ViDi: Descriptive Visual Data Clustering as Radiologist Assistant in COVID-19 Streamline Diagnostic

Sahithya Ravi[1 [0000−1111−2222−3333]], Samaneh Khoshrou[1 [0000−0002−6317−9453]], and Mykola Pechenizkiy[1 [0000−0003−4955−0743]]

Eindhoven University of Technology, Eindhoven, Netherlands
{s.ravi,s.khoshrou,m.pechenizkiy}@tue.nl

Abstract. In the light of the COVID-19 pandemic, deep learning methods have been widely investigated in detecting COVID-19 from chest X-rays. However, a more pragmatic approach to applying AI methods to a medical diagnosis is designing a framework that facilitates human-machine interaction and expert decision making. Studies have shown that categorization can play an essential rule in accelerating real-world decision making. Inspired by descriptive document clustering, we propose a domain-independent explanatory clustering framework to group contextually related instances and support radiologists' decision making. While most descriptive clustering approaches employ domain-specific characteristics to form meaningful clusters, we focus on model-level explanation as a more general-purpose element of every learning process to achieve cluster homogeneity. We employ DeepSHAP to generate homogeneous clusters in terms of disease severity and describe the clusters using favorable and unfavorable saliency maps, which visualize the class discriminating regions of an image. These human-interpretable maps complement radiologist knowledge to investigate the whole cluster at once. Besides, as part of this study, we evaluate a model based on VGG-19, which can identify COVID and pneumonia cases with a positive predictive value of 95% and 97%, respectively, comparable to the recent explainable approaches for COVID diagnosis.

Keywords: Descriptive clustering · Covid-19 · Chest X-ray · Explainable learning.

1 Introduction

The COVID-19 outbreak is the biggest challenge of 2020, and even with all the efforts, the tolls are still rising around the world. Jam-packed hospital wards make efficient triage of patients with COVID-19 requisite [6]. However, despite the rapid advancement of Artificial Intelligence (AI), AI in medicine is no silver bullet, and in most situations, domain experts still are sole decision-makers; COVID-related decisions are not an exception. The top two reasons that make a complete AI-powered solution out of sights are first, *"lack of representative and*

curated datasets", which may lead to train not only less effective but also unfair and biased models. Second, more data alone does not lead to a more practical model; as humans, we must understand and interpret the process as well as the outcome of an AI system [16]. *"Lack of transparency and interpretability"* makes systems less trustworthy to be deployed in real-world situations [11]. While a *complete AI-powered system entrusted with full responsibility acting autonomously on its own* is currently onerous, studies have shown that deployment of AI in human-in-the-loop fashion aiming to assist not substitute health-workers (especially radiologists) leading to better outcomes for patients [1]. Today influx of COVID-19 patients and acute shortage of medical resources (i.e., staff and equipment) have highlighted the need for a collaborative human-AI diagnosis framework more than ever.

Main Contributions In this work, we propose a visual data descriptive clustering method (ViDi) that aims to group CXR images with equal severity level and similar geographic extent of the infection together. The pipeline of ViDi starts with preprocessing and augmentation of more than 5000 CXR images, of which 200 are covid chest x-rays, followed by evaluation the success of the state-of-the-art VGGNet architecture [8] in a transfer learning (TL) setting, on their proficiency in chest-xray detection. Then, class-discriminating saliency maps are generated using DeepSHAP[10]. ViDi helps the clinicians in two main ways: 1) Single image exploration: the framework provides two saliency maps. The *favorable saliency map* highlights the regions that positively contribute to a particular prediction and the *glum map* represents the areas that contribute negatively. b) Cluster exploration: the framework groups CXR images by explanation similarity. Specifically, for COVID-19 dataset images with similar predicted severity scores and the infection's geographic extent are placed together. The clustering allows the clinician to compare CXR images based on different qualitative and quantitative measures and could be used to prioritize and adjust the treatment process, especially in overwhelmed circumstances. Besides, it provides a scalable solution for large scale CXR annotation. While existing works focus on predicting the severity scores from multiple data sources, including CXR images, here we focus on a framework that complements the findings from classic AI-based approaches and potentially facilitates and improve the decision-making process by radiologists or expert annotators.

2 Background

With a few exceptions [5,7,9,17], most AI-powered CXR analysis literature has focused on the accuracy of the prediction [12,14,15,19,14] rather than enlightening the roots of the prediction. Ghoshal et al. [7] propose a Bayesian Convolutional Neural network to estimate the diagnosis uncertainty, which potentially yields more reliable prediction and can alert radiologists on false prediction. DeepCovidExplainer [9] highlights the class discriminative regions using

[1] `https://medium.com/@zp489/f06e7daaee5`

a gradient-guided class activation maps; a heatmap is depicted to provide a human-interpretable explanation of the prediction. Since successive pooling layers have significantly decreased the deep activation maps' resolution, the class activation maps do not output a very precise localization. Cohen et al. [5] develop Chester, a web-delivered locally computed disease prediction system that aims to help clinicians understand a deep learning prediction. They use the gradient saliency map to explain network prediction. Practically speaking, the most discriminative (i.e., generally the gradient is high.) regions are depicted in red, while transparent regions have a negligible impact on the prediction. One issue with gradient-based interpreting approaches is that gradient is high at predictive regions and at locations that condition another region to impact, resulting in misguidance. In another study, Wang [17] proposed COVID-Net to detect distinctive abnormality in CXR images. The net employs GSInquire to identify the most critical regions (highlighted in red); poor decision visualization is one of the main limitations of this approach. All the mentioned explainable frameworks share the same characteristics: they try to shed some light on network prediction in a single x-ray image to help clinicians in the decision making process. In contrast to most gradient-based methods, using a difference-from-reference, ViDi uses DeepLIFT, which allows propagating a quality signal even in situations where the gradient is zero and avoids artifacts caused by discontinuities in the gradient. In addition to highlighting the critical favorable and glum regions for a single image, our framework provides a cluster of similar images, which we believe has a great potential in streamline diagnosis of COVID-19. Note that these methods complement radiologists' knowledge, and in fact, the domain expert is still the primary decision-maker.

3 Experimental Design

3.1 Datasets & Scenarios

We conduct our experiments on two public datasets: a) The COVID-19 x-ray image [6], including 392 COVID-19 postero-anterior chest x-ray images annotated with severity level. More information on the severity score assignment is available at [18]. b) the chest x-ray database from Kaggle [1], consisting of 5863 chest x-ray images chosen from retrospective cohorts of pediatric patients graded by two expert physicians. We define three scenarios to evaluate the effectiveness of proposed framework:

1. Binary classification of *pneumonia vs normal* images of the pneumonia dataset.
2. Binary classification of *covid vs normal* images of the covid dataset.
3. Multi class classification of *mild vs medium vs severe* images of the covid dataset.

The third task aims to classify the severity of the lung involvement based on the total opacity scores, ranging from 0 to 6. Similar to the grouping of lung involvement scores in COVID patients published by the Radiological Society of North America (RSNA) [2], we derive three categories mild, severe, and medium that refers to severity score below two, above four and within this range, respectively.

3.2 Model, Pre-training, and Training

Since the COVID-19 image dataset is still small and evolving, we employ a transfer learning strategy to utilize the knowledge from previously learned models and apply them to our problem. We trained VGG and DenseNet Architectures, followed by a prediction performance check. In this study, we choose VGG-19 model due to its success in the ImageNet Large Scale Visual Recognition Challenge [13] and COVID detection from chest x-rays as well as lighter computational complexity [8]. For the pre-training phase, involving pneumonia vs. normal classification (scenario 1), we initialized the VGG-19 architecture with ImageNet weights and modified the classification layer. By training the model on the pneumonia dataset, it learns common representations about lungs, which would be hard to achieve on the small set of COVID-19 images. The images used in both datasets were resized to 224 by 224 pixels and normalized using ImageNet standards.

For the training phase, the pre-processed images of the COVID-19 dataset were randomly split into a train and test set in the ratio 70/30 for scenario 2 (covid vs normal) and 60/40 for scenario 3(severity classifier) such that the classes are stratified. In order to enhance the generalization capacity of the model, we employ image augmentation strategies such as random horizontal flip with a 50% probability and randomized image rotation. The proposed model pretrained on the pneumonia dataset is then trained on the COVID19 dataset for both scenarios 2 and 3. The Adam optimizer a learning rate of 5e-4 was used for optimizing the weights. The experiments run for 50 epochs and batch size was set to 32.

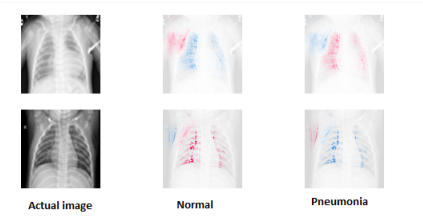


Fig. 1: Glum and favorable saliency map using DeepSHAP

4 Descriptive Visual Data Clustering

"Descriptive clustering consists of automatically organizing data instances into clusters and generating a descriptive summary for each cluster. [4]". The description should inform a user about each cluster's contents to speed up the decision-making process. There is no universal definition for a good description, and the selection of descriptions often relies on heuristic rules concerning the application. We model descriptive clustering as categorizing similar CXR images with regard to the most critical regions (pixels) in network prediction.

The subset of features used to predict a class serves as its description in two ways: 1) Favorable regions that push the model towards a positive prediction. 2) Unfavorable (glum) pixels which contribute negatively to the prediction.

Once our model is trained, we employ DeepSHAP [10] to understand and visualize model predictions using saliency maps. We chose SHAP explanation, since SHAP values prove more consistent with human intuition than other methods [10] and they output a model agnostic explanation. DeepSHAP is an enhanced version of the DeepLIFT algorithm [3], which measures feature contributions by calculating how a target(t) changes as the input(x) changes from the baseline. For a given input neuron x with difference-from-baseline Δx , and target neuron t with difference-from-baseline Δt that we wish to compute the contribution to, DeepLIFT calculates the slope or multiplier given by, $m_{\Delta x \Delta t} = \frac{C\Delta x \Delta t}{\Delta x}$, where the DeepLIFT contribution scores are given by, $C\Delta x \Delta t = \Delta t$. DeepSHAP defines multipliers in terms of shapely values and works by recursively passing DeepLIFT multipliers via backpropagation. Figure 1 illustrates an example saliency maps for two patients. For the first image of a pneumonia patient, blue regions in the lung with high opacity contribute negatively to the normal class. However, the same pixels contribute positively to pneumonia prediction (highlighted in red). While in the second image relating to a healthy patient, the red or positive shapely values correspond to transparent regions that contribute to the normal class.

Since shapely values try to isolate individual feature's effect, they are a good indicator of the similarity between instances. Clustering of images in shapely space can result in homogeneous clusters, with each cluster is dominated by instances with similar features. This is the main rationale behind our clustering algorithm. We pick K-means++ clustering due to its careful seeding method, simplicity, and speed in practice. The generated clusters are then visually inspected and assessed using clustering criteria in Sec. 5.

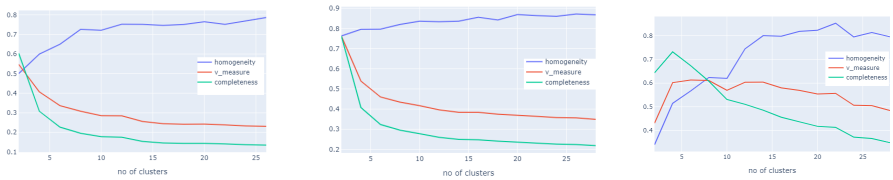
5 Results & Discussion

5.1 Model performance

The performance of the VGG-19 for three different scenarios is shown in Table 1. The achieved recall and F1-score slightly outperforms [9] with recall of 0.935 and F1-score of 0.928, and [8] with a recall of 0.93. Out 65 COVID-19 patient samples, only 3 were misclassified as normal, which is analogous to the misclassifcation ratio obtained in [9].

Table 1: Accuracy of the CNN on different settings

Dataset	Model	Accuracy	F1 score	Precision	Recall
pneumonia 2-class	VGG-19 pretrained on ImageNet	0.94	0.89	0.83	0.97
covid 2-class	VGG-19 pretrained on pneumonia	0.98	0.97	1	0.95
covid severity 3-class	VGG-19 pretrained on pneumonia	0.91	0.9	0.91	0.89



(a) Pneumonia vs normal (b) Covid vs normal (c) Covid severity

Fig. 2: Cluster quality assessment as a function of number of clusters (k)

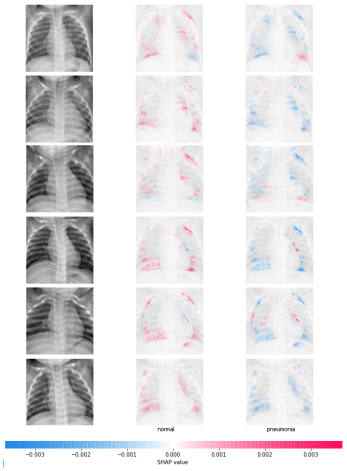


Fig. 3: Normal pneumonia cluster

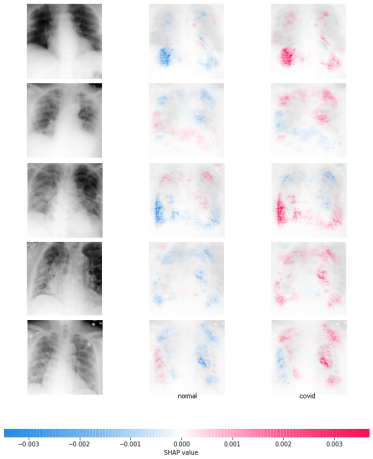


Fig. 4: Covid cluster

5.2 Clustering Analysis

We evaluate the clustering algorithm based on homogeneity, completeness, and v-measure. A clustering satisfies homogeneity if all of its clusters contain instances of a single class exclusively, whereas a clustering satisfies completeness when all instances of a given class are elements of a single cluster and not divided among many clusters. The v-measure is the harmonic mean of homogeneity and completeness. The trade-off between homogeneity and completeness can be observed in all three scenarios, as shown in Fig 2. While choosing the number of clusters, we prioritize homogeneity over completeness, as the number of clusters and the number of images per cluster at this stage are not too large to be handled by a domain expert. We choose k as 25, 20, and 22 for the three scenarios above, respectively.

5.3 Cluster Visualization

In this section, we visualize examples of homogeneous clusters generated using the descriptive clustering algorithm for all three scenarios:

1. *Pneumonia vs Normal* Figure 3 exemplifies a homogeneous normal cluster and the corresponding saliency maps. The favorable (red) regions on the lungs denote the transparent regions contributing positively to normal class. For the pneumonia class, the same regions contribute negatively (blue).
2. *Covid vs Normal*: Figure 4 illustrates a sample COVID-infected cluster. The favorable saliency map for COVID class and glum map for the normal class complement each other. For instance, in the first image, the opaque regions at the bottom of the lung contribute negatively to normal class and positively to the COVID class. Figure 5 shows an impure cluster generated from the COVID vs. normal scenario. All the cluster images are COVID positive, except for the fourth image from the top, which is a normal image. To achieve a higher number of pure clusters, it is thus important to determine the appropriate number of clusters (k).
3. *Covid severity* Figure 6, 8 , and 7 demonstrate a homogeneous mild, moderate and severe cluster respectively. Our approach provides two channels of information, why the model makes this prediction, and which regions push the model towards the other classes, which we believe would be more helpful than just providing a positive contribution. For example, in the severe cluster Figure 7 , we have three maps highlighting the feature contribution in every three possible classes. It is a cluster with maximum favorable (red) regions for severe class and prominent negative (blue) regions for mild class. Comparing these maps facilitates false positives identification. Also, looking at the whole cluster potentially accelerates decision making.

6 Conclusion and Future Direction

In this work, we present ViDi, a novel approach for model-based descriptive clustering with the help of explanation similarity. Rather than trying to replace a radiologist, ViDi has the potential to act as a bridge between AI and the expert, thus giving AI in a more convenient form to the medical community. In order to facilitate a better understanding of the generated clusters, we highlight both favorable and glum regions, which provide further insight into the behavior of the neural network. Further, the categorization of clusters into mild, moderate, and severe can act as a guiding mechanism for deciding the treatment option - at home, hospital, or ICU. We achieve high homogeneity of up to 80% with the current version of the COVID-19 dataset, but deployment in a medical setting requires further enhancement of the framework. Since the current version of the dataset is small, a more in-depth analysis requires further data, especially in the severity front. As future work, cdvc zdv

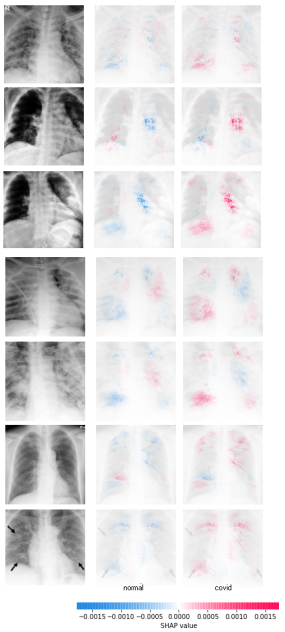


Fig. 5: Impure covid cluster

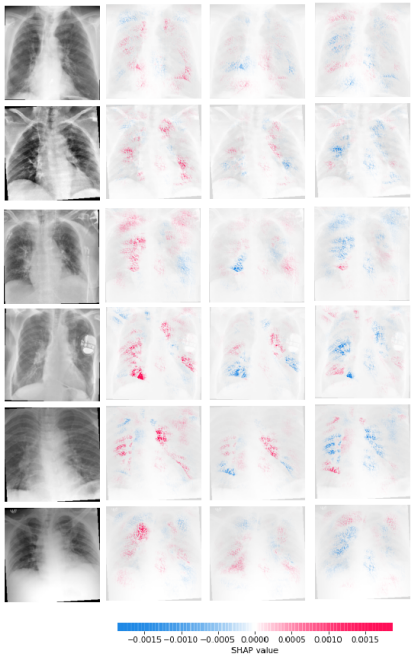


Fig. 6: Mild covid cluster

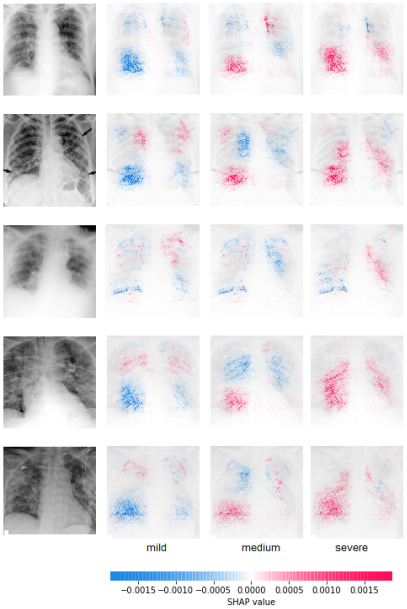


Fig. 7: Severe covid cluster

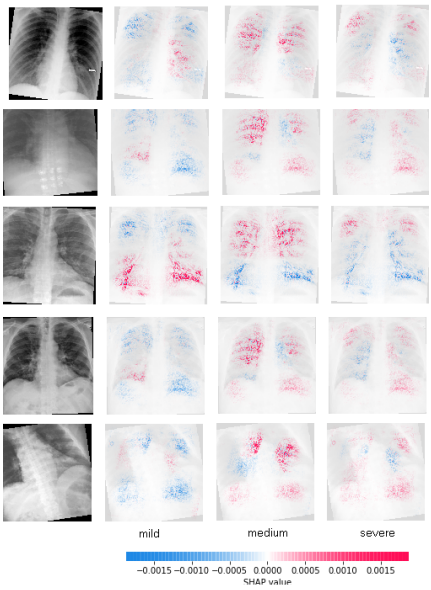


Fig. 8: Moderate covid cluster

References

1. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
2. A, B., X, M., M, H., et al: Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection. Radiology. 2020;200463 (2020), https://pubs.rsna.org/doi/pdf/10.1148/radiol.2020200463
3. Avanti Shrikumar, P.G., Kundaje, A.: Learning important features through propagating activation differences. NIPS proceedings (2019), https://arxiv.org/pdf/1704.02685.pdf
4. Brockmeier, A.J., Mu, T., Ananiadou, S., Goulermas, J.Y.: Self-tuned descriptive document clustering using a predictive network. IEEE Transactions on Knowledge and Data Engineering **30**(10), 1929–1942 (2018). https://doi.org/10.1109/tkde.2017.2781721
5. Cohen, J.P., Bertin, P., Frappier, V.: Chester: A web delivered locally computed chest x-ray disease prediction system. CoRR **abs/1901.11210** (2019), http://arxiv.org/abs/1901.11210
6. Cohen, J.P., Morrison, P., Dao, L.: Covid-19 image data collection. arXiv 2003.11597 (2020), https://github.com/ieee8023/covid-chestxray-dataset
7. Ghoshal, B., Tucker, A.: Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. CoRR **abs/2003.10769** (2020), https://arxiv.org/abs/2003.10769
8. Ioannis D. Apostolopoulos, T.A.M.: Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine 43:635-40;2020 (2020), https://link.springer.com/content/pdf/10.1007/s13246-020-00865-4.pdf
9. Karim, M.R., Döhmen, T., Rebholz-Schuhmann, D., Decker, S., Cochez, M., Beyan, O.: Deepcovidexplainer: Explainable COVID-19 predictions based on chest x-ray images. CoRR **abs/2004.04582** (2020), https://arxiv.org/abs/2004.04582
10. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. NIPS Proceedings (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
11. Marcus, G., Davis, E.: Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon Books, USA (2019)
12. Narin, A., Kaya, C., Pamuk, Z.: Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks. CoRR **abs/2003.10849** (2020), https://arxiv.org/abs/2003.10849
13. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
14. Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D.: Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. IEEE Reviews in Biomedical Engineering pp. 1–1 (2020)
15. Tang, Z., Zhao, W., Xie, X., Zhong, Z., Shi, F., Liu, J., Shen, D.: Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. CoRR **abs/2003.11988** (2020)
16. Vellido, A.: Societal issues concerning the application of artificial intelligence in medicine. Kidney Diseases **5**(1), 11–17 (2018). https://doi.org/10.1159/000492428
17. Wang, L., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. CoRR **abs/2003.09871** (2020), https://arxiv.org/abs/2003.09871

18. Wong, A., Lin, Z.Q., Wang, L., Chung, A.G., Shen, B., Abbasi, A., Hoshmand-Kochi, M., Duong, T.Q.: Covidnet-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 (2020)
19. Zhang, J., Xie, Y., Li, Y., Shen, C., Xia, Y.: COVID-19 screening on chest x-ray images using deep learning based anomaly detection. CoRR **abs/2003.12338** (2020)