An Initial Study into the Feasibility of Deep Learning-Based COVID-19 Severity Classification using Point-of-Care Ultrasound Imaging

Alexander MacLean — Vision and Image Processing Lab, University of Waterloo
Ashkan Ebadi — National Research Council Canada, Montreal
Adrian Florea — Department of Emergency Medicine, McGill University
Pengcheng Xi — National Research Council Canada, Ottawa
Alexander Wong — Vision and Image Processing Lab, University of Waterloo
Email: {alex.maclean, a28wong}@uwaterloo.ca, {ashkan.ebadi, pengcheng.xi}@nrc-cnrc.gc.ca, adrian.florea@mail.mcgill.ca

Abstract

Integral to the treatment of patients suffering from severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is assessment of the severity of the illness, allowing clinicians to more effectively apply care and devise a plan of treatment. Since the workload of clinicians is high at the best of times, let alone during a global pandemic, much work has gone into creating computer-aided clinical decision support systems, often enabled by deep learning tools. Previous work has investigated the ability to identify COVID-19 positive patients from point-of-care ultrasound (POCUS) images, but decision support systems for POCUS-based COVID-19 severity stratification have not yet been presented. In this study, we examine the feasibility of using a deep learning neural network architecture to classify POCUS images from an open source repository into distinct severity levels based on annotations from an experienced doctor of emergency medicine, hopefully leading to the implementation of such a system into a real-world clinical workflow.

1 Introduction

During the course of the COVID-19 pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), many attempts have been made to apply technology, and in particular deep learning-based Machine Learning (ML), to aid clinicians in diagnosing and treating the disease. Much work has gone into analyzing chest x-ray (CXR) [1] and computed tomography (CT) [2] imaging modalities, but recent work has begun to expand into ultrasound (US) imaging as well [3]. In particular, point-of-care ultrasound (POCUS) has started to create significant impact in care for COVID-19, largely due to its ease-of-use, cost, and lack of ionizing radiation in comparison to the other aforementioned imaging modalities, allowing it to be of use in lower resource contexts [4, 5].

One challenge encountered when dealing with POCUS images in the context of creating ML tools is the quality of the data, and in particular confidence in the consistency of the labels. Unlike the cases of CXR and CT, which often have large datasets released by administrative bodies where not only are images generally collected using consistent protocols, the publicly available dataset used in recent work [3], COVIDx-US [6], is based off of a collection of smaller datasets from research publications, releases from POCUS device companies, and radiology education sites brought together into a single location.

A recent update to the COVIDx-US dataset added a new set of labels for the entirety of the data available through that source that simultaneously adds more detail to the dataset by providing severity scores, instead of simple COVID-positive and -negative labelling, and ensures that the labels are applied to all cases via a consistent, structured method [6]. Severity assessment is critical to the workflow of clinicians, enabling proper allocation of resources and development of effective treatment plans, and previous work has shown the effectiveness of deep learning-based ML systems in performing such assessment via severity level classification in CXR images [7, 8].

The labelling was performed based on a COVID-19 Lung Ultrasound Severity (LUSS) method defined in literature [9]. The described LUSS score has a range of 0 to 3, with 0 corresponding to "normal" and 1-3 being various levels of severity of artifacts observed, and is based on signs and markers seen in the lung ultrasound images of infected patients, such as breaking of the pleural line, the presence and extent of consolidations in the lung tissue (darkened regions in the images) and general whitening of lung tissue seen below the pleural line [9]. Our contributing clinician (A.F.) who provided the scoring is an Assistant Professor in the department of Emergency Medicine and the ultrasound co-director for undergraduate medical students at McGill University. He is practicing Emergency Medicine full-time at Saint Mary's Hospital in Montreal. Examples of cases corresponding to each LUSS score from the COVIDx-US dataset are presented in Figure 1.

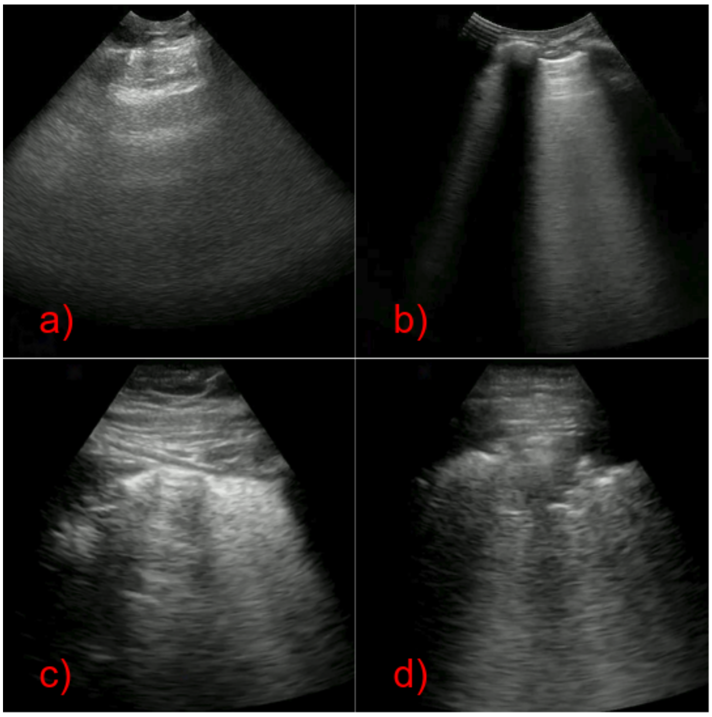

Fig. 1: Examples of lung POCUS images from the COVIDx-US dataset for different LUSS scores [6]. From top left, the samples shown represent examples with LUSS scores of a) 0, normal, b) 1, low severity, c) 2, medium severity, and d) 3, high severity.

This study will examine the feasibility of using deep learning neural networks to analyze lung POCUS images to perform classification of patient into LUSS severity scores. The overall performance of the system will be evaluated quantitatively and qualitatively to investigate whether such a decision support system could eventually be implemented to aid clinicians in decision making, and the specific successes and challenges of the system will be interpreted in the context of the problem space.

2 Methodology

As described previously, to examine the performance of deep learning neural networks to classify lung POCUS images into LUSS scores, the most updated dataset of COVIDx-US was used [6]. To ensure consistency in the dataset, cases without a valid LUSS score assigned to them (N/A) were discarded, as were cases of neck or cardiac ultrasound images and those captured with linear ultrasound probes. This data processing resulted in 133 lung POCUS videos captured with a convex probe with a valid LUSS score, which contained 17,578 image frames after processing done following methods from COVIDx-US. These images were split into training, validation, and test splits of 11,337, 3,051, and 3,190 images respectively, ensuring that videos from the same patients were kept in the same splits to avoid data leakage. The distributions of LUSS scores across both valid videos and valid images, since the number of images usable from each video is inconsistent, are shown in Figure 2.

In this preliminary investigation, we leveraged a residual deep neural network architecture [10] to train a model to classify images into one of the four LUSS score classes (0-3). The model was

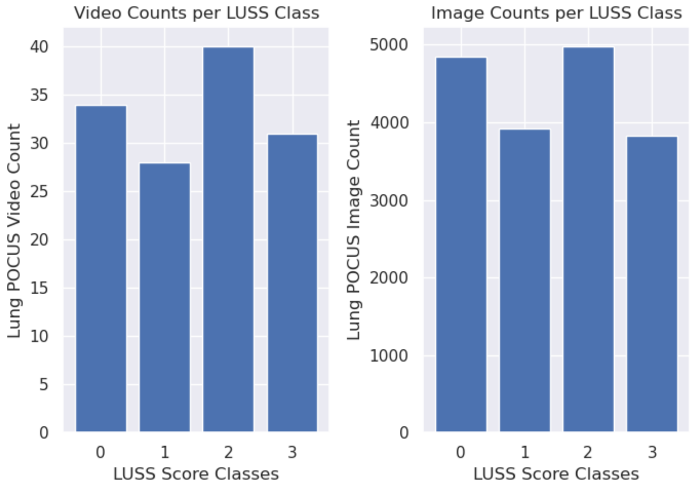

Fig. 2: The distribution of the number of videos (left) and images (right) corresponding to the 4 classes of LUSS scores. While separating the videos into individual usable images does slightly change the relative proportions in each class, the general trend is consistent, with classes 0 (no abnormalities seen) and 2 (medium severity) containing more samples than classes 1 and 3 (low and high severity respectively).

trained for 20 epochs following an exponential learning rate decay scheme with an initial learning rate of 5×10^{-4} using the Adam optimizer.

3 Results and Discussion

Results from the currently best performing model are as follows. Figure 3 contains the confusion matrix of the results from the the classification model on the testing set. Table 1 presents sensitivity and specificity metrics for the four LUSS classes as well as total accuracy across the entire test set.

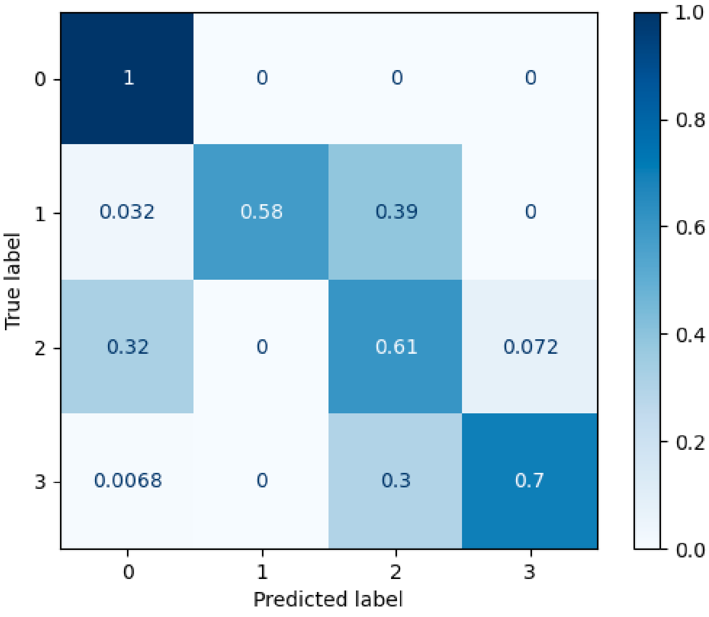

Fig. 3: Confusion matrix for results of LUSS score classification model on the testing set, presented normalized based on the number of images in each class of the ground truth labels.

At the current point in this study, the results are certainly promising. Since there are no other studies using the COVIDx-US dataset for severity classification, the overall accuracy score of 0.692 cannot be placed in a wider context, but it is supportive of the feasibility of the system to place images into the correct severity classes, yet leaves plenty of room for improvement.

Metric	LUSS 0	LUSS 1	LUSS 2	LUSS 3
Sensitivity	1.000	0.580	0.606	0.696
Specificity	0.735	1.000	0.470	0.907
Accuracy	0.692			

Table 1: Sensitivity, specificity, and overall accuracy of LUSS scores for the trained classification model.

Each of the 4 classes has a sensitivity of at least 0.580, suggesting that this neural network architecture is able to find patterns that can differentiate the classes from one another. Class 0 (normal) in fact has no misclassified samples, although the ability to perform well in this case in not unexpected, as previous work did show that neural networks can be successful at identifying COVID-19 and non-COVID-19 images [3]. However, class 0 does only have a specificity of 0.735, showing that many samples (mostly from class 2) are being falsely identified as class 0, thus work will need to be done to alleviate that issue.

It is also not too surprising that most of the errors seen in the confusion matrix in Figure 3 are between the various levels of LUSS scores that correspond to different severity classes. In these cases, the images were labeled based on the contributing clinician's interpretation of the patient case corresponding to the labeling protocol's description of signs and markers, and it is most likely the scale or extent of the signs that differentiate the cases themselves. The differences in the images will likely be a lot less easily identified, and the model at this stage in experimentation is likely to not be able to make those decisions well. In particular, the model seems to make many errors corresponding to class 2, both in terms of false negatives and false positives, leading to the lowest specificity of any of the classes at 0.470. Future work will focus on reinforcing the successes of the neural network model in classifying the normal, class 0, cases, while reducing the errors seen within the COVID-positive classes with scores 1-3.

At the point of this study, experimentation has not yet been performed on the architecture of the neural network itself, and tailoring neural networks to tasks in such a manner has been shown to be critical for increased performance in similar tasks in the past [3, 7]. Thus, such an exploration is the next area for work on this task, and is expected to significantly help the classification ability of future neural networks that will be built.

4 Conclusion

Initial performance of this deep learning neural network in identifying various COVID-19 severity levels from lung POCUS images are promising, especially when examining specifically its ability to distinguish between COVID-19 and non-COVID-19 cases. However, while still effective, the neural network certainly encounters challenges when tasked with separating the various severity levels of COVID-positive cases from each other. This suggests that further work, especially in terms of investigating a variety of network architectures beyond that used in this study, could improve the performance of the deep learning neural network architecture to levels that could assist clinicians in their work diagnosing and treating COVID-19 during the rest of the pandemic. Additionally, we hope that this work can act as a proof-of-concept that could lead to an acceleration of other work using machine learning in tandem with POCUS imaging in further clinical contexts.

Acknowledgments

We thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs program. All methods were carried out in accordance with University of Waterloo ethics guidelines and regulations.

References

[1] L. Wang, Z. Q. Lin, and A. Wong, "COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, Nov. 2020.

[2] H. Gunraj, L. Wang, and A. Wong, "COVIDNet-CT: A tailored deep convolutional neural network design for detection of

COVID-19 cases from chest CT images," *Frontiers in Medicine*, vol. 7, Dec. 2020.

[3] A. MacLean, S. Abbasi, A. Ebadi, A. Zhao, M. Pavlova, H. Gunraj, P. Xi, S. Kohli, and A. Wong, "Covid-net us: A tailored, highly efficient, self-attention deep convolutional neural network design for detection of covid-19 patient cases from point-of-care ultrasound imaging," 2021.

[4] O. Gehmacher, G. Mathis, A. Kopf, and M. Scheier, "Ultrasound imaging of pneumonia," *Ultrasound in Medicine & Biology*, vol. 21, no. 9, pp. 1119–1122, Jan. 1995. [Online]. Available: https://doi.org/10.1016/0301-5629(95)02003-9

[5] Y. Amatya, J. Rupp, F. M. Russell, J. Saunders, B. Bales, and D. R. House, "Diagnostic use of lung ultrasound compared to chest radiograph for suspected pneumonia in a resource-limited setting," *International Journal of Emergency Medicine*, vol. 11, no. 1, Mar. 2018. [Online]. Available: https://doi.org/10.1186/s12245-018-0170-2

[6] A. Ebadi, P. Xi, A. MacLean, S. Tremblay, S. Kohli, and A. Wong, "Covidx-us – an open-access benchmark dataset of ultrasound imaging data for ai-driven covid-19 analytics," 2021. [Online]. Available: https://arxiv.org/abs/2103.10003

[7] A. Wong, Z. Q. Lin, L. Wang, A. G. Chung, B. Shen, A. Abbasi, M. Hoshmand-Kochi, and T. Q. Duong, "Covid-net s: Towards computer-aided severity assessment via training and validation of deep neural networks for geographic extent and opacity extent scoring of chest x-rays for sars-cov-2 lung disease severity," 2021.

[8] H. Aboutalebi, M. Pavlova, M. J. Shafiee, A. Sabri, A. Alaref, and A. Wong, "Covid-net cxr-s: Deep convolutional neural network for severity assessment of covid-19 cases from chest x-ray images," 2021.

[9] G. Soldati, A. Smargiassi, R. Inchingolo, D. Buonsenso, T. Perrone, D. F. Briganti, S. Perlini, E. Torri, A. Mariani, E. E. Mossolani, F. Tursi, F. Mento, and L. Demi, "Proposal for international standardization of the use of lung ultrasound for patients with COVID -19," *Journal of Ultrasound in Medicine*, vol. 39, no. 7, pp. 1413–1419, Apr. 2020. [Online]. Available: https://doi.org/10.1002/jum.15285

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2016. [Online]. Available: https://doi.org/10.1109/cvpr.2016.90