CIS 522 – Final Project – Technical Report

Team DL^2

April 2022

Team Members:

- Lavnik Balyan; lavnikb; Email: `lavnikb@seas.upenn.edu`
- Dan Gallagher; gallagd; Email: `gallagd@seas.upenn.edu`
- Leontij Potupin; potupin; Email: `potupin@wharton.upenn.edu`

Abstract

The influx of patients resulting from the Covid-19 pandemic has overloaded hospitals around the globe. A key piece of information for both treating and triaging Covid patients is the progression of infection in the lungs, which can be examined via chest CT. Traditionally, this has been marked by radiologists, but is a time-consuming process for already overworked doctors. We investigate the efficacy of different deep learning models to automate the segmentation of COVID-19 infection in the lungs. We utilize both the commonly-seen U-Net architecture and less common SegNet architecture, attempting to improve upon prior literature by increasing model depth, using Bayesian optimization to tune hyperparameters, and augmenting images. We find that both U-Net and SegNet far outperform k-means for segmentation, and that both seem to improve from a deeper architecture. Additionally, while a basic SegNet architecture outperforms a basic U-Net, a deeper U-Net architecture resulted in our highest performing model. Our results show that relatively simple modifications can improve the capacity of relatively common architectures to segment COVID-19 infection in the lungs. This could help lay the groundwork for a tool to ease the burden on overworked radiologists and assist hospitals in establishing proper management plans for waves of Covid patients.

1 Introduction

In the past two years, the COVID-19 Pandemic has ravaged countless lives and has overloaded hospital systems across the globe. Doctors who were, in many cases, already overworked, were now tasked with handling a massive influx of patients in a highly infectious and stressful environment. Even now, though many governments are relaxing COVID-related restraints, hospitals are still being pushed to the brink. Deep learning, which excels at automating the interpretation of data that might take humans arduous amounts of time to process, might be able to ease this burden. One such way to do so is through the automation of infection segmentation. By examining a CT of the lungs, a doctor can get a sense of how severe a case of COVID is, how it might progress, and what treatment options may work best. However, actually marking which areas of the lung are infected is a tedious and time-consuming process. By instead using deep learning to segment infected and non-infected areas of the lungs, we can greatly reduce the stress placed on radiologists.

2 Related Work

Much of the literature focusing on the segmentation of medical images in the aftermath of the Covid-19 pandemic has focused on the U-Net architecture after its first introduction (Ronneberger et al., 2015), as suggested by Müller et al. (2021) or Diniz et al. (2021). However, Saood and Hatem (2021) found that, for the binary segmentation problem, a SegNet architecture outperformed U-Net, potentially due to its strength in handling imbalanced class representations. However, their implementation of SegNet left several areas of potential improvement, such as the use of image augmentation and Bayesian hyperparameter tuning as well as the introduction of different encoder and decoder architectures. Our goal is to use these changes to see if we can improve the performance of SegNet on segmenting COVID infections in lung CTs, and explore if such changes are enough to allow SegNet to surpass U-Net in the multi-class segmentation problem in addition to the binary one.

Furthermore, prior research has shown that biological and non-biological factors such as patient age, slide preparation date, slide origin, and scanner type can create batch effects that can cause problems for the training of DL models (Schmitt et al., 2021), as they fail to capture the true underlying structure of the CT scan. The possibility of batch effects highlights the importance of preventive methods like normalization, prepossessing and augmentation, especially when working with a limited data set.

3 Dataset and Features

Our dataset, sourced from MedSeg, consists of 100 axial CT scans from 48 COVID-positive patients. The scans themselves originate from the Italian Society of Medical and Interventional Radiology, and were pre-processed and converted to NIFTI files based on a process by radiologist Dr. Håvard Bjørke Jenssen. Each image came with a corresponding mask, on which a radiologist marked the presence of three different types of abnormalities that are often related to COVID-infection: ground-glass, consolidation, and pleural effusion.

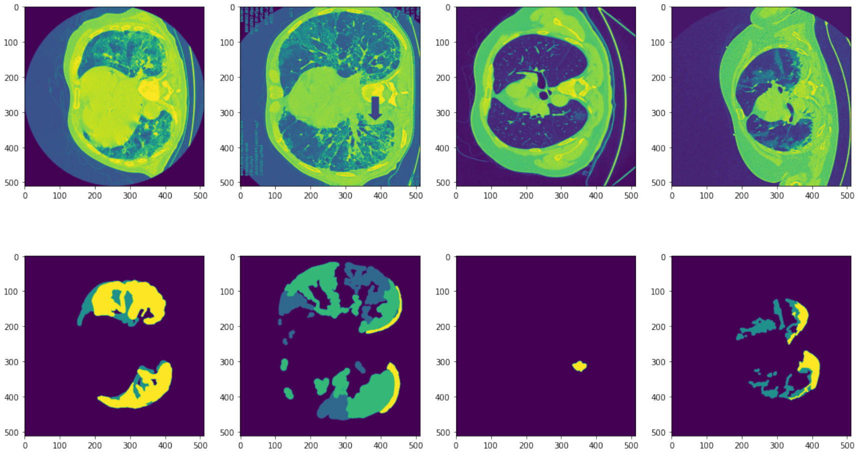

Figure 1: Raw data and corresponding masks

The images, which had lost a part of their upper-intensity range in the conversion from DICOM to JPG, needed to be normalized to the Hounsfield Unit scale, which is a dimensionless scale used to express the radiodensity of different parts of a CT in a standardized form. After being resized to 512 x 512 pixels and greyscaled, the images were normalized back to the Hounsfield Unit scale by taking RGB-values from areas of air (either externally from the patient or in the trachea), which was normalized to -1000, and fat (subcutaneous fat from the chest wall or pericardial fat), which was normalized to -100.

Afterward, the images were cropped by 20 pixels on each side, both to reduce the amount of irrelevant information (often present on the borders of CTs) and to reduce data imbalance, as cropping removed mostly non-infected pixels. The images were then resized again to 256 x 256 pixels, both to increase computation speed and align more closely with Saood & Hatem (2021) for more accurate comparisons. Utilizing a trick from Müller, Soto-Rey, and Kramer (2021), we then clipped pixel intensity values to fall between -1250 and +250 HU. This was an attempt to exploit the fact that we are only interested in certain HU ranges- infected regions (+50 to +100 HU) and lung regions (1000 to 700 HU), and thus could make our data easier for our model to manage without losing any relevant information. Finally, the images were normalized to a [0,1] range to avoid very large numbers being passed in our gradients.

The image masks were converted into both binarized masks (all abnormalities being grouped into a single label to produce a segmentation of infected vs non-infected portions of the lung) and a one-hot encoded mask of all four classes.

This is because we were interested in both classification schemes. Firstly, examining both schemes allows us to more accurately compare with Saood and Hatem (2021). Secondly, the different schemes can be useful in different situations, with the binary scheme providing doctors a quick and simple overview of where in the lung COVID may be spreading and the multi-class scheme providing a more in-depth but potentially less accurate view of the types of abnormalities in the lung, which can be useful in determining proper management strategies.

Finally, the data was split according to Saood and Hatem (2021), once again in order to more accurately compare results. This meant we randomly split our data into 72 training images, 18 test images, and 10 validation images.

Before running our models, the training data was augmented in an attempt to reduce overfitting and artificially increase the size of our small dataset. Augmentation was applied in a manner similar to Müller, Soto-Rey, and Kramer (2021). We separated our augmentations into spatial and non-spatial transforms, where the former were applied identically to the images and corresponding masks and the latter were applied only to images, as non-spatial transformations would fundamentally alter the meaning of various label values. Our spatial transformations included rotations, horizontal flips, scaling, and elastic deformations, while our non-spatial transformations consisted of gamma alterations and the addition of Gaussian noise. These transformations were applied in an on-the-fly style during training with various probabilities, which was meant to allow our model to still ”see” many more images while also reducing the probability of the model seeing the same image twice, and thus correspondingly reducing the risk of overfitting (Isensee, Jaeger, Kohl, Petersen, & Klaus, 2020). Surprisingly, this step was seemingly omitted from Saood Hatem’s implementation of SegNet, so we hope to use it to further bring out the architecture’s strengths.

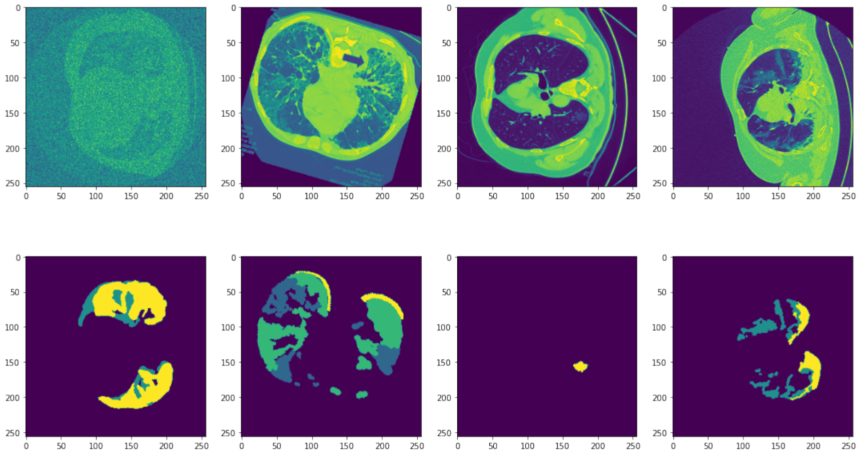

Figure 2: Augmented data and corresponding masks

4 Methodology

Our non-deep learning benchmark for the medical image segmentation task was the K-Means algorithm which partitions n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

As our base and advanced Deep Learning models, we decided to use the standard UNet and SegNet architectures and then modifications/improvements to these models. We decided to do this because UNet is an industry/research standard and SegNet has been recently proposed as an improvement over UNet in some scenarios. A few papers have also mentioned that these models could be improved by making them deeper, a study of this forms the bulk of our advanced work.

This improvement likely occurs because CTs of COVID-infected lungs present an environment with a highly imbalanced representation of classes; the non-infected and background pixels are highly dominant. SegNet, originally designed for segmenting driving environments for use in self-driving cars, operates on similarly imbalanced data, with the images often dominated by road and sky. The architecture's ability to handle an imbalanced segmentation environment may thus allow it to surpass U-Net in the segmentation of certain medical images, like COVID lung CTs.

UNet is a convolutional neural network architecture that was developed for biomedical image segmentation. UNet consists of a contracting path and an expansive path- the contracting path is a typical convolutional network that consists of repeated application of convolutions, each followed by ReLU and max pooling whereas in the contraction part, the spatial information is reduced while feature information is increased. The expansive pathway combines the feature and spatial information through a sequence of up-convolutions and concatenations with high-resolution features from the contracting path.

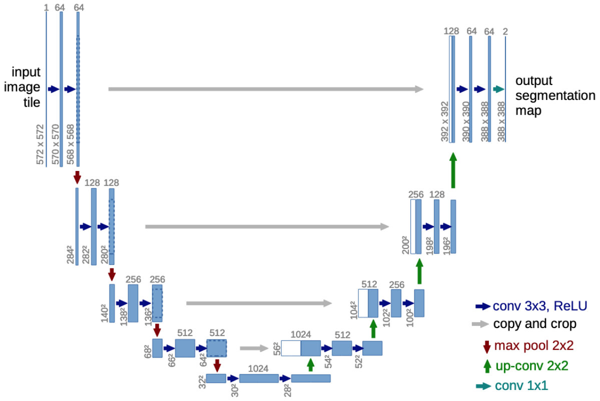

Figure 3: U-Net Architecture

SegNet is a deep fully convolutional neural network architecture for semantic pixel-wise segmentation that was initially developed for self-driving vehicle applications but has also been used for medical image segmentation in more recent research work. SegNet consists of an encoder network, a corresponding decoder network followed by a pixel-wise classification layer. The architecture of the encoder network is structurally similar to the VGG16 network. The role of the decoder network is to map the low resolution encoder feature maps to full input resolution feature maps for pixel-wise classification.

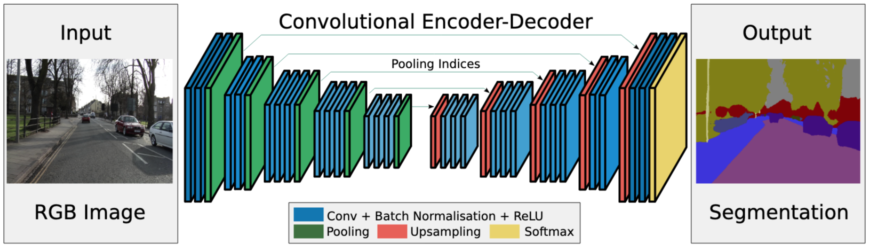

Figure 4: SegNet Architecture

We modify the base U-Net and SegNet models by adding additional encoder and decoder blocks to both models in the hopes that the addition of these blocks would lead to the model being able to better segment these CT scans. Our reasoning was that additional encoder and decoder blocks would allow our models to develop a more powerful and nuanced representation of the images, which would in turn allow the models to circumvent the ”simple” option of classifying everything as the extreme majority and instead learn the actual underlying patterns distinguishing non-infected and infected lung.

To supplement our models, we also included several innovations in training.

To address the imbalance in our classes, we modified Focal Tversky Loss to use as our loss function. Focal Tversky Loss, defined as

$$(1 - \frac{TP}{TP + \alpha FN + \beta FP})^{\gamma},$$

is designed to better capture minority classes in imbalanced datasets (Abraham Khan, 2019). By setting alpha ¿ beta (in our case ¿ .5 as alpha + beta always equal 1), we can penalize false negatives more strictly, which should hopefully help prevent our model from learning to simply classify everything as non-infected.

However, when applying this loss to the multiclass case, we noticed that, while we saw marked improvements in our models' sensitivty toward at least one infection class, it would still often fail to classify multiple types of infection classes. In an attempt to be able to further distinguish how we weighted each class, we attempted a relatively simple modification: calculate a Tversky loss for each class separately as a binary problem with its own alpha weight, then average the losses together. This allows us to set separate penalties on false negatives for each class, which could hopefully give us a greater sense of control in steering our model to predict minority classes. While some COVID segmentation uses traditional Tversky Loss, none that we could find extended it to a more modular multi-class setting, which we think could be critically important as our data contains not only an imbalance between non-infection and infection classes, but an imbalance between the different infected classes.

Additionally, rather than using GridSearch to tune our hyperparameters, we opted to use a Bayesian tuning scheme (Koehrsen, 2018). Instead of looping through a pre-selected array of values, Bayesian tuning samples from a distribution of values and tries to find the optimal hyperparameter within that distribution. It does so by attempting to create a probability model (surrogate) that maps hyperparameter values to probabilities of scores on the objective function. We maximize our expected improvement on this surrogate rather than the objective function itself, and evaluate the surrogate on the actual data. Thus, we use the knowledge of how our previous surrogate did to help build our new surrogate, and essentially help guide our search for optimal hyperparameters. The advantages of this are twofold. First, by using our past knowledge to guide our current hyperparameter choices, we can converge on an optimal hyperparameter faster, since we can call the objective function- which, for our advanced neural networks, are quite computationally expensive- fewer times. Secondly, our options for hyperparameter values becomes much more granular than the set points of a GridSearch. As far as we could tell, this method has not been used in current COVID segmentation literature, and we hope that the increased granularity of this tuning method will further improve upon prior literature.

We focused mainly on tuning initial learning rate, as it could have a monumental impact on where our models converged. We initially also tuned the loss function option between BCEDiceLoss and Tversky Loss to see which would most improve our ultimate dice score.

5 Results

Our main metric to evaluate performance of our models was the Sørensen–Dice coefficient, represented by

$$\frac{2|X \cap Y|}{(|X| + |Y|)}.$$

This coefficient measures the similarity between two sets, and is thus often used to evaluate image segmentation.

However, given the imbalanced nature of our data, it is important to evaluate not just the level of similarity between prediction and ground truth, but the sensitivity of our model to the minority classes (and, similarly, the specificity to the majority non-infected class). Sensitivity is defined as (True Positive)/(True Positive + False Negative), while Specificity is (True Negative)/(True Negative + False Positive). In other words, Sensitivity measures the proportion of correctly classified positives and Specificity the proportion of correctly classified negatives. Ideally, in addition to a high dice coefficient, a good model will have a relatively high Specificity for non-infected portions of the lung as well as relatively high Sensitivity for each infection class, as otherwise it may simply be predicting everything as non-infected.

We begin our results with our non-deep benchmark of K-Means clustering. Since K-Means is a unsupervised learning algorithm, and due to known issues with the algorithm like trouble clustering data where clusters are of varying sizes and density, as well as the dependence on initial values, we predicted poor prediction performance. Consistent with our initial hypothesis, the results of K-Means were non-significant with a dice score of close to zero.

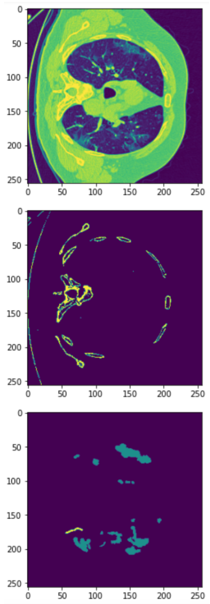

Figure 5: K-Means prediction with medical image, K-Means segmentation and ground truth from top to bottom

As we can see, the k-means algorithm clustered prevalent differences in the images, such as color, but due to lack of supervision did not identify the patterns in which we held actual interest. In other words, while k-means succeeded in finding structure, its unsupervised nature meant the structure it found was largely irrelevant to our task.

We expected the base UNet model to perform better in terms of the validation and test dice scores (Sørensen–Dice coefficient) for the multi-class problem, as was mentioned in Saood Hatem (2021). Additionally, we expected that the deeper U-Net and SegNet models would perform better in terms of the dice scores than the base models.

We ran our models using the Multi Focal Tversky Loss (to try and help with the class imbalance in our data) and using a custom RAdam optimiser.
Due to limited access to GPUs (Colab cuts access after a certain amount of use), we were only able to run a limited number of experiments.

The results for the experiments below were run on augmented data and tuned using Bayesian tuning, with subsequent versions of each model being deeper than the former.

Dice scores and background specificity

Model	Validation Dice Score	Test Dice Score	Non-infected Specificity
UNet base	0.878	0.904	0.305
UNet2	0.726	0.750	0.317
UNet3	0.875	0.896	0.653
SegNet base	0.882	0.909	.042
SegNet2	0.835	0.861	.031
SegNet3	0.875	0.896	.470

Sensitivity toward different classes

Model	Non-infected	Ground Glass	Consolidation	Pleural Effusion
UNet base	0.995	0.370	0.103	0.000
UNet2	0.997	0.193	0.326	0.000
UNet3	0.973	0.544	0.573	0.000
SegNet base	0.984	0.000	0.000	0.003
SegNet2	0.996	0.000	0.000	0.109
SegNet3	0.986	0.558	0.000	0.000

The following are examples of how each model performed, using the first 4 images of our test set, with the prediction on top and the corresponding ground truth mask below. Purple represents background/non-infection, yellow ground glass, sea-green consolidations, and the rare dark green pleural effusions.

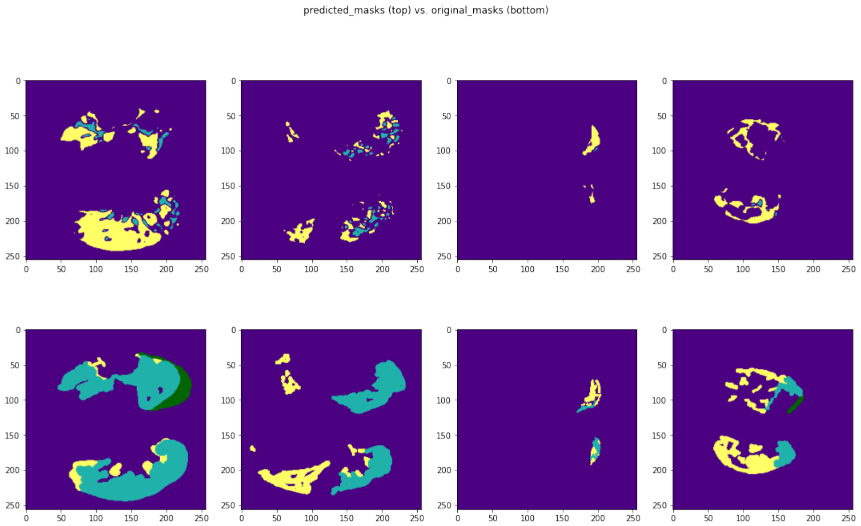

Figure 6: UNet Baseline Result

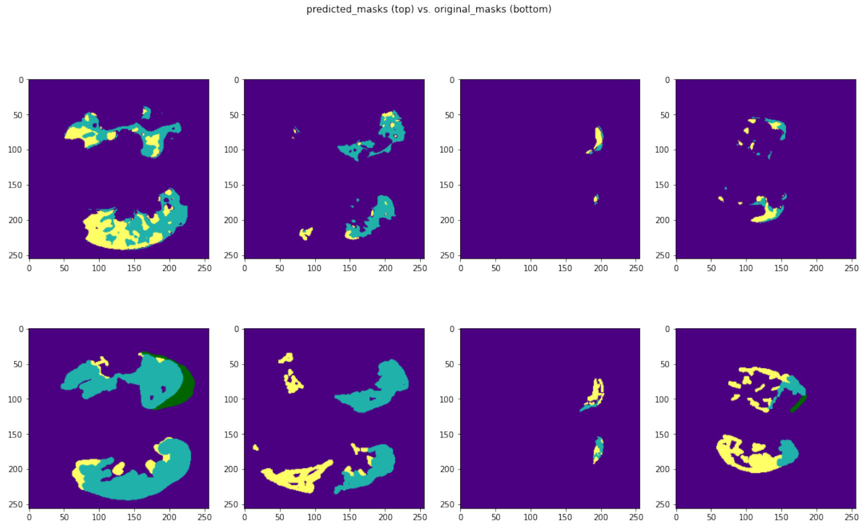

Figure 7: UNet2 Result

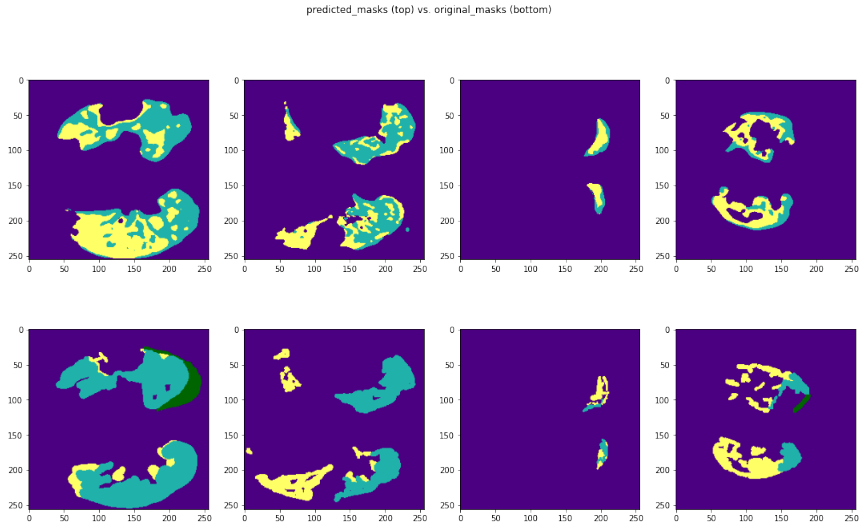

Figure 8: UNet3 Result

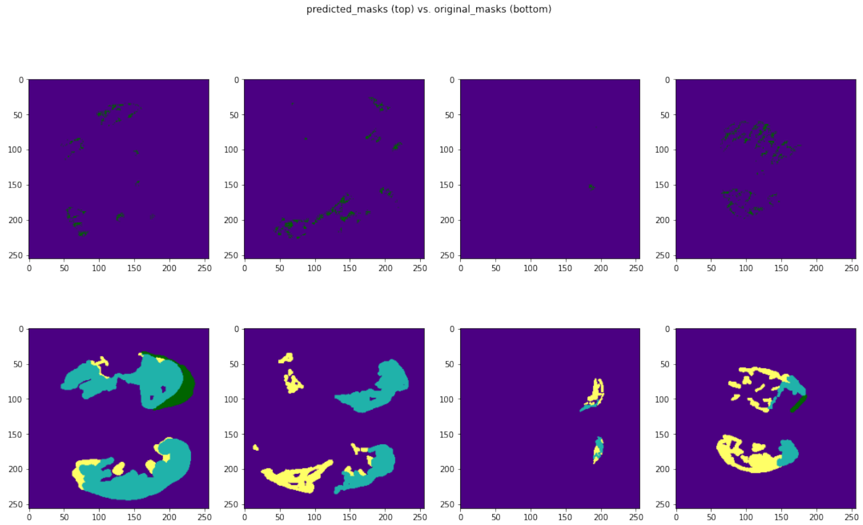

Figure 9: SegNet Baseline Result

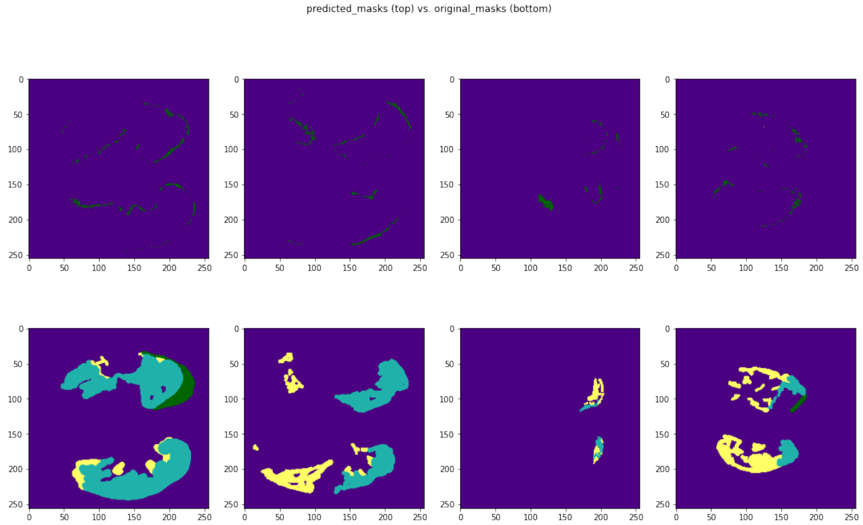

Figure 10: SegNet2 Result

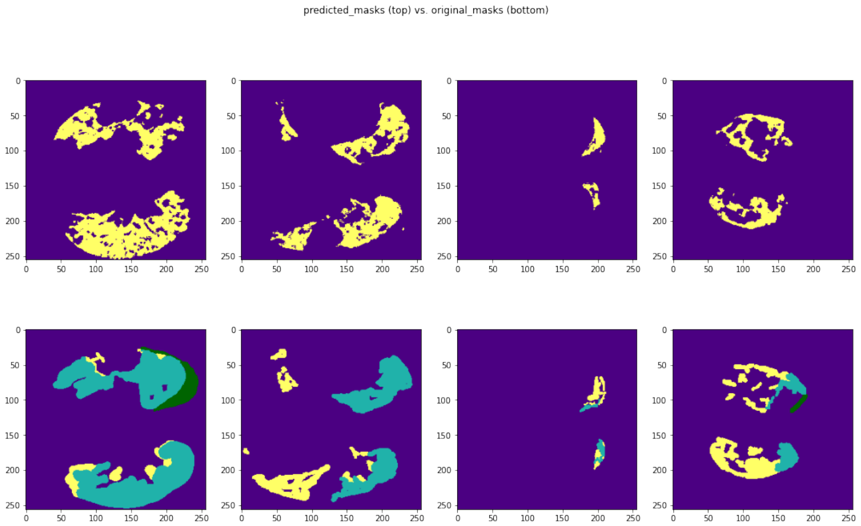

Figure 11: SegNet3 Result

Simply in terms of Dice scores, the SegNet generally outperforms UNet. However, upon examination of non-infection specificity and of infected-class sensitivity, it becomes clear that SegNet generally fails to overcome the imbalanced nature of the data, and simply predicts most pixels as non-infected.

In a similar fashion, the baselines of both models outperform their modified counterparts' dice scores, but once again perform worse in identifying actual infection. Both adding additional depth to the UNet and adding additional encoder and decoder blocks to the SegNet generally increase the models' abilities to identify infected portions of lung. The most well rounded model was our deepest version of UNet. Most of our SegNet models generally failed to capture most infected classes, though our deepest SegNet at least had more effectiveness capturing Ground Glass abnormalities.

Overall, UNet outperforming SegNet on the multiclass problem was consistent with prior literature, though given SegNet's history, we were quite surprised by its seeming inability to handle the imbalanced nature of the data. Interestingly, SegNet does seem slightly more willing than UNet to classify pixels as pleural effusions, which was an extreme minority class (UNet seems to completely ignore this class). Perhaps, with some tweaking, SegNet can be used to more effectively capture extreme minority classes.

It is also worth noting that while our best model performed fairly well, there is certainly much room for improvement.

6 Discussion

So far, we have built our Datasets with all the augmentations and other work, we have defined our loss functions and optimizers, trained and tested our models and gotten examples of our model predictions.

6.1 Findings

Our K-Means work (non-DL) was as unfruitful as we expected; the algorithm simply does not have the power to learn anything about the data and create any meaningful conclusions or predictions.

Overall, UNet generally outperformed SegNet, which often classified most of the image as non-infected. The modifications we performed on these models (making them deeper by adding more encoder and decoder blocks) tended to help improve the ability of the models to break through the class imbalance and classify infectious areas of the lungs. and this is probably because adding additional encoder and decoder blocks provides the models with greater representational power and ability to distinguish between disease vs non-disease lung areas but also between different classes of disease.

Two results were generally surprising. The first was the ineffectiveness of most of the SegNet models in classifying infectious lung. Given that, as we added more encoder and decoder blocks, the model became able to classify at least one class of infection fairly well, the issue might have simply been that our baseline SegNet was not complex enough to capture the nuances of the CTs. It is also possible that, given a binary problem of non-infection vs. infection, SegNet generally would have performed far better (as in Saood Hatem, 2021). It was also interesting that SegNet seemed more capable than UNet at classifying pleural effusions, suggesting that SegNet may perhaps be biased toward classyfing extreme minority classes. This reconciles with what we know about its origins as a classifier in extremely imbalanced environments, and perhaps suggests that, if sufficiently tweaked, can actually become more powerful than UNet by correctly segmenting highly infrequent abnormalities which UNet generally seems to ignore.

The second surprising result was how poor the Dice coefficient could be as an evaluator of segmentation in imbalanced class environments. While most papers that evaluate COVID-19 segmentation include measures of sensitivity and specificity, the metric listed at the forefront of the paper is often the Dice coefficient. However, we found that even models that perform truly horrendously (being essentially unable to make any meaningful classification) had high Dice scores, simply due to the fact that the Dice coefficient only cares about overall similarity between prediction and ground truth. Researchers working in imbalanced segmentation environments should perhaps reconsider the emphasis on the Dice coefficient, saving it for the appendix in the same way general accuracy is often calculated but not prioritized. Or perhaps the community is simply in need of a better metric to evaluate a model's overall performance, that combines the similarity measure provided by Dice with sensitivity and specificity scores for minority and majority classes, respectively.

Our work matters because medical image segmentation is a crucial area of deep learning research that has profound implications. Our work is on the cutting edge of research work being performed at the moment and the core work we are doing is based on recommendations from certain research papers. While we did not beat benchmarks, we showed that it is possible to improve upon the base UNet and SegNet models that are used by most papers in the field and we provide future steps that could potentially be used to beat the benchmarks (along with using greater computational resources).

The social impact of this work is fairly straightforward. This work could help alleviate the burden on highly specialised doctors to spend hours over single images trying to segment them by hand, where our models would spend less than a second on the same images. It would also help deliver consistent and quick results that could help with quicker diagnosis and analysis of the prognosis of certain diseases. Especially in the context of COVID-19, a lot of patients have severe lung damage and will require long-term care. Such patients would need to get regular medical imaging work done to help doctors better understand the extent of these patients' recovery, and these computational methods can help to augment doctors' work and help them deal with this large increase in volume of medical imaging.

6.2 Limitations and Ethical Considerations

There were several shortcomings to our models and experiments. To start, we found major discrepancies between the dice metric and the effectiveness of our models. In many cases, we found the dice metric did not account for the extreme imbalance in our data; models with high dice coefficients often showed staggeringly low (sometimes 0) sensitivity for some of the infected classes. Though we also included per class sensitivity scores among other metrics to account for this imbalance, the ineffectiveness of the dice metric made it difficult to compare models to each other and to other studies.

Furthermore, even after modifying our loss function specifically to handle imbalance, we still saw low sensitivity for certain infected classes across many models. The Multiclass Tversky loss function allowed us to capture places of infection (which allowed us to feel comfortable not using other methods to deal with imbalance, such as cropping, in the final models), it only partially helped us distinguish between infection types. Indeed, pleural effusions, a class with an extreme minority (on average representing a mere .001 percent of pixels) was essentially never classified.

There also may have been potential unaccounted-for batch effects introduced by differences in the specific CT machine used for each image. These may have worsened performance by introducing irrelevant information and additional biases into our data. Furthermore, since we did not attempt to harmonize our data, it is hard to say how our models will generalize, as they will then face all new batch effects from CT manufacturers and environments unseen in our small and Italy-localized dataset.

Finally, our test set was unfortunately quite small, which makes it harder to attest to the ability of our models to generalize. Some researchers have used larger datasets of volumetric CTs to bolster the size of their data, but we opted not to go for that option due to the fact that volumetric CTs are not nearly as commonly used for COVID-lung segmentation, and we did not want to base our model off a type of image that would most often not be used in practice.

These concerns lead us to a serious ethical consideration of our models: when we are wrong, it can cost people their health or even their lives. Indeed, a doctor may think a patient is healthier than they are and deny important treatment, or may simply think the type of infection is different than reality and create an incorrect management plan. The need to be sure that our model will work in a variety of environments is far greater in a medical environment than others, as the cost of failure is far larger. Given our models' limitations and the uncertainty regarding some of the evaluation metrics, we think more work needs to be done before rolling them out to actual doctors.

6.3 Future Research Directions

Many of our experiments' limitations resulted from a lack of time, compute power, and data. In fact, working with Colab GPUs meant that there was a limit to the number of experiments we could run before being locked out of GPU access. This constrained both the complexity of our models as well as the variety of our experiments.

Given more time and compute power, we would first want to get a better handle on our models' sensitivity to all three infection classes. This issue actually seems to be endemic to much of the literature. A variety of published studies either do not have class-wise sensitivity analyses or fail to achieve strong sensitivity scores for the infection classes. Indeed, Saood Hatem (2021) do not even list pleural effusions as a potential class in their sensitivity analysis due to its extreme minority. Of course, being able to accurately segment all of the various COVID-related abnormalities can be critical in determining a doctor's management plan, meaning it is important for future work to make a concerted effort to tackle the problem of our extremely imbalanced data. One option would be to further examine the weights of the multi-class Focal Tversky loss function in an attempt to increase the sensitivity of our model to minority classes, potentially using Bayesian tuning with the goal of maximizing some weighted sensitivity score (with the smaller classes being weighted higher).

We would also want to better account for batch effects. Reynolds (2022) suggests using harmonization to bring the different batches of data to the same space and remove non-biological effects. Given adequate compute power, it may be worthwhile to train a generative adversarial networks (GAN) to convert our multi-domain data into a single domain. We did not possess the resources to do this on top of our existing models, but for teams with greater resources, it is an intriguing possibility.

Another resource limitation was, of course, data. As with many subdomains within the medical imaging space, access to large quantities of segmented COVID-19 data is slim. Perhaps one of the most worthwhile contributions to the field is simply curating larger, more varied datasets. Furthermore, given the prevalence of Radiopaedia's volumetric CT dataset (Bell, 2022) among research in the field, it would be worthwhile to examine the generalizability of models based on volumetric CTs to non-volumetric CT images, which are far more likely to be used in practice.

Finally, due to time constraints, we were unable to fully examine the binary-classification problem of infection vs non-infection. While being able to identify the type of abnormality present in different parts of the lung can help doctor's better identify proper management plans, the extreme imbalance present in some infection classes makes the multi-class segmentor far less reliable. A binary segmentor would still help doctors assess how far COVID infection may be progressing in the lungs with a lower likelihood of incorrect classification. Given more time, we would conduct a thorough examination of how our innovations impacted SegNet's and UNet's performance in a binary problem, similar to the evaluations done by Saood & Hatem (2021).

7 Conclusions

In summary, we showed that deep learning methods significantly outperform conventional ML approaches like K-Means in medical image segmentation, with important applications for the detection of anomalies related to Covid-19 infections and associated social impact of alleviating the burden for radiologists. We demonstrated that our baseline UNet outperforms our baseline SegNet, while adding more decoder and encoder blocks to the UNet leads to an optimal overall performance. This finding suggests that these model architectures can be further improved to achieve better predictive performance.

8 References

1. Saood, A., & Hatem, I. (2021). COVID-19 lung CT image segmentation using deep learning methods: U-Net versus SegNet. BMC Medical Imaging, 21(1), 1-10.

2. Müller, D., Soto-Rey, I., & Kramer, F. (2021). Robust chest CT image segmentation of COVID-19 lung infection based on limited data. Informatics in medicine unlocked, 25, 100681.

3. Koehrsen, W. (2018, July 2). A conceptual explanation of Bayesian hyperparameter optimization for Machine Learning. Medium. Retrieved April 28, 2022, from https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f

4. Reynolds, M. (2022, April 5). Batch effects. Medium. Retrieved April 28, 2022, from https://towardsdatascience.com/batch-effects-c71c886ca9c5

5. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.

6. Schmitt, M., Maron, R. C., Hekler, A., Stenzinger, A., Hauschild, A., Weichenthal, M., ... & Brinker, T. J. (2021). Hidden variables in deep learning digital pathology and their potential to cause batch effects: prediction model study. Journal of medical Internet research, 23(2), e23436.

7. Ronneberger, O., Fischer, P., Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

8. Diniz, J., Quintanilha, D., Santos Neto, A. C., da Silva, G., Ferreira, J. L., Netto, S., Araújo, J., Da Cruz, L. B., Silva, T., da S Martins, C. M., Ferreira, M. M., Rego, V. G., Boaro, J., Cipriano, C., Silva, A. C., de Paiva, A. C., Junior, G. B., de Almeida, J., Nunes, R. A., Mogami, R., … Gattass, M. (2021). Segmentation and quantification of COVID-19 infections in CT using pulmonary vessels extraction and deep learning.

9. Abraham, N., Khan, N. M. (2019, April). A novel focal tversky loss function with improved attention u-net for lesion segmentation. In 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019) (pp. 683-687). IEEE.

Data from http://medicalsegmentation.com/covid19/