Explainable image analysis for decision support in medical healthcare

Roberto Corizzo
Department of Computer Science
American University
Washington, DC, USA
rcorizzo@american.edu

Yohan Dauphin
CPE Lyon
Lyon, France
yohan.dauphin@numericable.fr

Colin Bellinger
Digital Technologies
National Research Council of Canada
Ottawa, Canada
colin.bellinger@nrc-cnrc.gc.ca

Eftim Zdravevski
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
eftim.zdravevski@finki.ukim.mk

Nathalie Japkowicz
Department of Computer Science
American University
Washington, DC, USA
japkowic@american.edu

***Abstract*—Recent advances in medical imaging and deep learning have enabled the efficient analysis of large databases of images. Notable examples include the analysis of computed tomography (CT), magnetic resonance imaging (MRI), and X-ray. While the automatic classification of images has proven successful, adopting such a paradigm in the medical healthcare setting is unfeasible. Indeed, the physician in charge of the detailed medical assessment and diagnosis of patients cannot trust a deep learning model's decisions without further explanations or insights about their classification outcome. In this study, rather than relying on classification, we propose a new method that leverages deep neural networks to extract a representation of images and further analyze them through clustering, dimensionality reduction for visualization, and class activation mapping. Thus, the system does not make decisions on behalf of physicians. Instead, it helps them make a diagnosis. Experimental results on lung images affected by Pneumonia and Covid-19 lesions show the potential of our method as a tool for decision support in a medical setting. It allows the physician to identify groups of similar images and highlight regions of the input that the model deemed important for its predictions.**

***Index Terms*—Deep Learning, Clustering, XAI, Healthcare, COVID-19**

I. Introduction

Machine learning algorithms have been extensively used for various medical applications ranging from the interpretation of radiographs, computer-aided diagnosis, cancer detection, and the analysis of images with a variety of modalities such as computed tomography (CT) and ultrasound [1]. Even though these approaches were not substantially novel from a methodological perspective, the advances in computational power and Graphical Processing Units (GPUs) in the recent decade led to a dramatic increase of various deep learning architectures for medical image analysis.

One class of medical applications of deep learning is for image segmentation, particularly organ and substructure segmentation, including bladder segmentation [2] and cardiac substructures [3]. There are many benefits of these image segmentation tasks, including calculating clinical parameters such as volume, as well as defining the search region for computer-aided detection tasks [1]. Another similar task, yet more difficult, is lesion segmentation because the object being segmented can have varying shapes and sizes [4], [5].

Another popular class of problems is organ or lesion detection, as well as region and landmark localization [6], [7]. A very labor-intensive task for clinicians is detecting objects of interest or lesions in images, which is a crucial part of the diagnosis. The ultimate goal of these methods is improving the detection accuracy or decreasing the reading time of human experts.

Likewise, automated characterization of diseases has been quite a popular research subject for decades leading to various computer-aided diagnosis (CADx) systems [1]. The disease characterization often involves analysis of regions of medical images around potential disease sites. Traditionally, these methods have been applied for detecting cancer and tumor on Magnetic resonance imaging (MRI), X-Ray, and CT images [8]. However, this task is often related to prognosis aiming to recognize the stage of the disease, molecular subtype, or genomics to yield information on prognosis and potential treatment options [9].

Inspired by such applications of medical image analysis and pressed by the harsh reality of the current COVID-19 pandemic, in this research, we focus on aiding physicians and clinicians to better analyze Chest X-Ray (CXR) images of patients with pneumonia caused by COVID-19 infections. As the pandemic progressed, deep learning-based approaches for COVID-19 classification based on chest X-ray images have been published [10]–[17]. These works differ on the exact deep learning architectures, data augmentation strategies, datasets, etc. Alternatively, some approaches such as [18] rely on non-invasive methods for detecting pneumonia progression based on non-invasive ambient sensors.

The feature extraction process for segmentation, clustering,

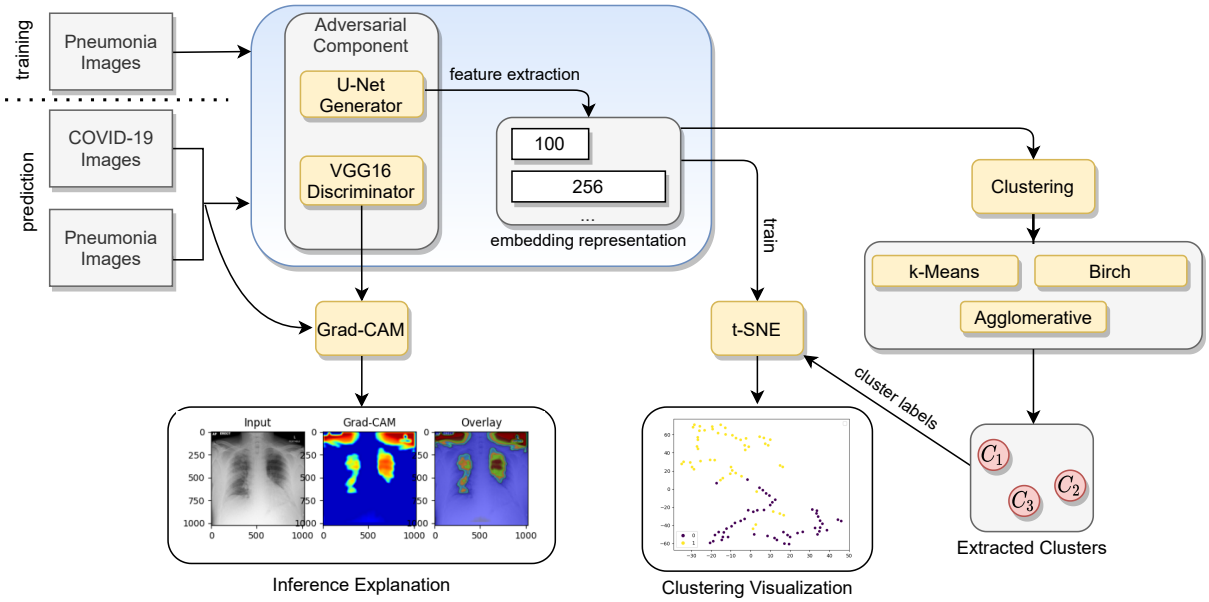


Fig. 1: Proposed architecture for explainable medical image analysis.

and classification based on image data often is most efficient when done with various deep learning approaches. There is a variety of architectures successfully applied for such tasks that leverage average pooling layers, convolutional layers, U-nets, Recurrent Neural Networks, or feature concatenation [19]–[22]. These methods often involve transfer learning from pre-trained Convolutional Neural Networks (CNNs) for feature extraction, and recently methods relying on adaptive learning rates have shown promising results to reduce overfitting [23] and shorten the training process. As dataset sizes grow, big data architectures become necessary for efficient and timely processing of it [24]–[26]. In turn, it entails the use of efficient algorithms for cluster-size and cost optimization [27], [28].

In this paper, we present a method for explainable image analysis to support decisions in the medical healthcare field. Instead of attempting to make decisions on behalf of physicians, we aim to identify groups of similar images and highlight regions of interest so that physicians can make better decisions. Differently than existing deep learning methods focused on the classification of medical images, our method leverages a synergy of a deep neural network architecture composed by a U-Net generator and a custom VGG16 discriminator, clustering, dimensionality reduction, and visualization techniques. An experimental evaluation on lung images containing Pneumonia and COVID-19 lesions shows our method's potential in identifying groups of patients with different characterizations of the disease. In particular, the combination of clustering and activation mappings, which provide a visual explanation of the results, presents valuable insights for the physician which may be of support for the design of different medical treatments. Ultimately, this should also help in the triage protocols to detect patients with low severity.

The remainder of the article is structured as follows. Section II describes the employed methods in this study, particularly describing the adversarial learning component, clustering, visualization, and inference explanations. Next, Section III explains the experimental setup, whereas Section IV discusses the obtained results and their practical applications. Finally, Section V concludes the paper, summarizing our contributions and pointing directions for future work.

II. Methods

In this section, we discuss in detail our proposed method for medical image analysis. The method leverages adversarial learning and clustering to generate output visualizations for the extracted clusters, and gradient-weighted class activation mappings of each image for inference explanation. A general overview of the method is shown in Figure 1 and described as pseudo-code in Algorithm 1, where for simplicity, we assume the existence of utility functions to train each model and extract features and omit their details. Each component is described in detail in the following subsections.

A. Adversarial learning component

A Generative Adversarial Network (GAN) captures the distribution of two networks: the Generator (G) generates fake instances from noise z to fool the Discriminator (D) whose purpose is to discriminate between true and fake instances. The two components G and D feed off each other to improve until an equilibrium (convergence) is reached. The outcome value $V(D, G)$ for the mutual mini-max optimization process for D and G can be summarized by Equation 1.

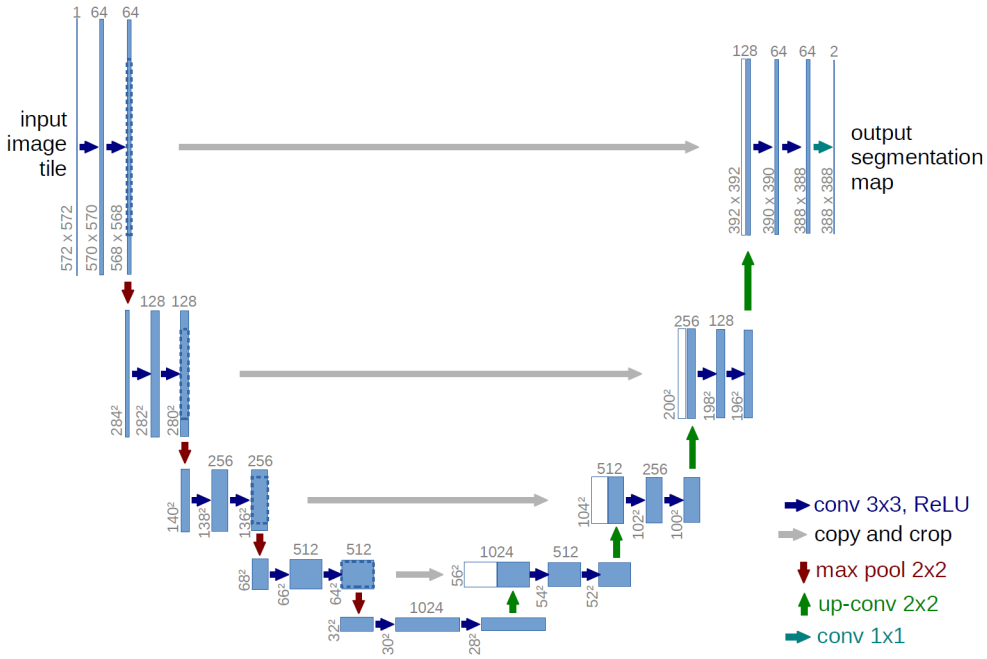


Fig. 2: U-Net deep neural network architecture for medical image segmentation [29].

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \\ + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

Our adversarial learning component consists of a U-Net generator and a custom VGG-16 discriminator with batch normalization. Their model architecture is shown in Figure 2 and 3.

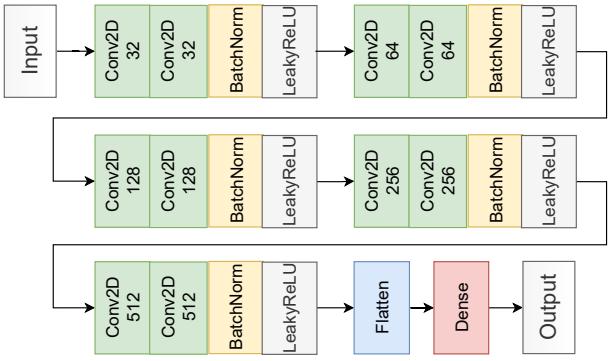


Fig. 3: Custom VGG16 discriminator architecture with batch normalization.

U-Net [29] recently became a popular model for medical image segmentation as it learns segmentation in an end-to-end setting even with few data samples. In the *contraction* phase, the number of feature channels is increased by a factor of 2 after each max-pooling operation. The sequence of convolution and max-pooling operations results in a spatial contraction that gradually becomes more focused or fine-grained in a smaller region (increases the "what" and decreases the "where"). The U-Net has an additional *expansion* phase, which consists of a sequence of up-convolutions and concatenations with high-resolution features resulting from the contracting phase. Convolutional operations and non-linear activation functions, combined with increasing channels, allow the model to extract a high-resolution segmentation map as the output. These characteristics led to the great success of this model architecture in medical image analysis.

However, it is essential to note that our use of adversarial learning is fundamentally different from how they are commonly used in the literature. In traditional work, a GAN's purpose is to model a distribution. The relationship between G and D is part of the minimax-based optimization strategy adopted by GANs to refine the model. Once trained, GANs are typically used on binary classification and data generation tasks. In our architecture, we adopt U-Net in two ways. First, as a generator trained one-class with Pneumonia images, aiming at generating high-quality segmentation maps. Second, as a feature extractor, generating vector-based features can be subsequently used for the clustering step. The second alternative provides the ability to feed a vector representation to the other components of the system.

The rationale of training our generator using one-class data stands in the opportunity to exploit a larger corpus of available images containing Pneumonia lesions, in contrast with the limited availability of COVID-19 images. Indeed, the one-class learning paradigm appears suitable for our application compared to traditional approaches, which use a single model and formulate the task as binary classification, which would be significantly penalized by the data imbalance [30] [31]. The interplay between the generator and discriminator leads to a

model that can yield high-accurate segmentations for Pneumonia images and accurately predict if an unseen image does or does not contain a Pneumonia lesion. However, we stress that the focus of our work is not purely on the classification aspect but on the opportunity to extract robust features during the inference stage and further analyze them, as explained in the following subsections.

COVID-19 prediction confidence: We perform an experiment to assess the confidence of the discriminator's predictions testing it using exclusively COVID-19 images. We generate t-SNE plots leveraging the vector representation of images extracted with the U-Net generator. We extract the values returned by the sigmoid activation function of the output layer of the discriminator and consider them as confidence scores. These values are used to color the resulting points of the t-SNE plots. The results reported in Figure 5 show that the discriminator has a varying confidence of its predictions. We consider this result as a confirmation that COVID-19 lesions are not equal and their characterization is worth further exploration and study.

COVID-19 sub-clusters: In order to further confirm the existence of sub-clusters in COVID-19 image data, we perform an experiment involving t-SNE generated plots with different vector sizes for the features extracted by the U-Net generator. The results reported in Figure 6 show a visible presence of sub-clusters in the plots. These results, in combination with the previously discussed results on the prediction confidence, provide additional evidence on the existence of multiple characterizations of COVID-19 in the data.

Agreement of clustering algorithms: In this experiment we test the ability of different clustering algorithms to discover the sub-concepts that we visually identified in the previous experiment. All clustering algorithms are executed on the vector representation of COVID-19 images extracted by the U-Net generator. Results in Figure 7 show that the clustering algorithms present a high level of agreement in terms of data points assignments to clusters. This information may be relevant for physicians who may consider the outcome of a single clustering algorithm not convincing or sufficient, and may consider the agreement of different clustering algorithms as an added confidence factor for their analyses.

B. Clustering and visualization

During the inference stage, the vector representation of unseen images (test set) is extracted by the U-Net generator with a feed-forward step that considers all the network layers preceding the up-convolution operations. This step yields a feature vector of 1024, as shown in Figure 2, but it can be customized to accommodate custom vector sizes. The goal of the clustering step is to identify concepts and sub-concepts in the data leveraging robust vector-based features. Subsequently, the vector representation is fed to different clustering algorithms: k-Means, Birch, and Agglomerative. Our rationale for involving different clustering algorithms in this component is twofold. First, offering a diverse and heterogeneous perspective of the results, since the algorithms work in substantially different ways [32] (e.g., k-Means is inherently centroid-based and spherical clusters, while Agglomerative and Birch are hierarchical and may facilitate the identification of finer-grained sub-clusters). Second, provide additional confidence on the reliability of the results in those cases when the different algorithms concur and, therefore, confirm them.

In addition to the clustering algorithms, the vector representation extracted by the U-Net generator is also provided as input to t-SNE [33], a popular technique for data visualization of high-dimensional data which leverages dimensionality reduction. Once the t-SNE model has been trained, and the clustering step has been completed, cluster labels returned by the different clustering algorithms are used to color data points, yielding multiple two-dimensional visualizations.

Gradient-weighted Class Activation Mapping: We select one image from each cluster and extract its Gradient-weighted Class Activation Mapping, highlighting the salient regions that the discriminator deemed significant for its prediction. Results in Figure 8 show that each cluster presents distinctive characteristics, such as the localization of the lesion and its degree of severity. This information is potentially beneficial for a physician, who can leverage the clustering to automatically identify similar groups of patients, and the Grad-CAM representations to gain insights into the characteristics of each cluster. Moreover, the extracted visualizations help the physician to uncover possible false positives in each cluster, i.e. images corresponding to patients who do not relate with the main characteristics of the cluster.

C. Inference explanation

The model architecture of the discriminator is exploited to extract a Gradient-weighted Class Activation Mapping (Grad-CAM) [34]. In our method, Grad-CAM leverages the gradients directed to the final convolutional layer of the discriminator to produce a localization map that highlights the salient regions of the image that led to the model's prediction. The resulting heatmap is used to complement the clustering results with a characterization of each image, which provides a form of inference explanation for the physician. Furthermore, our study argues that different clusters may belong to different forms or severity levels of COVID-19, which may demand a different treatment. Therefore, providing the physicians with the clustering results and the Grad-CAM representation of the images represents a way to support their decision of the medical protocols to follow for different sub-groups of patients.

III. Results

In this section we describe our experimental setup and dataset. The dataset analyzed in this study consists of 3876 images containing Pneumonia lesions, and 150 images containing COVID-19. Each image has a size of 512×512 pixels.

The model architecture is trained end-to-end: while the U-Net generator is trained using exclusively images containing Pneumonia, and learns to generate robust

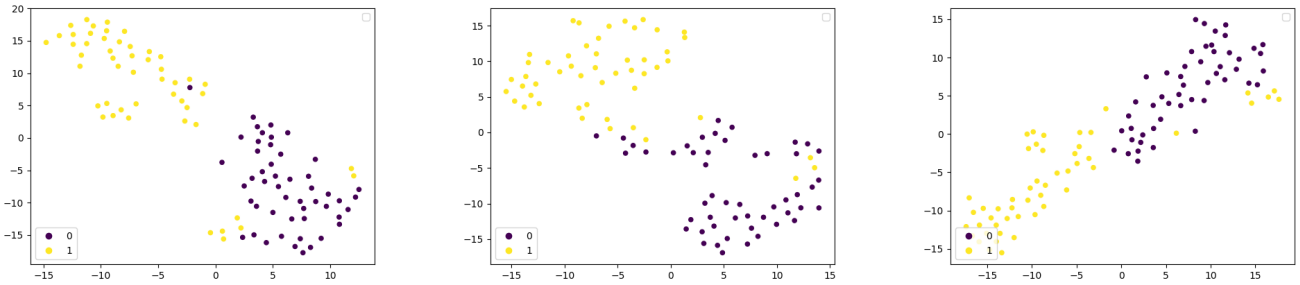


Fig. 4: Classification of Pneumonia and COVID-19 images with the discriminator component, after the training stage of the adversarial learning architecture. Vector features extracted with the U-Net generator are provided as input to t-SNE, and the resulting points are colored using discriminator's predictions. Experiments are repeated three times on random balanced splits of Pneumonia and COVID-19 images, yielding the three sub-plots.

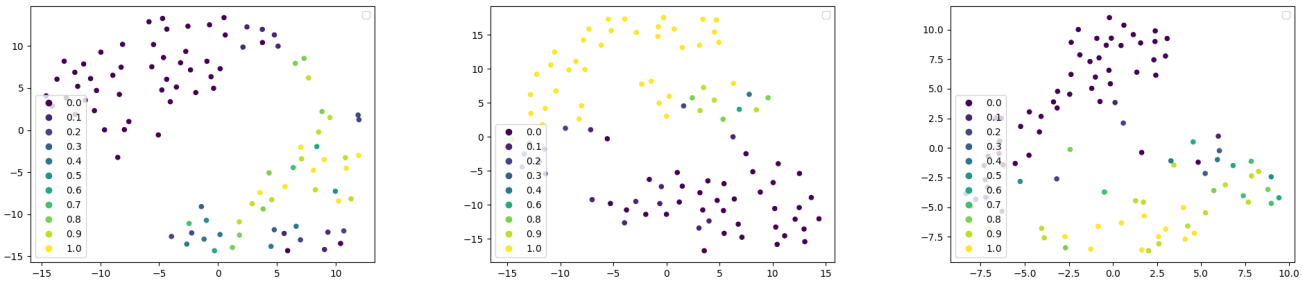
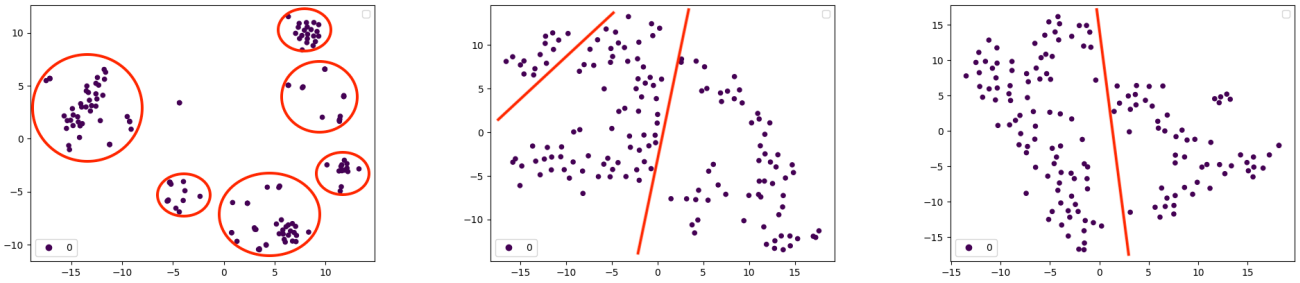
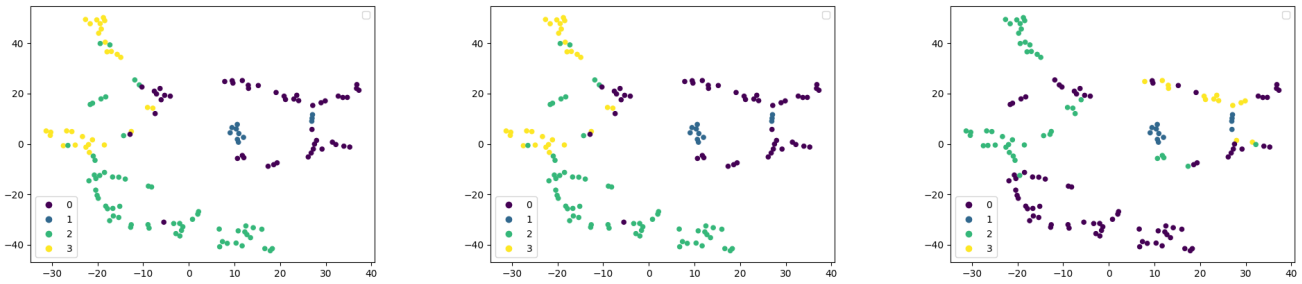


Fig. 5: Two-dimensional representation of the output activation values for the discriminator using only COVID-19 images. Vector features are extracted with the U-Net generator and are provided as input to t-SNE. The resulting points are colored according to the output values returned by the discriminator. Experiments are repeated three times on random splits containing COVID-19 images, yielding the three sub-plots. The distribution of data points and output activation values in all subplots emphasizes a varying confidence of predictions which likely correspond to different degrees of COVID-19.



(a) (b) (c)

Fig. 6: Two-dimensional representation of COVID-19 images. Vector features for the images are extracted with the U-Net generator and are provided as input to t-SNE. Experiments are performed with different vector lengths: 256 (a), 1024 (b), and 4096 (c). Even if all images correspond to patients affected by COVID-19, the distribution of data points in the plots clearly shows the existence of clusters and sub-clusters in the data, which may correspond to different forms or severity levels of COVID-19, which demand different medical treatments. We highlight this separation with manually annotated red circles and lines.



(a) (b) (c)

Fig. 7: Results of the clustering stage with COVID-19 images using Agglomerative (a), Birch (b), and k-Means (c) algorithms. Vector features for COVID-19 images are extracted with the U-Net generator and are provided as input to t-SNE. Resulting points are colored using the cluster number returned by the different algorithms. The results show that subconcepts can be automatically identified, and that there is a strong agreement of the cluster assignments returned by the different clustering algorithms, which may be used as an added confidence factor for the physician and his further analysis.

Input: training data T_R, testing data T_S, vector dimensionality l, number of clusters k
Result: clustering models C, t-SNE plots P_T, Grad-CAM visualizations P_G

// Adversarial training stage
$(G, D) = train(T_R,\ UNet,\ VGG16)$

// Inference stage
$V_{T_S} = extract\_features(T_S,\ G,\ l)$
$C = cluster(V_{T_S},\ k)$
$P_{T_S} = predict(T_S,\ D)$

// Visualization stage
$P_T = t\_sne(V_{T_S},\ C.labels)$
$P_G = grad\_cam(T_S,\ D,\ P_{T_S})$

Algorithm 1: Pseudo-code of the main workflow of the method

segmentation for such images, the discriminator learns to discriminate images according to a binary crossentropy loss function. The model is trained for 100 epochs with the Adam optimizer [35] using the following parameters values: $(bs = 32, l\_rate = 0.0001, beta\_1 = 0.5, beta\_2 = 0.999)$. For both the generator and the discriminator, we apply batch normalization after convolutional operations as a regularization technique, to prevent overfitting. The test folds used to extract our results are balanced and contain 50 Pneumonia images and 50 COVID-19 images. In our experiments, we use different custom values for the vector length (256, 512, and 1024) at this layer. The method is implemented in Python using TensorFlow and Keras libraries [1]. Experiments are performed on a workstation equipped with an Intel Xeon W-2145 (3.7GHz) CPU, 64GB of RAM (DDR4-2666), a 512GB SSD drive, and an NVIDIA Titan V GPU.

IV. Discussion

We present our results in the following subsections. First, we assess our discriminator's ability to classify Pneumonia and COVID-19 images. Second, we assess the confidence of the model about predictions on COVID-19 images. Third, we ascertain the ability of our clustering component to uncover sub-concepts in COVID-19 images. Fourth, we assess the agreement of the results generated by the different clustering algorithms. Fifth, we select images from the extracted clusters and verify if they correspond to different characterizations of COVID-19.

Discriminator assessment: We first consider the discriminator's ability to accurately classify Pneumnonia and COVID-19 images as a pre-condition to proceed with further analysis. Even if this aspect is the main one of interest for classification-based methods, it is not the main focus of our study. After the training stage, we provide balanced splits containing 50 Pneumonia and 50 COVID-19 images to the model. Experiments are repeated three times and results are averaged. The following values are observed: Precision: 0.840, Recall: 0.829, F-Score: 0.828. Subsequently, we extract two-dimensional t-SNE plots leveraging the vector representation of images extracted with the U-Net generator, and coloring the resulting points using the discriminator's prediction. The results represented in Figure 4 show the robustness of the discriminator for the classification task, confirming the success of the adversarial learning process.

[1]The source code and the datasets will be publicly released upon acceptance of the paper. This footnote will be replaced with the link of the GitHub repository.

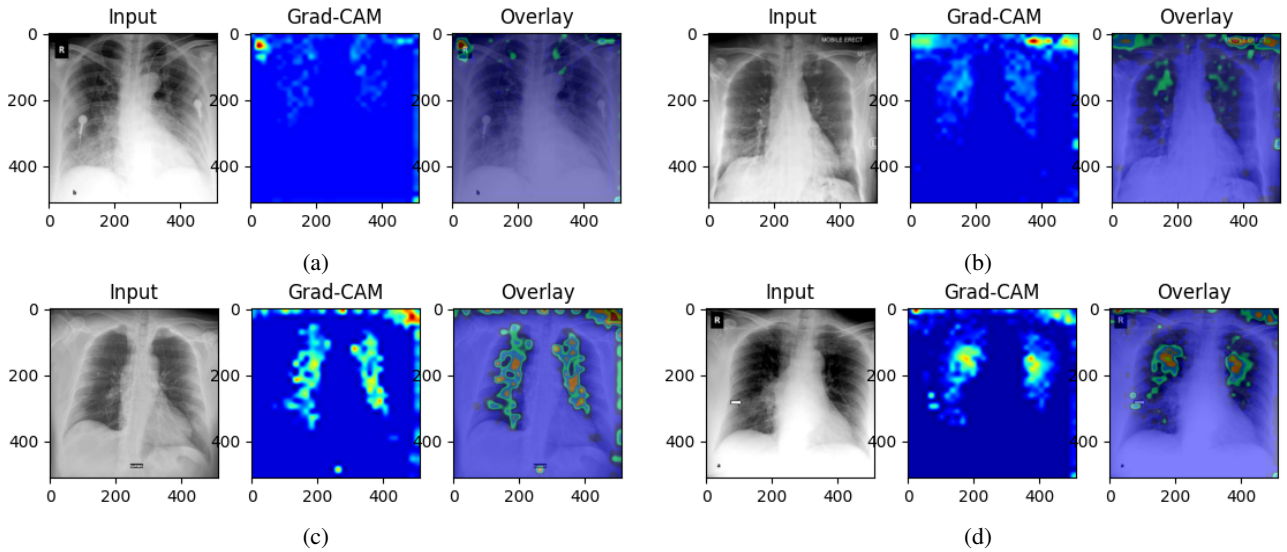


Fig. 8: Gradient-weighted Class Activation Mapping (Grad-CAM) visualization of representative images from four clusters (one image per cluster). The mapping of each image to the clusters can be linked to different characterizations of COVID-19: mild and dispersed (a), mild and diffused with some moderate intensity areas (b), severe and diffused (c), severe and diffused with some extremely severe areas (d). This result confirms the feasibility of our proposed method to identify sub-concepts in the data, which directly link to different severity levels of COVID-19 and may correspond to different medical treatments.

V. Conclusion

Recent innovations in medical imaging and deep learning led to an increased generation of medical data in digital form and emerging analytical models and tools for its analysis. However, most efforts have been dedicated to automated classification pipelines, which are often impractical in a medical healthcare context due to a lack of insights about their predictions. This study presented a hybrid method for decision support in medical healthcare that blends the powerful discriminatory ability of deep and adversarially trained neural networks with dimensionality reduction and clustering techniques. Our experiments with Pneumonia and COVID-19 images showed that our system could extract clusters of images corresponding to the different characterization of patients. Subsequently, Gradient-weighted Class Activation Mapping was used as a tool for inference explanation, with the potential to guide physicians in analyzing each cluster. We conclude that our system can generate insights that can reduce overhead for the physician, who can benefit from the analysis of groups of patients with similar characteristics and the design of better informed medical protocols tailored for the characteristics of each group. As future work, we aim to extend our method to analyze images of different body parts and enhance it with hierarchical modeling capabilities.

Acknowledgments

The work presented in this article was partially funded by DARPA through the project "Lifelong Streaming Anomaly Detection" (Grant N. A19-0131-003 and A21-0113-002). We also acknowledge the support of NVIDIA through the donation of a Titan V GPU.

References

[1] B. Sahiner, A. Pezeshk, L. M. Hadjiiski, X. Wang, K. Drukker, K. H. Cha, R. M. Summers, and M. L. Giger, "Deep learning in medical imaging and radiation therapy," *Medical physics*, vol. 46, no. 1, pp. e1–e36, 2019.

[2] K. H. Cha, L. Hadjiiski, R. K. Samala, H.-P. Chan, E. M. Caoili, and R. H. Cohan, "Urinary bladder segmentation in ct urography using deep-learning convolutional neural network and level sets," *Medical physics*, vol. 43, no. 4, pp. 1882–1896, 2016.

[3] J. Harms, Y. Lei, S. Tian, N. S. McCall, K. A. Higgins, J. D. Bradley, W. J. Curran, T. Liu, and X. Yang, "Automatic delineation of cardiac substructures using a region-based fully convolutional network," *Medical Physics*, 2021.

[4] V. Alex, K. Vaidhya, S. Thirunavukkarasu, C. Kesavadas, and G. Krishnamurthi, "Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation," *Journal of Medical Imaging*, vol. 4, no. 4, p. 041311, 2017.

[5] P. Korfiatis, T. L. Kline, and B. J. Erickson, "Automated segmentation of hyperintense regions in flair mri using deep learning," *Tomography*, vol. 2, no. 4, pp. 334–340, 2016.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[7] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using cnns," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 230–238.

[8] S. A. A. Ismael, A. Mohammed, and H. Hefny, "An enhanced deep learning approach for brain cancer mri images classification using residual networks," *Artificial intelligence in medicine*, vol. 102, p. 101779, 2020.

[9] S. S. Garapati, L. Hadjiiski, K. H. Cha, H.-P. Chan, E. M. Caoili, R. H. Cohan, A. Weizer, A. Alva, C. Paramagul, J. Wei *et al.*, "Urinary bladder cancer staging in ct urography using machine learning," *Medical physics*, vol. 44, no. 11, pp. 5814–5823, 2017.

[10] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J.-L. Suárez, J. Luengo, M. Valero-González *et al.*, "Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images," *IEEE Journal of*

Biomedical and Health Informatics, vol. 24, no. 12, pp. 3595–3605, 2020.

[11] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the covid-19 infected patients using densenet201 based deep transfer learning," *Journal of Biomolecular Structure and Dynamics*, pp. 1–8, 2020.

[12] K. Kamal, Z. Yin, M. Wu, and Z. Wu, "Evaluation of deep learning-based approaches for covid-19 classification based on chest x-ray images," *Signal, image and video processing*, pp. 1–8, 2021.

[13] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *Pattern Analysis and Applications*, pp. 1–14, 2021.

[14] M. Nishio, S. Noguchi, H. Matsuo, and T. Murakami, "Automatic classification between covid-19 pneumonia, non-covid-19 pneumonia, and the healthy on chest x-ray image: combination of data augmentation methods," *Scientific reports*, vol. 10, no. 1, pp. 1–6, 2020.

[15] J. E. Luján-García, M. A. Moreno-Ibarra, Y. Villuendas-Rey, and C. Yáñez-Márquez, "Fast covid-19 and pneumonia classification using chest x-ray images," *Mathematics*, vol. 8, no. 9, p. 1423, 2020.

[16] K. El Asnaoui, Y. Chawki, and A. Idri, "Automated methods for detection and classification pneumonia based on x-ray images using deep learning," in *Artificial Intelligence and Blockchain for Future Cybersecurity Applications*. Springer, 2021, pp. 257–284.

[17] A. U. Ibrahim, M. Ozsoz, S. Serte, F. Al-Turjman, and P. S. Yakoi, "Pneumonia classification using deep learning from chest x-ray images during covid-19," *Cognitive Computation*, pp. 1–13, 2021.

[18] A. Dimitrievski, E. Zdravevski, P. Lameski, M. V. Villasana, I. Miguel Pires, N. M. Garcia, F. Flórez-Revuelta, and V. Trajkovik, "Towards detecting pneumonia progression in covid-19 patients by monitoring sleep disturbance using data streams of non-invasive sensor networks," *Sensors*, vol. 21, no. 9, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/9/3030

[19] N. Antropova, B. Q. Huynh, and M. L. Giger, "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets," *Medical physics*, vol. 44, no. 10, pp. 5162–5171, 2017.

[20] B. Petrovska, E. Zdravevski, P. Lameski, R. Corizzo, I. Štajduhar, and J. Lerga, "Deep learning for feature extraction in remote sensing: A case-study of aerial scene classification," *Sensors*, vol. 20, no. 14, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/14/3906

[21] J. Virmani, R. Agarwal *et al.*, "Deep feature extraction and classification of breast ultrasound images," *Multimedia Tools and Applications*, vol. 79, no. 37, pp. 27 257–27 292, 2020.

[22] W. Liu, C. Qin, K. Gao, H. Li, Z. Qin, Y. Cao, and W. Si, "Research on medical data feature extraction and intelligent recognition technology based on convolutional neural network," *IEEE Access*, vol. 7, pp. 150 157–150 167, 2019.

[23] B. Petrovska, T. Atanasova-Pacemska, R. Corizzo, P. Mignone, P. Lameski, and E. Zdravevski, "Aerial scene classification through fine-tuning with adaptive learning rates and label smoothing," *Applied Sciences*, vol. 10, no. 17, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/17/5792

[24] E. Zdravevski, P. Lameski, C. Apanowicz, and D. Slezak, "From big data to business analytics: The case study of churn prediction," *Applied Soft Computing*, vol. 90, p. 106164, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494620301046

[25] M. S. Kamal, N. Dey, and A. S. Ashour, "Large scale medical data mining for accurate diagnosis: A blueprint," in *Handbook of large-scale distributed Computing in smart healthcare*. Springer, 2017, pp. 157–176.

[26] H. Müller and D. Unay, "Retrieval from and understanding of large-scale multi-modal medical datasets: a review," *IEEE transactions on multimedia*, vol. 19, no. 9, pp. 2093–2104, 2017.

[27] M. Grzegorowski, E. Zdravevski, A. Janusz, P. Lameski, C. Apanowicz, and D. Slezak, "Cost optimization for big data workloads based on dynamic scheduling and cluster-size tuning," *Big Data Research*, vol. 25, p. 100203, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214579621000204

[28] E. Zdravevski, P. Lameski, A. Dimitrievski, M. Grzegorowski, and C. Apanowicz, "Cluster-size optimization within a cloud-based etl framework for big data," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 3754–3763.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[30] B. Krawczyk, C. Bellinger, R. Corizzo, and N. Japkowicz, "Undersampling with support vectors for multi-class imbalanced data classification," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021.

[31] C. Bellinger, R. Corizzo, and N. Japkowicz, "Remix: Calibrated resampling for class imbalance in deep learning," in *International Conference on Discovery Science*. Springer, 2021.

[32] R. Xu and D. Wunsch, *Clustering*. John Wiley & Sons, 2008, vol. 10.

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.