COVID-19 CT Images Segmentation

Piotr Grabysz
Capstone Project Report

June 28, 2025

Abstract: Segment radiological findings on axial slices of lungs, by using data from a Kaggle competition.

1 Introduction

A common stage in diagnosis for COVID-19 patients is computer tomography (CT). A radiologist is often asked to estimate the extent of damage with respect to lung volume. Manual segmentation of infected areas is time-consuming and labor-intensive for radiologists. This project aims to develop a deep learning model to automate the segmentation of two key radiological findings in axial CT slices: **ground glass opacities** and **consolidations**. The project is based on data from a Kaggle competition [4] and builds upon the U-Net architecture [7] for semantic segmentation.

1.1 Domain background

A computed tomography scan (CT scan), is a medical imaging technique used to obtain detailed internal images of the body. CT scanners use a rotating X-ray tube and a row of detectors placed in a gantry to measure X-ray attenuations by different tissues inside the body. The multiple X-ray measurements taken from different angles are then processed on a computer using tomographic reconstruction algorithms to produce tomographic (cross-sectional) images (virtual "slices") of a body [10].

1.2 Problem statement

In this project I want to design a model, which takes a CT scan as an input and outputs masks corresponding to two classes: *ground glass* and *consolidation*. According to Wikipedia, *ground glass* is

> an area of increased attenuation due to air displacement by fluid, airway collapse, fibrosis, or a neoplastic process. When a substance other than air fills an area of the lung it increases that area's density. On both x-ray and CT, this appears more grey or hazy as opposed to the normally dark-appearing lungs. Although

> it can sometimes be seen in normal lungs, common pathologic causes include infections, interstitial lung disease, and pulmonary edema [11].

whereas the *consolidation* means

> a region of normally compressible lung tissue that has filled with liquid instead of air. The condition is marked by induration (swelling or hardening of normally soft tissue) of a normally aerated lung.
>
> Consolidated tissue is more radio-opaque than normally aerated lung parenchyma, so that it is clearly demonstrable in radiography and on CT scans. Consolidation is often a middle-to-late stage feature/complication in pulmonary infections [12].

The process of creating binary masks is called image segmentation. Segmenting CT scans into *ground-glass* and *consolidation* is the goal of the Kaggle competition [4].

2 Dataset

2.1 Dataset description

The data comes from a Kaggle competition **COVID-19 CT Images Segmentation** [4]. At the time of writing, the competition is still active.

The dataset is built from two parts. The first part is a dataset of 100 axial CT images from more than 40 patients with COVID-19. The scans were labelled using a tool called Medseg [1]. The second part contains segmented 9 axial volumetric CTs from Radiopaedia [6]. It includes whole volumes and therefore both positive and negative slices (373 out of the total of 829 slices have been evaluated by a radiologist as positive and segmented). Sample scans and masks are presented in the Figure 1.

Both parts consists of greyscale 512 × 512 images and binary masks with 4 channels: 0 – *ground glass*, 1 – *consolidations*, 2 – *lungs other*, 3 – *background*. Only the first 2 channels are relevant for this task. There are also 10 test images provided for the competition.

Images and masks are stored in arrays in numpy format. There are 5 files given in total:

- `images_medseg.npy`
- `masks_medseg.npy`
- `test_images_medseg.npy`
- `images_radiopedia.npy`
- `masks_radiopedia.npy`

2.2 Exploratory data analysis

An exploratory analysis was performed to better understand the characteristics of the dataset.

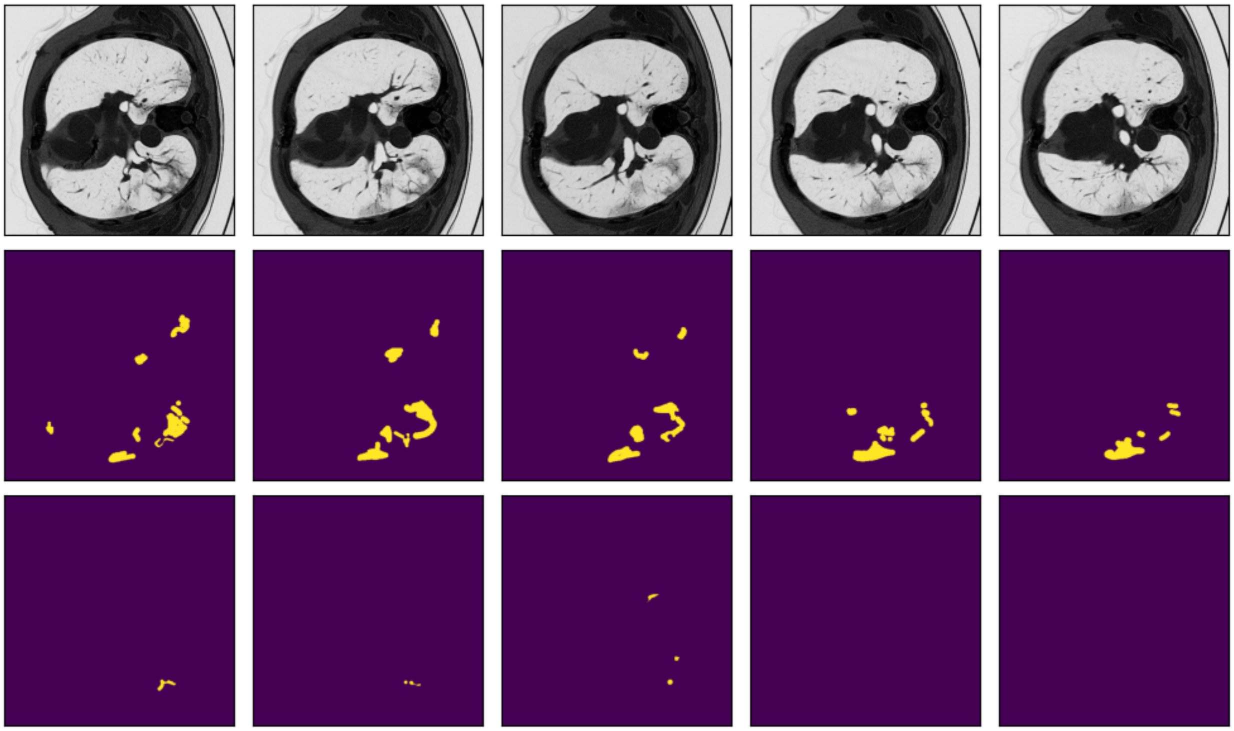


Figure 1: Sample images from Radiopaedia. The first row shows CT scans, second row: *ground-glass* mask and the third row shows *consolidation* mask.

2.2.1 Class distribution

A significant class imbalance was observed within the dataset. In particular, many slices from the Radiopaedia subset do not contain any positive annotations. Of the 829 total slices, only 373 have at least one non-zero region in either of the target channels (ground glass or consolidations). Ground glass opacities were found to be more prevalent than consolidations.

The lack of positive annotations comes from the fact, that not all CT scans show lungs area, but other parts of human chest, so there was nothing to annotate (please refer to the Figure 3).

This imbalance highlights the importance of careful sampling during model training and validation, as well as the potential need for augmentation or weighting strategies during loss computation.

2.2.2 Visual inspection

The samples vary in how they look like, even for an untrained eye. The MedSeg part of images looks like the scans were taken in different conditions. They scale and shape of lungs is different for different samples. The lungs might fill almost entire image, or only around half of it. The sample of MedSeg images in shown in Figure 2.

The Radiopaedia scans look like taken in exactly same conditions, because each scan covers approximately the same portion of an image. However, majority of them doesn't show any lungs area, which can be seen in the Figure 3. Because they don't contain lungs,

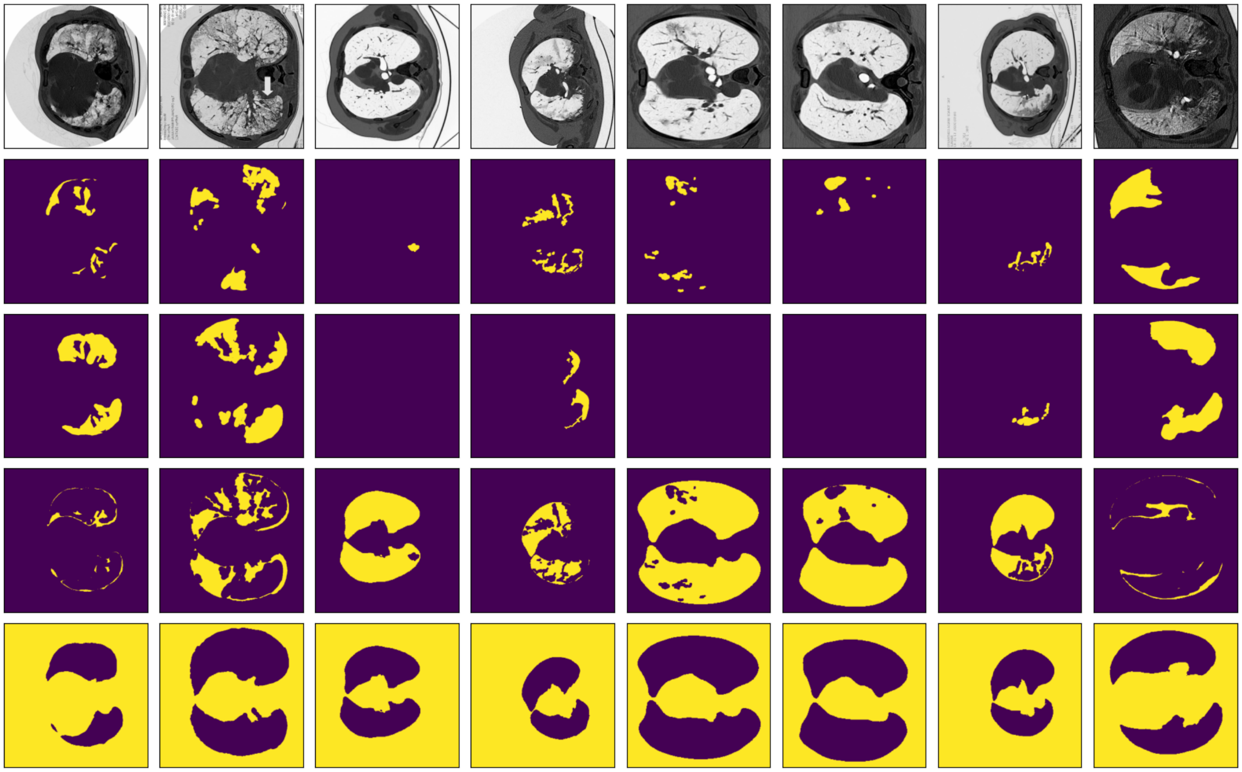


Figure 2: Sample images from MedSeg. The order of rows is the same as in every other next figure: CT scans, ground-glass, consolidation, lungs other and background. A reader can see that the scans vary in shape and scale.

they also can't contain target classes annotations.

On the other hand, samples from Radiopaedia which do contain some positive annotations, look much more *clean* than MedSeg images. The samples look like taken in identical conditions, which can be seen in Figure 4.

Lastly, let us see test images, which will be used to get the score on Kaggle leaderboard. There are only 10 of them, so it is not difficult to investigate them all. All the test images come from MedSeg part of the dataset and unfortunately they inherit all the issues from this dataset: they vary in shape, rotation and even contrast (Figure 5).

3 Evaluation metrics

The metric is imposed by the Kaggle competition and it is *pixel-wise F1 score*. You can think about image segmentation as a binary classification for each pixel: it might either belong to a given class or not. F1 score is a metric used in classification and it is a harmonic mean of precision and recall:

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2TP}{2TP + FP + FN}$$

where TP means True Positive, FP: False Positive and FN: False Negative.

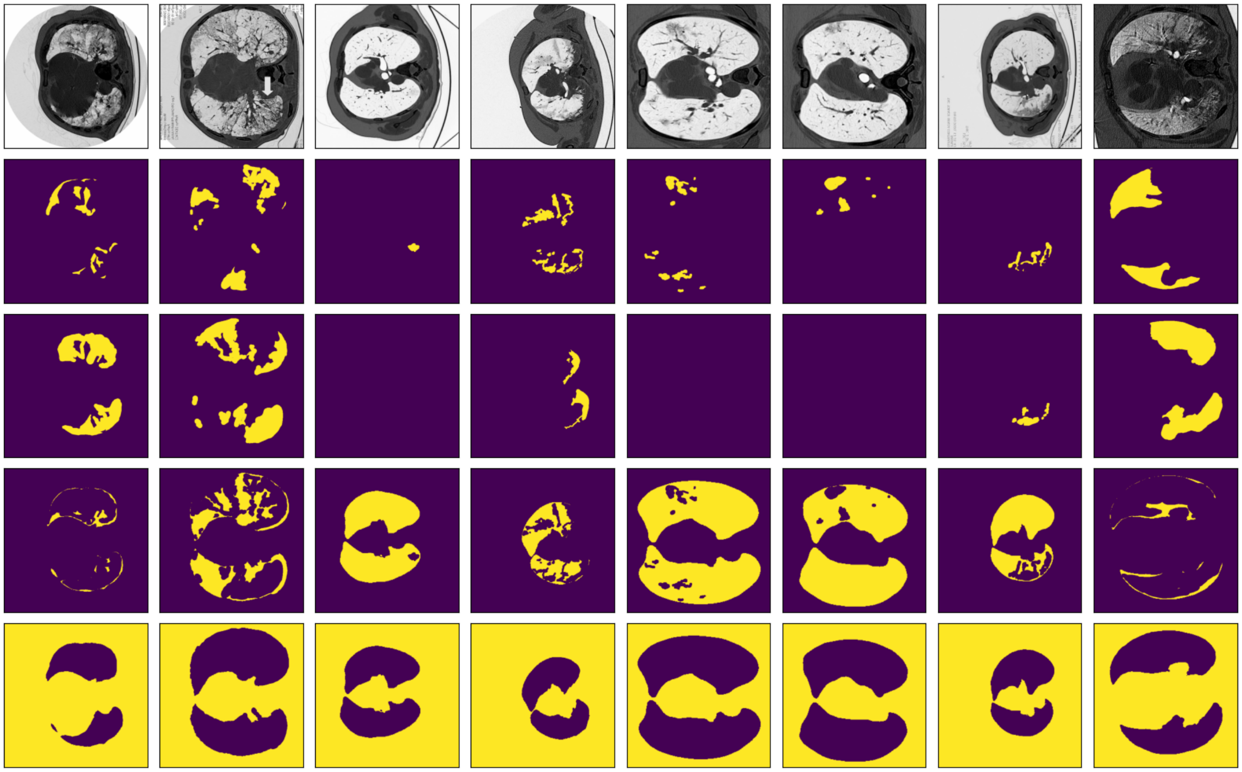


Figure 3: Sample images from Radiopaedia, which doesn't contain any lungs, or contain only very small portion of it. It can be seen that masks corresponding to *lungs other* or completely empty for first 4 samples.

In the context of image segmentation, the pixel-wise F1 score is sometimes called a Dice coefficient (or overlap index) and can be alternatively defined as

$$DICE = \frac{2|S_g \cap S_p|}{|S_g| + |S_p|}$$

where S_g is the ground truth image and S_p is the predicted image. The Dice coefficient is the most used metric in validating medical volume segmentations [8].

Because the model needs to predict two classes (*ground glass* and *consolidation*), the score is averaged for both of them. The 10 test images will be used to calculate the Kaggle score and position myself in the Kaggle leaderboard.

3.1 Benchmark model

Because this is a Kaggle competition, then quite a natural benchmark is to compare my solution to the competition leaderboard. At the time of writing, the best model achieved a score of 0.73301. Of course, I cannot claim at this moment that I will beat the best model in the competition.

In the competition's "Code" section there is also a notebook showing a baseline model written in Keras [5]. It trains an out-of-the-box U-Net model. It achieves a score of 0.64561.

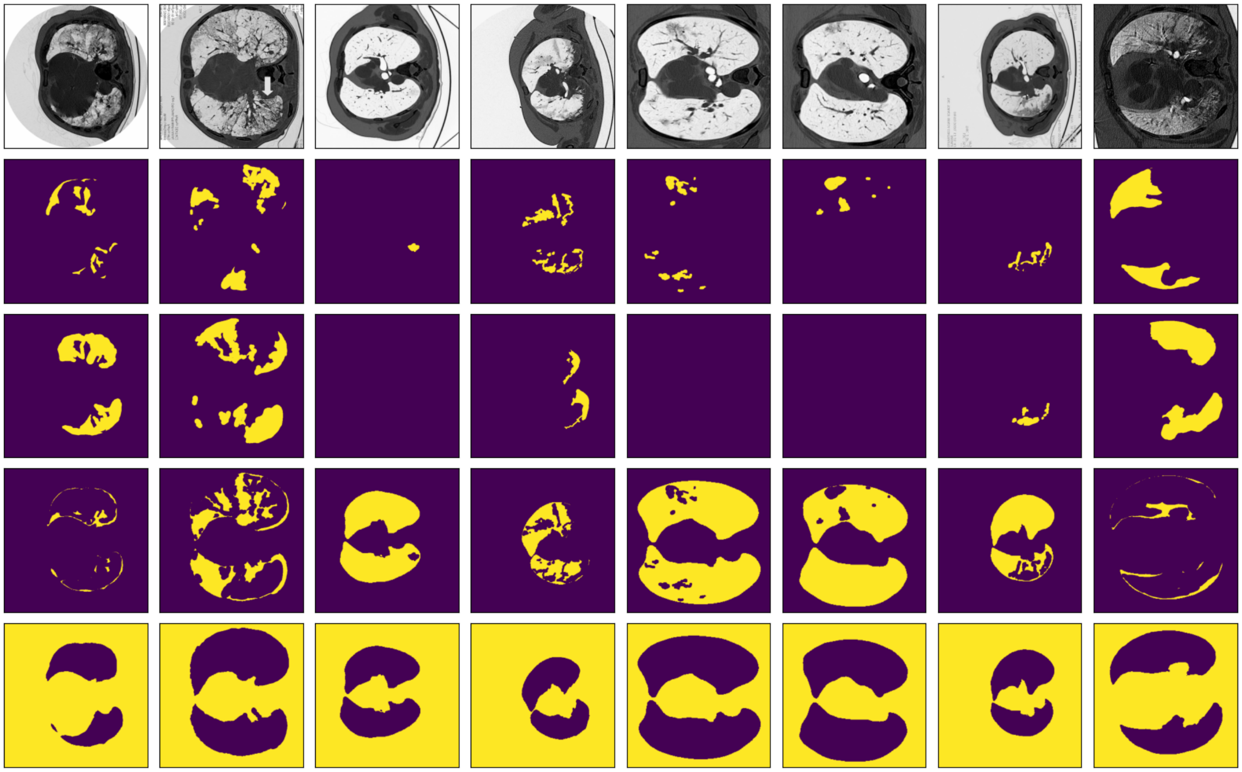


Figure 4: Sample images from Radiopaedia, which does contain positive annotations.

4 Proposed algorithm to tackle the problem

In order to predict masks for computed tomography scans I will train a deep learning computer vision model. Because this is a problem of image segmentation, I will use U-Net architecture. This architecture was introduced in the paper *U-Net: Convolutional Networks for Biomedical Image Segmentation* [7] for segmentation of neuronal structures in electron microscopic stacks. The authors showed that such a network can be trained end-to-end from very few images. They used training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently. It is also vital in this project, as I only have a few hundreds CT scans.

The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. See Figure 6.

5 Solution

5.1 Data preparation

5.1.1 Removing irrelevant samples

As described in the section *Class distribution*, not all samples show any lung area nor annotated target classes. I drop all such samples which do not contain any *ground glass* or *consolidation*. This leaves me with 472 samples.

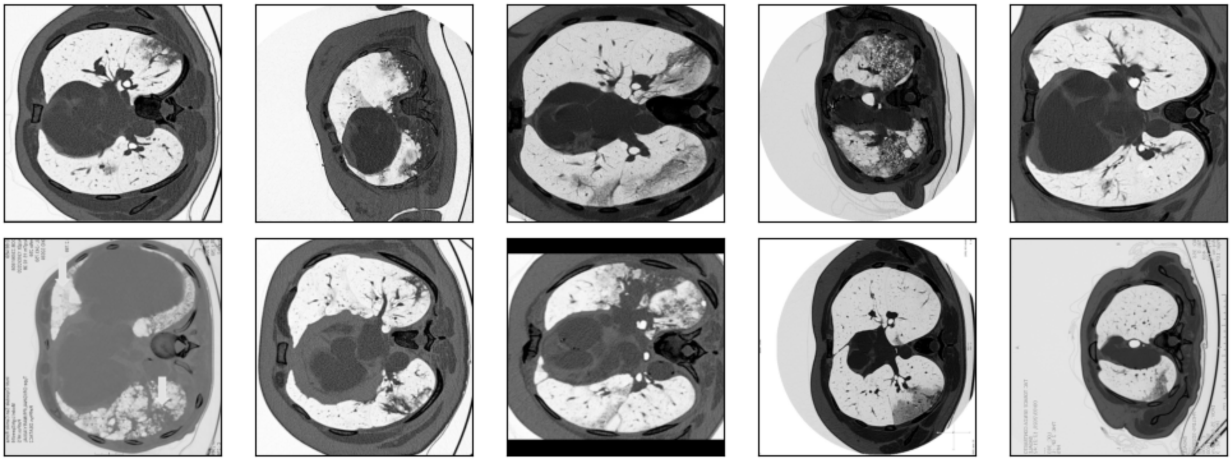


Figure 5: Test images in the Kaggle competition.

5.1.2 Data split

To ensure an unbiased evaluation of model performance, the dataset was partitioned into training and validation subsets. As described later, the validation set will be used to reduce a learning rate on plateau and to early stop a model which is not improving.

Given that the Kaggle test set is derived from the MedSeg dataset, the validation set was also sampled exclusively from MedSeg, to maintain consistency between the validation and test sets.

5.1.3 Image normalization

The pixel values were originally in Hounsfield units [3], from -1606 to 598. The Hounsfield units describe radiodensities of different materials, like air, water or bones. CT image intensities were normalized to $[0, 1]$ values. As medical CT scans vary in brightness and contrast, normalization helps standardize the input distribution. Before normalization, the values above 500 were clipped to 500 and values below -1500 were clipped to -1500. This process is suggested in the competition's baseline process and helps to focus on radiodensities relevant to the problem of detecting Covid-infected lungs areas.

5.1.4 Uploading the data

After processing, the data is saved as numpy arrays `.npz` and uploaded to an S3 bucket.

5.2 Implementation

5.2.1 The model

As mentioned in *Proposed algorithm to tackle the problem* I use the U-Net architecture, because it is a popular choice for CT scan segmentation, including lung segmentation in COVID-19 patients [2]. However, in contrast to the very first U-net described in [7], I use a network with a pre-trained backbone. Such a solution is already implemented

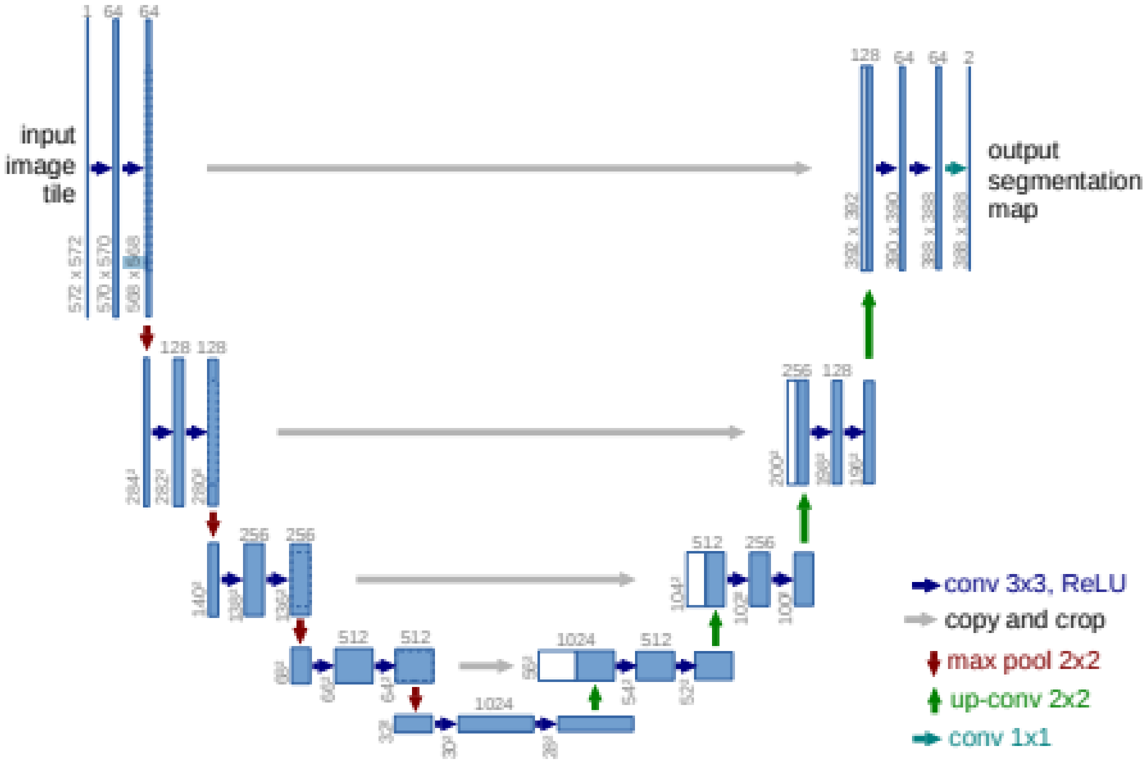


Figure 6: The original U-Net architecture [7]

in a PyTorch framework called Segmentation Models (`https://github.com/qubvel-org/segmentation_models.pytorch`). Additionally, I use `Lighthing` library, which facilitates working with PyTorch code by removing some boiler plate code.

Using a pre-trained model should allow for faster training and possibly also better results, given a small amount of training data. I use the `efficientnet-b0` [9] as the backbone, because it is a small yet powerful network.

Loss function requires some special consideration. One can think about image segmentation as a binary classification for each pixel. Because there are multiple binary masks here, it's actually a multi-label classification. On the other hand, the metric which is used in the competition is F1 pixel wise score, also called Dice. That is why I decided to use hybrid loss:

$$L_{total} = \alpha \cdot L_{BCE} + (1 - \alpha) \cdot L_{Dice} \tag{1}$$

I use $\alpha = 0.5$, which means that I give equal weight to Binary Cross Entropy loss and Dice loss.

5.2.2 Data augmentation

Because the dataset is small, I use a lot of data augmentation. Data augmentation is often used in computer vision when there is a limited amount of data and also the authors of the original U-Net paper claimed that data augmentation helped them to achieve strong results on medical data [7]. I used the following transformations:

- horizontal flipping,
- elastic deformation (as used in the paper [7]),
- affine transformations (scale, rotate, translate, shear),
- random changes of brightness and contrast,
- adding Gaussian noise,
- random cropping and scaling.

5.3 Refinement

I dedicated substantial effort to refining the model architecture and training configuration. The development process involved iterative experimentation and evaluation across several segmentation models and hyperparameter settings.

I used and evaluated the following architectures:

- U-Net (baseline model),
- U-Net++, which introduces nested and dense skip connections,
- Feature Pyramid Network (FPN), designed to enhance semantic representation across scales.

I tested each model with EfficientNet-B0 and EfficientNet-B1 backbones. I used a learning rate schedule, which halves the learning rate on plateau. I tried different learning rates (from 0.01 to 0.0001) and different α[1]: 0.25, 0.5 and 0.75. Also, as mentioned above, extensive data augmentation was also implemented.

These outcomes were logged with a Tensorboard logger.

6 Results

6.1 Final Model Performance

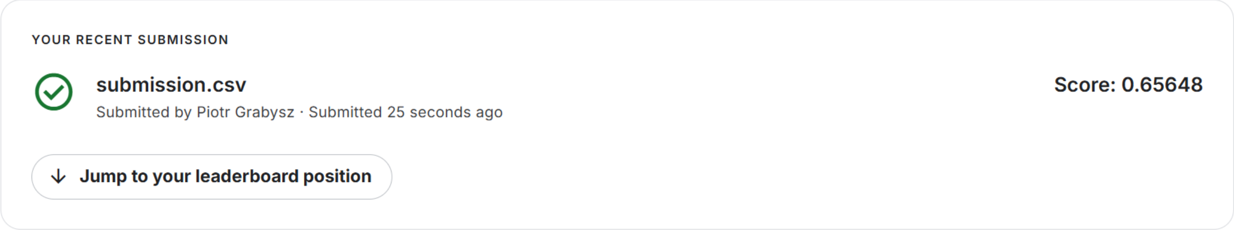
My best performing model achieved 0.65648 Dice score (averaged pixel-wise F1 score) on the Kaggle leaderboard. This model used:

- Feature Pyramid Network (FPN) architecture
- EfficientNet-B1 encoder based backbone
- learning rate = 0.001
- $\alpha = 0.5$

[1] Weighting impact of Binary Cross Entropy Loss and Dice loss

I set the number of epochs to 50, but the training stopped after 30 epochs because of early stopping configured to stop after 5 epochs without improvement.

The results on the validation set of the model's best epoch are shown in the Table 1. The Figure 8 shows the Dice score on training and validation sets over epochs. The Figure 9 shows ground-glass and consolidation masks predicted for the test data.



YOUR RECENT SUBMISSION

submission.csv

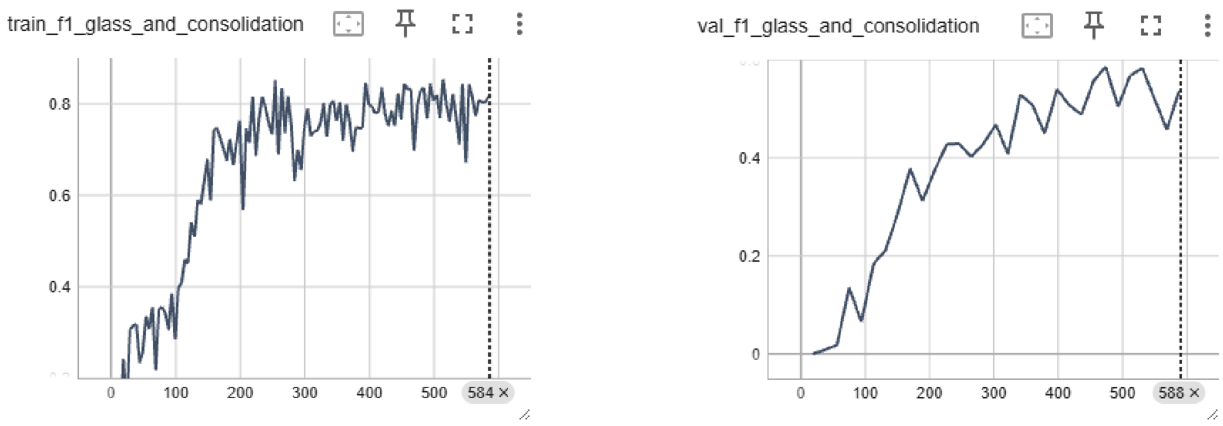
Submitted by Piotr Grabysz · Submitted 25 seconds ago

Score: 0.65648

Figure 7: My submission to Kaggle competition.

Table 1: Performance metrics for each target classes on the validation set.

Class	F1 Score	Precision	Recall
Ground-glass	0.6638	0.6079	0.7309
Consolidation	0.5080	0.6333	0.4242
Ground-glass + Consolidation (avg.)	0.5859	0.6206	0.57755



(a) Training Score

(b) Validation Score

Figure 8: Training and validation Dice score across epochs.

6.2 Interpretation and significance

The final model outperforms the "keras baseline" [5], which achieved score of 0.64561, only by a tiny margin. However, this is still an improvement. Moreover, the project contributes meaningfully by:

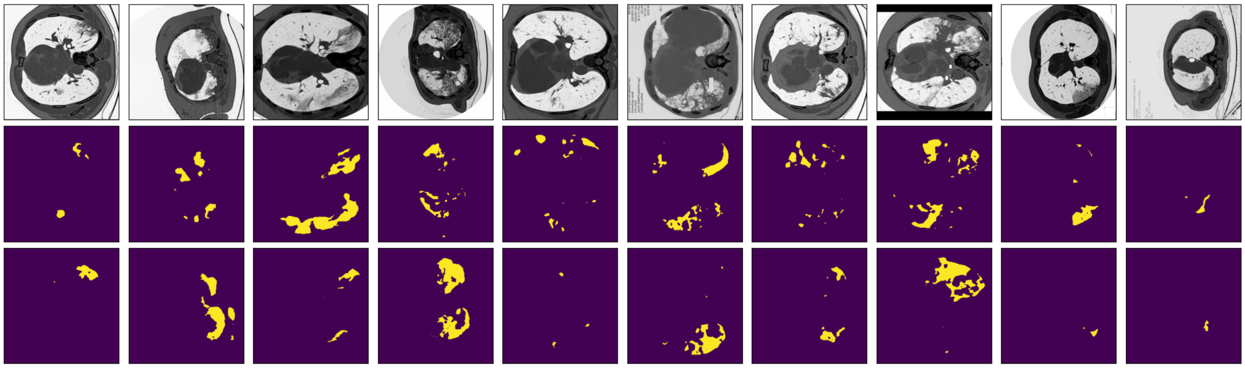


Figure 9: Ground-glass and consolidation masks predicted for the test data.

- Demonstrating a reproducible and modular segmentation pipeline,
- Exploring and evaluating multiple architectures,
- Identifying key limitations in the dataset and training stability.

The solution can be considered a solid proof of concept. Future improvements may benefit from transfer learning from related medical datasets or annotating additional datasets.

References

[1] https://medium.com/@hbjenssen/covid-19-radiology-data-collection-and-preparation-for-artificial-intelligence-4ecece97bb5b.

[2] Zhang F. "Application of machine learning in CT images and X-rays of COVID-19 pneumonia". In: *Medicine (Baltimore)* (2021). DOI: 10.1097/MD.0000000000026855.

[3] *Hounsfield unit*. https://radiopaedia.org/articles/hounsfield-unit. Radiopaedia.

[4] Igor.Slinko. *COVID-19 CT Images Segmentation*. https://kaggle.com/competitions/covid-segmentation. Kaggle. 2020.

[5] *keras baseline*. https://www.kaggle.com/code/yellowduck/keras-baseline. Kaggle.

[6] https://radiopaedia.org/articles/covid-19-4?lang=us.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: https://arxiv.org/abs/1505.04597.

[8] Hanbury A Taha AA. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool." In: *BMC Med Imaging* (2015). DOI: 10.1186/s12880-015-0068-x.

[9] Mingxing Tan and Quoc V. Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020. arXiv: 1905.11946 [cs.LG]. URL: https://arxiv.org/abs/1905.11946.

[10] https://en.wikipedia.org/wiki/CT_scan.

[11] https://en.wikipedia.org/wiki/Ground-glass_opacity.

[12] https://en.wikipedia.org/wiki/Pulmonary_consolidation.