UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

MESTRADO ACADÊMICO EM ENGENHARIA DE TELEINFORMÁTICA

PEDRO CROSARA MOTTA

COMPUTER-AIDED DETECTION AND SEGMENTATION SYSTEM OF LESIONS OF COVID-19 AND COMMUNITY-ACQUIRED PNEUMONIA AND THEIR EXTENSION IN CT LUNG IMAGES

FORTALEZA

2023

PEDRO CROSARA MOTTA

COMPUTER-AIDED DETECTION AND SEGMENTATION SYSTEM OF LESIONS OF COVID-19 AND COMMUNITY-ACQUIRED PNEUMONIA AND THEIR EXTENSION IN CT LUNG IMAGES

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Teleinformática do Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Orientador: Prof. Dr. Paulo Cesar Cortez

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

M875c Motta, Pedro.
Computer-Aided Detection and Segmentation System of lesions of COVID-19 and Community-Acquired Pneumonia and their extension in CT lung images / Pedro Motta. – 2023.
75 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2023.
Orientação: Prof. Dr. Paulo Cesar Cortez.

1. COVID-19. 2. Sistema de auxilio ao diagnóstico por computador. 3. CNN. 4. Segmentação. 5. Classificação. I. Título.

CDD 621.38

PEDRO CROSARA MOTTA

COMPUTER-AIDED DETECTION AND SEGMENTATION SYSTEM OF LESIONS OF COVID-19 AND COMMUNITY-ACQUIRED PNEUMONIA AND THEIR EXTENSION IN CT LUNG IMAGES

Dissertação apresentada ao Curso de Mestrado Acadêmico em Engenharia de Teleinformática do Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Aprovada em: 18 de Setembro de 2023

BANCA EXAMINADORA

Prof. Dr. Paulo Cesar Cortez (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Victor Hugo Costa de Albuquerque
Universidade Federal do Ceará (UFC)

Prof. Dr. Alexandre Augusto da Penha Coelho
Universidade Federal do Ceará (UFC)

Prof. Dr. Marcelo Alcântara Holanda
Universidade Federal do Ceará (UFC)

Prof. Dr. Eduardo Tavares Costa
Universidade de Campinas (UNICAMP)

Prof. Dr. Wagner Coelho de Albuquerque Pereira
Universidade Federal do Rio de Janeiro (UFRJ)

AGRADECIMENTOS

Aos meus pais, Marcelo Motta e Raquel Crosara, à minha irmã, Ana Sofia e à minha avó Shirley por todo carinho, amor e apoio ao longo de toda a minha vida;

Ao meu Orientador Prof. Dr. Paulo César Cortez, por toda a orientação desde o início da minha graduação;

Aos amigos de laboratório, Bruno Riccelli e Débora Ferreira, pelo apoio durante a graduação e mestrado;

Aos amigos e amigas, que ajudaram de forma direta ou indireta, com apoio e incentivo, destacando os amigos médicos Dharien Oliveira e Matheus Sombra, pela paciência e disponibilidade para sanar eventuais dúvidas;

Ao Laboratório de Engenharia de Sistemas de Computação (LESC), à Universidade Federal do Ceará (UFC) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq);

"You're unlikely to discover something new without a lot of practice on old stuff, but further, you should get a heck of a lot of fun out of working out funny relations and interesting things."

(Richard Feynman)

RESUMO

"Mesmo com mais de 80% da população vacinada contra a COVID-19, a doença continua causando vítimas. Por isso, é crucial na luta contra essa epidemia ter um sistema de auxílio ao diagnóstico por computador que possa ajudar de forma eficiente na identificação da COVID-19 em exames de tomografia computadorizada e determinar o nível de cuidado necessário, bem como se a doença está progredindo ou regredindo, especialmente na Unidade de Terapia Intensiva. Para criar tal ferramenta, combinaram-se bancos de dados públicos da literatura para treinar modelos de segmentação de pulmão e lesões de diferentes distribuições. Em seguida, foram treinados oito modelos de CNN para classificação de COVID-19 e pneumonia comum. Por fim, se o exame for classificado como COVID-19, as lesões são quantificadas e a severidade da TC é avaliada. Para validação externa utilizou-se o banco de dados SPGC, com Resnext101 Unet++ e MobileNet Unet para segmentação de pulmão e lesão, respectivamente, obtendo uma acurácia de 98,05%, um F1-Score de 98,70%, uma precisão de 98,7%, uma recall de 98,7% e uma especificidade de 96,05%, precisando de apenas 19,70 segundos por varredura completa do CT. Finalmente, para classificar as lesões detectadas como COVID-19 ou pneumonia comum, o Densenet201 alcançou uma precisão de 90,47%, um F1-Score de 93,85%, uma precisão de 88,42%, uma recall de 100,0% e uma especificidade de 65,07%. Os resultados mostraram que nosso sistema detectou e segmentou lesões de COVID-19 e de pneumonia adquirida comum em varreduras de CT corretamente, diferenciando essas duas classes de exames normais."

Palavras-chave: COVID-19; Sistema de auxilio ao diagnóstico por computador; CNN; Segmentação; Classificação; Imagens médicas; exame TC; Validação externa.

ABSTRACT

Even with more than 80% of the population wholly vaccinated for COVID-19, the disease still claims victims. Thus, having a Computer Aided Diagnostic system that can securely assist in identifying COVID-19 and determining the level of care required and if the disease is progressing or digressing, particularly in the Intensive Care Unit, is crucial in the fight against this epidemic. To create such tool, we first merged public datasets from the literature to train Lung and Lesion segmentation models from different distributions. Then we trained eight CNN models for COVID-19 and Common Acquired Pneumonia classification. Finally, if the exam is classified as COVID-19, we quantified the lesions and evaluated the severity of the full CT Scan. For external validation on SPGC Dataset, using Resnext101 Unet++ and MobileNet Unet for lung and lesion segmentation, respectively, we achieved an accuracy of 98.05%, an F1-score of 98.70%, a precision of 98.7%, a recall of 98.7%, and a specificity of 96.05%, needing only 19.70 seconds per full CT scan. Finally, when classifying these detected lesions, Densenet201 reached an accuracy of 90.47%, an F1-score of 93.85%, a precision of 88.42%, a recall of 100.0%, and a specificity of 65.07%. The results showed that our pipeline correctly detected and segmented lesions from COVID-19 and Common Acquired Pneumonia in CT scans, differentiating these two classes from Normal exams.

Keywords: COVID-19; Computer Aided Diagnostic; CNN; Segmentation; Classification; Medical image; CT Scan; External validation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

COVID-19	Coronavirus Disease 2019
CT	Computed Tomography
GGO	Ground-Glass Opacity
CAP	Community Acquired Pneumonia
CAD	Computer Aided Diagnostic
ICU	Intensive Care Unit
CNN	Convolutional Neural Network
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
WHO	World Health Organization
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
RT-PCR	Reverse Transcriptase-PCR
ANN	Artificial Neural Network
FPN	Feature Pyramid Network
MAnet	Multi-Scale Attention Network
AI	Artificial Intelligence
NLP	Natural Language Processing
DICOM	Digital Imaging and Communications in Medicine
NIFTI	Neuroimaging Informatics Technology Initiative
PNG	Portable Network Graphics
HD	Hausdorff Distance
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

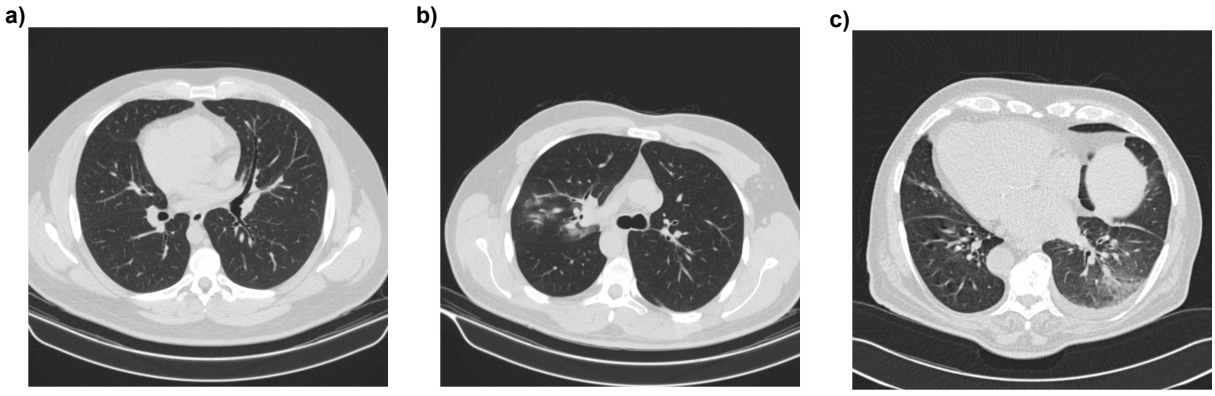
SUMÁRIO

1 INTRODUCTION

Even with more than 80% of the population wholly vaccinated for Coronavirus Disease 2019 (COVID-19) and the development of knowledge for treating it, this disease still claims victims (DONG *et al.*, 2020; NIH, 2023). Moreover, the COVID-19 pandemic has caused several global economic, social, environmental, and healthcare impacts (RUME; ISLAM, 2020; MATTIOLI *et al.*, 2020).

Computed Tomography (CT) scans of the chest can effectively aid in diagnosing individuals suspected of having COVID-19, as pneumonia is a frequently observed symptom of COVID-19 (ZHAO *et al.*, 2020a). In the CT analysis, the main characteristics present in patients with COVID-19 are: Ground-Glass Opacity (GGO) (88.0%), bilateral involvement (87.5%), peripheral distribution (76.0%), and multilobar involvement (78.8%) (SALEHI *et al.*, 2020). GGO, seen on CT images as increased density in lung tissue, can be caused by various factors, including partial filling of the alveoli, increased blood flow, or a combination of both. While it is a common finding in CT scans of individuals diagnosed with COVID-19, it is not exclusive due to the virus. Other conditions, such as influenza, cytomegalovirus, Community Acquired Pneumonia (CAP), and pulmonary edema, can also cause it, as is shown in Figure 1. Therefore, relying solely on detecting and segmenting ground-glass opacity is insufficient for diagnosing COVID-19 (MATOS MARINA JUSTI ROSA DE; ROSA, 2021).

Figure 1 – (a) CT slice without COVID-19 nor CAP. (b) CT slice with COVID-19. (c) CT slice with CAP



Source: Author (2023)

Computer Aided Diagnostic (CAD) systems, which may use machine learning methods to aid in the diagnostic process, can assist medical doctors in pinpointing specific areas of concern in medical images. These identified regions can then be used to detect illnesses and

provide numerical data. Physicians can analyse this data to assess the progression or regression of the disease (VALENTE *et al.*, 2016).

Using a CAD system mainly indicates an improvement in at least medical doctors' sensitivity, specificity, or speed of diagnosis. This improvement is most noticeable for junior and resident radiologists. Additionally, providing class activation maps for the experts' radiologists can help them examine the involved regions (YOUSEFZADEH *et al.*, 2021).

Thus, having a CAD system that can securely assist physicians in identifying COVID-19 and determining the level of care required by the patient and if the disease is progressing or digressing, particularly in the Intensive Care Unit (ICU), is crucial in the fight against this disease (PARAH *et al.*, 2021).

1.1 Objectives

The main objective of this work is to develop a ready-to-use system that segments ground-glass opacity and consolidation lesions on full CT scans, classifies exams with COVID-19 or CAP, and quantifies the severity of lesions on full COVID-19 CT scans.

1.2 Specific Objectives

The specific objectives of this work are listed below:

- to develop a system for segmenting lungs and lesions, detecting COVID-19 and CAP, and calculating COVID-19 severity using machine learning;
- to realise an extensive segmentation architecture statistical analysis on a combination of datasets with Healthy, COVID-19, and Other Diseases patients for lung and lesion detection;
- to validate with a cross-dataset approach, aiming for a better model generalisation through an external validation dataset;

This work is organised as follows. Chapter 2 provides a theoretical foundation and a revision of related works in the literature. In Chapter 3, we describe our applied methodology. The results and discussions are presented in Chapter 4. Finally, the conclusions of this work are detailed in Chapter 5.

1.3 Scientific Contributions

Currently, the content of this dissertation is published or is under review with the following bibliographic information:

Journal Papers

JP1. MOTTA, PEDRO CROSARA; CORTEZ, PAULO CÉSAR ; SILVA, BRUNO R. S. ; YANG, GUANG ; ALBUQUERQUE, VICTOR HUGO C. DE. Automatic COVID-19 and Common-Acquired Pneumonia Diagnosis Using Chest CT Scans. Bioengineering-Basel, v. 10, p. 529, 2023 Disponível em: http://dx.doi.org/10.3390/bioengineering10050529.

Book Chapters

BC1. MOTTA, P. C.; CORTEZ, P. C. ; MARQUES, J. A. L. COVID-19 Classification Using CT Scans with Convolutional Neural Networks. In: Joao Alexandre Lobo Marques; Simon James Fong. (Org.). Computerized Systems for Diagnosis and Treatment of COVID-19. 1ed.Cham: Springer International Publishing, 2023, v., p. 99-116. Disponível em: http://dx.doi.org/10.1007/978-3-031-30788-1_7.

BC2. MARQUES, J. A. L. ; MACEDO, D. S. ; MOTTA, P. C. ; SILVA, B. R. S. ; CARVALHO, F. H. C. ; KEHDI, R. C. ; CAVALCANTE, L. R. L. ; VIANA, M. S. ; LOS, D. ; FIORENZA, N. G. Exploratory Data Analysis on Clinical and Emotional Parameters of Pregnant Women with COVID-19 Symptoms. In: Joao Alexandre Lobo Marques; Simon James Fong. (Org.). Computerized Systems for Diagnosis and Treatment of COVID-19. 1ed.Cham: Springer International Publishing, 2023, v., p. 179-209. Disponível em: http://dx.doi.org/10.1007/978-3-031-30788-1_11.

BC3. SILVA, B. R. S. ; CORTEZ, P. C. ; MOTTA, P. C. ; MARQUES, J. A. L. Covid-19 Detection Based on Chest X-Ray Images Using Multiple Transfer Learning CNN Models. In: João Alexandre Lobo Marques; James Fong. (Org.). Computerized Systems for Diagnosis and Treatment of COVID-19. 1ed.Cham: Springer International Publishing, 2023, v., p. 45-63. Disponível em: http://dx.doi.org/10.1007/978-3-031-30788-1_4.

2 THEORETICAL FOUNDATION AND RELATED WORKS

In this chapter, we will present concepts incrementally, starting with basic Convolutional Neural Network (CNN) for classification, then showing novel CNN architectures for optimization, and finally, the modifications made in CNNs for image segmentation. Then, we summarize novel works published in journals of relevant impact that presented similarity to our work in materials, such as datasets and architectures, or scope and methodology.

2.1 Theoretical foundation

2.1.1 Chest Computed Tomography

In radiography exams, subtle differences in subject contrast below about 5 percent are not visible in the image due to several limitations. These include the projection of 3D anatomy onto a 2D image receptor, which obscures differences in X-ray transmission for structures parallel to the beam, and the inability of traditional image receptors to resolve minor intensity differences in incident radiation. Large-area X-ray beams also produce significant scattered radiation, further hindering the display of subtle contrast differences. Computed tomography (CT) overcomes these limitations, revealing slight differences in subject contrast. While CT has lower spatial resolution than conventional radiography, it excels in visualizing subject contrast and allows for cross-sectional imaging, making it highly valuable for anatomical visualization in various body regions (HENDEE; RITENOUR, 2002).

The CT produces a series of images by a tomographic method. Each image is derived from a particular slice. It involves a rotating X-ray source and a detector array that synchronizes with the X-ray source. As they rotate around the patient, these components capture various X-ray images from various angles. A computer then processes these images to reconstruct detailed cross-sectional slices of the body (HOUNSFIELD, 1973).

Hounsfield Units (HU) is a standardized measurement scale used in CT imaging to assess the density of materials within the human body. They help distinguish between different tissues based on density, with reference points of -1000 HU for air and 0 HU for water. HU values are crucial in clinical practice, aiding in identifying various structures, diagnosis of medical conditions, and treatment planning (HOUNSFIELD, 1973).

2.1.2 COVID-19

In December 2019, an outbreak of pneumonia cases with an unknown cause occurred in Wuhan, Hubei, China, attracting global attention. Then, a new coronavirus originating from bats, called the 2019 novel coronavirus, was identified through deep sequencing analysis (WHO, 2020c; REN *et al.*, 2020). This virus, named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) (WHO, 2020a), is genetically similar to the bat SARS-like coronavirus but belongs to a distinct clade, indicating the spread of a novel coronavirus (REN *et al.*, 2020). On January 31, 2020, the World Health Organization (WHO) declared the outbreak of COVID-19 a Public Health Emergency of International Concern (WHO, 2020b). Moreover, the COVID-19 pandemic caused several global economic, social, environmental, and healthcare impacts (RUME; ISLAM, 2020; MATTIOLI *et al.*, 2020).

The rapid transmission of the virus reminded the previous outbreaks of SARS-CoV and MERS-CoV in the 21st century. SARS-CoV-2 has efficient human-to-human transmission capability. As a result, the number of confirmed COVID-19 cases surged, although the mortality rate of COVID-19 is lower than SARS-CoV and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) (ZHAO *et al.*, 2020b).

COVID-19 symptoms mainly include fever, chills, cough, sore throat, breathing difficulty, myalgia or fatigue, nausea, vomiting, and diarrhoea. More severe cases can lead to cardiac injury, respiratory failure, acute respiratory distress syndrome, and even death. Older people along with people with other medical conditions have a high risk of mortality (RUME; ISLAM, 2020; HUANG *et al.*, 2020; WANG *et al.*, 2020a; HOLSHUE *et al.*, 2020; CHEN *et al.*, 2020).

Real-time Reverse Transcriptase-PCR (RT-PCR) is the primary method to detect SARS-CoV-2 as it is a specific and straightforward qualitative assay and adequate sensitivity to early diagnosis. However, an issue with the real-time RT-PCR test is the risk of false negatives (TAHAMTAN; ARDEBILI, 2020). Some patients had CT findings even when the RT-PCR results were negative (SALEHI *et al.*, 2020; HUANG *et al.*, 2020; XIE *et al.*, 2020) Thus, an RT-PCR negative result does not exclude the possibility of COVID-19 infection and should not be used as the only criterion for treatment or patient management decisions (TAHAMTAN; ARDEBILI, 2020).

In the CT analysis, the main characteristics present in patients with COVID-19 are: GGO (88.0%), bilateral involvement (87.5%), peripheral distribution (76.0%), and multilobar

involvement (78.8%). Isolated GGO or a combination of GGO and consolidated opacities were some of the most common CT findings. Other CT findings as interlobular septal thickening, bronchiectasis, pleural thickening, and subpleural involvement Pleural effusion, pericardial effusion, lymphadenopathy, cavitation, CT halosign, and pneumothorax were less common or rare. (SALEHI *et al.*, 2020).

However, GGO is a non-specific finding that can represent thickening of the interstitium, partial filling of the alveoli, or partial collapse of the alveoli, increased blood supply, or even a combination of these findings. Radiographically, it is defined as an increase in the density of the lung parenchyma while preserving the bronchovascular markings, which differs from consolidation. While it is a common finding in CT scans of individuals diagnosed with COVID-19, it is not exclusive due to the virus. Other conditions, such as influenza, cytomegalovirus, CAP, and pulmonary edema, can also cause it, as is shown in Figure 1. Therefore, relying solely on detecting and segmenting ground-glass opacity is insufficient for diagnosing COVID-19 (MATOS MARINA JUSTI ROSA DE; ROSA, 2021).

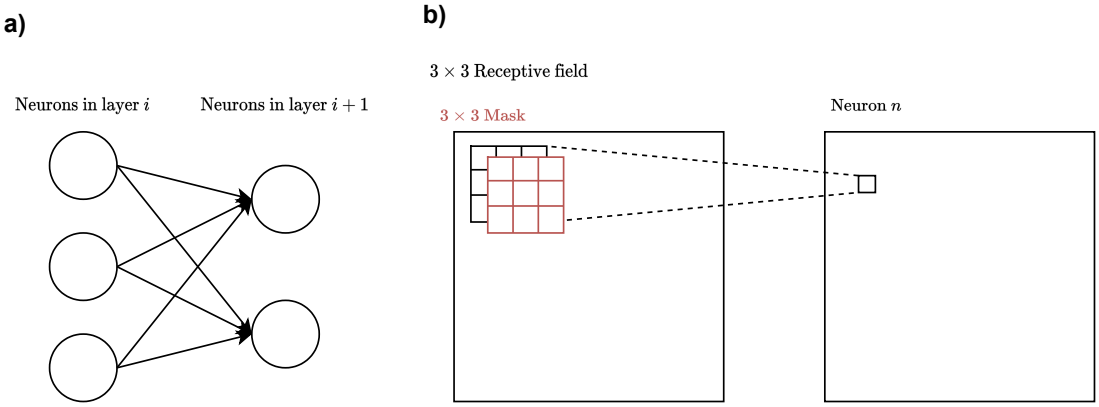
2.1.3 Convolutional Neural Networks - CNNs

Artificial Neural Network (ANN) for classification were designed to receive a 1-D input data vector from a sample, perform computations, and determine which class the sample belongs to. Therefore, if one wishes to use an ANN to classify a 2-D image, the image must first be represented as a vector by extracting only the most essential characteristics from the image. With these characteristics (in a 1-D vector), it is possible to feed the ANN and classify the image. However, this feature selection and extraction process occurs before the ANN learning step, needing a human to discover which characteristics are most important in representing a group of images. This human dependency can be a time-consuming and tedious activity, and in the end, even with good results, there may be another combination of characteristics that better represents the classes of an image that humans have not tested. The advantage of the CNN is that a 2-D image can be directly provided as input to the CNN, automatically learning which features to use and classifying the image.

Unlike an ANN, which have all its neurons fully connected (Figure 2.a), meaning the output of each neuron in a layer is connected to the input of all neurons in the next layer, in the CNN. In the CNN, it is possible to analogize that a neuron receives as input only a single value, calculated by the convolution of a mask with a neighbourhood of pixels from the output

image of the previous layer (Figure 2.b) (GONZALEZ; WOODS, 2018).

Figure 2 – a) ANN neurons. a) CNN neurons.



Source: Author (2023)

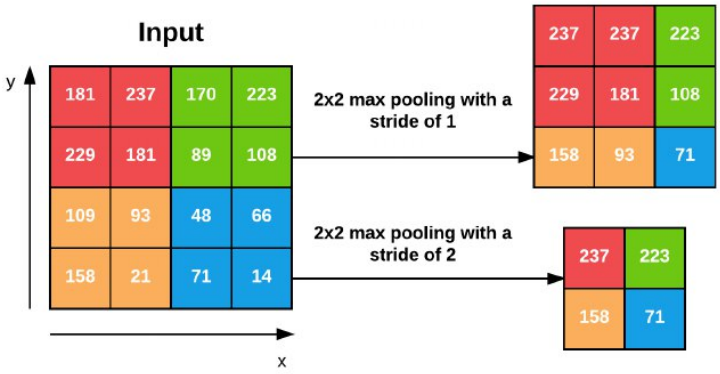
This pixel neighbourhood is called a receptive field and represents where the CNN is "looking at"each moment (the part of the image that will be convoluted with the mask). The receptive field walks through the entire image, being able to move from pixel to pixel or even "skip"some pixels, reducing the number of convoluted values and consequently reducing the output image dimension (subsampling). The stride defines the number of pixels the receptive field will skip (GONZALEZ; WOODS, 2018).

In each convolution result, a bias is added, which goes through an activation function and returns a value. This value is placed in its corresponding position (x, y) in the next layer's input (which we previously called a neuron). By repeating this process, using the same mask and the same bias for all input image pixels (except those skipped depending on the stride), a 2-D image called a feature map is obtained. Applying different masks and biases, we obtain several feature maps that ideally enhance essential image features. It is these masks and biases that the CNN will learn (GONZALEZ; WOODS, 2018).

Also, to reduce the dimension of the feature maps, one can use a pooling layer, where either a maximum filter (max-pooling) or an average filter (mean-pooling), usually of size 2x2, is applied to the image. For each 2x2 neighbourhood of the image, the maximum pixel value of the neighbourhood is calculated, for example, in max-pooling, which will represent it in the resulting feature map (Figure 3) (ROSEBROCK, 2018).

Finally, we convert the feature maps generated by the last pooling layer into vectors and feed them into a fully connected neural network. As usual, the fully connected network has one output for each class to which the image can belong. The class with the highest output value

Figure 3 – **Left:** 4*x*4 input image. **Right:** subsampled output with a pooling of 2*x*2 and stride of 1 and 2.

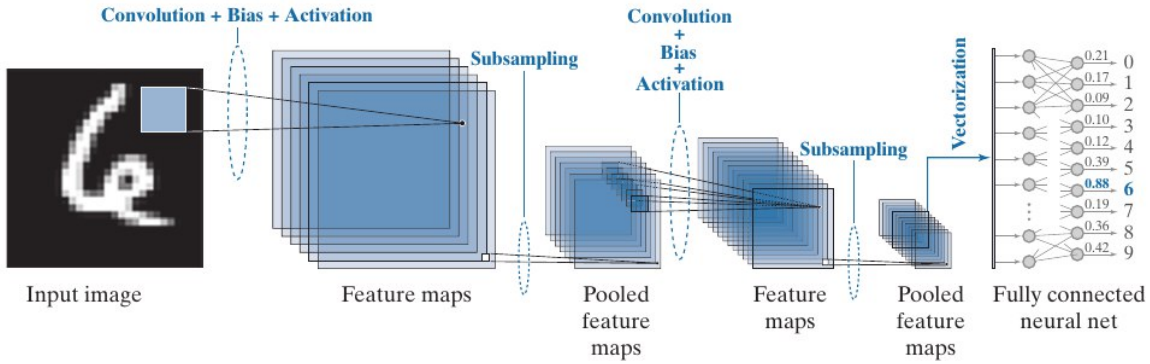


Source: (ROSEBROCK, 2018).

is assigned to the image (GONZALEZ; WOODS, 2018). All of this process is shown in Figure 4, where:

- it starts with an image of the handwritten number 6;
- the image goes through a convolutional layer and an activation layer, forming several feature maps;
- the feature maps are reduced (subsampling) and forwarded to the next convolutional layer, which will form new feature maps;
- these feature maps are reduced again, vectorized, and fed into the fully connected neural network;
- the fully connected network returns one value for each class (in this case, the numbers from 0 to 9). As the class with the highest value is "6", the image is classified as "6".

Figure 4 – Usual CNN architecture.



Source: (GONZALEZ; WOODS, 2018).

2.1.3.1 MobileNetV2

MobileNetV2 is a CNN architecture designed for mobile and resource-constrained environments, such as limited computational resources and memory. Their paper presents two main contributions that improve the original MobileNet architecture:

1. Inverted Residuals: The inverted residual block reverses the traditional residual block. Instead of increasing the number of channels in the middle of the block, as in a traditional residual block, the inverted residual block first reduces the number of channels and then increases them. This is done by applying a 1x1 convolution layer with fewer channels before the main 3x3 depthwise convolution layer, followed by another 1x1 convolution layer that expands the channels to the original size. This design significantly reduces computation and memory usage while maintaining accuracy;
2. Linear Bottlenecks: MobileNetV2 also introduces linear bottlenecks, which refers to using a linear activation function in the bottleneck layers of the network. The authors found that non-linear activation functions such as ReLU in the bottleneck layers caused information loss and reduced accuracy. Using a linear activation function, MobileNetV2 can preserve information and improve accuracy.

MobileNetV2 also includes other optimizations, such as using a width multiplier to adjust the number of channels in the network based on the available computational resources and using a new type of batch normalization that reduces the number of parameters and computations (SANDLER *et al.*, 2018a).

2.1.3.2 ResNet50

The ResNet architectures have a new type of residual block called a bottleneck block, which enables the construction of very deep convolutional neural networks with improved accuracy.

A residual block is a building block in a CNN that allows the network to learn a residual mapping instead of directly trying to fit a desired mapping. This is achieved by adding a shortcut connection that bypasses one or more layers in the block. The bottleneck block is a residual block that includes three layers: a 1x1 convolution layer that reduces the number of input channels, a 3x3 convolution layer that performs the main computation, and another 1x1 convolution layer that increases the number of output channels back to the original size. The

authors found that this bottleneck structure reduces the parameters and computation required to train the network while maintaining or improving accuracy.

The authors also found that as the depth of the network increases, the accuracy first improves but then saturates and eventually degrades. To address this issue, they introduced a new training method called deep residual learning, which includes residual connections between layers several hops away from each other. This allows the network to learn more effective features and reduces the vanishing gradient problem, which can occur in very deep networks (HE *et al.*, 2016a).

2.1.3.3 DenseNet201

The Densely Connected Convolutional Networks (DenseNet) is a new type of CNN architecture connecting all layers directly. In a traditional CNN, the output of each layer is fed only to the next layer. In contrast, a DenseNet connects each layer to every subsequent layer. Feature maps of all previous layers are concatenated as inputs to the current layer. This design creates a dense connectivity pattern, hence the name DenseNet.

The authors found that the dense connectivity pattern reduces the number of parameters required to achieve a certain level of accuracy and improves gradient flow, which can help alleviate the vanishing gradient problem in deep networks. The dense connectivity pattern also promotes feature reuse and enhances the flow of information through the network. DenseNet comprises several dense blocks, a series of dense layers followed by a transition layer that reduces the number of channels. The authors found that this design leads to better performance than traditional CNNs with a similar number of parameters (HUANG *et al.*, 2017a).

2.1.3.4 ResNeXt101

Aggregated Residual Transformations for Deep Neural Networks (ResNeXt) introduce a new CNN architecture that achieves state-of-the-art performance on image classification tasks. The main idea behind ResNeXt is to create a network that combines the benefits of two popular techniques in deep learning: residual connections and cardinality.

Residual connections are a way to make it easier for a neural network to learn by allowing information from earlier layers to bypass some of the later layers. On the other hand, cardinality refers to the number of paths that information can take through the network. ResNeXt uses both techniques by creating a block consisting of multiple paths through the network, each

with its own set of weights. These paths are then aggregated by concatenation to produce the output of the block.

ResNeXt can achieve state-of-the-art performance with a relatively small number of parameters compared to other models, making it more efficient to train and use in practice (XIE *et al.*, 2016).

2.1.3.5 SqueezeNet

SqueezeNet introduces a new CNN architecture that achieves high accuracy on image classification tasks while using significantly fewer parameters than previous state-of-the-art models. It achieves this by focusing on "squeeze"operations, which reduce the number of input channels to a convolutional layer. SqueezeNet uses a combination of three main techniques to achieve its efficiency:

1. It replaces 3x3 filters with 1x1 filters, which reduces the number of parameters in the model.
2. It uses a technique called "fire modules", which consist of a squeeze layer followed by an expand layer. The squeeze layer has 1x1 filters and reduces the number of input channels, while the expand layer has 1x1 and 3x3 filters to increase the number of output channels.
3. SqueezeNet uses aggressive down-sampling to reduce the spatial size of the feature maps and further reduce the number of parameters. Despite its small size, SqueezeNet achieves accuracy comparable to AlexNet, a much larger and more complex model. In addition, SqueezeNet has a model size of less than 0.5 MB, making it well-suited for deployment on resource-constrained devices such as smartphones and IoT devices.

SqueezeNet demonstrates that high accuracy on image classification tasks can be achieved with significantly fewer parameters than previous state-of-the-art models, making it a promising architecture for efficient deep learning (IANDOLA *et al.*, 2016).

2.1.3.6 EfficientNet

EfficientNet proposes a new approach to model scaling for CNNs. The authors argue that previous approaches to model scaling have been ad hoc and that there is a need for a more principled approach that balances accuracy and computational efficiency.

To achieve this, a compound scaling method is introduced that systematically scales the depth, width, and resolution of the CNN. Specifically, a scaling formula that allows them

to scale up the CNN balanced while keeping the number of parameters and computational cost under control is proposed.

A family of CNN models, called EfficientNets, that achieve state-of-the-art accuracy on the ImageNet dataset while using fewer parameters and less computation than previous state-of-the-art models, is developed. They also show that EfficientNets generalize well to other computer vision tasks, such as object detection and segmentation. EfficientNet models achieve this efficiency by incorporating several design choices, such as a new convolutional layer called a "swish"activation function and a new type of scaling called "compound scaling"that systematically balances depth, width, and resolution (TAN; LE, 2019).

2.1.3.7 ShuffleNet

ShuffleNet is a new CNN architecture designed for mobile devices with limited computational resources. The authors argue that existing CNN architectures are too computationally expensive for use on mobile devices and that there is a need for a more efficient architecture.

ShuffleNet achieves its efficiency by using several techniques. First, it uses group convolutions, which divide the input and output channels into groups and perform convolution separately in each group. This reduces the number of parameters in the model and improves its efficiency. Second, ShuffleNet uses a channel shuffle operation that exchanges information between groups. This operation is designed to enhance the model's accuracy and maintain its efficiency. The authors show that ShuffleNet achieves state-of-the-art accuracy on the ImageNet dataset using significantly fewer parameters and less computation than previous state-of-the-art models (ZHANG *et al.*, 2017).

2.1.3.8 GhostNet

GhostNet is a new CNN designed to be efficient and accurate. The authors argue that existing models design approaches have focused too heavily on increasing the number of parameters while neglecting the importance of efficient computation. To address this, GhostNet introduces a new type of block called a "ghost bottleneck", designed to extract more features from cheap operations. The ghost bottleneck consists of a combination of cheap operations, such as depthwise separable convolutions, and a new operation called a "ghost module", which creates a low-rank approximation of the input feature maps.

The authors show that GhostNet achieves state-of-the-art accuracy on the ImageNet

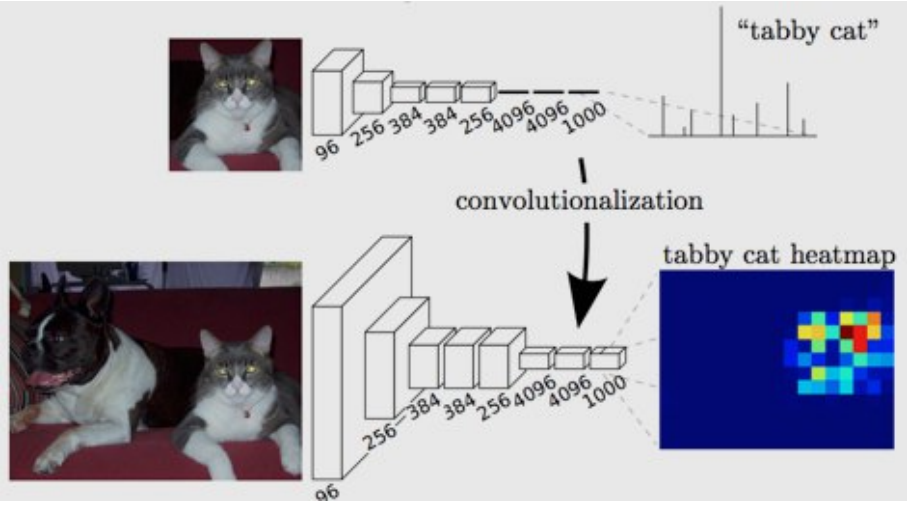
dataset using fewer parameters and less computation than previous state-of-the-art models. In addition to its efficiency, GhostNet introduces several other novel ideas, such as a new activation function called "SiLU", which outperforms existing activation functions on various tasks (HAN *et al.*, 2019).

2.1.4 CNNs for Segmentation

For some analyses, image classification is insufficient and additional information is necessary. For example, spatial information about COVID-19 lesions, such as the size of the lesion or in which lung lobe it is located, can aid in the prognosis of the disease. Therefore, segmenting the object after detecting it is often a necessary step.

In addition to traditional image segmentation techniques such as thresholding and edge detection, CNNs have been adapted to perform more robust and complex segmentations, allowing for efficient segmentation of objects even in low-quality images or images with variations in lighting or camera position. As a result, several CNN architectures are being developed for medical image segmentation and analysis. Classification CNNs take fixed-size input images and produce output without dimensionality, since fully connected layers do not return spatial information. However, replacing these layers with other convolutional layers is possible. This converts the output into a classification map (Figure 5), where each pixel is assigned to a class, such as "background"or "tabby cat"(LONG *et al.*, 2014).

Figure 5 – CNN for segmentation.

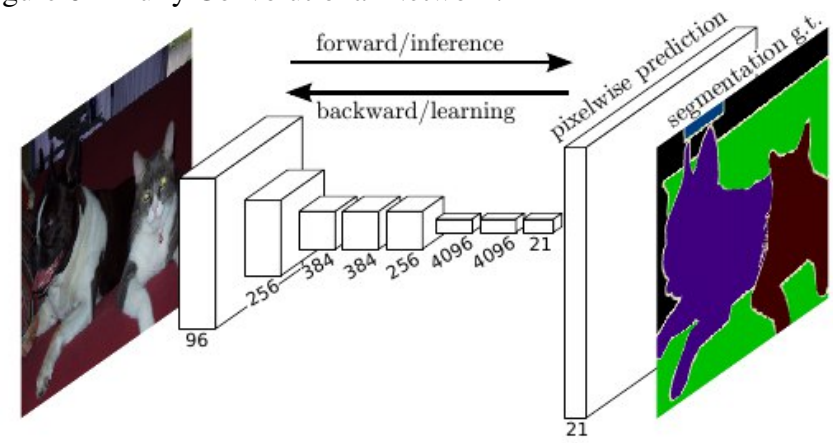


Source: https://www.jeremyjordan.me/semantic-segmentation

Single-stage decoding (Figure 6) from small feature maps to a classification map of

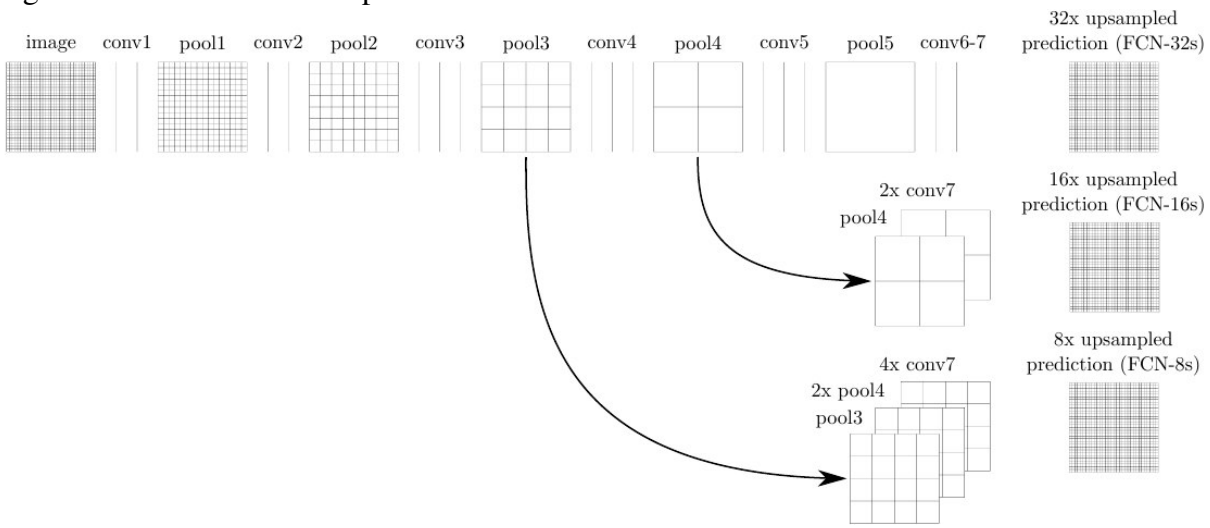
the original image size, can provide inaccurate segmentation. Long *et al.* (2014) addressed this issue by decoding the image in stages with skip connections (Figure 7) from earlier layers, where feature maps are not as small, to deeper layers. These skip connections provide the necessary details to form more refined segmentation boundaries.

Figure 6 – Fully Convolutional Network.



Source: (LONG *et al.*, 2014)

Figure 7 – Network with skip connections.



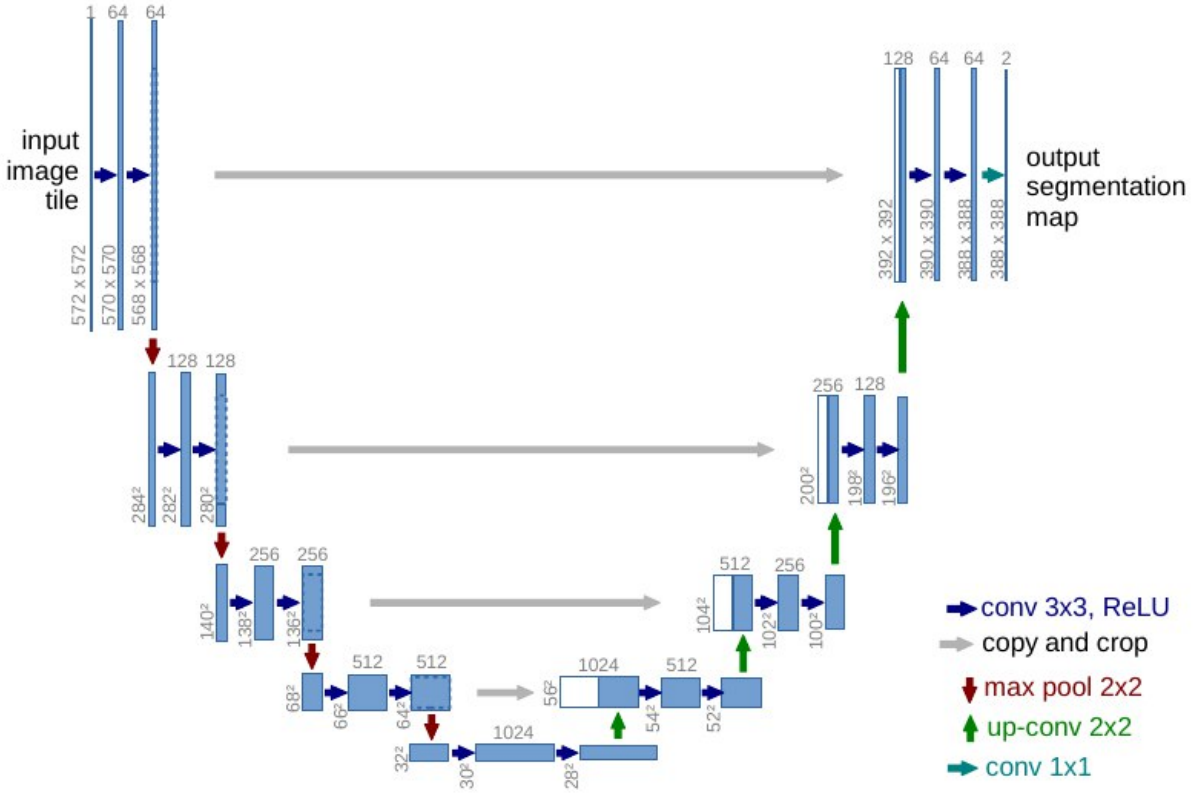
Source: (LONG *et al.*, 2014)

2.1.4.1 Unet

To improve segmentation architectures, Ronneberber *et al.* (2015) enhanced the decoder, making it approximately symmetric to the encoder. In other words, in the encoding step of the architecture shown in Figure 8, each block consists of two 3*x*3 convolutions, each followed by an activation function and a 2*x*2 max pooling operation with stride 2. In the decoding

step, each block expands the feature maps, applies a 2*x*2 convolution, concatenates it with the corresponding reduced feature map, and applies two 3*x*3 convolutions followed by an activation function. Finally, a 1*x*1 convolution is used to map the feature vectors to the desired number of classes (RONNEBERBER *et al.*, 2015).

Figure 8 – U-Net architecture.



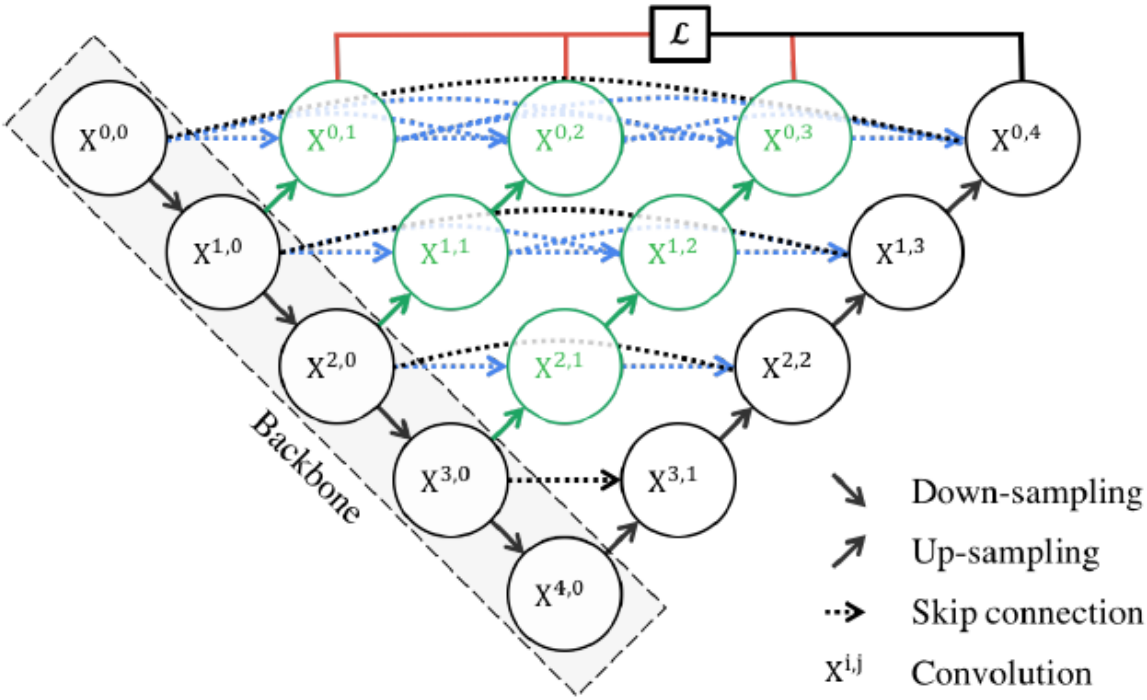
Source: (RONNEBERBER *et al.*, 2015)

2.1.4.2 Unet++

Zhou *et al.* (2018) formulated a new architecture, arguing that the model could capture finer details of objects more efficiently if the high-resolution feature maps from the encoding part were gradually enriched before being combined with the semantically rich feature maps from the decoding part. The difference between U-Net and U-Net++ is shown in Figure 9. The skip pathways (in green and blue) that connect the encoding and decoding networks have been reformulated, and deep supervision is applied (red connections) (ZHOU *et al.*, 2018).

In UNet, the encoder's feature maps are sent directly to the decoder, while in U-Net++, they first go through a dense convolution block where the number of convolutions depends on their level in the pyramid. For example, the skip pathway between nodes $X^{0,0}$ and

Figure 9 – U-Net++ architecture.



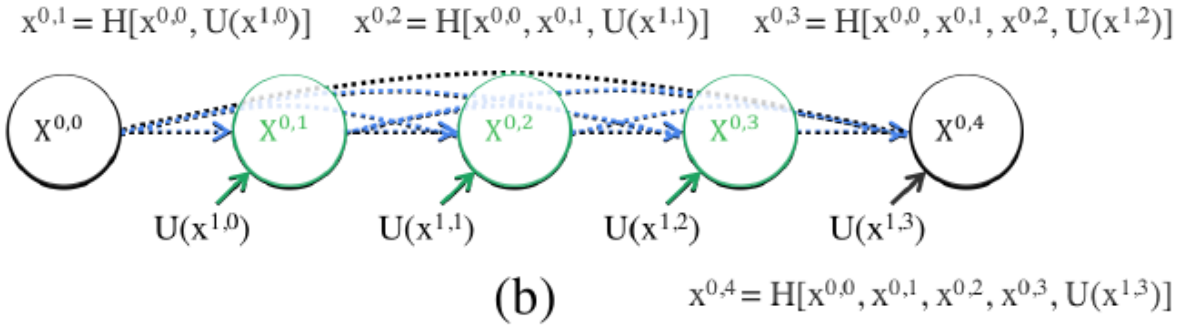
Source: (ZHOU *et al.*, 2018)

$X^{0,3}$ consists of a convolution block with three convolution layers, with a concatenation layer to combine the outputs of the previous convolution layers with the expanded output of the dense block below in the pyramid. These operations bring the semantic level of the encoder's feature maps closer to the level of the feature maps in the decoder. For example, $x^{i,j}$ is the output of node $X^{i,j}$, where i determines the down-sampling layer in the encoder and j the convolution layer of the dense block in the skip pathway. The set of feature maps represented by the output $x^{i,j}$ is computed by

$$x^{i,j} = \begin{cases} H(x^{i-1,j}), & j = 0. \\ H\left(\left[[x^{i,k}]_{k=0}^{j-1}, U(x^{i+1,j-1})\right]\right), & j > 0. \end{cases} \tag{2.1}$$

Where $H(.)$ is a convolution operation followed by an activation function, $U(.)$ is an up-sampling layer, and $[.]$ is the concatenation layer. As shown in Figure 10, nodes at level $j = 0$ receive only one input from the previous layers; nodes at level $j = 1$ receive two inputs; and nodes at levels $j > 1$ receive $j + 1$ inputs, where j are the outputs of the previous j nodes in the same skip pathway. The last input is the up-sampled output from the skip pathway below in the pyramid (ZHOU *et al.*, 2018).

Figure 10 – Skip pathway example.



Source: (ZHOU *et al.*, 2018)

In deep supervision, multiple segmentation maps are formed at different levels of resolution, and then either an average of all maps is taken or one of the maps is selected and designated as the final map. For example, figure 11 shows the difference in architecture depending on which map is selected.

2.1.5 FPN

Feature Pyramid Network (FPN) is a novel approach to improve the performance of object detection algorithms. It aims to leverage multiscale features from deep CNNs for object detection.

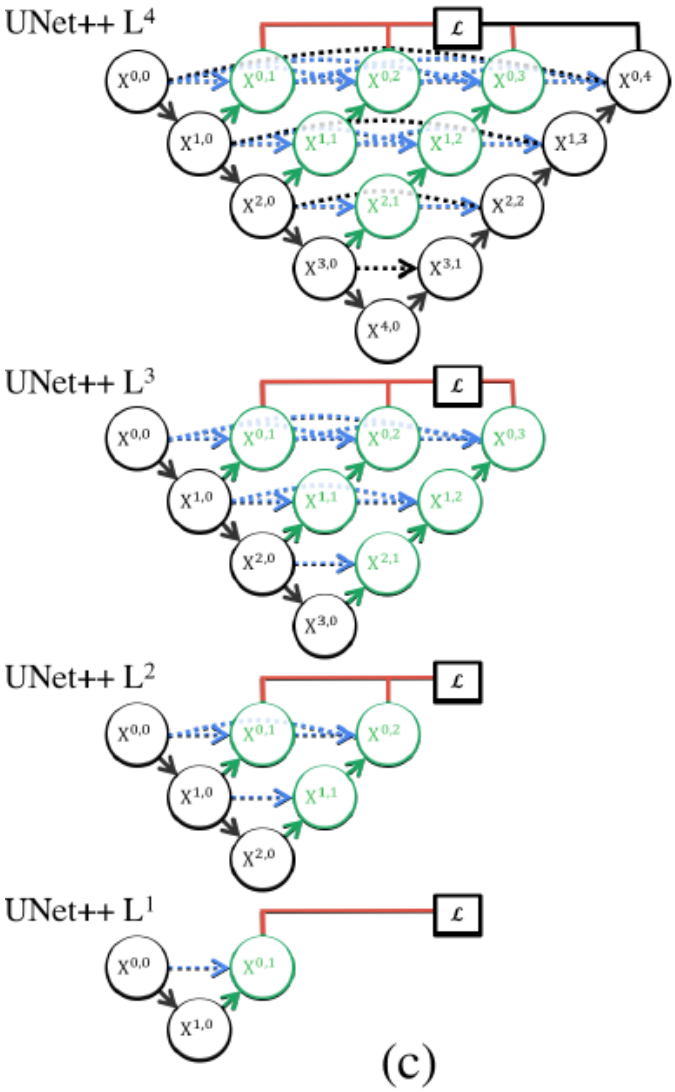
FPN builds on top of a base network, which consists of a backbone network that produces a rich hierarchy of features at different spatial resolutions. The FPN architecture introduces a top-down pathway and lateral connections to the backbone network, which enables the fusion of high-level semantic information with fine-grained features. The resulting multiscale feature pyramid detects objects of different sizes and scales (LIN *et al.*, 2017a).

2.1.5.1 MAnet

Multi-Scale Attention Network (MAnet) is a novel deep-learning architecture for segmenting liver and liver tumours from CT scans. The model is based on the encoder-decoder architecture with attention mechanisms to capture multiscale contextual information from the input image.

MAnet consists of a multiscale feature extraction module and an attention-guided decoder module. The feature extraction module extracts multiscale features using convolutional layers with different dilation rates. The attention-guided decoder module generates segmen-

Figure 11 – Different architectures examples depending on deep supervision.



Source: (ZHOU *et al.*, 2018)

tation masks by fusing multiscale features with attention maps highlighting the input image's informative regions (FAN *et al.*, 2020a).

2.1.6 Related Works

Because of the rapid manifestations of COVID-19 and the significant number of disease cases, many Artificial Intelligence (AI) studies have been conducted to aid medical diagnosis with medical data in the areas of disease classification, and lung and lesion segmentation. We selected novel works published in journals of relevant impact that presented similarity to our work in materials, such as datasets and architectures, or scope and methodology.

Natural Language Processing (NLP) models can efficiently extract information from

clinical reports, providing a comprehensive view of a patient's symptoms and medical history. NLP models can be helpful in scenarios where radiographic images are unavailable or difficult to obtain. Moreover, NLP models can be trained on relatively small datasets, which may be beneficial when data availability is limited. On the other hand, image analysis with CNNs can provide more direct and accurate information about the presence of COVID-19 in radiographic images. CNNs have shown great promise in accurately detecting COVID-19 in chest X-rays and CT scans. However, CNNs require large datasets to be trained effectively, and interpreting the results may not always be straightforward.

Some authors applied NLP methods to extract text information from medical reports to identify evidence of COVID-19 (MALDEN *et al.*, 2022) by analysing symptoms such as fever, cough, headache, fatigue, dyspnea, and others in 359,938 patients with laboratory tests positive for SARS-CoV-2. Others performed text classification based on radiology or CT scan reports (QOMARIYAH *et al.*, 2022; LÓPEZ-ÚBEDA *et al.*, 2020) to classify COVID-19 and non-COVID-19 patients.

The use of machine learning in COVID-19 detection with X-ray imaging has been explored in scientific research. Many papers have proposed using algorithms, such as traditional machine learning, Convolutional Neural Networks (CNNs), and transfer learning, to analyse X-rays to detect the disease. These studies have shown promising results, demonstrating the potential of machine learning in aiding clinicians in their screening process and improving the speed and accuracy of COVID-19 diagnosis.

Some works have proposed different methods to detect COVID-19 using X-ray images. For example, while Ohata *et al.* (OHATA *et al.*, 2021) and Basha *et al.* (BASHA *et al.*, 2022) used machine learning methods for feature extraction and classification, Hu *et al.* (HU *et al.*, 2022) employed transfer learning and pre-trained models. Despite the promising results obtained by these studies, some limitations could still be addressed. For example, the studies employed relatively small datasets, which may limit their generalizability. Nonetheless, all three papers are limited by the resolution of the X-ray images, which can affect detection accuracy.

Machine learning has also been used to detect COVID-19 in CT images. This approach is considered more sensitive than traditional methods such as X-rays and PCR, as CT scans provide high-resolution images more suited to analysis using machine learning algorithms (ZHAO *et al.*, 2020a; NG *et al.*, 2020). Furthermore, using machine learning in CT images also aids human interpretation, which can be prone to errors and subjectivity. Hence,

combining machine learning to assist in COVID-19 diagnosis with CT images is a promising development in the fight against the pandemic.

Overall, previous papers demonstrated the potential of deep learning models for detecting and classifying COVID-19 using CT scans. They used different architectures, pre-processing techniques, and datasets to achieve their results, showing promising results in distinguishing COVID-19 from healthy or CAP patients. Some developed new architectures, such as AH-Net, ReCOV-101, and COVNet (HARMON *et al.*, 2020; ROHILA *et al.*, 2021; LI *et al.*, 2020), while others used transfer learning techniques (HASAN *et al.*, 2021; ABDEL-BASSET *et al.*, 2021). However, these studies also had limitations, as they only classified CT scans, or even single slices, in classes such as normal and COVID-19; normal, COVID-19, and CAP; COVID-19 and non-COVID-19; and normal and COVID-19 severity. They lacked the usage of an external validation dataset. In addition, some used explainability algorithms to interpret the classifications made by the models (HARMON *et al.*, 2020; LI *et al.*, 2020), which are still unreliable according to medical doctors (NAUDÉ, 2020; LI *et al.*, 2023), and none returned quantitative values.

Several papers proposed deep learning techniques to segment and classify COVID-19 pneumonia lesions in CT scans. Zhang et al. (ZHANG *et al.*, 2020) adapted 3D ResNet-18 to segment lesions. Amyar et al. (AMYAR *et al.*, 2020) developed a Multi-Task Learning (MTL) architecture based on COVID-19 classification, lesion segmentation, and image reconstruction. Qiblawey *et al.* (QIBLAWEY *et al.*, 2021) used encoder–decoder CNNs, UNet, and FPN to segment the lungs and COVID-19 lesions, achieving high COVID-19 detection performance. Wang *et al.* (WANG *et al.*, 2020b) proposed a noise-robust Dice loss function and a self-ensembling framework for COVID-19 lesion segmentation. Finally, Zhou *et al.* (ZHOU *et al.*, 2020) used a CT scan simulator for COVID-19 and a deep learning algorithm to segment and quantify the infection regions.

These works mainly segmented lesions and classified exams as COVID-19 or normal, providing more quantitative results than classification models and explainability algorithms. However, if CAP exams were provided to their models, these exams were wrongfully classified as COVID-19 (QIBLAWEY *et al.*, 2021; WANG *et al.*, 2020b; ZHOU *et al.*, 2020). Amyar *et al.* conducted lesion segmentation and classification, but did not validate their methods on an external dataset (AMYAR *et al.*, 2020).

Zhang et al. (ZHANG *et al.*, 2020) segmented the whole CT scan for both lungs and lesions and then forwarded the full CT scan to a 3D ResNet, a 3D network that takes longer to

train and evaluate than our 2D approach, which selects one slice from the full CT scan to classify. Moreover, they did not make their full dataset publicly available, which makes validation and comparison difficult.

Diagnosis involves identifying a disease or condition based on signs, symptoms, and diagnostic tests, while prognosis involves predicting the likely course of a disease or condition and its possible outcomes. In various medical applications, deep learning models have been used in diagnosis and prognosis tasks. Some works focused on developing deep learning models for accurate disease diagnosis (OHATA *et al.*, 2021; BASHA *et al.*, 2022; HU *et al.*, 2022; HARMON *et al.*, 2020; LI *et al.*, 2020; AMYAR *et al.*, 2020; WANG *et al.*, 2020b). In contrast, other works focused on predicting the prognosis of a disease (ROHILA *et al.*, 2021; ZHANG *et al.*, 2020; QIBLAWEY *et al.*, 2021; ZHOU *et al.*, 2020), such as estimating the likelihood of survival or disease progression. Finally, some papers combined both diagnosis and prognosis tasks. Our work aims to do both, i.e., diagnosing the disease as COVID-19 or CAP, and if the disease is classified as COVID-19, giving the prognosis of the severity of the disease. While deep learning models have shown promise in both diagnosis and prognosis tasks, it is essential to recognise the limitations of these models and use them in conjunction with other clinical information and expertise (ROBERTS *et al.*, 2021; DRIGGS *et al.*, 2021).

A brief comparison between CT scan-related papers and this work can be found in Table 1.

Tabela 1 – Related works summary.

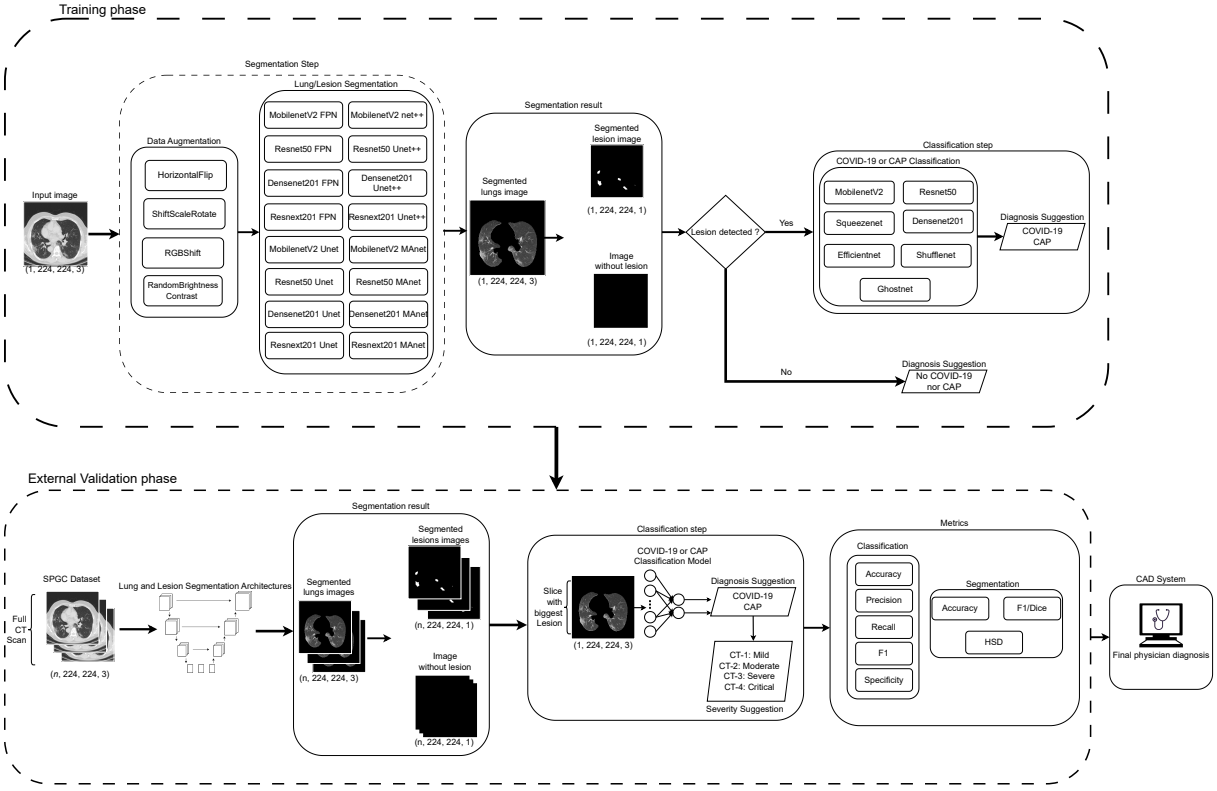
Work	Segmentation	Classification	Datasets	External validation	Metrics
(HARMON *et al.*, 2020)	AH-Net	Densenet121-based	private	**X**	Accuracy, Sensitivity and Specificity
(ROHILA *et al.*, 2021)	Threshhold, Region growing	ResNet50, ResNet101, DenseNet169, DenseNet201	MosMedData	**X**	Accuracy
(LI *et al.*, 2020)	U-net	Resnet50-based	private	**X**	Sensibility, Specificity and AUC
(HASAN *et al.*, 2021)	Threshhold, Morphological operations	3D CNN-based proposed	MosMedData	**X**	AUC
(ABDEL-BASSET *et al.*, 2021)	U-net-based	EfficientNet-B7-based	COVID-CT-MD	**X**	Accuracy, DSC (F1), Jaccard index, AUC
(ZHANG *et al.*, 2020)	U-net, DRUNET, FCN, SegNet	3D ResNet-18-based			Accuracy, AUROC
(AMYAR *et al.*, 2020)	Encoder-Decoder-based	Alexnet, VGG-16, VGG-19, ResNet50, DenseNet169, InceptionV3, Inception-ResNet v2 and Efficient-Net	COVID-CT-MD, HBCC, MedSeg	**X**	Accuracy, DSC, Sensibility, Specificity, AUC
(QIBLAWEY *et al.*, 2021)	ED-CNNs, UNet, FPN	Lesion segmentation Threshold	COVID-CT, CTDATA (Kaggle), MosMedData	[scale=0.4](0,.35) – (.25,0) – (1,.7) – (.25,.15) – cycle;	DSC, IoU, Sensitivity, Specificity
(WANG *et al.*, 2020b)	Encoder-decoder-based	**X**	private	**X**	DSC, RVE, HD95
(ZHOU *et al.*, 2020)	proposed	**X**	Harbin, private	**X**	DSC, Recall
This work	Resnext101 Unet++ and MobilenetV2 Unet	Densenet201	Coronacases, Kaggle, Medical Segmentation, MosMedData, COVIDxCT, SPGC	[scale=0.4](0,.35) – (.25,0) – (1,.7) – (.25,.15) – cycle;	Accuracy, F1 (DSC), HD, Precision, Recall, Specificity

Source: author (2023).

3 MATERIALS AND METHODS

This chapter provides a description of the workflow for this work. We first merged public datasets from the literature to train Lung and Lesion segmentation models from different distributions. Then we trained classification CNN models on a subset of the COVIDxCT dataset, containing only COVID-19 and CAP classes. Finally, if the exam is classified as COVID-19, we quantified the lesions and evaluated the severity of the exam using MosMedData. We applied our system to the SPGC dataset, dividing it into Normal and Lesion exams and secondly into Covid-19 and Common Acquired Pneumonia Lesions. The workflow flowchart can be seen in Figure 12.

Figure 12 – Fluxogram of proposed system employed in this work. We first train models for lung and lesion segmentation and for COVID-19 or CAP classification. Then, we externally validate our models.



Source: Author (2023)

As CT scans can have n different numbers of slices, we apply our segmentation models to all slices. First, the exam is classified as Normal if no lesion is detected on the slices. Next, the exam is classified as with Lesion if lesions are detected. Then, the slice with the biggest lesion area is used to classify the whole exam as COVID-19 or CAP.

3.1 Datasets

We employed a combination of three public datasets for the Lung Segmentation task, resulting in a total of 3677 images from Chest CT scans and their corresponding lung masks (JUN *et al.*, 2020; SEGMENTATION, 2020; KAGGLE, 2017). For the Lesion Segmentation task, we utilized a combination of four public datasets, yielding 6493 images from Chest CT scans and their lesion masks (JUN *et al.*, 2020; SEGMENTATION, 2020; KAGGLE, 2017; MOROZOV *et al.*, 2020). Both tasks employed 10-fold cross-validation, with an 80% split for training and a 20% split for testing, and 10% of the training data was allocated for validation. To ensure consistency, we transformed all images from Digital Imaging and Communications in Medicine (DICOM) or Neuroimaging Informatics Technology Initiative (NIFTI) format to Portable Network Graphics (PNG) in the Hounsfield Unit range of 0-255 using a window of -500 and a width of 750.

For the Classification task, we utilized the COVIDxCT dataset, which included 294,552 images from COVID-19 positive cases and 62,966 images from Common Acquired Pneumonia cases for training, 8147 and 8008, respectively, for validation, and 7965 and 7894 for testing. As the COVIDxCT dataset already provided a set train/validation/test split, we did not use k-fold cross-validation to allow comparison with benchmarks (GUNRAJ *et al.*, 2022).

MosMedData provided 50 COVID-19-positive CT scans with lesions segmentation golden standard. We randomly selected 50 COVID-19 negative exams to add 100 MosMedData exams to our training set. Then, we used the remaining 1010 exams to validate our lesion quantification and disease severity step. MosMedData has an average of 42 slices per exam. COVID-19 scans are divided into four classes: CT -1 to CT-4, with increasing severity, and CT-0, the COVID-19 negative class. Samples are distributed as CT-0 –254, CT-1 –684, CT-2 – 125, CT-3 – 45, CT-4 – 2 (MOROZOV *et al.*, 2020).

Finally, we employed the SPGC dataset for external validation, which included 307 full CT scans, 76 from normal patients, 60 from Common Acquired Pneumonia, and 171 from COVID-19. Each exam has an average of 150 slices. (AFSHAR *et al.*, 2021). Table 2 summarizes all datasets used in this work and the task they are used for.

Tabela 2 – Datasets used for each task.

Database	Task	# COVID-19 Exams	# CAP Exams	# No COVID-19 nor CAP Exams	# Total Images
Coronacases	Lung Segmentation	10	0	0	2581
Kaggle	Lung/Lesion Segmentation	0	0	n/a	267
Medical Seg.	Lung Segmentation	9	0	0	829
Coronacases	Lesion Segmentation	10	0	0	2156
Medical Seg	Lesion Segmentation	9	0	0	713
Mosmed Seg	Lesion Segmentation	50	0	50	3357
Mosmed Seg	Validation	806	0	50	42,224
COVIDxCT	Image Classification	3731	932	0	353,536
SPGC	External Validation	171	60	71	46,024

Source: author (2023).

3.2 Data Augmentation

To expand the generalization capabilities of our models and produce more images with lesions, we used Data Augmentation methods on our training sets such as: randomly flipping the image horizontally; randomly translating, scaling, and rotating the image; randomly shifting values for each channel of the input RGB image; and randomly change brightness and contrast of the image (BUSLAEV *et al.*, 2020). Table 3 shows a summary of the techniques and parameters.

Tabela 3 – Data Augmentation techniques and parameters.

Method	Task	Parameters
HorizontalFlip	Lung/Lesion Segmentation	p=0.5
ShiftScaleRotate	Lung/Lesion Segmentation	shift limit=0.05, scale limit=0.1, rotate limit=15, p=0.5
RGBShift	Lung Segmentation	r shift limit=25, g shift limit=25, b shift limit=25, p=0.5
RandomBrightnessContrast	Lung/Lesion Segmentation	brightness limit=0.3, contrast limit=0.3, p=0.5

Source: author (2023).

3.3 Grid Search

As lesion segmentation is more complex than lung segmentation, we used Grid Search for 200 runs to optimize our hyperparameters and obtain better results for each architecture

(BIEWALD, 2020). Table 4 summarizes optimized hyperparameters and their parameters.

Tabela 4 – Grid Search Parameters.

Hyperparameters	Task	Parameters
Batch Size	Lesion Segmentation	[8, 16, 32, 64]
Epochs	Lesion Segmentation	[25,50,75]
Learning Rate	Lesion Segmentation	[0.001, 0.0001, 0.00001]
Encoder	Lesion Segmentation	[mobilenet, resnet50, densenet201, resnext101]
Decoder	Lesion Segmentation	[FPN, Unet, Unet++, MAnet]
Patience	Lesion Segmentation	[5, 10, 15]
Loss	Lesion Segmentation	[Lovasz, Dice, Tversky]
Tversky Beta	Lesion Segmentation	[0.3, 0.4, 0.6, 0.7, 0.8, 0.9]
Optimizer	Lesion Segmentation	[Adam, RMSprop]

Source: author (2023).

3.4 Segmentation models

We utilized well-known state-of-the-art and novel encoders and decoders to analyse various structures for lung, and lesion segmentation (IAKUBOVSKII, 2019). We tested sixteen combinations of encoders and decoders, combining methods with different sizes and techniques, as displayed in Table 5. The encoders utilized in this evaluation were MobilenetV2, Resnet50, Densenet201, and Resnext101 (SANDLER *et al.*, 2018b; HE *et al.*, 2016b; HUANG *et al.*, 2017b; XIE *et al.*, 2017). The decoders used were Unet, FPN, Unet++, and MAnet (RONNEBERGER *et al.*, 2015; LIN *et al.*, 2017b; ZHOU *et al.*, 2020; FAN *et al.*, 2020b).

The chosen loss function for lung segmentation was Lovasz, and the learning rate was set to 0.001 with Adam optimization. The batch size was 64, and the maximum number of epochs was 50. For lesion segmentation, we optimized each hyperparameter for the F1-Score metric with the grid search, presented the values in Table 5.

Tversky loss is a loss function that is commonly used in machine learning for binary classification problems where the classes may not be balanced, such as our lesion segmentation task. It is a generalization of the Dice loss. It is the only loss from the selected ones with the possibility of defining a β value choice to tune a desired trade-off between false positives and false negatives (FAWCETT, 2006).

Tabela 5 – Lesion Segmentation Optimized Hyperparameters.

Architecture	Batch size	Epochs	Loss	Beta	LR	Optimizer	Patience
MobilenetV2 FPN	16	75	Dice	n/a	0.0001	Adam	15
Resnet50 FPN	64	50	Dice	n/a	0.0001	RMSprop	15
Densenet201 FPN	16	75	Dice	n/a	0.0001	RMSprop	15
Resnext101 FPN	64	75	Tversky	0.9	0.001	Adam	10
MobilenetV2 Unet	64	50	Tversky	0.3	0.001	Adam	15
Resnet50 Unet	64	50	Tversky	0.7	0.0001	RMSprop	10
Densenet201 Unet	32	75	Tversky	0.3	0.00001	Adam	15
Resnext101 Unet	64	75	Lovasz	n/a	0.0001	Adam	10
MobilenetV2 Unet++	32	50	Dice	n/a	0.0001	Adam	15
Resnet50 Unet++	32	75	Lovasz	n/a	0.00001	RMSprop	10
Densenet201 Unet++	32	75	Tversky	0.3	0.00001	Adam	15
Resnext101 Unet++	16	50	Lovasz	n/a	0.0001	Adam	15
MobilenetV2 MAnet	32	25	Tversky	0.7	0.0001	RMSprop	5
Resnet50 MAnet	64	50	Lovasz	n/a	0.00001	RMSprop	5
Densenet201 MAnet	32	75	Dice	n/a	0.00001	RMSprop	15
Resnext101 MAnet	16	75	Lovasz	n/a	0.0001	Adam	15

Source: author (2023).

3.5 Lesion quantification

MosMedData does not provide a quantified computed approach for pulmonary commitment analysis; expert physicians qualitatively evaluate COVID-19 severity. First, we calculate the area of the left and right lungs and lesions to approximate the disease severity analysis. Then, we divided the area of the lesions by the area of the lung they are inside to obtain a percentage of parenchymal involvement for each lung. Lungs with lesion/lung percentage of ≤ 25 are categorized as CT-1. Percentages between > 25 and ≤ 50 are categorized as CT-2. Percentages between > 50 and ≤ 75 are categorized as CT-2. And, percentages ≥ 75 are categorized as CT-4.

3.6 Classification models

We tested eight state-of-the-art models, including MobilenetV2, Resnet50, Densenet201, Resnext101, SqueezeNet, EfficientNet, ShuffleNet, and GhostNet, all pre-trained on ImageNet. The loss function utilized was Cross Entropy, the learning rate was set at 0.0001, and Adam optimization was used. The batch size was 64, the maximum number of training epochs was 20, and patience of 5 epochs.

3.7 Evaluation Metrics

Segmentation models were evaluated using Accuracy, F1-Score (DSC score), Hausdorff Distance (HD), and training and testing time. Classification models were evaluated using Accuracy, F1-Score, Precision, Recall, Specificity, and Confusion Matrix.

For the segmentation tasks, True Positive (TP) refers to correctly segmented lesion pixels, True Negative (TN) to correctly segmented background pixels, False Positive (FP) to background pixels wrongfully classified as lesion pixels, and False Negative (FN) to lesion pixels wrongfully classified as background pixels.

For the classification tasks, TP refers to correctly classified exams with lesions, TN to correctly classified exams without lesions, FP to exams without lesions wrongfully classified as with lesions, and FN to exams with lesions wrongfully classified as without lesions.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.1}$$

A higher accuracy mainly indicates better performance. However, accuracy is not always the best metric to evaluate a model, mainly because we are dealing with unbalanced data, and misclassifications have different consequences. For example, it is worse to classify a COVID-19 exam as a non-COVID-19 exam than the other way around.

$$F1 = DSC = \frac{2TP}{2TP+FP+FN} \tag{3.2}$$

F1-score, on the other hand, is a metric that considers both Precision, where a high precision indicates that the model is accurately identifying positive cases:

$$P = \frac{TP}{TP+FP} \tag{3.3}$$

and Recall, where a high Recall indicates that the model is accurately identifying most positive cases, even if it also misclassifies some negative cases as positives:

$$R = \frac{TP}{TP+FN} \tag{3.4}$$

F1-score is the harmonic mean of Precision and Recall, and provides a balance between the two metrics. In cases where the data is imbalanced, F1-score can provide a more informative evaluation of the model's performance because it penalizes models that only predict the majority class. Therefore, as our goal is to identify COVID-19-positive cases with high precision, F1-score may be a more appropriate metric than accuracy.

Specificity refers to the ability of a model to correctly identify the negative cases, i.e., those that do not have COVID-19. A high specificity indicates that the model can accurately identify people who do not have the virus, which is essential to avoid false positives:

$$S = \frac{TN}{TN+FP} \tag{3.5}$$

It is important to note that a model with low specificity, but high recall, identifies many true positive cases but also has many false positives. Finally, segmentation models were also evaluated by the Hausdorff Distance (HD):

$$d(\mathscr{X},\mathscr{Y}) = \sup\left\{\sup_{x\in\mathscr{X}}\inf_{y\in\mathscr{Y}} d(x,y), \sup_{y\in\mathscr{Y}}\inf_{x\in\mathscr{X}} d(x,y)\right\}. \tag{3.6}$$

The Hausdorff Distance is a metric on the space of compact, non-empty sets. The Hausdorff metric between two sets, X and Y, is defined as the maximum of two values: the Hausdorff distance from X to Y and the Hausdorff distance from Y to X. The Hausdorff metric is commonly used in computer vision, image processing, and pattern recognition. It compares the similarity of shapes, images, or other data types. For this work, X and Y are the segmented images returned by our architectures and the ground truth images, respectively.

3.8 Statistical Tests

We used boxplots for visualizing and comparing the distributions of numerical data. They provide a quick summary of the data's central tendency, spread, and skewness and can be particularly useful for identifying outliers and skewness in the data.

To better understand the significance of our results and meaningfully analyse the best models, we used the following steps (Figure 13) for statistical analysis:

1. All columns are checked with the Shapiro-Wilk test for normality;
2. If all columns are normal, we use Bartlett's test for homogeneity, otherwise we use Levene's test;
3. If all populations are normal and homoscedastic, we use repeated measures ANOVA with Tukey's HSD as post-hoc test;
4. If at least one population is not normal or the populations are heteroscedastic, we use Friedman's test with the Nemenyi post-hoc test.

We used the Shapiro-Wilk test to test the normality assumption (SHAPIRO; WILK, 1965). Then, we applied the Bartlett or Levene test, depending on Shapiro-Wilk's results.

Bartlett's test is a homogeneity test of variances to determine if the variances of the metrics of the architectures are equal. It tests the null hypothesis that the variances of all groups are similar (BARTLETT; FOWLER, 1937).

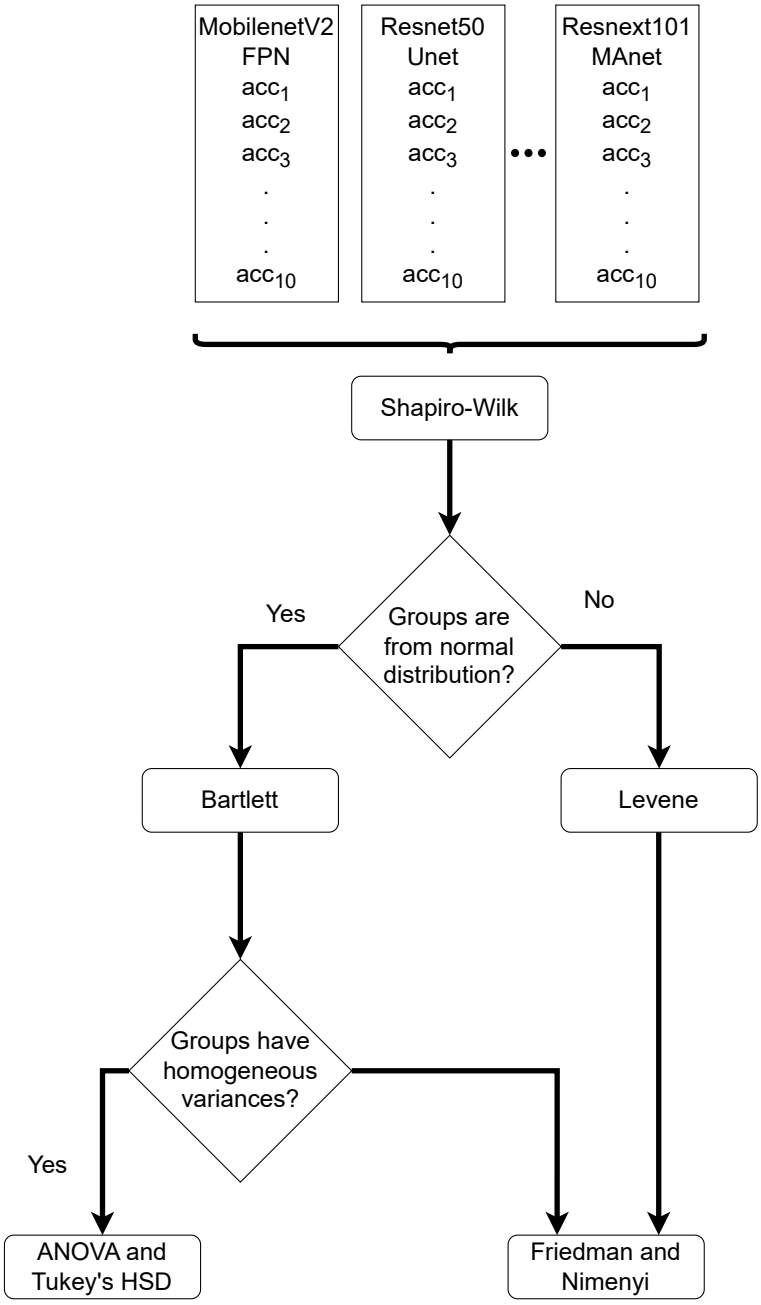
Levene test assesses the assumption of equal variances before conducting a test to compare the means of the metrics of the architectures. It provides a way to determine if the variances of the groups (each group is a 10-fold result for a metric) are equal, which is an essential assumption for the ANOVA test (BROWN; FORSYTHE, 1974).

We conducted the repeated measures ANOVA test to determine if there was a significant difference in means between the metrics of the architectures. The repeated measures ANOVA test calculates a statistic and provides a p-value, which can be used to determine if the differences between the group means are significant (GREENHOUSE; GEISSER, 1959).

We performed the Friedman non-parametric test to determine if there is a significant difference between the metrics of the architectures. It tests the null hypothesis that the population medians of all groups are equal. The Nemenyi posthoc test is a multiple comparison test that we used to identify which groups are significantly different from each other after a significant result from the Friedman test. Every group pair has its difference between the medians calculated, and if this difference is bigger than a critical distance (CD), the groups are significantly different. The CD is a threshold that helps determine which group comparisons are statistically significant, and it is calculated based on the number of groups being compared and the overall significance level chosen ($\alpha = 0.05$ in our case) (FRIEDMAN, 1937; NEMENYI, 1963).

The Tukey HSD test is a multiple comparison test used to compare all possible pairs

Figure 13 – Fluxogram of the statistical tests used to understand the significance of our results. These steps are repeated for Accuracy, F1-Score, and HD. Each column is the 10-fold metric output for each segmentation metric.



Source: Author (2023)

of means in the set of metrics of the architectures. We used the Tukey HSD test to identify which specific pairs of metrics are significantly different from each other, considering the multiple comparisons. We utilized the Bartlett test to assess the assumption of equal variances before conducting the Tukey HSD test. If the Bartlett test shows that the variances are equal, then the Tukey HSD test can be used to compare the means of the metrics of the architectures (TUKEY, 1949).

3.9 Development Environment

For the development of this work, we utilized several cutting-edge tools and technologies to ensure the best possible outcome. We employed open-source libraries such as PyTorch, Pytorch Lightning, Segmentation Models Pytorch (SMP), Autorank, and WandB. Our hardware setup included an NVIDIA GeForce RTX 3060 12 GB graphics card and a 12th Gen Intel Core i7-12700KF x 20 processor, along with 64 GB of memory.

4 RESULTS AND DISCUSSION

This section presents the results concerning the methodology employed for lung and lesion segmentation, COVID-19 and CAP classification in CT exams. We compared state-of-the-art models through Accuracy, Precision, Recall, F1-Score, Specificity, Hausdorff Distance, and processing time.

4.1 Lung segmentation

The first task was to segment the lungs from the background on raw CT slices to remove artefacts for COVID-19 and CAP detection. We summarize the results for this step in Table 6.

Tabela 6 – Lung Segmentation Results.

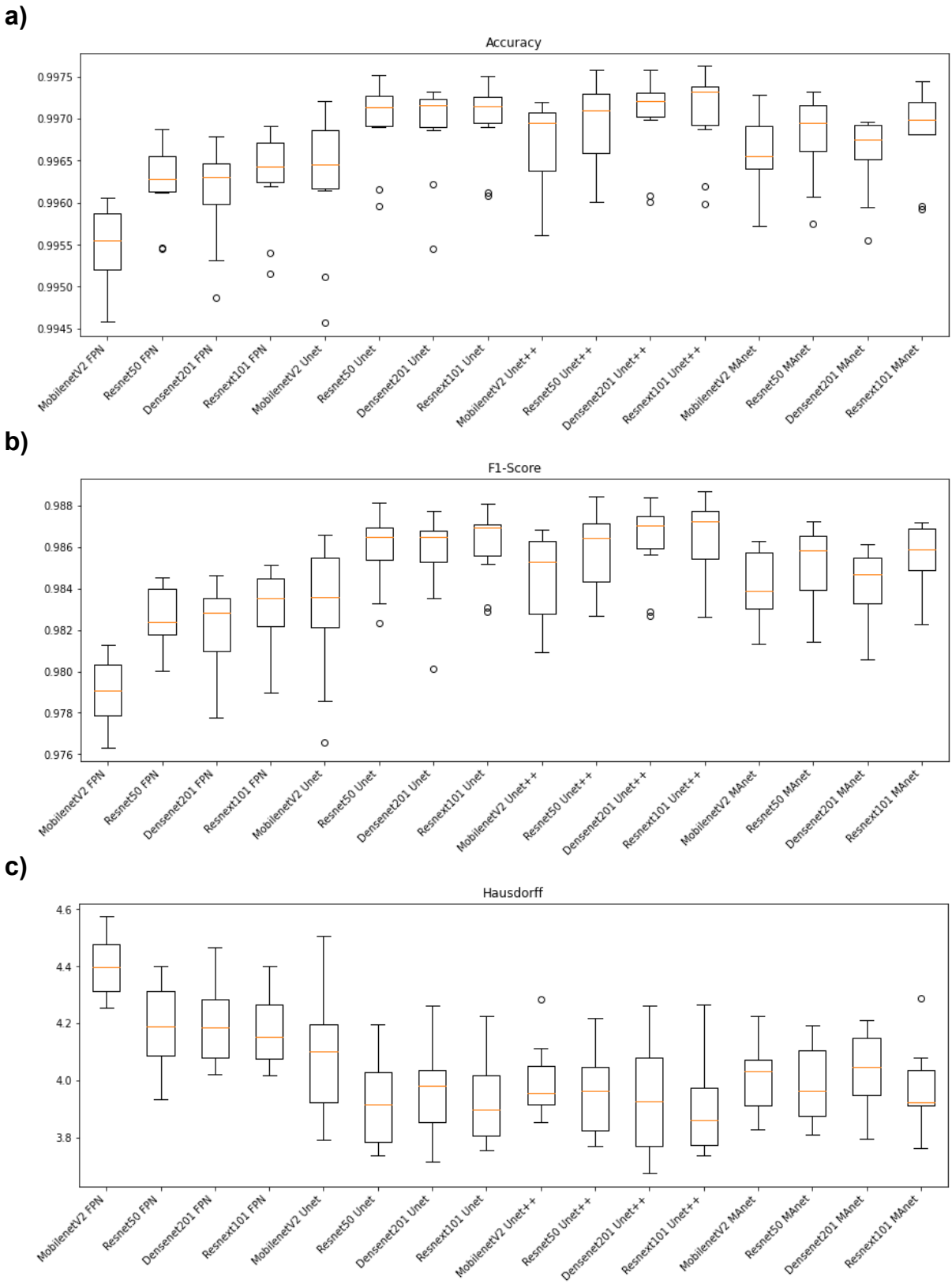
Architecture	Acc (%)	F1 (DSC) (%)	HD
MobilenetV2 FPN	99.55 ± 0.05	97.9 ± 0.16	4.4 ± 0.1
Resnet50 FPN	99.62 ± 0.04	98.25 ± 0.15	4.19 ± 0.15
Densenet201 FPN	99.61 ± 0.06	98.21 ± 0.2	4.2 ± 0.14
Resnext101 FPN	99.63 ± 0.06	98.29 ± 0.2	4.18 ± 0.12
MobilenetV2 Unet	99.63 ± 0.08	98.29 ± 0.31	4.1 ± 0.21
Resnet50 Unet	99.7 ± 0.05	98.59 ± 0.17	3.92 ± 0.15
Densenet201 Unet	99.69 ± 0.06	98.56 ± 0.21	3.96 ± 0.17
Resnext101 Unet	99.7 ± 0.05	98.61 ± 0.17	3.92 ± 0.15
MobilenetV2 Unet++	99.67 ± 0.05	98.46 ± 0.2	4.0 ± 0.12
Resnet50 Unet++	99.69 ± 0.05	98.58 ± 0.18	3.95 ± 0.14
Densenet201 Unet++	99.7 ± 0.05	98.63 ± 0.19	3.93 ± 0.18
Resnext101 Unet++	99.71 ± 0.05	98.64 ± 0.19	3.9 ± 0.16
MobilenetV2 MAnet	99.66 ± 0.05	98.41 ± 0.17	4.01 ± 0.12
Resnet50 MAnet	99.68 ± 0.05	98.51 ± 0.18	3.99 ± 0.14
Densenet201 MAnet	99.66 ± 0.05	98.42 ± 0.17	4.03 ± 0.13
Resnext101 MAnet	99.69 ± 0.05	98.54 ± 0.17	3.96 ± 0.14

Source: author (2023).

In general, all architectures presented state-of-the-art results regarding Accuracy, F1-Score (DSC), and Hausdorff Distance. Resnext101 Unet++ outperforms the other architectures for all metrics, achieving 99.71 ± 0.05%, 98.64 ± 0.19%, and 3.9 ± 0.16 for Accuracy, F1-Score (DSC), and Hausdorff Distance, respectively. However, all architectures presented a similar performance for the three metrics. In the following sections, we analyse the significance of our results through statistical tests, aiming to confirm their relevance.

Figure 14 illustrates the segmentation metrics boxplots applied for lung segmentation: Accuracy, F1-Score (DSC), and Hausdorff Distance.

Figure 14 – Boxplots of segmentation metrics applied in this work. a) Accuracy, b) F1-Score (DSC) and c) Hausdorff Distance.



Source: Author (2023)

Concerning the accuracy metric, we can see by the y-axis that all algorithms perform similarly, as accuracy variates from 0.994 to 0.997. First, however, we remark on some important aspects when comparing our segmentation architectures. For instance, Resnet50 Unet, Densenet201 Unet, Resnext101 Unet, Densenet201 Unet++, and Resnext101 Unet++ presented the best accuracy medians (Figure 14.a), lying higher than other algorithm boxes. Moreover, the interquartile ranges of these algorithms are smaller than the others, indicating that the accuracy values are less dispersed with a left-skewed distribution. On the other hand, MobilenetV2 FPN presented the lowest accuracy with more dispersed data and a soft left-skewed distribution. The remaining algorithms presented competitive accuracies but with dispersed and skewed values. In addition, only MobilenetV2 FPN, MobilenetV2 Unet++, Resnet50 Unet++, and MobilenetV2 MAnet have no outliers.

In general, F1-Score behaviour is similar. For example, Resnet50 Unet, Densenet201 Unet, Resnext101 Unet, Densenet201 Unet++, and Resnext101 Unet++ again presented the best median values (Figure 14.b), with the left-skewed distribution. However, Resnet101 Unet++ had a more dispersed data distribution.

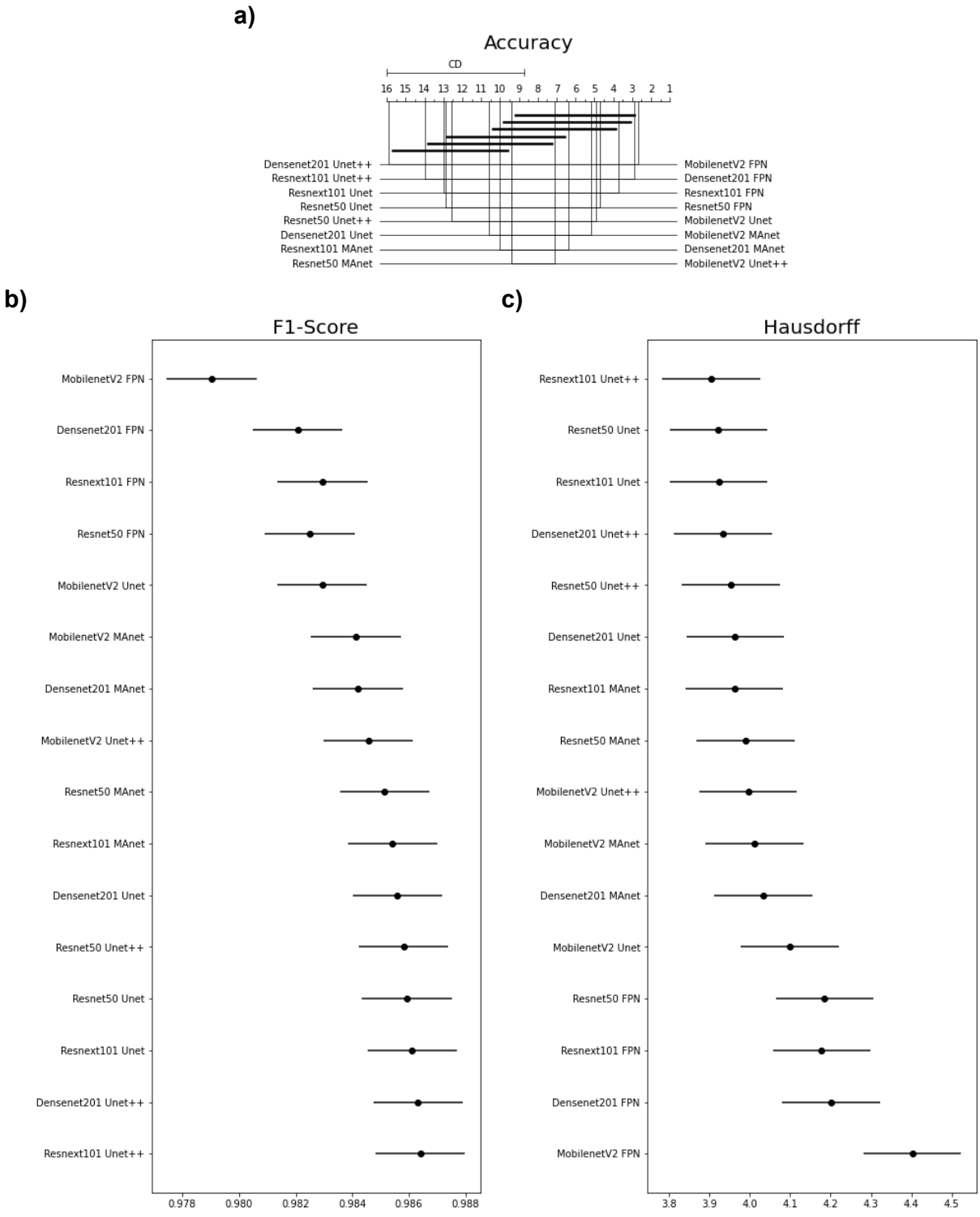
The architectures had more dispersed data for the Hausdorff metric (Figure 14.c). For example, Resnext101 Unet++ had the lowest median, with a right-skewed distribution, and MobilenetV2 FPN presented the highest Hausdorff median.

Because one accuracy population is not normal (Densenet201 Unet), we applied Friedman's test with the Nemenyi post-hoc test to analyse whether the accuracies' distributions differ. We presented the test results in Figure 15.a. Remembering that differences are significant if the distance between the mean ranks is greater than the CD.

We failed to reject the null hypothesis that the population is normal for all F1-Score populations. Therefore, we assume that all F1-Score populations are normal. We applied Bartlett's test for homogeneity and failed to reject the null hypothesis that the data is homoscedastic. Thus, we assume that our data is homoscedastic. Because we have more than two populations and all populations are normal and homoscedastic, we use repeated measures ANOVA as an omnibus test to determine any significant differences between the mean values of the populations. As results from the ANOVA test were significant, we used the post-hoc Tukey HSD test to infer which differences are significant. Populations are significantly different if their confidence intervals are not overlapping in Figure. 15.b.

None of the architectures significantly differed in accuracy, as they had a mean rank

Figure 15 – Statistical tests for our metrics for the lung segmentation task. a) Accuracy, b) F1-Score (DSC), and c) Hausdorff Distance..



Source: Author (2023)

distance smaller than the critical distance for at least one other evaluated architecture (Figure 15.a). Nonetheless, the architecture that had the most different accuracy from the others was MobilenetV2 FPN.

Most confidence maps overlap (Figure 15.b), except for MobilenetV2 FPN, the fastest architecture for training and testing (Figure 16). When selecting an architecture, we can choose MobilenetV2 FPN for a fast architecture with a slight loss in F1-Score. On the other hand, suppose we decide on an architecture with a higher F1-Score. In that case, we can choose any other architecture because F1-Score differences are insignificant. Thus, the best choice would be Resnet50 Unet++, the second-fastest architecture and, as shown by the test, does not significantly differ F1-Score from other slower architectures.

Hausdorff Distance results are generally similar (Figure 15.c). Again, MobilenetV2 FPN had the most significant difference, while other architectures had no significant difference for Hausdorff Distance.

The fastest model for training and testing was MobilenetV2 FPN, and the slowest was Resnext101 Unet++. However, even if the difference for the shortest training time (513.5 seconds) was more than ten times faster than the longest (5304.3 seconds), the fastest test time was of 1.9 seconds, and the slowest was 8.5 for evaluating on 3677 images, or an average of 0.51×10^{-3} and 2.3×10^{-3} seconds per image, respectively. As complexity rose, other models followed a linear training and testing time growth. We present this behaviour in Figure 16.

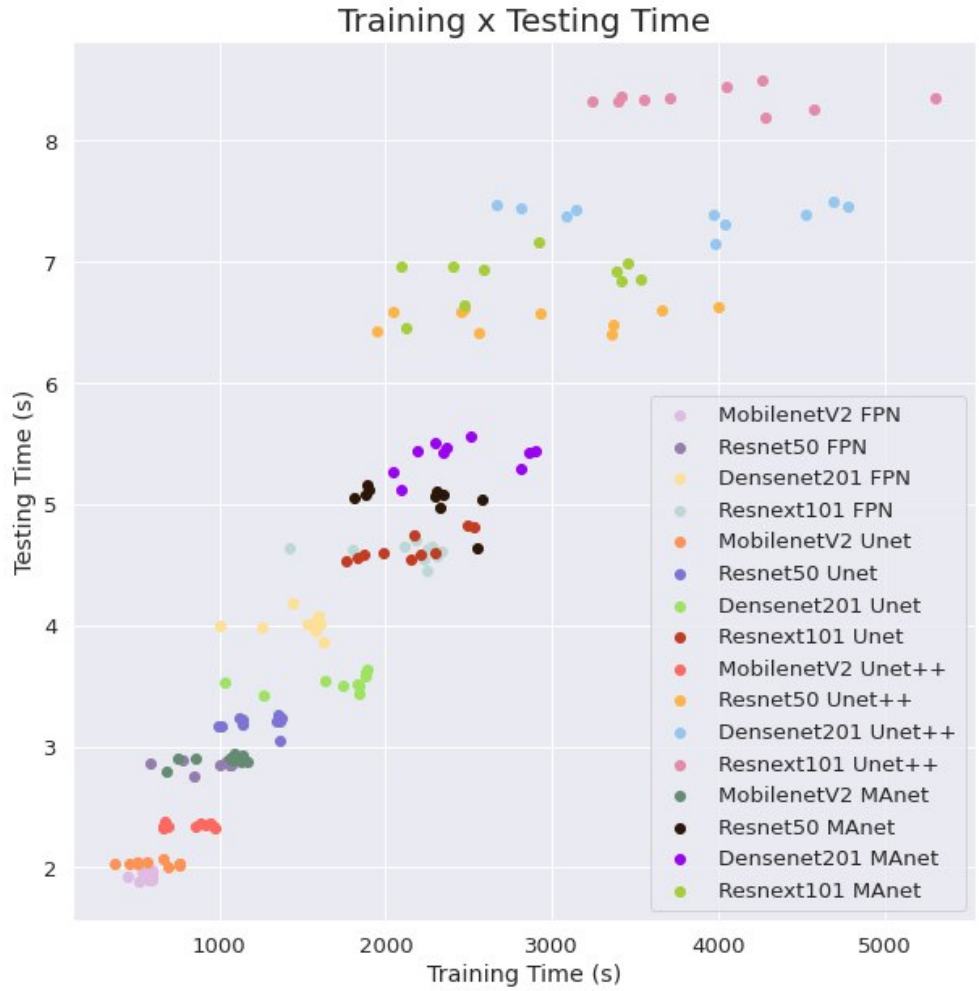
4.2 Lesion Segmentation

The second task was to segment lesions inside the lungs from the previously segmented CT slices for COVID-19 and CAP detection. We summarize the results for this step in Table 7.

All architectures presented excellent results regarding Accuracy, F1-Score (DSC), and Hausdorff Distance. Densenet201 Unet, Resnet50 Unet++, and Resnext101 Unet++ outperform the other architectures for Accuracy, achieving $99.87 \pm 0.01\%$. Densenet201 Unet++ obtains the highest F1-Score (DSC) for all architectures, achieving $85.16 \pm 1.13\%$. However, all architectures presented a similar performance for the three metrics. In the following sections, we analyse the significance of our results through statistical tests, aiming to confirm their relevance. However, MobilenetV2 FPN, the fastest architecture, obtained the smallest HD of 2.86 ± 0.12.

The accuracies are high because most of the Ground Truth image is mainly composed of black pixels, with only a small percentage of the image being white lesions pixels. When we calculate the Accuracy of our models, these black pixels raise all accuracies, reducing the metric credibility.

Figure 16 – Training and Testing time for lung segmentation.



Source: Author (2023)

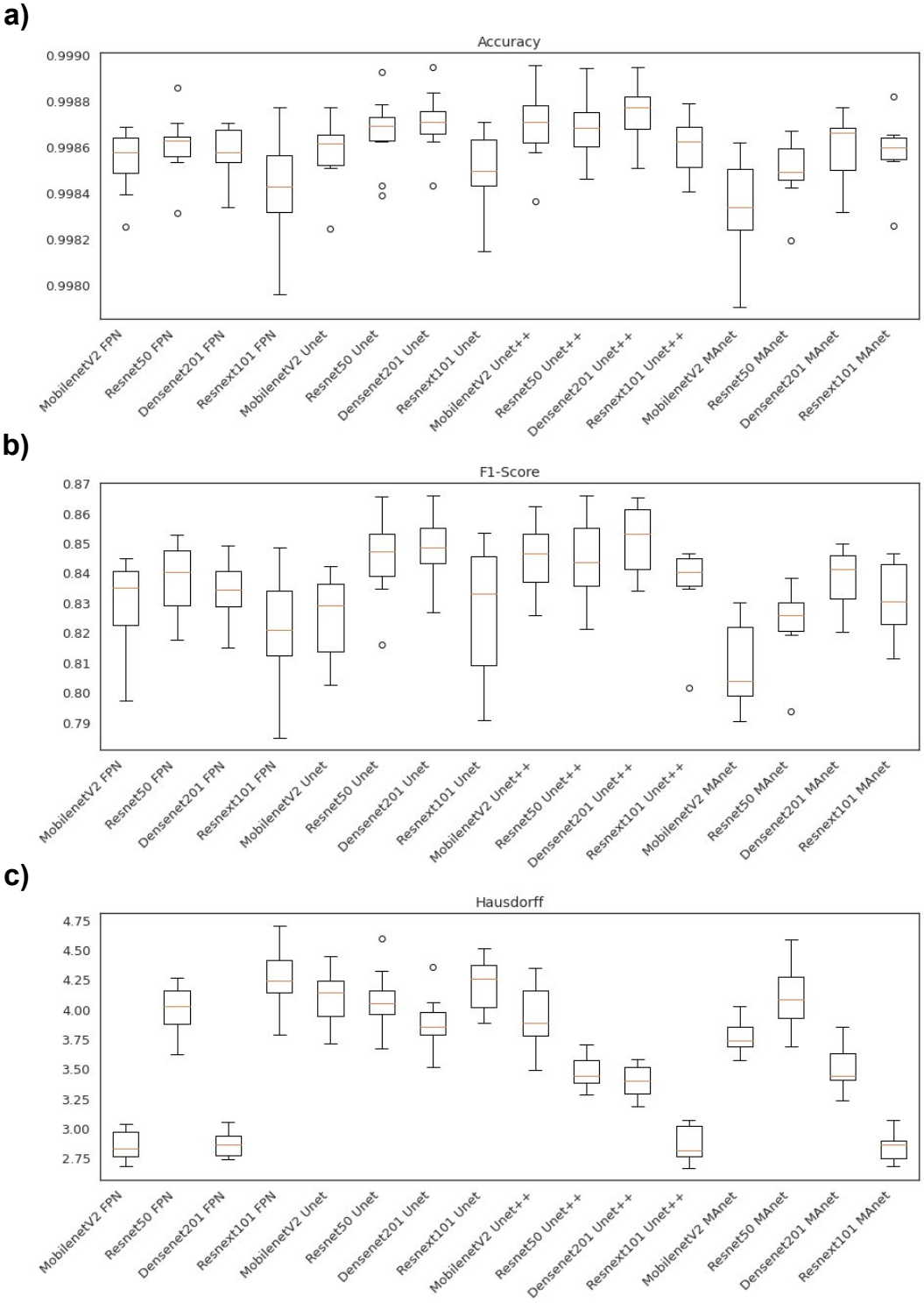
Tabela 7 – Lesion Segmentation Results.

Architecture	**Acc (%)**	**F1 (DSC) (%)**	**HD**
MobilenetV2 FPN	99.85 ± 0.01	82.95 ± 1.45	**2.86 ± 0.12**
Resnet50 FPN	99.86 ± 0.01	83.84 ± 1.13	4.0 ± 0.2
Densenet201 FPN	99.86 ± 0.01	83.47 ± 1.01	2.87 ± 0.1
Resnext101 FPN	99.84 ± 0.02	82.17 ± 1.71	4.28 ± 0.25
MobilenetV2 Unet	99.86 ± 0.01	82.59 ± 1.32	4.1 ± 0.21
Resnet50 Unet	99.87 ± 0.02	84.55 ± 1.32	4.06 ± 0.26
Densenet201 Unet	**99.87 ± 0.01**	84.8 ± 1.1	3.88 ± 0.22
Resnext101 Unet	99.85 ± 0.02	82.75 ± 2.1	4.21 ± 0.2
MobilenetV2 Unet++	99.87 ± 0.02	84.51 ± 1.08	3.95 ± 0.26
Resnet50 Unet++	**99.87 ± 0.01**	84.41 ± 1.32	3.48 ± 0.12
Densenet201 Unet++	**99.87 ± 0.01**	**85.16 ± 1.13**	3.4 ± 0.13
Resnext101 Unet++	99.86 ± 0.01	83.72 ± 1.26	2.87 ± 0.14
MobilenetV2 MAnet	99.83 ± 0.02	80.9 ± 1.34	3.77 ± 0.13
Resnet50 MAnet	99.85 ± 0.01	82.37 ± 1.14	4.11 ± 0.27
Densenet201 MAnet	99.86 ± 0.01	83.81 ± 1.01	3.52 ± 0.19
Resnext101 MAnet	99.86 ± 0.01	83.18 ± 1.16	2.86 ± 0.13

Source: author (2023).

Figure 17 illustrates the segmentation metrics boxplots applied for lesion segmentation: Accuracy, F1-Score (DSC), and Hausdorff Distance.

Figure 17 – Boxplots of segmentation metrics applied in this work for lesion segmentation. a) Accuracy, b) F1-Score(DSC) and c) Hausdorff Distance.



Source: Author (2023)

Concerning the accuracy metric, we can see by the y-axis that all algorithms per-

formed very similarly, as accuracy variates from 0.9980 to 0.9990. First, however, we remark on some important aspects when comparing our segmentation architectures for this metric. For instance, Resnet50 Unet, Densenet201 Unet, Densenet201 Unet++, and Resnext101 Unet++ presented higher accuracy medians (Figure 17.a). Moreover, the interquartile ranges of these algorithms are smaller than the others, indicating that the accuracy values are less dispersed with a left-skewed distribution.

On the other hand, MobilenetV2 MAnet presented the lowest accuracy with more dispersed data and a soft left-skewed distribution. The remaining algorithms presented competitive accuracies but with more dispersed and skewed values. In addition, only Densenet201 FPN, Resnext101 Unet, Resnet50Unet++, DenseNet 201 Unet++, Resnext101 Unet++, and Densenet201 MAnet have no discrepant values.

Concerning F1-Score, the Unet decoders (ResNet 50 Unet, Densenet201 Unet, MobilenetV2 Unet++, Resnet50 Unet++, and Densenet201 Unet++) presented higher median values with lower dispersion (Figure 17.b). On the other hand, Resnext101 Unet had a more dispersed data distribution. Moreover, only Resnet50 Unet, Resnext101 Unet++, and Resnet50 MAnet architectures presented discrepant values.

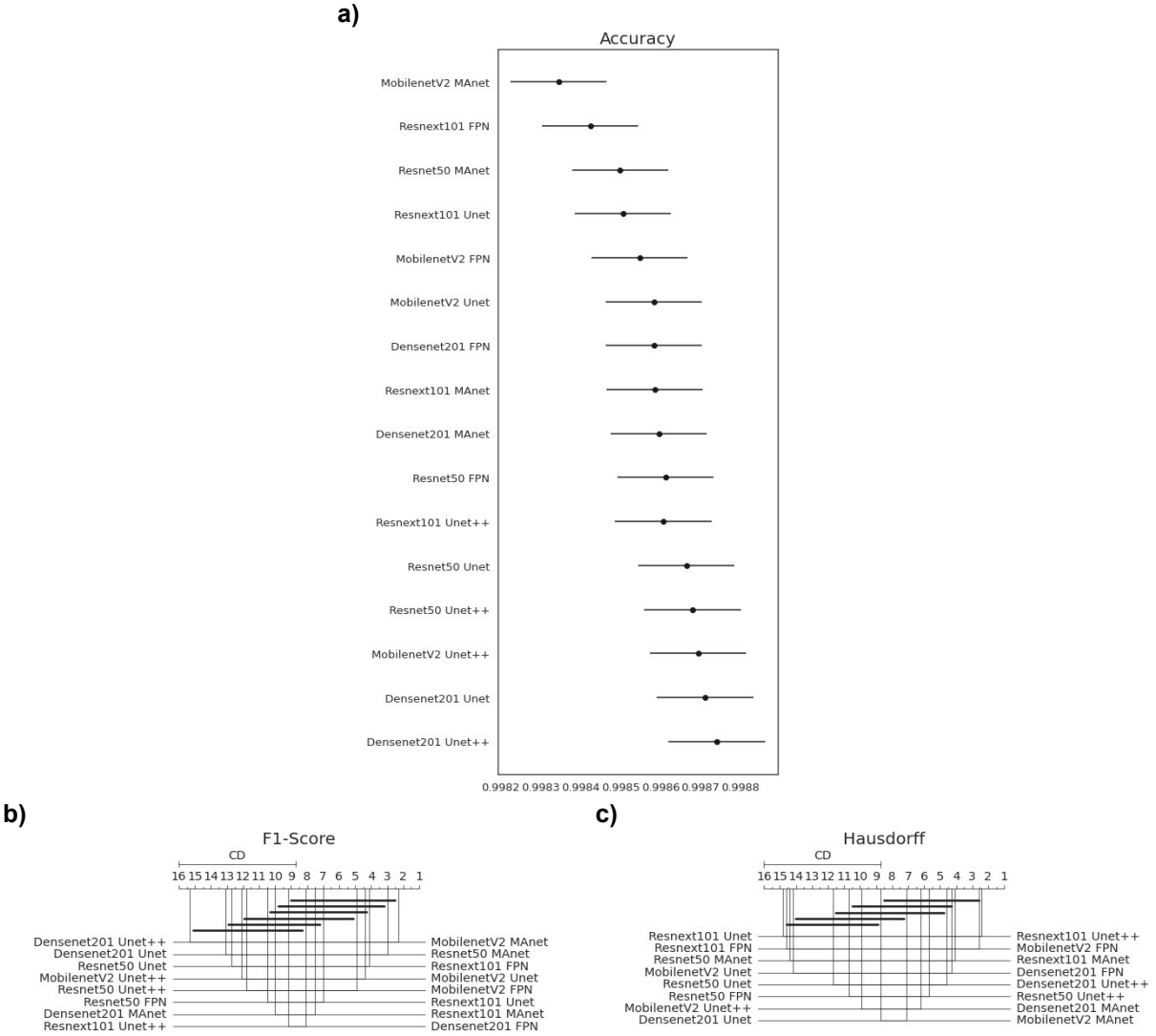
In general, the architectures had less dispersed data for the Hausdorff metric (Figure 17.c). For example, Resnext101 Unet++ had the lowest median, with a right-skewed distribution, and MobilenetV2 FPN, Densenet201 FPN, Resnext101 Unet++, and Resnext101 MAnet presented the lowest Hausdorff median.

We failed to reject the null hypothesis that the population is normal for all Accuracy populations. Therefore, we assume that all Accuracy populations are normal. We applied Bartlett's test for homogeneity and failed to reject the null hypothesis that the data is homoscedastic. Thus, we assume that our data is homoscedastic. Because we have more than two populations and all populations are normal and homoscedastic, we use repeated measures ANOVA as an omnibus test to determine any significant differences between the mean values of the populations. As results from the ANOVA test were significant, we used the post-hoc Tukey HSD test to infer which differences are significant. Populations are significantly different if their confidence intervals are not overlapping in Figure 18.a.

Because one F1-Score and one HD population is not normal (Resnext101 Unet++), we applied Friedman's test with the Nemenyi post-hoc test to analyse if there is a difference between the accuracies' distributions. We presented the test results in Figure 18.b-c. Differences

are significant if the distance between the mean ranks is greater than the critical distance (CD).

Figure 18 – Statistical test results for our metrics for the lesion segmentation task. a) Accuracy, b) F1-Score, and c) Hausdorff Distance.



Source: Author (2023)

Most confidence maps overlap (Figure 18.a), except for MobilenetV2 MAnet, that overlaps mainly with Resnext101 FPN and Resnet50 MAnet. Resnext101 FPN and Resnet50 MAnet have similar results for all metrics and similar training and testing times. However, MobilenetV2 MAnet is faster in training and testing with a small decrease in Accuracy (Figure 19). Thus, when selecting an architecture, we can choose MobilenetV2 MAnet for a fast architecture with a slight loss in Accuracy. On the other hand, suppose we decide on an architecture with a higher Accuracy. In that case, we can choose any other architecture because F1-Score differences are insignificant. Thus, the best choice would be Resnet50 Unet++ again, the second-fastest architecture and, as shown by the test, does not significantly differ F1-Score

from other slower architectures.

None of the architectures significantly differed from the others for F1-Score, as they had a mean rank distance smaller than the critical distance for at least one other evaluated architecture (Figure 18.b). Nonetheless, the architecture that had the most different F1-Score from the others was MobilenetV2 MAnet.

In general, Hausdorff Distance results are similar (Figure 18.c). MobilenetV2 FPN, Resnext101 Unet++, and Resnext101 MAnet had the most significant difference, while other architectures had no significant difference for Hausdorff Distance.

The fastest model for training was MobilenetV2 MAnet, and for testing was MobilenetV2 FPN. However, MobilenetV2 MAnet converged faster, needing only 25 epochs. The slowest for the train were Densenet201 Unet++ and Resnext101 Unet++, and the slowest for the test was Resnext101 Unet++. However, even if the difference for the fastest training time (573.0 seconds) was more than thirty times faster than the slowest (19164.5 seconds), the fastest test time was of 2.8 seconds, and the slowest was 12.6 for evaluating on 6493 images, or an average of 0.43×10^{-3} and 1.9×10^{-3} seconds per image, respectively. As complexity rose, other models followed a linear training and testing time growth. We present this behaviour in Figure 18.

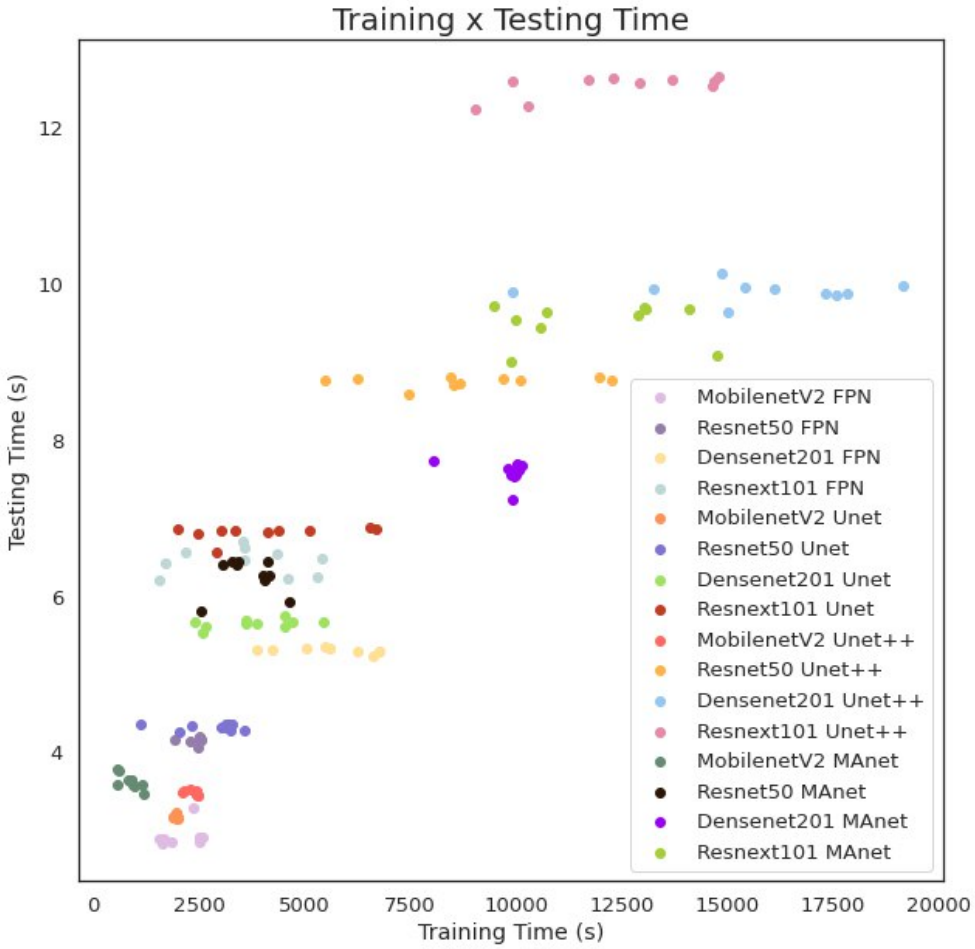
4.3 Lesion detection

We first applied our architectures on the other 1010 full CT scans of MosMedData to validate our system in a 3D scenario to detect and segment all lesions in an exam and then classify the exam as "with lesion"if any lesion is found or "without lesion"otherwise. The results are summarized in Table 8.

All architectures had similar and competitive results for MosMedData. MobileNet Unet had the highest Accuracy, F1-Score, and Recall with 94.36%, 96.5%, and 97.39%, respectively. However, it only achieved a specificity of 82.35%. Densenet201 MAnet obtained the highest Precision and Specificity, with 97.23% and 90.2%, respectively. But it achieved a lower accuracy of 87.82% and recall of 87.22%.

These metrics indicate that MobileNet Unet had the smallest number of false negatives (21 exams or 2.60%) but a higher number of false positives (36 exams or 17.65%). Therefore, as losing a positive exam over a negative is more critical, MobileNet Unet might be an efficient option to detect COVID-19 on MosMedData.

Figure 19 – Training and Testing time for lesion segmentation.



Source: Author (2023)

Then, to evaluate our architecture's robustness, we performed an external validation on the SPGC Dataset, which was not on the training/validation/test sets, thus having a different distribution from our original images. Furthermore, SPGC Dataset has CAP exams, which were added to the "with lesion"class. Table 9 presents the results of all evaluated architectures in this work.

All architectures had similar and competitive results for external validation on the SPGC dataset. MobileNet Unet had the highest Accuracy and F1-Score, with 98.05% and 98.7%, respectively.

MobileNet Unet is an intermediate architecture with a small encoder of only 3.4 million parameters and a decoder of 32 million parameters. Its size might have aided it in learning the task without overfitting samples with the same distribution from train/validation/test sets.

It is worth mentioning that external validation plays a vital role when comparing

Tabela 8 – COVID-19 Lesion detection external validation on MosMedData.

Architecture	Acc (%)	F1 (%)	Prec (%)	Rec (%)	Spec (%)	Time per exam (s)
MobilenetV2 FPN	91.88	94.94	94.36	95.53	77.45	12.69
Resnet50 FPN	90.40	94.0	93.71	94.29	75.0	**12.42**
Densenet201 FPN	90.89	94.43	92.20	96.77	67.65	15.05
Resnext101 FPN	91.39	94.68	93.37	96.03	73.04	14.68
MobilenetV2 Unet	**94.36**	**96.5**	95.62	**97.39**	82.35	12.88
Resnet50 Unet	92.48	95.31	94.84	95.78	79.41	12.79
Densenet201 Unet	90.40	94.08	92.56	95.66	69.61	13.61
Resnext101 Unet	91.09	94.55	92.32	96.9	68.14	12.54
MobilenetV2 Unet++	92.18	95.05	95.95	94.17	84.31	12.44
Resnet50 Unet++	90.40	93.87	95.62	92.18	83.33	13.03
Densenet201 Unet++	91.78	94.94	93.40	96.53	73.04	12.98
Resnext101 Unet++	91.88	94.99	93.61	96.40	74.02	14.94
MobilenetV2 MAnet	90.99	94.38	93.97	94.79	75.98	14.05
Resnet50 MAnet	87.92	92.49	91.81	93.18	67.16	14.34
Densenet201 MAnet	87.82	91.96	**97.23**	87.22	**90.2**	17.43
Resnext101 MAnet	91.88	94.94	94.47	95.41	77.94	16.87

Source: author (2023).

CNNs, because it simulates real-world situations, allowing us to choose the architecture that best generalizes for new samples.

4.4 COVID-19 and CAP Classification

We trained eight deep-learning models on COVIDxCT to differentiate between COVID-19 and CAP CT slices. This classification distinguishes previously segmented lesions from these two diseases, as our segmentation models can not distinguish between COVID-19 and CAP lesions. We present our results in Table 10.

Our results for classifying CT slices as COVID-19 or CAP from COVIDxCT using eight different deep-learning models are competitive. All the models have achieved high accuracy,

Tabela 9 – Lesion detection external validation on SPGC Dataset.

Architecture	Acc (%)	F1 (%)	Prec (%)	Rec (%)	Spec (%)	Time per exam (s)
MobilenetV2 FPN	97.39	98.28	97.85	98.70	93.42	**19.03**
Resnet50 FPN	95.11	96.82	95.0	98.7	84.21	19.49
Densenet201 FPN	96.42	97.64	96.61	98.70	89.47	21.71
Resnext101 FPN	97.39	98.28	97.85	98.70	93.42	21.45
MobilenetV2 Unet	**98.05**	**98.70**	98.7	98.7	96.05	19.70
Resnet50 Unet	97.39	98.28	97.45	**99.13**	92.11	21.45
Densenet201 Unet	94.79	96.61	94.61	98.70	82.89	21.71
Resnext101 Unet	91.53	94.63	90.51	**99.13**	68.42	22.72
MobilenetV2 Unet++	97.72	98.47	99.12	97.84	**97.37**	19.22
Resnet50 Unet++	95.77	97.12	**99.55**	94.81	98.68	23.01
Densenet201 Unet++	97.39	98.28	97.45	**99.13**	92.11	23.29
Resnext101 Unet++	94.79	96.61	94.61	98.70	82.89	23.29
MobilenetV2 MAnet	95.11	96.82	95.0	98.70	84.21	20.26
Resnet50 MAnet	96.42	97.58	99.11	96.10	**97.37**	22.93
Densenet201 MAnet	94.14	96.22	93.47	**99.13**	78.95	25.33
Resnext101 MAnet	95.44	97.03	95.02	**99.13**	84.21	21.45

Source: author (2023).

Tabela 10 – COVID-19 and CAP Classification Results for COVIDxCT.

Architecture	Acc (%)	F1 (%)	Prec (%)	Rec (%)	Spec (%)
Mobilenet	95.85	95.91	94.12	97.78	93.94
Resnet50	95.79	95.90	93.09	98.88	92.73
Densenet201	96.43	96.48	94.66	98.37	94.50
Resnext101	**96.79**	**96.84**	94.71	99.07	94.52
Squeezenet	94.84	95.00	91.75	98.49	91.22
Efficientnet	96.18	96.20	**95.36**	97.04	**95.32**
Shufflenet	95.69	95.78	93.35	98.35	93.05
Ghostnet	96.18	96.28	93.42	**99.32**	93.06

Source: author (2023).

F1-score, precision, recall, and specificity. Among the models, Resnext101 has achieved the highest overall performance, with an accuracy of 96.79%, F1-score of 96.84%, precision of 94.71%, recall of 99.07%, and specificity of 94.52%. The performance of the other models is

also noteworthy, with accuracy ranging from 94.84% to 96.79%. Finally, it is worth pointing out that the models' specificity varies considerably, ranging from 91.22% to 95.32%.

Then, we externally validate these eight deep-learning models on the slices with the most extensive lesions detected on SPGC Dataset, which can have lesions caused by COVID-19 or CAP. Finally, we summarize the results in Table 11.

Tabela 11 – COVID-19 and CAP Classification external validation on SPGC Dataset.

Architecture	Acc (%)	F1 (%)	Prec (%)	Rec (%)	Spec (%)
Mobilenet	87.44	91.87	86.77	97.61	60.31
Resnet50	86.58	91.50	84.77	99.40	52.38
Densenet201	**90.47**	**93.85**	**88.42**	**100.0**	**65.07**
Resnext101	88.31	92.43	87.30	98.21	61.90
Squeezenet	86.14	91.20	84.69	98.80	52.38
Efficientnet	89.17	93.03	87.43	99.40	61.90
Shufflenet	87.87	92.26	86.08	99.40	57.14

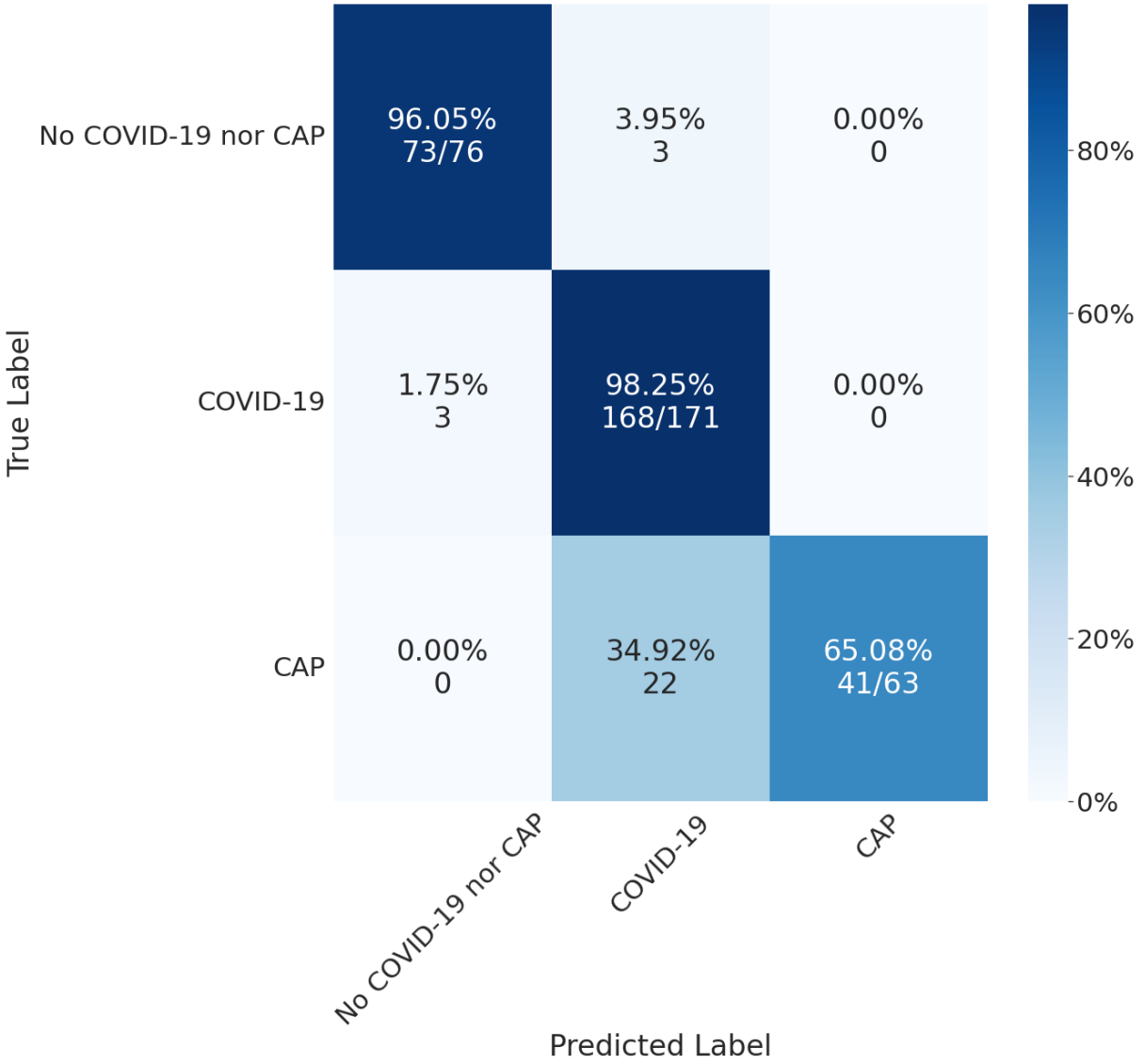
Source: author (2023).

These results indicate that the eight deep-learning models we evaluated have promising potential for distinguishing COVID-19 and CAP using CT images. Overall, Densenet201 achieved the best performance with the highest accuracy, F1-score, and specificity. However, it is worth noting that the relatively low specificity for CAP means that the models may be more prone to false negatives for this class. This is an important consideration, as accurate detection of Common Acquired Pneumonia is also critical for the appropriate treatment and management of patients. It is important to note that these results were obtained by externally validating the models on a single slice from each CT scan from the SPGC dataset. Because the SPGC Dataset has a smaller sample size than the COVIDxCT dataset used for model training, further evaluation on larger and more diverse datasets is needed to fully assess the generalizability and robustness of the models. Furthermore, to use these 2D deep-learning models and gain processing time, the three-dimensionality of SPGC Dataset CT scans is discarded, which also causes a loss of information.

When merging the segmentation, detection, and classification tasks, we obtained the confusion matrix in Figure 20. For lung segmentation, we applied Resnext101 Unet++; for lesion segmentation, we applied MobilenetV2 Unet; and for COVID-19 or CAP classification, we used Densenet201. These architectures were selected by their overall results, focusing mainly on a low false negative rate.

The confusion matrix shows that the classifier performed well in the COVID-19 class,

Figure 20 – Final results using MobilenetV2 Unet for lesion detection and Densenet201 for COVID-19 or CAP classification.



Source: Author (2023)

with a high number of true positives (168) and a low number of false positives (3). However, there are some misclassifications, as 35% of CAP exams were classified as COVID-19. These results suggest that our classification models could not differentiate between the two classes, or that there was insufficient information on the CT slice to differentiate.

4.5 COVID-19 Severity

In order to provide numerical data about the segmented COVID-19 lesions, we calculated the severity of the disease based on the compromised area of the lungs. Then, we applied this methodology on MosMedData and compared results. A summary is presented at

Table 12.

Tabela 12 – COVID-19 severity for MosMedData.

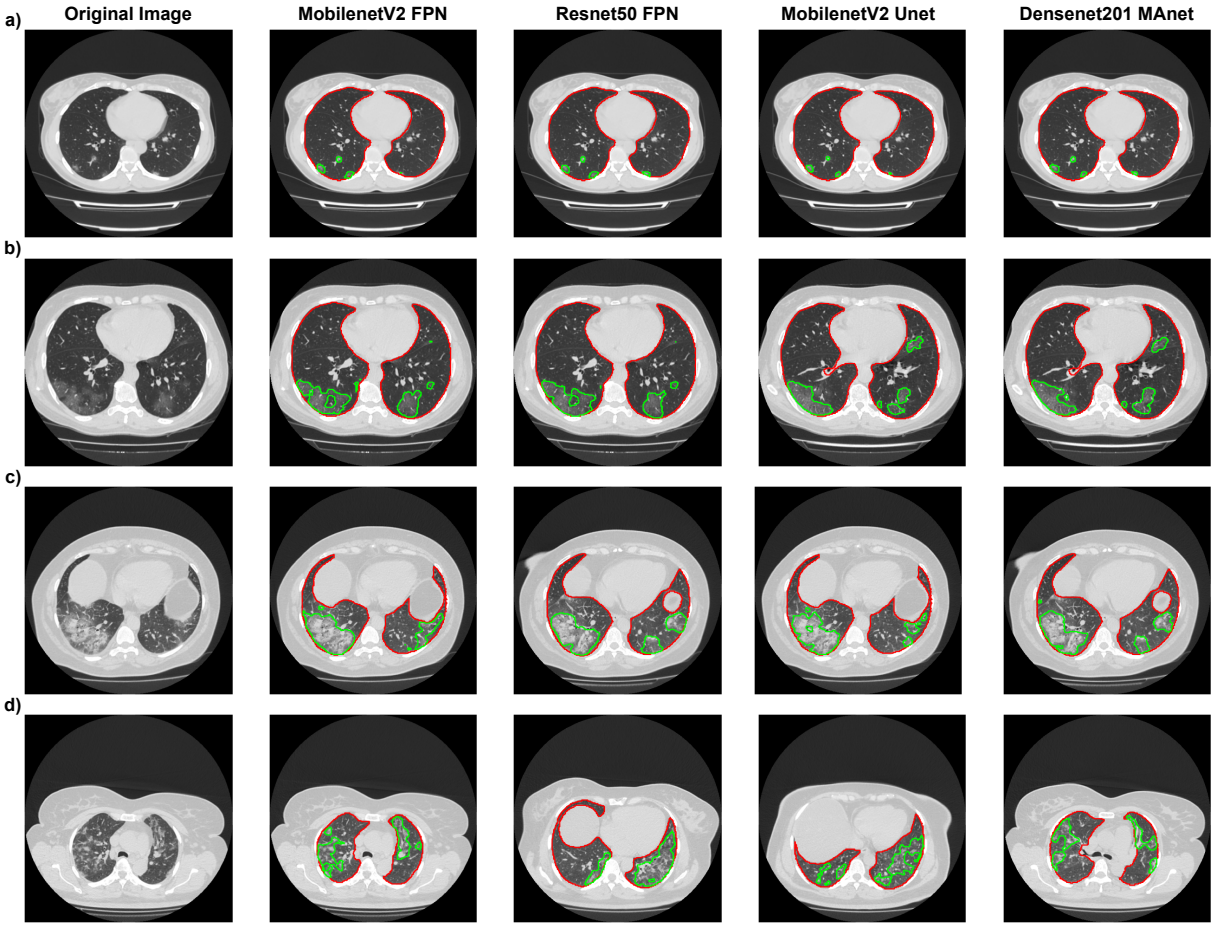
Architecture	Acc (%)	F1 (%)	Prec (%)	Rec (%)
MobilenetV2 FPN	69.9	69.95	70.27	69.9
Resnet50 FPN	71.98	70.77	70.52	71.98
Densenet201 FPN	67.13	67.81	69.07	67.13
Resnext101 FPN	69.5	69.68	70.27	69.5
MobilenetV2 Unet	**75.05**	**73.26**	**72.67**	**75.05**
Resnet50 Unet	72.67	71.08	70.25	72.67
Densenet201 Unet	66.93	67.49	68.42	66.93
Resnext101 Unet	69.41	69.07	69.58	69.41
MobilenetV2 Unet++	72.97	71.09	70.3	72.97
Resnet50 Unet++	69.21	68.7	68.79	69.21
Densenet201 Unet++	72.18	70.86	70.83	72.18
Resnext101 Unet++	70.89	70.4	70.44	70.89
MobilenetV2 MAnet	70.89	70.25	70.29	70.89
Resnet50 MAnet	66.93	66.26	65.99	66.93
Densenet201 MAnet	66.44	66.12	67.28	66.44
Resnext101 MAnet	71.09	70.58	70.47	71.09

Source: author (2023).

Again, Mobilenet Unet obtained the highest results, with an accuracy of 75.05%, an F1-Score of 73.26%, a Precision of 72.67%, and a Recall of 75.05%. Even if metrics are not as high as for binary classification ("without lesion"or "with lesion"), Figure 21 shows for four architectures (Resnet50 FPN, MobileNet FPN, MobileNet Unet, and DenseNet MAnet) that our system correctly segmented most lesions presented on the CT scans. The lower metrics might happen because of the qualitative analysis made when labelling MosMedData, which we could not replicate with quantitative values.

We presented the images where each model found the most extensive lesion area for that specific exam. The experiment found that all architectures could locate lesions in the same lung areas, indicating consistent performance. However, some architectures were unable to identify certain lesion areas accurately. Specifically, the MobilenetV2 FPN architecture failed to locate a small lesion in the right lung in the presented image (21.a), while the other three architectures correctly identified it. These findings suggest that while all architectures performed similarly overall, there were still differences in their ability to accurately identify certain lesion areas, highlighting the importance of selecting the most suitable architecture for a specific task. These difficulties in detecting certain lesion areas could have been one of the factors that

Figure 21 – Segmentation results for Resnet50 FPN, Mobilenet FPN, Mobilenet Unet, and Densenet MAnet on MosMedData. Lungs segmentation is represented as the red contours, and lesion segmentation is represented as the green contours; a) Image from an exam of class 1. b) Image from an exam of class 2. c) Image from an exam of class 3. d) Image from an exam of class 4.



Source: Author (2023)

worsened the results presented in the confusion matrices in Table 13. Returning a segmentation smaller than the actual lesion might compromise medical analysis, as the lung commitment might be worse than our system informs. However, our system correctly detected lesions, and doctors could analyze each slice and notice that lesions were more extensive than it seemed.

Despite the success of our models in differentiating COVID-19 from non-COVID-19 cases (as shown in Table 8), we still observe a high degree of error when it comes to distinguishing between different severity classes of COVID-19 on MosMedData. This error may be due to several factors, such as incorrect segmentation of lesions on CT scans by our models or the lack of a quantified evaluation on MosMedData, as specialists qualitatively evaluated severity. This may have affected our results even when lesions were correctly segmented.

According to our results, Resnet50 FPN was the fastest architecture for MosMedData

Tabela 13 – Confusion Matrix results for Resnet50 FPN, Mobilenet FPN, Mobilenet Unet, and Densenet MAnet on MosMedData.

True Class	Classified as	MobilenetV2 FPN	Resnet50 FPN	MobilenetV2 Unet	Densenet201 MAnet
CT-0	CT-0	158 (77%)	153 (70%)	168 (82%)	184 (90%)
	CT-1	46 (23%)	50 (25%)	36 (18%)	20 (10%)
	CT-2	0	1	0	0
	CT-3	0	0	0	0
	CT-4	0	0	0	0
CT-1	CT-0	35 (6%)	45 (7%)	21 (3%)	100 (16%)
	CT-1	506 (80%)	534 (84%)	559 (88%)	452 (71%)
	CT-2	59 (9%)	37 (6%)	36 (6%)	52 (52%)
	CT-3	14 (2%)	5 (1%)	8 (1%)	12 (2%)
	CT-4	20 (3%)	13 (2%)	10 (2%)	18 (3%)
CT-2	CT-0	1 (1%)	1 (1%)	0	2 (1%)
	CT-1	80 (63%)	82 (65%)	94 (74%)	83 (66%)
	CT-2	34 (28%)	35 (28%)	26 (21%)	25 (20%)
	CT-3	8 (6%)	5 (4%)	4 (3%)	10 (8%)
	CT-4	3 (2%)	3 (2%)	2 (2%)	6 (5%)
CT-3	CT-0	0	0	0	1 (2%)
	CT-1	21 (48%)	18 (41%)	22 (50%)	18 (41%)
	CT-2	9 (20%)	15 (34%)	13 (30%)	10 (23%)
	CT-3	7 (16%)	3 (7%)	4 (9%)	9 (20%)
	CT-4	7 (16%)	8 (18%)	5 (11%)	6 (14%)
CT-4	CT-0	0	0	0	0
	CT-1	0	0	0	0
	CT-2	0	0	1 (50%)	1 (50%)
	CT-3	1 (50%)	0	0	0
	CT-4	1 (50%)	2 (100%)	1 (50%)	1 (50%)

Source: author (2023).

while Densenet201 MAnet was the slowest. Specifically, the average time taken by Resnet50 FPN to segment lesions from all slices in MosMedData was 12.42 seconds. On the other hand, Resnext101 Unet++ took 17.43 seconds to segment lesions. For the SPGC Dataset, the fastest architecture was MobileNet FPN, with an average time of 12.42 seconds to segment lesions. On the other hand, Densenet MAnet was the slowest, with an average time of 25.33 seconds to segment lesions. However, despite the speed difference, both models are viable for real-life usage. This means highlighted models can be used effectively in clinical settings, where speed and accuracy are essential and computational resources might be limited. The choice of model will depend on the user's specific needs, such as the available computational resources.

4.6 Results Summary

Our lung segmentation results presented a statistically similar performance for all architectures evaluated. The main difference between the architectures was training and testing time, where MobilenetV2 FPN was the fastest with an Accuracy of 99.55 ± 0.05, DSC Score of 97.9 ± 0.16, and HD of 4.4 ± 0.1. Resnext101 Unet++ obtained the highest metrics, with an accuracy of 99.71 ± 0.05%, a DSC Score of 98.64 ± 0.19%, and an HD of 3.9 ± 0.16%.

We also obtained a statistically similar performance for lesion segmentation for all architectures. Now, the fastest architecture for training was MobilenetV2 MAnet, and for testing was MobilenetV2 FPN. MobilenetV2 MAnet obtained an Accuracy of 99.83 ± 0.02, a DSC Score of 80.9 ± 1.34, and an HD of 3.77 ± 0.13. MobilenetV2 FPN obtained an Accuracy of 99.85 ± 0.01, a DSC Score of 82.95 ± 1.45, and an HD of 2.86 ± 0.12. Densenet201 Unet++ achieved the highest metrics, with an accuracy of 99.87 ± 0.01%, a DSC Score of 85.16 ± 1.13%, and an HD of 3.4 ± 0.13%.

Then, we used our lesion segmentation architectures for COVID-19 lesion detection on MosMedData and lesion (COVID-19 or CAP) detection on SPGC Dataset.

For MosmedData, Mobilenet Unet had the highest Accuracy, F1-Score, Recall, and Specificity with 94.36%, 96.5%, and 97.39% of 82.35%. It also had the smallest number of false negatives (21 exams or 2.60%) but a high number of false positives (36 exams or 17.65%). Therefore, as losing a positive exam over a negative is more critical, MobileNet Unet might be an efficient option to detect COVID-19 on MosMedData.

Then, we performed an external validation on SPGC Dataset. All architectures had similar and competitive results. MobileNet Unet had the highest Accuracy and F1-Score, with 98.05% and 98.7%, respectively. It also obtained competitive values for Precision, Recall, and Specificity, such as 98.7%, 98.7%, and 96.05%, respectively.

For the classification between COVID-19 or CAP task on COVIDxCT Dataset, Resnext101 achieved the highest overall performance, with an Accuracy of 96.79%, F1-score of 96.84%, Precision of 94.71%, Recall of 99.07%, and Specificity of 94.52%. When externally validating on the SPGC Dataset, Densenet201 had the highest overall performance, reaching an accuracy of 90.47%, an F1-score of 93.85%, a precision of 88.42%, a recall of 100.0%, and a specificity of 65.07%.

Finally, for analysing COVID-19 severity based on lung and lesion segmentation, Mobilenet Unet obtained the highest results, with an accuracy of 75.05%, an F1-Score of 73.26%,

a Precision of 72.67%, and a Recall of 75.05%.

4.7 Limitations

The first limitation of this work is that all the architectures evaluated in this study were based on 2D images, whereas CT scans provide 3D information. Although using 2D images simplifies the computational complexity and reduces the training time required for the models, it may not fully capture the complexity and variations of the 3D structures. Consequently, the accuracy of the models in predicting and diagnosing various medical conditions using a 2D approach may be lower than with a 3D approach. Another limitation of this study is that analysing only the CT Scan of a patient may not be sufficient for a diagnosis. CT scans provide useful information about the body's internal structures, but they do not provide information about the patient's symptoms or medical history. Therefore, integrating the CT Scan analysis with clinical data processed by natural language models could improve the accuracy of the diagnosis. By combining image and language models, physicians can make more informed decisions and provide better patient treatment options.

While our approach showed encouraging results, other CT scan factors that may contribute to the differentiation between CAP and COVID-19 might not be captured when analysing only 2D slices separately. Moreover, it is known that there is a significant overlap in the imaging features of COVID-19 and other respiratory diseases, which makes differentiation challenging even with the use of advanced imaging techniques.

Despite these limitations, the findings of this study provide valuable insights into the potential applications of deep learning and computer vision techniques for medical image analysis. Future studies can build upon these findings and further explore using 3D imaging and language models to improve medical diagnosis and treatment accuracy and efficiency.

5 CONCLUSIONS

In this work, we proposed a Deep Learning-based approach for lung and lesion detection and segmentation, and COVID-19 and CAP classification using full CT scans. The results showed that our system correctly detected and segmented lesions from COVID-19 and CAP in CT scans, differentiating these two classes from Normal exams. For the classification task, we achieved competitive results for accuracy, precision, recall, F1-score, and specificity for COVIDxCT dataset. However, our metrics dropped when externally validating on the SPGC dataset, but are still competitive. From our results, our system can aid medical doctors in analysing COVID-19 patients, mainly by providing quantitative values for lesion and lung sizes. We also statistically demonstrated that most state-of-the-art CNN models for segmentation achieve similar results, with insignificant differences between results from one or another.

Nevertheless, this work did not exhaust the possibilities of researching COVID-19 and CAP detection, segmentation, and classification. In future works, one might evaluate the tradeoff between processing time and accuracy using 3D segmentation and classification architectures. In addition, clinical data can be used to aid in differentiating COVID-19 and CAP CT exams.

REFERENCES

ABDEL-BASSET, M.; HAWASH, H.; MOUSTAFA, N.; ELKOMY, O. M. Two-stage deep learning framework for discrimination between covid-19 and community-acquired pneumonia from chest ct scans. **Pattern Recognition Letters**, v. 152, p. 311–319, 2021. ISSN 0167-8655. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167865521003871.

AFSHAR, P.; HEIDARIAN, S.; ENSHAEI, N.; NADERKHANI, F.; RAFIEE, M. J.; OIKONOMOU, A.; FARD, F. B.; SAMIMI, K.; PLATANIOTIS, K. N.; MOHAMMADI, A. Covid-ct-md, covid-19 computed tomography scan dataset applicable in machine learning and deep learning. **Scientific Data**, v. 8, n. 1, p. 121, Apr 2021. ISSN 2052-4463. Disponível em: https://doi.org/10.1038/s41597-021-00900-3.

AMYAR, A.; MODZELEWSKI, R.; LI, H.; RUAN, S. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. **Comput Biol Med**, United States, v. 126, p. 104037, out. 2020.

BARTLETT, M. S.; FOWLER, R. H. Properties of sufficiency and statistical tests. **Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences**, v. 160, n. 901, p. 268–282, 1937. Disponível em: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1937.0109.

BASHA, S. M.; NETO, A. V. L.; ALSHATHRI, S.; ELAZIZ, M. A.; MOHISIN, S. H.; ALBUQUERQUE, V. H. C. D. Multithreshold segmentation and machine learning based approach to differentiate COVID-19 from viral pneumonia. **Comput Intell Neurosci**, United States, v. 2022, p. 2728866, ago. 2022.

BIEWALD, L. **Experiment Tracking with Weights and Biases**. 2020. Software available from wandb.com. Disponível em: https://www.wandb.com/.

BROWN, M. B.; FORSYTHE, A. B. Robust tests for the equality of variances. **Journal of the American Statistical Association**, [American Statistical Association, Taylor Francis, Ltd.], v. 69, n. 346, p. 364–367, 1974. ISSN 01621459. Disponível em: http://www.jstor.org/stable/2285659.

BUSLAEV, A.; IGLOVIKOV, V. I.; KHVEDCHENYA, E.; PARINOV, A.; DRUZHININ, M.; KALININ, A. A. Albumentations: Fast and flexible image augmentations. **Information**, v. 11, n. 2, 2020. ISSN 2078-2489. Disponível em: https://www.mdpi.com/2078-2489/11/2/125.

CHEN, N.; ZHOU, M.; DONG, X.; QU, J.; GONG, F.; HAN, Y.; QIU, Y.; WANG, J.; LIU, Y.; WEI, Y.; XIA, J.; YU, T.; ZHANG, X.; ZHANG, L. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. **The Lancet**, v. 395, n. 10223, p. 507–513, 2020. ISSN 0140-6736. Disponível em: https://www.sciencedirect.com/science/article/pii/S0140673620302117.

DONG, E.; DU, H.; GARDNER, L. An interactive web-based dashboard to track COVID-19 in real time. **Lancet Infect Dis**, United States, v. 20, n. 5, p. 533–534, fev. 2020.

DRIGGS, D.; SELBY, I.; ROBERTS, M.; GKRANIA-KLOTSAS, E.; RUDD, J. H. F.; YANG, G.; BABAR, J.; SALA, E.; SCHöNLIEB, C.-B. a. Machine learning for covid-19 diagnosis and prognostication: Lessons for amplifying the signal while reducing the noise. **Radiology: Artificial Intelligence**, v. 3, n. 4, p. e210011, 2021. Disponível em: https://doi.org/10.1148/ryai.2021210011.

FAN, T.; WANG, G.; LI, Y.; WANG, H. Ma-net: A multi-scale attention network for liver and tumor segmentation. **IEEE Access**, v. 8, p. 179656–179665, 2020.

FAN, T.; WANG, G.; LI, Y.; WANG, H. Ma-net: A multi-scale attention network for liver and tumor segmentation. **IEEE Access**, v. 8, p. 179656–179665, 2020.

FAWCETT, T. An introduction to roc analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006. ISSN 0167-8655. ROC Analysis in Pattern Recognition. Disponível em: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the American Statistical Association**, [American Statistical Association, Taylor Francis, Ltd.], v. 32, n. 200, p. 675–701, 1937. ISSN 01621459. Disponível em: http://www.jstor.org/stable/2279372.

GONZALEZ, R.; WOODS, R. **Digital Image Processing, Global Edition**. Pearson Education, 2018. ISBN 9781292223070. Disponível em: https://books.google.com.br/books?id=P8AoEAAAQBAJ.

GREENHOUSE, S. W.; GEISSER, S. On methods in the analysis of profile data. **Psychometrika**, v. 24, n. 2, p. 95–112, Jun 1959. ISSN 1860-0980. Disponível em: https://doi.org/10.1007/BF02289823.

GUNRAJ, H.; SABRI, A.; KOFF, D.; WONG, A. Covid-net ct-2: Enhanced deep neural networks for detection of covid-19 from chest ct images through bigger, more diverse learning. **Frontiers in Medicine**, v. 8, p. 729287, 2022. ISSN 2296-858X. Disponível em: https://www.frontiersin.org/articles/10.3389/fmed.2021.729287.

HAN, K.; WANG, Y.; TIAN, Q.; GUO, J.; XU, C.; XU, C. Ghostnet: More features from cheap operations. **CoRR**, abs/1911.11907, 2019. Disponível em: http://arxiv.org/abs/1911.11907.

HARMON, S. A.; SANFORD, T. H.; XU, S.; TURKBEY, E. B.; ROTH, H.; XU, Z.; YANG, D.; MYRONENKO, A.; ANDERSON, V.; AMALOU, A.; BLAIN, M.; KASSIN, M.; LONG, D.; VARBLE, N.; WALKER, S. M.; BAGCI, U.; IERARDI, A. M.; STELLATO, E.; PLENSICH, G. G.; FRANCESCHELLI, G.; GIRLANDO, C.; IRMICI, G.; LABELLA, D.; HAMMOUD, D.; MALAYERI, A.; JONES, E.; SUMMERS, R. M.; CHOYKE, P. L.; XU, D.; FLORES, M.; TAMURA, K.; OBINATA, H.; MORI, H.; PATELLA, F.; CARIATI, M.; CARRAFIELLO, G.; AN, P.; WOOD, B. J.; TURKBEY, B. Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. **Nature Communications**, v. 11, n. 1, p. 4080, Aug 2020. ISSN 2041-1723. Disponível em: https://doi.org/10.1038/s41467-020-17971-2.

HASAN, M. K.; JAWAD, M. T.; HASAN, K. N. I.; PARTHA, S. B.; MASBA, M. M. A.; SAHA, S.; MONI, M. A. Covid-19 identification from volumetric chest ct scans using a progressively resized 3d-cnn incorporating segmentation, augmentation, and class-rebalancing. **Informatics in Medicine Unlocked**, v. 26, p. 100709, 2021. ISSN 2352-9148. Disponível em: https://www.sciencedirect.com/science/article/pii/S235291482100191X.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S. l.*: *s. n.*], 2016. p. 770–778.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S. l.*: *s. n.*], 2016. p. 770–778.

HENDEE, W.; RITENOUR, E. Computed tomography. In: ____. **Medical Imaging Physics**. John Wiley Sons, Ltd, 2002. cap. 15, p. 251–264. ISBN 9780471221159. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471221155.ch15.

HOLSHUE, M. L.; DEBOLT, C.; LINDQUIST, S.; LOFY, K. H.; WIESMAN, J.; BRUCE, H.; SPITTERS, C.; ERICSON, K.; WILKERSON, S.; TURAL, A.; DIAZ, G.; COHN, A.; FOX, L.; PATEL, A.; GERBER, S. I.; KIM, L.; TONG, S.; LU, X.; LINDSTROM, S.; PALLANSCH, M. A.; WELDON, W. C.; BIGGS, H. M.; UYEKI, T. M.; PILLAI, S. K. First case of 2019 novel coronavirus in the united states. **New England Journal of Medicine**, v. 382, n. 10, p. 929–936, 2020. PMID: 32004427. Disponível em: https://doi.org/10.1056/NEJMoa2001191.

HOUNSFIELD, G. N. Computerized transverse axial scanning (tomography). 1. description of system. **Br J Radiol**, England, v. 46, n. 552, p. 1016–1022, dez. 1973.

HU, Q.; GOIS, F. N. B.; COSTA, R.; ZHANG, L.; YIN, L.; MAGAIA, N.; de Albuquerque, V. H. C. Explainable artificial intelligence-based edge fuzzy images for covid-19 detection and identification. **Applied Soft Computing**, v. 123, p. 108966, 2022. ISSN 1568-4946. Disponível em: https://www.sciencedirect.com/science/article/pii/S1568494622003064.

HUANG, C.; WANG, Y.; LI, X.; REN, L.; ZHAO, J.; HU, Y.; ZHANG, L.; FAN, G.; XU, J.; GU, X.; CHENG, Z.; YU, T.; XIA, J.; WEI, Y.; WU, W.; XIE, X.; YIN, W.; LI, H.; LIU, M.; XIAO, Y.; GAO, H.; GUO, L.; XIE, J.; WANG, G.; JIANG, R.; GAO, Z.; JIN, Q.; WANG, J.; CAO, B. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. **The Lancet**, v. 395, n. 10223, p. 497–506, 2020. ISSN 0140-6736. Disponível em: https://www.sciencedirect.com/science/article/pii/S0140673620301835.

HUANG, G.; LIU, Z.; MAATEN, L. V. D.; WEINBERGER, K. Q. Densely connected convolutional networks. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S. l.*: *s. n.*], 2017. p. 2261–2269.

HUANG, G.; LIU, Z.; MAATEN, L. V. D.; WEINBERGER, K. Q. Densely connected convolutional networks. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S. l.*: *s. n.*], 2017. p. 2261–2269.

HUANG, P.; LIU, T.; HUANG, L.; LIU, H.; LEI, M.; XU, W.; HU, X.; CHEN, J.; LIU, B. Use of chest CT in combination with negative RT-PCR assay for the 2019 novel coronavirus but high clinical suspicion. **Radiology**, United States, v. 295, n. 1, p. 22–23, fev. 2020.

IAKUBOVSKII, P. **Segmentation Models Pytorch**. [*S. l.*]: GitHub, 2019. https://github.com/qubvel/segmentation_models.pytorch.

IANDOLA, F. N.; MOSKEWICZ, M. W.; ASHRAF, K.; HAN, S.; DALLY, W. J.; KEUTZER, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. **CoRR**, abs/1602.07360, 2016. Disponível em: http://arxiv.org/abs/1602.07360.

JUN, M.; CHENG, G.; YIXIN, W.; XINGLE, A.; JIANTAO, G.; ZIQI, Y.; MINQING, Z.; XIN, L.; XUEYUAN, D.; SHUCHENG, C.; HAO, W.; SEN, M.; XIAOYU, Y.; ZIWEI, N.; CHEN, L.; LU, T.; YUNTAO, Z.; QIONGJIE, Z.; GUOQIANG, D.; JIAN, H.

COVID-19 CT Lung and Infection Segmentation Dataset. Zenodo, abr. 2020. Disponível em: https://doi.org/10.5281/zenodo.3757476.

KAGGLE. Finding and measuring lungs in ct data. available online: https://www.kaggle.com/kmader/finding-lungs-in-ct-data (accessed on 8 november 2021). 2017.

LI, L.; QIN, L.; XU, Z.; YIN, Y.; WANG, X.; KONG, B.; BAI, J.; LU, Y.; FANG, Z.; SONG, Q.; CAO, K.; LIU, D.; WANG, G.; XU, Q.; FANG, X.; ZHANG, S.; XIA, J.; XIA, J. Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: Evaluation of the diagnostic accuracy. **Radiology**, v. 296, n. 2, p. E65–E71, 2020. PMID: 32191588. Disponível em: https://doi.org/10.1148/radiol.2020200905.

LI, M.; FANG, Y.; TANG, Z.; ONUORAH, C.; XIA, J.; SER, J. D.; WALSH, S.; YANG, G. Explainable covid-19 infections identification and delineation using calibrated pseudo labels. **IEEE Transactions on Emerging Topics in Computational Intelligence**, v. 7, n. 1, p. 26–35, 2023.

LIN, T.-Y.; DOLLáR, P.; GIRSHICK, R.; HE, K.; HARIHARAN, B.; BELONGIE, S. Feature pyramid networks for object detection. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S. l.*: *s. n.*], 2017. p. 936–944.

LIN, T.-Y.; DOLLáR, P.; GIRSHICK, R.; HE, K.; HARIHARAN, B.; BELONGIE, S. Feature pyramid networks for object detection. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S. l.*: *s. n.*], 2017. p. 936–944.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. **arXiv**, 2014. Disponível em: https://arxiv.org/pdf/1411.4038.

LÓPEZ-ÚBEDA, P.; DÍAZ-GALIANO, M. C.; MARTÍN-NOGUEROL, T.; LUNA, A.; UREÑA-LÓPEZ, L. A.; MARTÍN-VALDIVIA, M. T. Covid-19 detection in radiological text reports integrating entity recognition. **Computers in Biology and Medicine**, Elsevier, v. 127, p. 104066, 2020.

MALDEN, D. E.; TARTOF, S. Y.; ACKERSON, B. K.; HONG, V.; SKARBINSKI, J.; YAU, V.; QIAN, L.; FISCHER, H.; SHAW, S. F.; CAPAROSA, S. *et al.* Natural language processing for improved characterization of covid-19 symptoms: Observational study of 350,000 patients in a large integrated health care system. **JMIR Public Health and Surveillance**, JMIR Publications Inc., Toronto, Canada, v. 8, n. 12, p. e41529, 2022.

MATOS MARINA JUSTI ROSA DE; ROSA, M. E. E. B. V. M. A. L. T. W. B. G. L. F. E. K. U. N. C. R. C. P. R. B. D. S. M. M. A. Y. P. S. N. R. T. G. B. d. S. S. M. C. B. d. S. G. Diagnósticos diferenciais de opacidade em vidro fosco aguda na tomografia computadorizada de tórax: ensaio pictórico. **einstein (São Paulo)**, v. 19, n. 5, p. 1072–1077, 2021. Disponível em: https://doi.org/10.31744/einstein_journal/2021RW5772.

MATTIOLI, A. V.; PUVIANI, M. B.; NASI, M.; FARINETTI, A. Covid-19 pandemic: the effects of quarantine on cardiovascular risk. **European journal of clinical nutrition**, Nature Publishing Group, v. 74, n. 6, p. 852–855, 2020.

MOROZOV, S. P.; ANDREYCHENKO, A. E.; PAVLOV, N. A.; VLADZYMYRSKYY, A. V.; LEDIKHOVA, N. V.; GOMBOLEVSKIY, V. A.; BLOKHIN, I. A.; GELEZHE, P. B.;

GONCHAR, A. V.; CHERNINA, V. Y. **MosMedData: Chest CT Scans With COVID-19 Related Findings Dataset**. arXiv, 2020. Disponível em: https://arxiv.org/abs/2005.06465.

NAUDÉ, W. Artificial intelligence vs COVID-19: limitations, constraints and pitfalls. **AI Soc**, Germany, v. 35, n. 3, p. 761–765, abr. 2020.

NEMENYI, P. **Distribution-free Multiple Comparisons**. Princeton University, 1963. Disponível em: https://books.google.com.br/books?id=nhDMtgAACAAJ.

NG, M.-Y.; LEE, E. Y. P.; YANG, J.; YANG, F.; LI, X.; WANG, H.; LUI, M. M.-S.; LO, C. S.-Y.; LEUNG, B.; KHONG, P.-L.; HUI, C. K.-M.; YUEN, K.-Y.; KUO, M. D. Imaging profile of the COVID-19 infection: Radiologic findings and literature review. **Radiol Cardiothorac Imaging**, United States, v. 2, n. 1, p. e200034, fev. 2020.

NIH. **COVID-19 Treatment Guidelines Panel. Coronavirus Disease 2019 (COVID-19) Treatment Guidelines**. 2023. Disponível em: https://www.covid19treatmentguidelines.nih.gov.

OHATA, E. F.; BEZERRA, G. M.; CHAGAS, J. V. S. d.; NETO, A. V. L.; ALBUQUERQUE, A. B.; ALBUQUERQUE, V. H. C. d.; FILHO, P. P. R. Automatic detection of covid-19 infection using chest x-ray images through transfer learning. **IEEE/CAA Journal of Automatica Sinica**, v. 8, n. 1, p. 239–248, 2021.

PARAH, S. A.; KAW, J. A.; BELLAVISTA, P.; LOAN, N. A.; BHAT, G. M.; MUHAMMAD, K.; ALBUQUERQUE, V. H. C. de. Efficient security and authentication for edge-based internet of medical things. **IEEE Internet of Things Journal**, v. 8, n. 21, p. 15652–15662, 2021.

QIBLAWEY, Y.; TAHIR, A.; CHOWDHURY, M. E. H.; KHANDAKAR, A.; KIRANYAZ, S.; RAHMAN, T.; IBTEHAZ, N.; MAHMUD, S.; MAADEED, S. A.; MUSHARAVATI, F.; AYARI, M. A. Detection and severity classification of covid-19 in ct images using deep learning. **Diagnostics**, v. 11, n. 5, 2021. ISSN 2075-4418. Disponível em: https://www.mdpi.com/2075-4418/11/5/893.

QOMARIYAH, N. N.; ARAMINTA, A. S.; REYNALDI, R.; SENJAYA, M.; ASRI, S. D. A.; KAZAKOV, D. Nlp text classification for covid-19 automatic detection from radiology report in indonesian language. In: IEEE. **2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)**. [*S. l.*], 2022. p. 565–569.

REN, L.-L.; WANG, Y.-M.; WU, Z.-Q.; XIANG, Z.-C.; GUO, L.; XU, T.; JIANG, Y.-Z.; XIONG, Y.; LI, Y.-J.; LI, X.-W. *et al.* Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. **Chinese medical journal**, Chinese Medical Journals Publishing House Co., Ltd. 42 Dongsi Xidajie . . . , v. 133, n. 09, p. 1015–1024, 2020.

ROBERTS, M.; DRIGGS, D.; THORPE, M.; GILBEY, J.; YEUNG, M.; URSPRUNG, S.; AVILES-RIVERO, A. I.; ETMANN, C.; MCCAGUE, C.; BEER, L.; WEIR-MCCALL, J. R.; TENG, Z.; GKRANIA-KLOTSAS, E.; RUGGIERO, A.; KORHONEN, A.; JEFFERSON, E.; AKO, E.; LANGS, G.; GOZALIASL, G.; YANG, G.; PROSCH, H.; PRELLER, J.; STANCZUK, J.; TANG, J.; HOFMANNINGER, J.; BABAR, J.; SÁNCHEZ, L. E.; THILLAI, M.; GONZALEZ, P. M.; TEARE, P.; ZHU, X.; PATEL, M.; CAFOLLA, C.; AZADBAKHT, H.; JACOB, J.; LOWE, J.; ZHANG, K.; BRADLEY, K.; WASSIN, M.; HOLZER, M.; JI, K.; ORTET, M. D.; AI, T.; WALTON, N.; LIO, P.; STRANKS, S.; SHADBAHR, T.; LIN, W.; ZHA, Y.; NIU, Z.; RUDD, J. H. F.; SALA, E.; SCHÖNLIEB, C.-B.; AIX-COVNET. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using

chest radiographs and ct scans. **Nature Machine Intelligence**, v. 3, n. 3, p. 199–217, Mar 2021. ISSN 2522-5839. Disponível em: https://doi.org/10.1038/s42256-021-00307-0.

ROHILA, V. S.; GUPTA, N.; KAUL, A.; SHARMA, D. K. Deep learning assisted covid-19 detection using full ct-scans. **Internet of Things**, v. 14, p. 100377, 2021. ISSN 2542-6605. Disponível em: https://www.sciencedirect.com/science/article/pii/S2542660521000214.

RONNEBERBER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. **arXiv**, 2015. Disponível em: https://arxiv.org/pdf/1505.04597.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: NAVAB, N.; HORNEGGER, J.; WELLS, W. M.; FRANGI, A. F. (Ed.). **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015**. Cham: Springer International Publishing, 2015. p. 234–241. ISBN 978-3-319-24574-4.

ROSEBROCK, A. **Deep Learning for Computer Vision with Python**. [*S. l.*]: Pearson, 2018. v. 1ª Edição.

RUME, T.; ISLAM, S. D.-U. Environmental effects of covid-19 pandemic and potential strategies of sustainability. **Heliyon**, v. 6, n. 9, p. e04965, 2020. ISSN 2405-8440. Disponível em: https://www.sciencedirect.com/science/article/pii/S2405844020318089.

SALEHI, S.; ABEDI, A.; BALAKRISHNAN, S.; GHOLAMREZANEZHAD, A. Coronavirus disease 2019 (covid-19): A systematic review of imaging findings in 919 patients. **American Journal of Roentgenology**, v. 215, n. 1, p. 87–93, 2020. PMID: 32174129. Disponível em: https://doi.org/10.2214/AJR.20.23034.

SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [*S. l.*: *s. n.*], 2018. p. 4510–4520.

SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [*S. l.*: *s. n.*], 2018. p. 4510–4520.

SEGMENTATION, M. Covid-19 ct segmentation dataset. available online: http://medicalsegmentation.com/covid19/ (accessed on 8 november 2021). 2020.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples)†. **Biometrika**, v. 52, n. 3-4, p. 591–611, 12 1965. ISSN 0006-3444. Disponível em: https://doi.org/10.1093/biomet/52.3-4.591.

TAHAMTAN, A.; ARDEBILI, A. Real-time RT-PCR in COVID-19 detection: issues affecting the results. **Expert Rev Mol Diagn**, England, v. 20, n. 5, p. 453–454, abr. 2020.

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. **CoRR**, abs/1905.11946, 2019. Disponível em: http://arxiv.org/abs/1905.11946.

TUKEY, J. W. Comparing individual means in the analysis of variance. **Biometrics**, [Wiley, International Biometric Society], v. 5, n. 2, p. 99–114, 1949. ISSN 0006341X, 15410420. Disponível em: http://www.jstor.org/stable/3001913.

VALENTE, I. R. S.; CORTEZ, P. C.; NETO, E. C.; SOARES, J. M.; de Albuquerque, V. H. C.; TAVARES, J. M. R. Automatic 3d pulmonary nodule detection in ct images: A survey. **Computer Methods and Programs in Biomedicine**, v. 124, p. 91–107, 2016. ISSN 0169-2607. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169260715300298.

WANG, C.; PAN, R.; WAN, X.; TAN, Y.; XU, L.; HO, C. S.; HO, R. C. Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (covid-19) epidemic among the general population in china. **International Journal of Environmental Research and Public Health**, v. 17, n. 5, 2020. ISSN 1660-4601. Disponível em: https://www.mdpi.com/1660-4601/17/5/1729.

WANG, G.; LIU, X.; LI, C.; XU, Z.; RUAN, J.; ZHU, H.; MENG, T.; LI, K.; HUANG, N.; ZHANG, S. A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images. **IEEE Transactions on Medical Imaging**, v. 39, n. 8, p. 2653–2663, 2020.

WHO. **Naming the coronavirus disease (COVID-2019) and the virus that causes it**. 2020. Disponível em: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it.

WHO. **Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)**. 2020. Disponível em: https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov).

WHO. **WHO statement regarding cluster of pneumonia cases in Wuhan, China. Jan 9, 2020**. 2020. Disponível em: https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china.

XIE, S.; GIRSHICK, R.; DOLLáR, P.; TU, Z.; HE, K. Aggregated residual transformations for deep neural networks. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S. l.*: *s. n.*], 2017. p. 5987–5995.

XIE, S.; GIRSHICK, R. B.; DOLLÁR, P.; TU, Z.; HE, K. Aggregated residual transformations for deep neural networks. **CoRR**, abs/1611.05431, 2016. Disponível em: http://arxiv.org/abs/1611.05431.

XIE, X.; ZHONG, Z.; ZHAO, W.; ZHENG, C.; WANG, F.; LIU, J. Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: Relationship to negative RT-PCR testing. **Radiology**, United States, v. 296, n. 2, p. E41–E45, fev. 2020.

YOUSEFZADEH, M.; ESFAHANIAN, P.; MOVAHED, S. M. S.; GORGIN, S.; RAHMATI, D.; ABEDINI, A.; NADJI, S. A.; HASELI, S.; KARAM, M. B.; KIANI, A.; HOSEINYAZDI, M.; ROSHANDEL, J.; LASHGARI, R. ai-corona: Radiologist-assistant deep learning framework for covid-19 diagnosis in chest ct scans. **PLOS ONE**, Public Library of Science, v. 16, n. 5, p. 1–20, 05 2021. Disponível em: https://doi.org/10.1371/journal.pone.0250952.

ZHANG, K.; LIU, X.; SHEN, J.; LI, Z.; SANG, Y.; WU, X.; ZHA, Y.; LIANG, W.; WANG, C.; WANG, K.; YE, L.; GAO, M.; ZHOU, Z.; LI, L.; WANG, J.; YANG, Z.; CAI, H.; XU, J.; YANG, L.; CAI, W.; XU, W.; WU, S.; ZHANG, W.; JIANG, S.; ZHENG, L.; ZHANG, X.;

WANG, L.; LU, L.; LI, J.; YIN, H.; WANG, W.; LI, O.; ZHANG, C.; LIANG, L.; WU, T.; DENG, R.; WEI, K.; ZHOU, Y.; CHEN, T.; LAU, J. Y.-N.; FOK, M.; HE, J.; LIN, T.; LI, W.; WANG, G. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. **Cell**, United States, v. 181, n. 6, p. 1423–1433.e11, maio 2020.

ZHANG, X.; ZHOU, X.; LIN, M.; SUN, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. **CoRR**, abs/1707.01083, 2017. Disponível em: http://arxiv.org/abs/1707.01083.

ZHAO, W.; ZHONG, Z.; XIE, X.; YU, Q.; LIU, J. Relation between chest ct findings and clinical conditions of coronavirus disease (covid-19) pneumonia: A multicenter study. **American Journal of Roentgenology**, v. 214, n. 5, p. 1072–1077, 2020. PMID: 32125873. Disponível em: https://doi.org/10.2214/AJR.20.22976.

ZHAO, W.; ZHONG, Z.; XIE, X.; YU, Q.; LIU, J. Relation between chest ct findings and clinical conditions of coronavirus disease (covid-19) pneumonia: A multicenter study. **American Journal of Roentgenology**, v. 214, n. 5, p. 1072–1077, 2020. PMID: 32125873. Disponível em: https://doi.org/10.2214/AJR.20.22976.

ZHOU, L.; LI, Z.; ZHOU, J.; LI, H.; CHEN, Y.; HUANG, Y.; XIE, D.; ZHAO, L.; FAN, M.; HASHMI, S.; ABDELKAREEM, F.; EIADA, R.; XIAO, X.; LI, L.; QIU, Z.; GAO, X. A rapid, accurate and machine-agnostic segmentation and quantification method for ct-based covid-19 diagnosis. **IEEE Transactions on Medical Imaging**, v. 39, n. 8, p. 2638–2652, 2020.

ZHOU, Z.; SIDDIQUEE, M. M. R.; TAJBAKHSH; LIANG, J. Unet++: A nested u-net architecture for medical image segmentation. **arXiv**, 2018. Disponível em: https://arxiv.org/pdf/1807.10165v1.

ZHOU, Z.; SIDDIQUEE, M. M. R.; TAJBAKHSH, N.; LIANG, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. **IEEE Transactions on Medical Imaging**, v. 39, n. 6, p. 1856–1867, 2020.