FEATURES OF ICU ADMISSION IN X-RAY IMAGES OF COVID-19 PATIENTS

Douglas P. S. Gomes[⋆] *Anwaar Ulhaq*[⋆] *Manoranjan Paul*[⋆] *Michael J. Horry*[⋆ ⋆]
Subrata Chakraborty[⋆ ⋆] *Manash Saha*[⋆ ⋆ ⋆] *Tanmoy Debnath*[⋆] *D.M. Motiur Rahaman*[⋆ ⋆ ⋆ ⋆]

[⋆] Machine Vision and Digital Health (MAVIDH) Research group, School of Computing and Mathematics, Charles Sturt University, NSW, Australia

[⋆ ⋆] Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Information, Systems, and Modelling
Faculty of Engineering and IT, University of Technology Sydney, NSW, Australia

[⋆ ⋆ ⋆]Manning Rural Referral Hospital, Taree, NSW, Australia

[⋆ ⋆ ⋆ ⋆]National Wine and Grape Industry Centre, Charles Sturt University, NSW, Australia

ABSTRACT

This paper presents an original methodology for extracting semantic features from X-rays images that correlate to severity from a data set with patient ICU admission labels through interpretable models. The validation is partially performed by a proposed method that correlates the extracted features with a separate larger data set that does not contain the ICU-outcome labels. The analysis points out that a few features explain most of the variance between patients admitted in ICUs or not. The methods herein can be viewed as a statistical approach highlighting the importance of features related to ICU admission that may have been only qualitatively reported. In between features shown to be over-represented in the external data set were ones like 'Consolidation' (1.67), 'Alveolar' (1.33), and 'Effusion' (1.3). A brief analysis on the locations also showed higher frequency in labels like 'Bilateral' (1.58) and Peripheral (1.28) in patients labelled with higher chances to be admitted in ICU. To properly handle the limited data sets, a state-of-the-art lung segmentation network was also trained and presented, together with the use of low-complexity and interpretable models to avoid overfitting.

Index Terms— Covid-19, deep learning, ICU, severity, X-ray.

1. INTRODUCTION

Chest medical imaging has proven to be useful in managing more serious COVID-19 infections since respiratory dysfunction is one of the primary sources of COVID-19 morbidity and mortality. Several researchers have shown Deep Learning to be useful in classifying COVID-19 cases from medical images including CXR [1], CT [2] with some research also achieving promising results with the Ultrasound imaging mode [3]. However, using Deep Learning as a diagnostic tool can be problematic as it is hard to assess biases, risk, potential overfitting, and ability to generalize in clinical settings [4]. Its value resides more on the prognosis and treatment

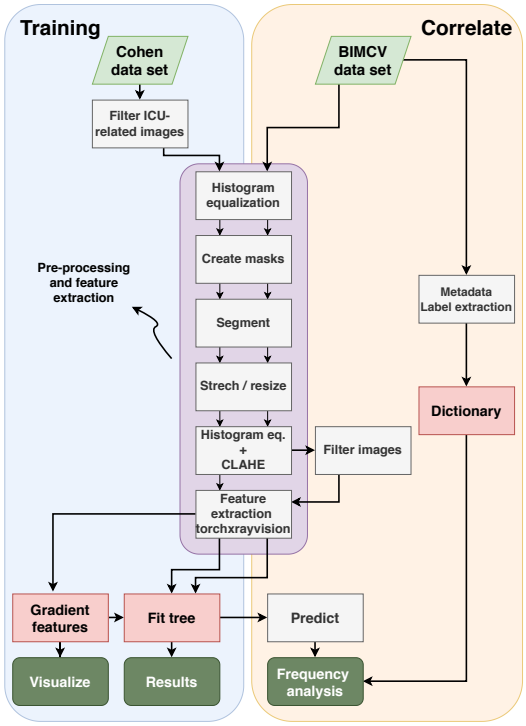


Fig. 1: ICU-related images are filtered from a limited data set, augmented, processed, and features are extracted via a specialized neural network. A few of the extracted features are used to fit a shallow decision tree, which is further compared to an external data set compose of text labels.

side than in actual diagnostic use [5]. Chest X-rays (CXR) and Computed Tomography (CT) imaging can be useful tools since these methods help clinicians to establish a baseline pulmonary status and identify underlying pulmonary conditions that may contribute to the patients' risk. Compared to CT, CXR is less expensive, more available, and require less technical expertise to perform and interpret than Ultrasound [6].

This paper therefore focuses on the potential disease pro-

gression by providing a methodology for extracting features that are correlated with patients that develop severe COVID-19 symptoms (admitted to ICUs). It leverages the existence of a data set with patient outcome labels of CXR images and the fact that some of these were collected before ICU admission. As such a data set is limited, significant measures were taken to aggressively limit potential overfitting such as image pre-processing, augmentation, lung segmentation and feature selection. To achieve significant results, a segmentation model achieving state-of-the-art results is also proposed and part of the contribution. Besides cross-validation, the found features are also correlated with an external data set to attest their significance. Such a data set has pathologies and localization labels, which also allows for insights relating to their locations in the lung.

In short, the contributions can be summarized: (i) presentation of a method that extracts semantic features explaining the variance between patients that severe or not, (ii) correlation with an external data set, improving validation, and (iii) a lung-segmentation model achieving state-of-the-art results.

2. METHODS

Given the limited amount of data to learn the severity-correlating features, a focus on limiting potential overfitting was central to most methodological decisions. Instead of disregarding such investigation simply because the data is limited, the idea proposed here is to use low-complexity, interpretable methods to test features for potential predictive value and correlate them to an external data set and literature. Fig. 1 is set to depict a high-level graphical summary of the methodology.

2.1. Data sets

The data set used to learn the features present in patients with higher chances of being admitted to ICU was a subset of the data presented by Cohen et al. [7]. It is one of the most popular data sets on the literature, favourited more than 2000 times on Github. One of its most positive characteristics is its rich metadata containing categories such as sex, age, location, patient condition, and outcome (ICU admission), allowing for the investigation presented here. For the images that had rich descriptors, two prominent labels were of interest when analysing this ICU data set: 'went-icu' and 'in-icu'. An image taken from a patient marked with 'Y' on the former label and 'N' on the latter is a sample from a patient that eventually developed severe symptoms before they were admitted in ICU. Therefore, one can reasonably form the hypothesis that there might be features in these images that are associated with patients that were eventually admitted in ICU. In total, 100 images containing these two labels were used in the analysis, which were further multiplied by a factor of approximately 10 (1040 images) by gentle affine random data augmentations: rotation, piecewise affine transformation, translation, and shear.

To expand the validation and address potential generalization concerns, a larger, external data set was also used at the testing phase. Its metadata does not contain labels regarding patient outcome (went to ICU or not), but the information it has could be correlated to results from the learned semantic features by frequency. This data set of X-ray images from COVID-19 patients, named BIMCV [8] and publicly available, contains doctor's annotations for each subject that were used here as discrete labels and descriptive terms. Besides numerous, such labels refer not only to the pathologies observed by the doctors in the patient's lungs but also points to their location. A total of 1312 images were labelled as valid, which were used to create a dictionary with the pathologies and locations labels and their respective frequency for each group (labelled with a high chance of going severe or not). The correlation is a measure of how the features relevant to the classification in the first data set are represented (frequency) in the different classes assigned in the second data set. The null hypotheses, in this case, is that if the images were sampled at random, the frequency of labels would be equal in both sets.

2.2. Processing pipeline

The selected images, augmented to a factor of 10, are assigned to a class (ICU or not) and fed through a pre-processing pipeline established to normalize and remove potential bias-inducing artifacts. The histogram equalization steps are set to impose a normalizing effect on the contrast of images by equalizing the distribution of pixel intensities that might be concentrated in a narrow range. Although there are more intricate equalization techniques, conventional and Contrast Limited Adaptive Histogram (CLAHE) were used for the normalizing steps. The choice was motivated by the direction of using simpler yet robust techniques to avoid introducing potential overfitting bias; conventional plays the role of global equalization while CLAHE does it locally. The lung-segmentation model is applied to each X-ray image, improving the signal-to-noise ratio for machine learning classification. Such model has a U-Net architecture with a ResNet backbone and is trained with images are from Montgomery and Shenzen datasets [9], resulting in a combination corpus of 1185 CXR image/mask pairs. Artefacts in the generated lung field masks were removed by a combination of morphological closing, contour filling and flood-filling, resulting in a set of automatically segmented lung-field images. As a final step, the images were also automatically cropped to fully contain the segmented lung field, resulting in a uniform image set to the downstream classifier and normalizing for different lung sizes. Fig. 2 illustrates an example of an images before and after pre-processing.

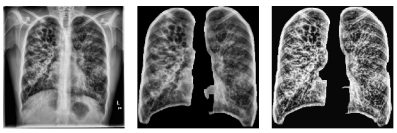


Fig. 2: Chest X-Ray images in different stages of the pre-processing pipeline: original images (left), segmented lungs without the normalization techniques (middle), and lungs segmented with normalization applied (right).

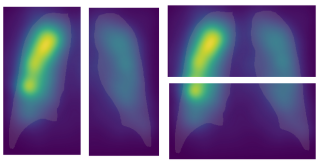


Fig. 3: Gradient features illustrations. Sectioning of gradient maps where the features are calculated: longitudinal (left) and transversal (right) cuts.

2.3. Feature extraction

The feature extraction model is a DenseNet pre-trained on 80,000+ lung X-ray images. The model, part of the TorchXRayVision library, was firstly presented in [10] and also later used in a work proposing a COVID-19 pneumonia severity score [11]. The semantic labelling was performed by adding an 18-node layer at the end of the network and training it to classify different pathology labels through a sigmoid activation layer. The labels that each of the nodes was trained to classify were the ones found in the large data sets used: *Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Enlarged Cardiomediastinum, Fibrosis, Fracture, Hernia, Infiltration, Lung Lesion, Lung Opacity, Mass, Nodule, Pleural Thickening, Pneumonia, Pneumothorax*.

In addition to the features extracted by the pre-trained, specialized network, an original feature was conceptualized to translate information regarding the location of the pathologies in the lung, called here entropy gradient features. The output features from the network, despite sufficient when attesting separability, do not translate any information regarding the location of the pathologies in the lung. The calculation of the gradients in the network graph leaves in respect to an input image was performed by the autograd torch class, which is often used to create saliency maps. First, the maps are construed with the energy of the accumulated gradients in the layers of the network and then sectioned in two cuts: longitudinal and transversal. For each of these cuts, an entropy measure is calculated; a single integer value based on the Shannon entropy, which can be seen as a measure of the spread of the activations in their respective cuts. Figure 3 is set to exemplify how the energy of the activations appear in the image and how it was sectioned to translate the information of the their locations.

2.4. Classification and validation

Given their interpretability and low-complexity, decision trees were chosen to analyse the extracted features. Besides being human interpretable, one can easily limit the effects of overfitting by mechanisms of pruning, such as setting the minimum amount of samples in the tree leaves. For the experiments performed here, for example, trees were pruned to contain at least 20 samples in each leaf in a maximum depth of 4 levels, greatly reducing complexity and model dimensionality. The results are presented for both the features extracted with the pre-trained network and the proposed entropy gradient features in a whole set and cross-validation approach. To avoid confusion, it is worth noting that classification here is used only as a method to assess separability between classes, rather than the optimal way to detect such patients.

3. RESULTS

Regarding the segmentation network trained on large chest X-rays image data sets, the designed architecture achieved a maximum validation dice similarity coefficient of 0.988 at epoch 93. The best performance observed in the literature was achieved using a complex CNN achieving a dice similarity coefficient of 0.980 [12], followed by Yang et al. [13] with 0.975 and Novikov et al. [14] with 0.974.

The separability (whole set) and cross-validation metrics were calculated for the extracted semantic features and the gradient entropy features. The former, features from the specialized network, resulted in 0.8 accuracy and 0.74 F1-score when fitted on the whole data set and 0.78 accuracy and 0.72 F1-score for the cross-validation scenario. If the features are ranked by samples under their respective node, it can result in interesting insights since they are trained to have semantic meaning. The three features with most samples, in this case, were ‘Effusion’, ‘Consolidation’, and ‘Cardiomegaly’. Fig. 4 depicts a 3-level version of the tree for better illustration purposes, which has performance closely equivalent to the 4-level version.

While the pathologies picked by the decision tree as best predictors of severity is illuminating, they do not translate information about the location of such pathologies. The simple method proposed here uses hand-engineered features on gradient maps to check for areas where the features are prevalent. Although not revealing as the pathology tree, the experiment shown to result in a relevant level of separability reaching accuracy of 0.78, F1-score of 0.73, and cross-validation accuracy of 0.73. Some relevant features by the number of samples were ‘Fibrosis’ on the lower transversal cut of the lung, and ‘Effusion’ and ‘Edema’ on the upper cut of the lung (see Fig. 3).

The tens of thousand labels contained in the BIMCV data set allowed for an interesting comparison between the classes

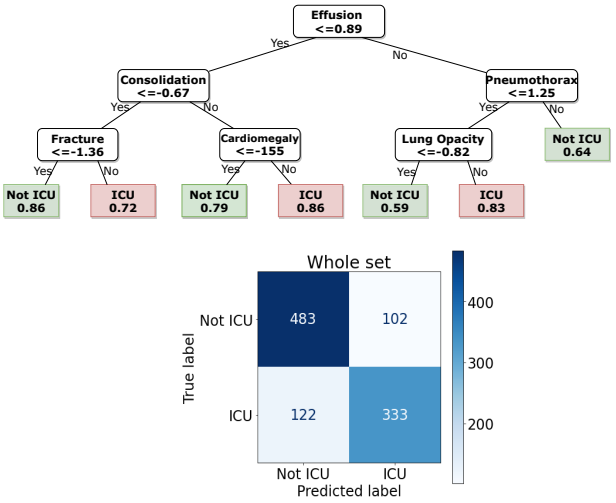


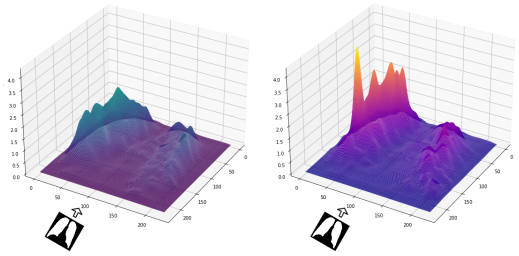
Fig. 4: Decision trees and confusion matrices from the experiment with semantic features.

after classification by the fitted model. The images had their respective text labels assigned to sets given by their predicted class (went severe or not). With many text labels for each image, two dictionaries were construed with the keys being the unique words in the set and the values representing the frequency of such labels. The comparison was made with the normalized ratio between the frequency of words in the potentially severe class (1) and not (0). Such a ratio was calculated only for words that had the minimal arbitrary number of appearances of 20 in each class. It may be worth noting that, if the samples were picked at random, such a ratio would be equal to 1, given the normalization. Table 1 presents the pathological and localization labels with normalized ratios higher than 1.2 and lower than 0.8, in descending order. The most interesting aspect from such comparison the fact that two of the most over-represented features in the images of the external data set were also the ones chosen as most important when fitting the decision tree in the data set used for training. This shows that the features used to separate the data set used for training are also consistently detected on the external data set.

The method presented here can then be interpreted as a way to express the statistical difference between features previously known but mostly qualitatively, described by practitioners. With this methodology, one uses only the specialized pre-trained network and a dictionary comparison with text-labeled data to infer such significance in patients that develop severe symptoms. When compared to other works, two over-represented features in the class of higher chances to go severe — 'Consolidation' and 'Bilateral' — are often cited as pathologies relating to the severity in the literature [15, 16]. For example, the authors in [17] reported that the evolution from ground-glass opacities to consolidation was present in some severe patients. They also pointed out that a systematic review on the subject showed that the prevalent locations were bilateral and peripheral (another over-represented term from results presented here). Lastly, another highly cited work [18] reported that most of the patients had bilateral involvements and that ICU-admitted patients showed bilateral multiple lobular and subsegmental areas of consolidation.

Table 1: Over- and under-represented pathology and localization labels and their frequency ratio

Pathology		Localization	
Feature	Ratio (C1/C0)	Feature	Ratio (C1/C0)
'Consolidation' (127)	1.67	**'Bilateral' (449)**	1.58
'Alveolar' (151)	1.34	'Middle' (326)	1.39
'Effusion' (60)	1.30	'Lower' (452)	1.31
'COVID' (443)	1.24	**'Peripheral' (489)**	1.28
'Pneumonia' (301)	1.24	'Upper' (305)	1.25
'Pleural' (67)	1.22	'Left' (549)	1.23
'Infiltrates' (184)	1.23	...	
'Interstitial' (223)	1.2	'Hilar' (99)	0.66
...		'Mediastinum' (84)	0.56
'Normal' (168)	0.48		



(a) (b)

Fig. 5: Surface plots of the average of gradient maps in each class: a) Non-ICU and b) ICU.

With the methodology adopted, an interesting way to visualize the spatial distribution of some features becomes possible. Instead of simply plotting the gradient maps, one can calculate the average of activations in relation to one feature for all images, then averaging these whole class activation for a representation of that particular feature. Figure 5 presents an example for the highly-ranked semantic features: 'Effusion'.

4. CONCLUSIONS

The results presented point out that selected semantic features correlate to patients who eventually develop severe symptoms and are admitted to ICU. The analysis included an original methodology that heavily mitigates potential overfitting while being interpretable, and proposes a method to compare them to an external data set of text labels. Another original contribution is are the lung segmentation network resulting in state-of-the-art results. A more detailed report and source code of experiments can be found in the in work repository [1].

[1] https://github.com/dougpsg/covid_mavidh_icufeatures_scoring

5. REFERENCES

[1] Yujin Oh, Sangjoon Park, and Jong Chul Ye, "Deep learning covid-19 features on cxr using limited training data sets," *IEEE Transactions on Medical Imaging*, 2020.

[2] Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke Wang, et al., "Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423 – 1433.e11, 2020.

[3] Michael J Horry, Subrata Chakraborty, Manoranjan Paul, Anwaar Ulhaq, Biswajeet Pradhan, Manas Saha, and Nagesh Shukla, "Covid-19 detection through transfer learning using multimodal imaging data," *IEEE Access*, vol. 8, pp. 149808–149824, 2020.

[4] Laure Wynants, Ben Van Calster, Marc MJ Bonten, Gary S Collins, Thomas PA Debray, Maarten De Vos, Maria C Haller, Georg Heinze, Karel GM Moons, Richard D Riley, et al., "Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal," *BMJ*, vol. 369, 2020.

[5] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv 2006.11988*, 2020.

[6] A. Ulhaq, J. Born, A. Khan, D. Gomes, S. Chakraborty, and M. Paul, "Covid-19 control by computer vision approaches: A survey," *IEEE Access*, vol. Early Access, pp. 1–1, 2020.

[7] Joseph Paul Cohen, Paul Morrison, and Lan Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020.

[8] Maria de la Iglesia Vayá, José Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al., "Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients.," *arXiv preprint arXiv:2006.01174*, 2020.

[9] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, pp. 475, 2014.

[10] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand, "On the limits of cross-domain generalization in automated x-ray prediction," in *Medical Imaging with Deep Learning*, 2020.

[11] Joseph Paul Cohen, Lan Dao, Paul Morrison, Karsten Roth, Yoshua Bengio, Beiyi Shen, Almas Abbasi, Mahsa Hoshmand-Kochi, Marzyeh Ghassemi, Haifang Li, et al., "Predicting covid-19 pneumonia severity on chest x-ray with deep learning," *arXiv preprint arXiv:2005.11856*, 2020.

[12] Sangheum Hwang and Sunggyun Park, "Accurate lung segmentation via network-wise training of convolutional networks," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 92–99. Springer International Publishing, 2017.

[13] Wei Yang, Yunbi Liu, Liyan Lin, Zhaoqiang Yun, Zhentai Lu, Qianjin Feng, and Wufan Chen, "Lung field segmentation in chest radiographs from boundary maps by a structured edge detector," *IEEE journal of biomedical and health informatics*, vol. 22, no. 3, pp. 842–851, 2017.

[14] A. A. Novikov, D. Lenis, D. Major, J. Hladůvka, M. Wimmer, and K. Bühler, "Fully convolutional architectures for multiclass segmentation in chest radiographs," *IEEE Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1865–1876, 2018.

[15] Adam Bernheim, Xueyan Mei, Mingqian Huang, Yang Yang, Zahi A Fayad, Ning Zhang, Kaiyue Diao, Bin Lin, Xiqi Zhu, Kunwei Li, et al., "Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection," *Radiology*, p. 200463, 2020.

[16] Wei Zhao, Zheng Zhong, Xingzhi Xie, Qizhi Yu, and Jun Liu, "Relation between chest ct findings and clinical conditions of coronavirus disease (covid-19) pneumonia: a multicenter study," *American Journal of Roentgenology*, vol. 214, no. 5, pp. 1072–1077, 2020.

[17] Ming-Yen Ng, Elaine YP Lee, Jin Yang, Fangfang Yang, Xia Li, Hongxia Wang, Macy Mei-sze Lui, Christine Shing-Yen Lo, Barry Leung, Pek-Lan Khong, et al., "Imaging profile of the covid-19 infection: radiologic findings and literature review," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, pp. e200034, 2020.

[18] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al., "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china," *The lancet*, vol. 395, no. 10223, pp. 497–506, 2020.