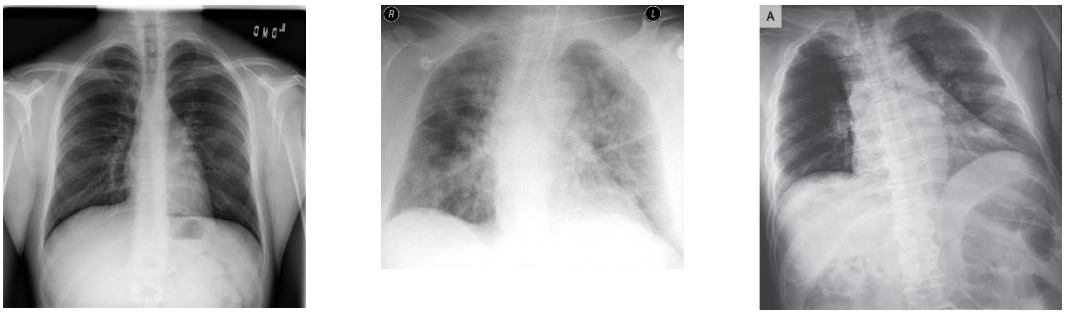

Figura 1 – Exemplos de imagens do banco de dados. (a) Sem pneumonia, (b) pneumonia não COVID-19, (c) COVID-19

Na Figura 1 são apresentadas três imagens de exemplo do banco de dados, das três classes: (a) sem pneumonia, (b) pneumonia não COVID-19 e (c) COVID-19.

2.2 Arquiteturas

Neste trabalho foram avaliadas quatro diferentes arquiteturas de Redes Neurais Convolucionais (CNNs) a fim de classificar as imagens de raio X de tórax. A função de ativação utilizada em todas as arquiteturas foi a ReLu e na camada de saída foi utilizada a Softmax. Em todas as arquiteturas foi utilizada uma camada densa para a predição, com 10 neurônios.

2.2.1 CNN Básica

Partindo de uma arquitetura mais simples, foi implementada a arquitetura apresentada na Figura 2. Utilizou-se duas camadas de convolução 2D com *max pooling*, e foi utilizado um Kernel fixo de 3x3.

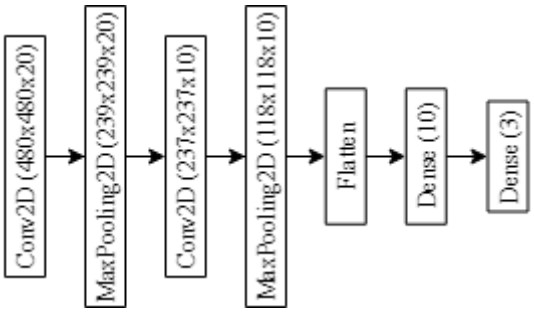

Figura 2 – Arquitetura da CNN básica

2.2.2 CNN Básica com Dropout

Para a segunda arquitetura implementada, apresentada na Figura 3, foi utilizado a idéia de *Dropout* para diminuir o *overfitting*. Para isso foi implementando uma camada de *Dropout* depois da camada Densa e antes da camada de saída. O percentual de *dropout* foi de 40% e a convolução também teve um Kernel fixo de 3x3.

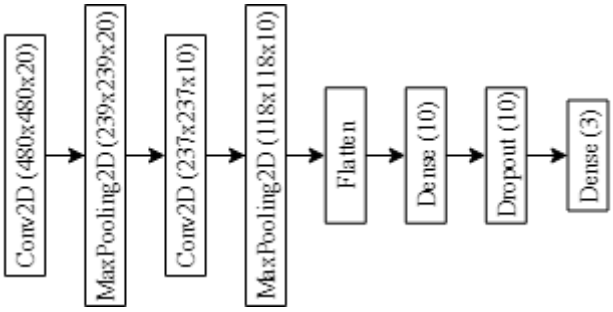

Figura 3 – Arquitetura CNN básica com Dropout

2.2.3 CNN Multicamadas com DepthWiseConvolution (DWC)

Na implementação da terceira arquitetura, Figura 4, seguindo a ideia apresentada no artigo base, foi feita uma rede mais profunda para aumentar o nível de granularidade. Para isso foi utilizado camadas de *Depth Wise Convolution* (DWC), que tem como finalidade aprender as características espaciais mantendo a capacidade de representatividade e diminuindo a complexidade computacional. Essas camadas foram precedidas e sucedidas por camadas de convolução 2D de kernel 1x1. Nas primeiras duas camadas foi utilizado Kernel de 7x7.

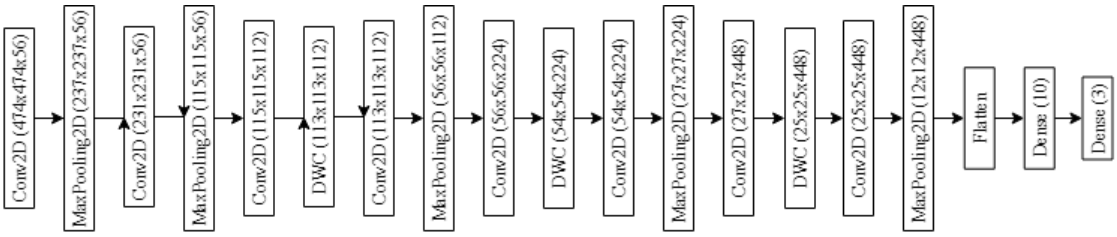

Figura 4 – Arquitetura CNN Multicamadas com DepthWiseConvolution

2.2.4 CNN Multicamadas com Batch Normalization

Continuando com a ideia de implementar uma rede mais profunda, foi feita uma quarta arquitetura (Figura 5). Aplicou-se uma camada de normalização de *batch* após a primeira camada de convolução e antes da camada de *max pooling*, com o intuito de diminuir a variabilidade entre os dados. Foi utilizado Kernel que começa em 7x7 e termina em 1x1.

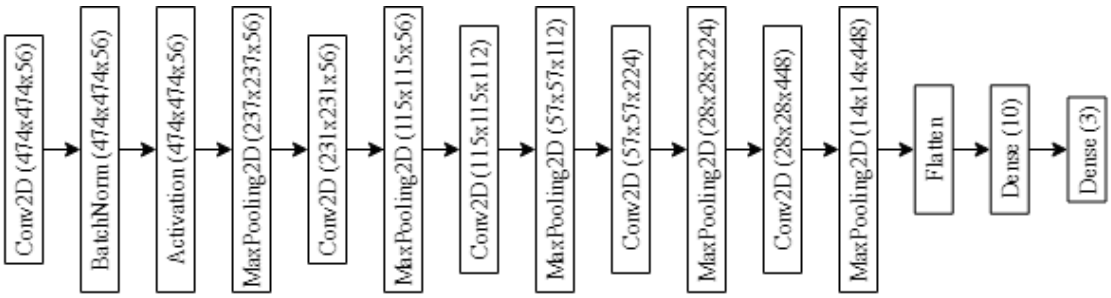

Figura 5 – Arquitetura CNN Multicamadas com Batch Normalization

3 Metodologia

Antes do processamento das imagens, foi feito o corte e redimensionamento das mesmas para 480 por 480 pixels. Nesse trabalho não foi feito um pré-treinamento das redes em um banco de dados de imagens. Na etapa de treinamento das redes, os dados foram separados em um grupo de treinamento e um de testes. Sendo que, devido ao nível de desbalanceamento dos dados, principalmente na classe covid, foi necessário fazer o balanceamento de *batches*.

Durante o treinamento das duas primeiras redes, utilizou-se a ideia de balanceamento de *batches* proposta em [1,2], porém ocorreram muitos problemas de convergência. Após uma análise minuciosa do problema, viu-se que a estratégia utilizada no artigo acabava tornando os *batches* desbalanceados em relação à classe de pneumonia não Covid. Devido a isso, foi formulada uma nova estratégia de balanceamento dos *batches*, a qual sanou o problema e foi utilizada no treinamento das duas últimas arquiteturas.

No treinamento, foi utilizado o *Data Augmentation* dos seguintes tipos:

- Translação;
- Rotação;
- Giro na horizontal;
- Ampliação;
- Mudança de intensidade.

No treinamento dos modelos, foi utilizado a validação cruzada de 5 *folds*, 40 épocas e *patience* de 4, utilizando a acurácia como métrica. Para o cálculo do *loss* foi utilizado a entropia cruzada categórica esparsa no otimizador Adam com os seguintes hiper parâmetros:

- Taxa de aprendizado = 0,001;
- $\beta 1 = 0,9$;
- $\beta 2 = 0,999$;

- $\varepsilon = 1e^{-7}$.

Para a avaliação dos resultados foram calculadas as seguintes métricas: *Loss*, Acurácia, Sensibilidade, Valor Preditivo Positivo (PPV) e *f1-score*. Dessa forma foi possível ver a taxa de acertos total de cada rede, além de sua capacidade de predição e sensibilidade em cada uma das classes.

Dado os resultados que serão apresentados a seguir foi vista a necessidade de aplicação de um teste estatístico não paramétrico pareado (Wilcoxon) para comparar os resultados de acurácia de duas arquiteturas, tendo como hipótese nula, a não diferença entre elas, e alpha de 0,05.

4 Resultados

A seguir são apresentados os resultados obtidos das arquiteturas testadas. Na Tabela 1 temos os resultados de *Loss*, acurácia e desvio padrão obtidos da validação cruzada na etapa de treinamento. Pode-se perceber que, diferente das demais, a arquitetura mais básica é a que possuí maior variabilidade nos valores de acurácia.

Tabela 1 – Resultados do treinamento

Arquitetura/Métrica	*Loss*	Acurácia (d.p.) (%)
Básica	0,6292	74,50 (6,30)
Básica Dropout	0,6185	75,90 (2,42)
Multi DWC	0,4657	81,92 (2,43)
Multi Batch Norm	0,3847	85,27 (2,20)

O teste estatístico de Wilcoxon pareado mostra que teve diferença estatística (p = 0,0431) na comparação da CNN Multicamadas DWC com a CNN Multicamadas *Batch Normalization*. Dessa forma, pode-se afirmar que a última arquitetura obteve o melhor resultado de acurácia.

Na Tabela 2 temos os resultados de sensibilidade e desvio padrão obtidos da validação cruzada na etapa de treinamento. Pode-se ver nesses resultados que a classe que possui o menor valor de sensibilidade é a de pneumonia não COVID19.

Tabela 2 – Resultados de sensibilidade do treinamento

	Sensibilidade (d.p.) (%)		
Arquitetura/Classe	**Normal**	**Não COVID19**	**COVID19**
Básica	76,19 (6,80)	71,21 (3,36)	76,34 (10,89)
Básica Dropout	81,73 (4,40)	66,54 (1,43)	76,03 (6,44)
Multi DWConv	82,34 (3,10)	76,73 (4,12)	86,68 (3,68)
Multi Batch Norm	86,89 (1,48)	80,88 (3,76)	87,65 (2,98)

Na Tabela 3 temos os resultados de valor preditivo positivo (PPV) e desvio padrão obtidos da validação cruzada na etapa de treinamento. Novamente pode-se perceber que a arquitetura básica é a que possui maior variabilidade dos resultados de PPV, em todas as classes. É interessante perceber que o PPV da classe pneumonia não COVID19 é a de menor valor comparado às outras classes para todas as arquiteturas testadas.

Tabela 3 – Resultados de PPV do treinamento

PPV (d.p.) (%)			
Arquitetura/Classe	**Normal**	**Não COVID19**	**COVID19**
Básica	78,01 (7,15)	67,52 (8,20)	78,86 (6,50)
Básica Dropout	74,93 (4,04)	70,78 (3,78)	82,61 (2,94)
Multi DWConv	85,54 (2,33)	76,67 (4,59)	83,06 (3,27)
Multi Batch Norm	86,65 (2,52)	78,78 (2,41)	90,54 (2,58)

Na Tabela 4 temos os resultados de f1-score e desvio padrão obtidos da validação cruzada na etapa de treinamento. Seus valores também são mais baixos na classe pneumonia não COVID19.

Tabela 4 – Resultados de f1-score do treinamento

f1-score (d.p.) (%)			
Arquitetura/Classe	**Normal**	**Não COVID19**	**COVID19**
Básica	76,98 (6,11)	69,13 (5,03)	77,43 (8,25)
Básica Dropout	78,03 (1,54)	68,53 (1,37)	78,98 (2,71)
Multi DWConv	83,88 (2,16)	76,56 (2,51)	84,79 (2,77)
Multi Batch Norm	86,76 (1,65)	79,78 (2,46)	89,04 (2,16)

A seguir serão apresentados os resultados obtidos, detalhados para cada arquitetura testada.

4.1 CNN Básica

O tempo computacional na validação cruzada foi de aproximadamente 8 horas em cada *fold*. E o tempo de treino com todos os dados do grupo de treinamento foi de aproximadamente 10 horas. Totalizando em 50 horas totais de processamento computacional. Na Figura 6, pode ser vista a matriz de confusão obtida no grupo de teste, nessa matriz as linhas são os valores verdadeiros e nas colunas os valores preditos pelo modelo.

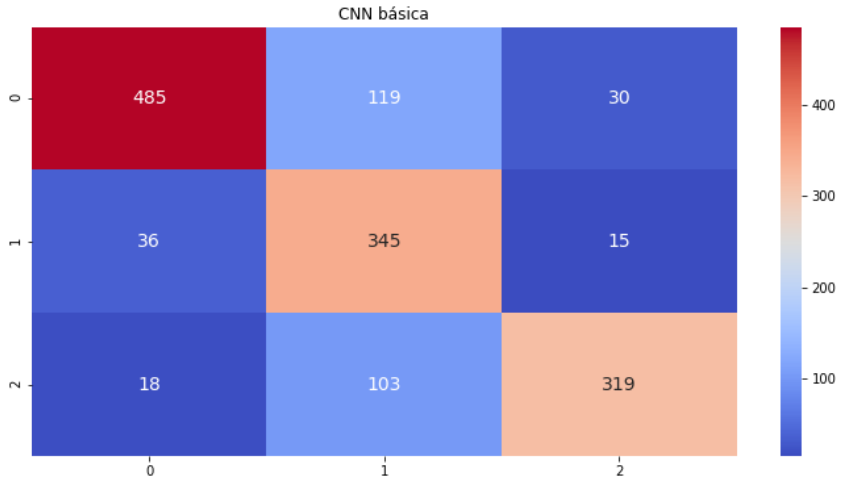

Figura 6 – Matriz de confusão da CNN básica no grupo de teste

Tabela 5 – Resultados do teste da CNN básica

	PPV (%)	Sensibilidade (%)	f1-score (%)	N
Normal	89,98	76,05	82,69	634
Não COVID19	60,85	87,12	71,65	396
Covid19	87,64	72,05	79,35	440
Acurácia (%)	78,16			1470
Loss	0,587			1470

Na Tabela 5 são apresentados os resultados obtidos no grupo de teste utilizando a CNN básica. Nesse caso podemos ver um baixíssimo valor (60,85%) de PPV na classe não COVID19 em relação aos demais.

4.2 CNN Básica com *Dropout*

O tempo computacional na validação cruzada foi de aproximadamente 8 horas em cada *fold*. E o tempo de treino com todos os dados do grupo de treinamento foi de aproximadamente 16 horas. Totalizando em 56 horas totais de processamento computacional. Na Figura 7, pode ser vista a matriz de confusão obtida no grupo de teste.

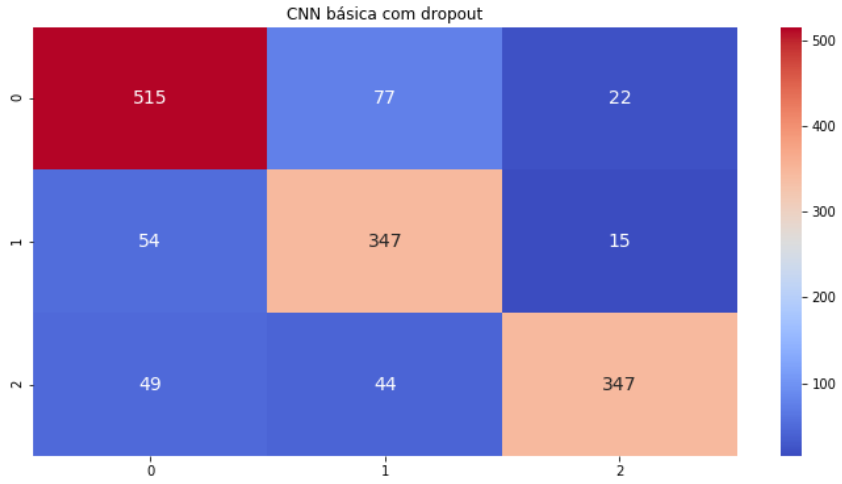

Figura 7 – Matriz de confusão da CNN Básica com Dropout

Tabela 6 – Resultados do teste da CNN Básica com Dropout

	PPV (%)	Sensibilidade (%)	f1-score (%)	N
Normal	83,33	83,88	83,60	614
Não COVID19	74,14	83,41	78,51	416
Covid19	90,36	78,86	84,22	440
Acurácia (%)	82,24			1470
Loss	0,503			1470

Os resultados obtidos no grupo de teste utilizando a CNN Básica com Dropout são apresentados na Tabela 6. Nessa arquitetura pode-se ver que novamente, o PPV na classe não COVID19 é a menor (74,14%) em relação aos demais. E a classe COVID19 possui o maior valor de PPV (90,36%), porém o menor de sensibilidade (78,86%).

4.3 CNN Multicamadas com *DepthWiseConvolution* (DWC)

O tempo computacional na validação cruzada foi de aproximadamente de 20 horas em cada *fold*. E o tempo de treino com todos os dados do grupo de treinamento foi de aproximadamente 25 horas. Totalizando em 125 horas totais de processamento computacional. Na Figura 8, pode ser vista a matriz de confusão obtida no grupo de teste.

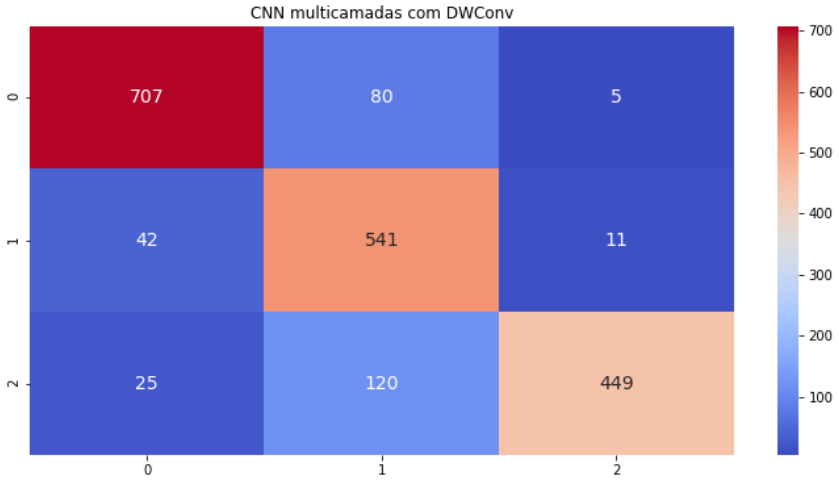

Figura 8 – Matriz de confusão da CNN Multicamadas com DWC

Tabela 7 – Resultados do teste da CNN Multicamadas com DWC

	PPV (%)	Sensibilidade (%)	f1-score (%)	N
Normal	91,34	89,27	90,29	792
Não COVID19	73,01	91,08	81,05	594
Covid19	96,56	75,59	84,80	594
Acurácia (%)	85,71			1980
Loss	0,376			1980

Na Tabela 7 são apresentados os resultados obtidos no grupo de teste utilizando a CNN Multicamadas com DWC. Nessa arquitetura pode-se ver que novamente, o PPV na classe não COVID19 é o menor (73,01%) em relação aos demais. Porém possui o maior valor de sensibilidade (91,08%).

4.4 CNN Multicamadas com *Batch Normalization*

O tempo computacional na validação cruzada foi de aproximadamente 16 horas em cada *fold*. E o tempo de treino com todos os dados do grupo de treinamento foi de aproximadamente 16 horas. Totalizando em 96 horas totais de processamento computacional. Na Figura 9, pode ser vista a matriz de confusão obtida no grupo de teste.

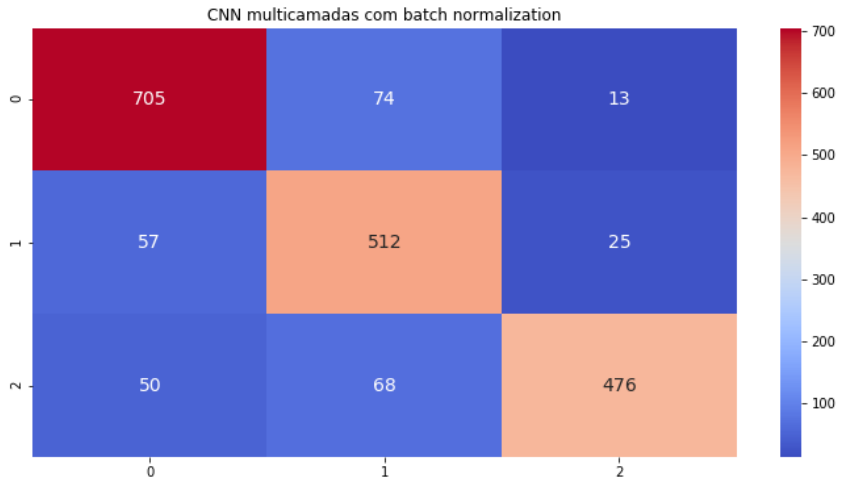

Figura 9 – Matriz de confusão da CNN Multicamadas com Batch Normalization

Tabela 8 – Resultados do teste da CNN Multicamadas com Batch Normalization

	PPV (%)	Sensibilidade (%)	f1-score (%)	N
Normal	86,82	89,02	87,91	792
Não COVID19	78,29	86,20	82,05	594
Covid19	92,61	80,13	85,92	594
Acurácia (%)	85,51			1980
Loss	0,374			1980

Os resultados obtidos no grupo de teste utilizando a CNN Multicamadas com *Batch Normalization* são apresentados na Tabela 8. Pode-se ver que a classe COVID19 possui o maior valor de PPV (92,6%) porém o menor de sensibilidade (80,13%).

5 Discussão e Conclusões

Os resultados da validação cruzada nas duas primeiras redes, mostram que a adição do *Dropout* reduz a variância entre os resultados de cada *fold*. Em média, a acurácia das duas redes é bem próxima, mas o alto desvio padrão da primeira rede sugere a presença de *overfitting*. Isso se confirma nos resultados obtidos no grupo de teste, que mostram que a rede com *Dropout*, obteve resultados melhores tanto de acurácia quanto das métricas em cada classe.

As duas últimas redes, que foram treinadas com a nova estratégia de balanceamento de *batches*, apresentaram resultados próximos na validação cruzada, porém a melhor foi a CNN com multicamadas e *Batch Normalization*, a qual obteve uma acurácia de 85,27%. Não foi possível comparar esses resultados com a rede proposta em [1], pois neste artigo não foi realizado nenhum tipo de validação.

Analisando os resultados obtidos no grupo de teste, a terceira arquitetura, que foi baseada na arquitetura da CovidNet, porém com menor profundidade, obteve uma acurácia de 85,71%, a qual não é muito distante da obtida em [1], de 93,3%. Isso sugere que o aumento da profundidade da rede pode melhorar o seu desempenho na classificação. Esta rede obteve um baixo valor de sensibilidade na classe Covid-19 e um alto valor de PPV. Analisando a métrica *f1-score*, o trabalho em [1] obteve um valor de 94,79% na classe de Covid-19, no passo que a rede CNN multicamadas com DWC obteve 84,8%.

Além de ser a rede com o melhor valor de acurácia durante a validação cruzada, a quarta arquitetura também obteve os melhores valores de predição pela métrica *f1-score*, onde apenas na primeira classe seu valor foi superado pelo obtido na terceira arquitetura. Tanto os resultados obtidos na validação cruzada quanto no teste sugerem que o aumento da profundidade da rede melhora os resultados tanto globalmente (aumento da acurácia) quanto em cada classe (aumento no valor do *f1-score*).

Para trabalhos futuros, pretende-se explorar mais as combinações das diferentes técnicas empregadas neste trabalho, aplicando os diferentes conceitos em redes mais especializadas. Além disso, a fim de se otimizar o treinamento, deseja-se realizar um pré-treino em bancos de dados de imagens existentes.

6 Referências

1. WANG, Linda; LIN, Zhong Qiu; WONG, Alexander. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports, v. 10, n. 1, p. 1-12, 2020.

2. Wang, L, Lin, Z Q, Wong, A. COVID-Net Open Source Initiative. https://github.com/lindawangg/COVID-Net (2020).

3. Cohen, J. P., Morrison, P. & Dao, L. COVID-19 image data collection. arXiv 2003.11597 (2020).

4. Chung, A. Figure 1 COVID-19 chest x-ray data initiative. https://github.com/agchung/Figure1-COVID-chestxray-dataset (2020).

5. Chung, A. Figure 1 COVID-19 chest x-ray data initiative. https://github.com/agchung/Figure1-COVID-chestxray-dataset (2020)

6. of North America, R. S. RSNA pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data (2019).

7. of North America, R. S. COVID-19 radiography database. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database (2019).