Interpretable COVID-19 Chest X-Ray Classification via Orthogonality Constraint *

Ella Y. Wang
BASIS Chandler
ellawang9@gmail.com

Anirudh Som
SRI International
Anirudh.Som@sri.com

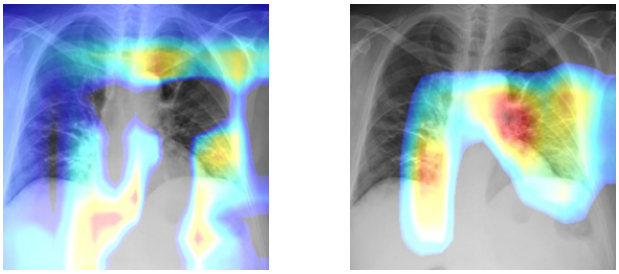
Ankita Shukla
Arizona State University
Ankita.Shukla@asu.edu

Hongjun Choi
Arizona State University
hchoi71@asu.edu

Pavan Turaga
Arizona State University
pturaga@asu.edu

Abstract

Deep neural networks have increasingly been used as an auxiliary tool in healthcare applications, due to their ability to improve performance of several diagnosis tasks. However, these methods are not widely adopted in clinical settings due to the practical limitations in the reliability, generalizability, and interpretability of deep learning based systems. As a result, methods have been developed that impose additional constraints during network training to gain more control as well as improve interpretabilty, facilitating their acceptance in healthcare community. In this work, we investigate the benefit of using Orthogonal Spheres (OS) constraint for classification of COVID-19 cases from chest X-ray images. The OS constraint can be written as a simple orthonormality term which is used in conjunction with the standard cross-entropy loss during classification network training. Previous studies have demonstrated significant benefits in applying such constraints to deep learning models. Our findings corroborate these observations, indicating that the orthonormality loss function effectively produces improved semantic localization via GradCAM visualizations, enhanced classification performance, and reduced model calibration error. Our approach achieves an improvement in accuracy of 1.6% and 4.8% for two- and three-class classification, respectively; similar results are found for models with data augmentation applied. In addition to these findings, our work also presents a new application of the OS regularizer in healthcare, increasing the post-hoc interpretability and performance of deep learning models for COVID-19 classification to facilitate adoption of these methods in clinical settings. We also identify the limitations of our strategy that can be explored for further research in future.



Baseline OS-constrained

Figure 1. GradCAM visualization obtained from DarkCovidNet model [17] (serves as our baseline model) and our approach, highlighting the region of interest learned by the corresponding networks.

1. Introduction

Deep learning techniques have been increasingly used as an adjunct tool in medical science for developing automated solutions for disease diagnosis. For example, they have been used to classify brain disease [13], segment lung and fundus images [24], and detect breast cancer [19]. More recently, due to the wide spread of COVID-19, deep networks have also shown to be useful in developing tools for automated detection of such cases from the chest X-ray images [17, 22, 14, 9]. Thus, providing assistance in accurate and rapid diagnosis that reduces the burden on doctors as well as overcome the limitations of time consuming methods like Reverse Transcription-Polymerase Chain Reaction (RT-PCR).

COVID-19 is often diagnosed with a Reverse Transcription-Polymerase Chain Reaction (RT-PCR)

*Accepted to ACM-CHIL 2021 workshop track.

using upper and lower respiratory specimens [21]. However, the low sensitivity of RT-PCR (60-70%), high false negative rates, long processing times, and shortages of testing kits hinder diagnosis and cause delays in starting treatment [10, 25]. In contrast, radiologic imaging such as computed tomography (CT) and X-ray are promising diagnostics for COVID-19. X-ray evaluations are relatively easy and fast to perform and achieve much higher sensitivity than RT-PCR, making them a more reliable and useful technology for early detection of COVID-19 [2]. CT is widely used in countries such as Turkey where testing kits are largely unavailable. Researchers have found that consolidation, ground-glass opacities, crazy paving pattern, and reticular pattern are common features in CT images of patients with COVID-19; Bernheim et al.[3] observed bilateral and peripheral ground-glass opacities (GGO) as key characteristics, and Li and Xia[12] identified GGO and consolidation as observations. However, such subtle irregularities can only be detected by radiology experts and require valuable time, delaying diagnosis and treatment.

Although deep learning models have achieved significant performance gains in medical tasks, they have not been readily adopted in clinical settings due to their limited reliability, generalizability and interpretability. This limits the practical application of deep learning in healthcare due to a lack of understanding in such methods. Therefore, in order to facilitate the adoption of deep learning models it is increasingly important to elucidate and confer trustworthiness in how these methods work.

Several deep learning approaches to automated detection of COVID-19 from chest X-ray classification have recently been developed [17, 22, 14, 9]. However, the post-hoc interpretability of these models is rather limited as regions of interest tend to be delocalized, resulting in less explainable interpretations of deep-classification networks in terms of semantic localization when input activation maps are visualized by the technique of Grad-CAM, which makes it difficult for radiologists to understand model decisions [18]. Furthermore, there is still much room for improvement in the overall performance and accuracy of existing models.

In this work, we aim to improve the performance of chest X-ray classification and also improve the interpretability to aid in identifying COVID cases. In the medical field, meaningful interpretability is especially important to ensure improved comprehension and explanation of model predictions for end users, such as radiologists. Interpretable deep learning models would assist healthcare personnel in driving more logical and data-driven actions, improving the quality of healthcare.

In this work, we make use of OS parameterization to effectively train deep neural network for automated detection and classification of COVID-19 in chest X-ray images. Our work is primarily driven by the findings of earlier works by Shukla et al.[20]. and Choi et al.[4] that have used OS constraints to improve the generalization of learned representations. Our implementation of OS constraints for chest X-ray image datasets[5, 23] yields improvements in classification performance and better localization and preservation of regions of interest in Grad-CAM heatmap visualizations compared to baseline models. Our OS-constrained model achieved slightly higher accuracy than baseline models [17], and we observe that the OS regularizer resulted in higher activation around lung areas and reduced focus on the background. These findings contribute to greater post-hoc interpretability and performance of deep learning models for detecting COVID-19. Our approach may also provide radiologists more insight into understanding classification decisions and lead to greater acceptance of deep learning models in clinical settings.

2. Related Works

Several studies and research works have been published on the diagnosis of COVID-19 from X-ray images. Hemdan et al.[9] proposed a COVIDX-Net model made up of seven CNN models to detect COVID-19. Minaee et al.[14] prepared a dataset of 5000 chest X-rays and trained four popular CNNs, reporting that ResNet18 and SqueezeNet obtained the best performance. Wang and Wong[22] proposed a COVID-Net deep learning model to diagnose COVID-19 from X-ray images, which achieved 92.4% accuracy in identifying healthy, non-COVID pneumonia-infected, and COVID-19-infected patients. However, these methods used limited data to develop the models. Most notably, T. Ozturk et al.[17] proposed the DarkCovidNet model, which has an end-to-end architecture without the need for manual feature extraction methods. Trained on a more extensive dataset of 1125 chest X-ray images[5, 23], the model achieved superior performance compared to other studies, obtaining 98.08% and 87.02% accuracy for two- and three-class classification, respectively.

In existing approaches, GradCAM[18] heat maps are used to visualize the parts of an image contributed towards the model's classification. We observe from the results of previous works that the heat maps generated from current deep learning models are highly varied. Many visualizations point to delocalized regions of interest outside the lungs, including the shoulder bone and lung bone, despite these areas being unaffected by COVID-19. Such varied heat maps are not meaningful for post-hoc interpretation by radiologists and provide unclear insight regarding which regions of an image contributed to the final prediction. Figure 1 shows GradCAM visualizations obtained from current models, which serve as a baseline, in comparison to our orthogonal spheres approach, in which the regions highlight with red and yellow colors are considered to be important in model decisions. Baseline models convey low interpretabil-

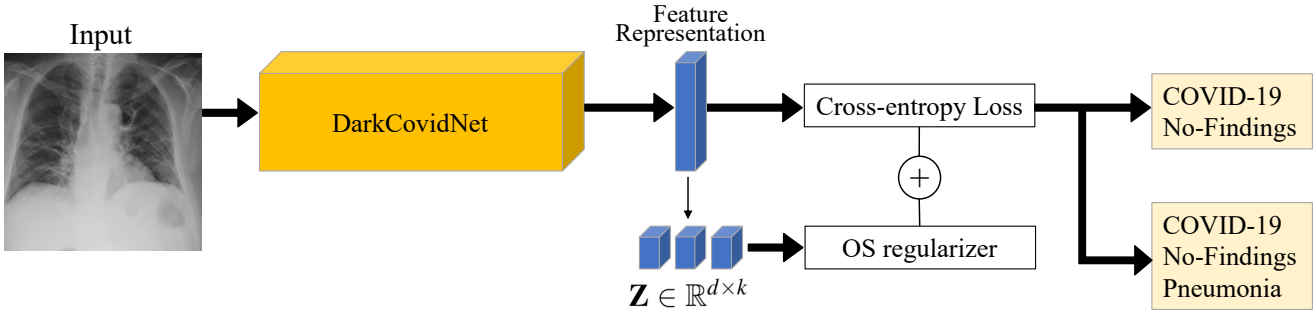


Figure 2. An overview of proposed network modification on DarkCovidNet model with orthogonal spheres constraint. The m dimensional fully connected layer representation is partitioned in k subsets, where each subset representation is of size $d = \frac{m}{k}$ and is stacked to represent a matrix $\mathbf{Z}$. The network is trained with cross entropy loss along with a OS regularizer that penalizes the deviation of $\mathbf{Z}$ from orthogonality condition.

ity of model decisions, but our OS-constrained model highlights more centralized and relevant areas in the lungs.

3. Background

In this section, we provide a brief overview of the two components that are used in developing our strategy for chest X-ray classification for identifying COVID-19 cases.

3.1. DarkCovidNet Model

Ozturk et al.[17] proposed DarkCovidNet model that classifies chest X-ray images into three classes - no-findings, pneumonia, and COVID-19. We used DarkCovidNet model as our baseline model due to its superior performance over existing methods and modify it to incorporate the OS constraint. The emergence of these methods is due to preference of X-ray imaging over CT scans due to their lower radiation dose. The DarkCovidNet model is shown to perform well with sufficient sensitivity in tasks such as detecting ground-glass opacities (GGO) in patients with COVID-19[27]. Further, the DarkCovidNet model was trained with a comparatively larger dataset when compared to other counterpart methods [22, 9, 15], developed for COVID-10 identification from Chest X-Ray images.

Input images are of shape $256x256x3$. The DarkCovidNet model consists of 17 convolution layers and 5 pooling layers. Each DarkNet layer consists of a convolution layer, batch normalization, and a LeakyReLu operation [26]. Batch normalization standardizes inputs, stabilizes the model, and reduces training time. LeakyReLU is a version of the ReLU operation [1] which has a small epsilon value to prevent dying neurons. In the DarkCovidNet model, max pooling is used in all of the pooling operations. The model ends with Flatten and Dense layers that produce the outputs.

The last convolutional layer of the DarkCovidNet model for three classes uses $3 \times 3 \times 1$ convolutional filter with height 3, width 3, and depth 1. With this setup, the baseline

Table 1. Details of DarkCovidNet [17] model for 3-class classification task.

Layer Number	Layer Type	Output Shape	Number of Trainable Parameters
1	Conv2D	[8, 256, 256]	216
2	Conv2D	[16, 128, 128]	1152
3	Conv2D	[32, 64, 64]	4608
4	Conv2D	[16, 66, 66]	512
5	Conv2D	[32, 66, 66]	4608
6	Conv2D	[64, 33, 33]	18,432
7	Conv2D	[32, 35, 35]	2048
8	Conv2D	[64, 35, 35]	18,432
9	Conv2D	[128, 17, 17]	73,728
10	Conv2D	[64, 19, 19]	8192
11	Conv2D	[129, 19, 19]	73,728
12	Conv2D	[256, 9, 9]	294,912
13	Conv2D	[128, 11, 11]	32,768
14	Conv2D	[256, 11, 11]	294,912
15	Conv2D	[128, 13, 13]	256
16	Conv2D	[256, 13, 13]	294,912
17	Conv2D	[3, 13, 13]	6915
18	Flatten	[338]	0
19	Linear	[3]	678

DarkCovidNet model has a total of 1,170,811 parameters. This convolutional layer is modified in our experiments to incorporate the OS constraint that requires the representation to be split into k feature blocks of equal dimensions.

3.2. Orthogonal Spheres

We make use of the OS parameterization proposed by Shukla et al. [20] in generative model setting and adapt it for our classification setting. For a given input image, let $\mathbf{Z} \in \mathbb{R}^m$ represent the output of a specific layer from the CNN model, where m is the feature dimension. We parti-

tion this representation in k feature blocks as $\mathbf{Z} \in \mathbb{R}^{d \times k} = [\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^k]$, where k represents the number of partitions and d is the dimension of each partition that is obtained as $d = \frac{m}{k}$. To make the matrix $\mathbf{Z} \in \mathbb{R}^m$ as orthogonal as possible, we regularize the off-diagonal elements in the matrix to be zero. Applying this orthogonality condition on the matrix $\mathbf{Z} \in \mathbb{R}^m$, we arrive at the simple orthonormality term shown below

$$L_{\text{OS}} = \left\| \mathbf{Z}^\top \mathbf{Z} - \mathbf{I} \right\|_F^2 \tag{1}$$

Here, L_{OS} represents the OS regularizer and $\mathbf{I}$ represents the $k \times k$ identity matrix, with $\|\cdot\|_F$ being the Frobenius norm. The OS regularizer is applied along with the standard cross-entropy loss function.

This OS constraint was recently employed by Choi et al. [4] and have that the network learns more diverse representations, reducing model calibration error while effectively improving the semantic localization. These improvements were shown on standard computer vision daatsets like CIFAR10[11], CIFAR100[11], SVHN[16], and tiny ImageNet datasets[6]. In this work, we explore and harness the capabilities of OS constraints for medical images to improve the intepretability of results, hence making them acceptable to medical practitioners.

4. Proposed Strategy

Deep networks are conventionally trained using the categorical-cross-entropy loss function for classification task. However, models obtained using this loss function tend to exhibit low interpretability, feature redundancy, and poor calibration. Instead, we approach this problem with orthogonal-sphere (OS) constraints. The OS parameterization discussed in subsection 2.3 is applied to output of the flatten layer following the last convolutional layer of the DarkCovidNet model. In doing so, we sought to reduce the number of correlated features learnt by deeper layers in the network. Our training pipeline for the proposed implementation of the OS regularizer is depicted in Figure 2.

The OS regularization function was used together with regular categorical cross-entropy loss. Thus, with L_{OS} representing the OS regularizer, our total loss function can be characterized as

$$L_{\text{Total}} = \lambda L_{\text{cross-entropy}} + (1 - \lambda) L_{\text{OS}}. \tag{2}$$

Here, $0 \leq \lambda \leq 1$ is a trade-off parameter.

5. Experimental Results

5.1. Dataset Description

Our experiments are conducted on the same dataset as used by Ozturk et al.[17]. The dataset has three classes: COVID-19 cases, pneumonia and healthy or no-finding. The images for COVID-19 class are obtained from an open source database of COVID-19 chest X-ray images collected by Cohen *et al.* [5]. This database is continuously updated with images submitted by researchers. Currently, there are 132 X-ray images of COVID-19 diagnosis in the database, out of which 125 are confirmed to be positive. We use these 125 images for the COVID-19 class in our experiments. In the healthy (no-findings) and pneumonia classes, 500 chest X-ray images for each class were obtained randomly from the ChestX-ray8 database collected by Wang et al. [23], making a total of 1125 images in the dataset.

5.2. Experimental Setup

We performed experiments to classify COVID-19 from chest X-ray images in two different scenarios. First, we trained the DarkCovidNet model (Baseline) and OS-constrained model (Baseline + OS) to classify X-ray images into three classes: COVID-19, Pneumonia, and No-Findings. Secondly, the performance of these two models was evaluated in a classification task with two classes: COVID-19 and No-Findings. The performance of the models are evaluated using 5-fold cross-validation - the models are evaluated for each fold, and the average classification performance of the model is calculated. We use a 80/20 split for training and testing.

5.3. Training Protocol and Hyper-parameter Settings

All the experiments are conducted using a NVIDIA Tesla P100 GPU and Python 3.7 with Tensorflow 2.3.0. Our models are trained for 100 epochs using the Adam optimizer, batch-size = 32, and initial learning-rate = 0.003. We used the default Adam momentum parameters: β_1 = 0.9 and β_2 = 0.999. Following the implementation of the DarkCovidNet model by T. Ozturk et al.[17], we apply exponential learning rate decay to decay every 1000 steps with a base of 0.7. We apply batch-normalization with leaky ReLu activation with α = 0.1. To account for the class imbalance due to smaller number of COVID-19 images, i.e. 125 samples compared to 500 in the No-Findings and Pneumonia classes, we assign COVID-19 class four times the weight of the other two classes. The baseline DarkCovidNet model is trained using categorical cross-entropy loss function, while the OS-constrained model is trained by augmenting this loss with the orthogonality loss. During network training, we use random horizontal flipping and slight vertical and horizontal image translation for data augmentation. When interpreting experimental results, the label "baseline" represents the original DarkCovidNet model trained with only cross-entropy loss; the label "+OS" signifies that OS-constraint applied on the baseline model, and so both cross-entropy loss and the OS regularizer are applied when training the

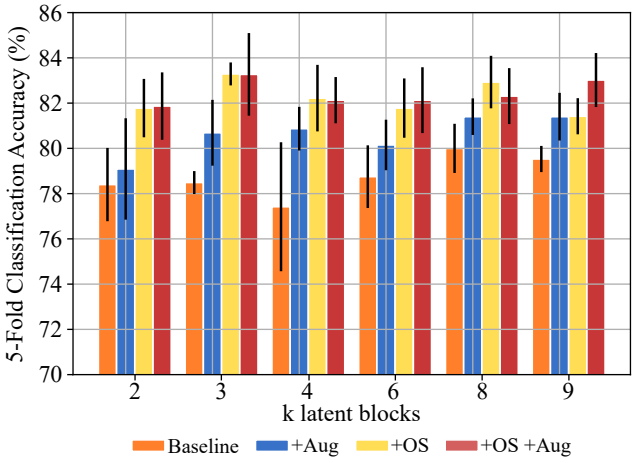


Figure 3. 5-fold average 3-class classification accuracy for baseline and OS-constrained models trained with and without data augmentation using different values of k. Standard deviation bars are included.

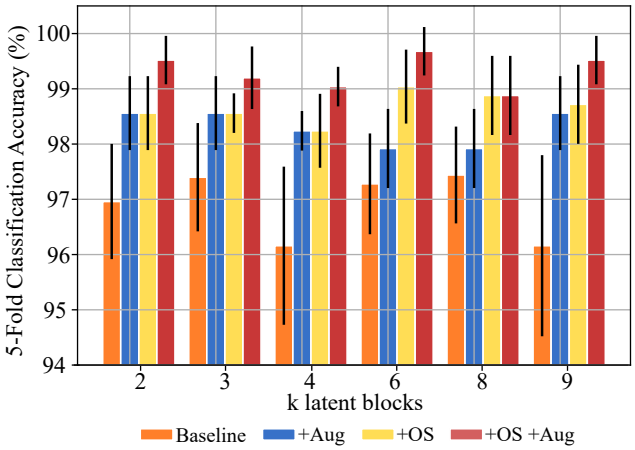


Figure 4. 5-fold average 2-class classification accuracy for baseline and OS-constrained models trained with and without data augmentation using different values of k.

model; the label "+Aug" signifies that data augmentation is used when training the model. It should be noted that Drak-CovidNet model does not use data augmentation during network training. In order to obtain flattened output with units divisible by k, the last layer filter in the model is modified to dimensions of kxkx1. For example, in experiments using k = 4, the last convolutional layer had a filter size of 4x4x1, resulting in a flattened output size of 676 units. Experiments were performed using k = 2, 3, 4, 6, 8, and 9.

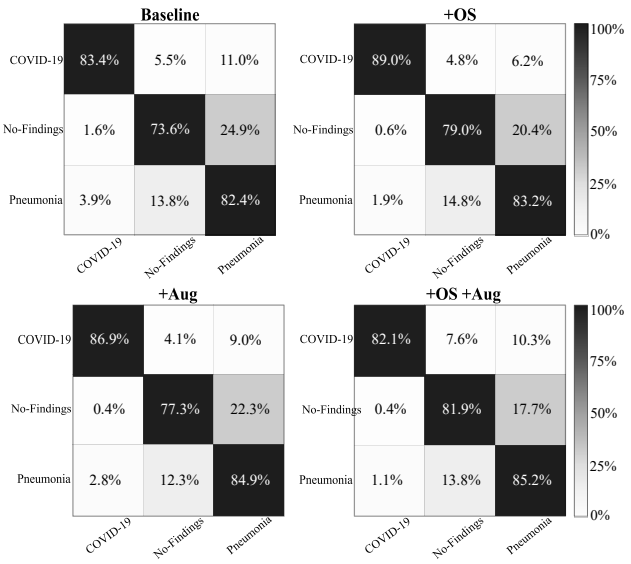


Figure 5. Confusion matrices for 3-class classification. The first row represents performance of regular baseline and OS-constrained models. The second row represents results obtained from the models with data augmentation applied.

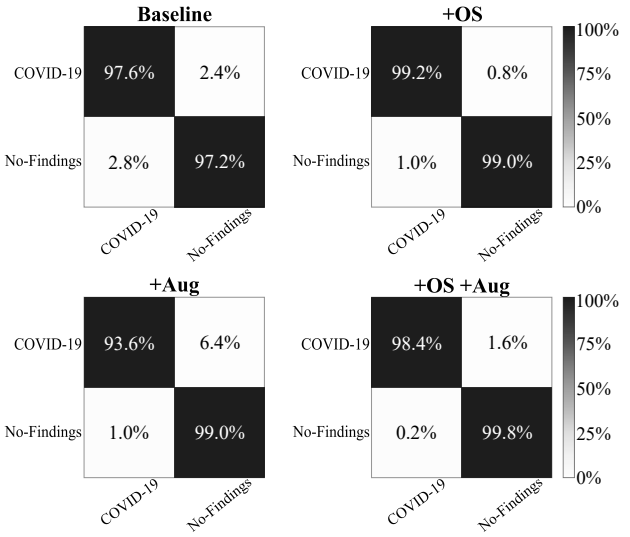


Figure 6. Confusion matrices for 2-class classification. The first row represents performance of baseline and OS-constrained models. The second row represents results obtained from the models with data augmentation applied.

6. Results

6.1. Optimizing k Value

To optimize the OS-constrained model, we performed experiments to determine the value of k that results in the highest classification accuracy. As mentioned previously, k

Table 2. Comparison of baseline and OS-constrained models with and without data augmentation in 3-class classification setting. The best results are reported in bold.

	Accuracy	Precision	Recall	F1-score
Baseline	78.49 ± 0.578	79.68 ± 1.31	79.54 ± 0.612	79.03 ± 0.492
+Aug	81.69 ± 1.95	83.21 ± 1.91	83.57 ± 1.95	83.63 ± 1.94
+OS	82.03 ± 0.798	**83.67** ± 0.879	84.13 ± 0.832	**84.82** ± 1.04
+OS +Aug	**83.29** ± 1.19	83.14 ± 0.935	**86.78** ± 1.03	84.32 ± 1.19

Table 3. Comparison of baseline and OS-constrained models with and without data augmentation for 2-class classification. The best results are highlighted in bold.

	Accuracy	Precision	Recall	F1-score
Baseline	97.28 ± 1.62	97.60± 1.41	89.71 ± 1.64	93.49 ± 1.60
+Aug	97.92± 0.669	93.60 ± 0.686	95.90 ± 0.663	94.74 ± 0.677
+OS	99.04 ± 0.716	**99.20** ± 0.711	96.12 ± 0.721	97.64 ± 0.713
+OS +Aug	**99.52** ± 0.438	98.4± 0.433	**99.19** ± 0.413	**98.8** ± 0.444

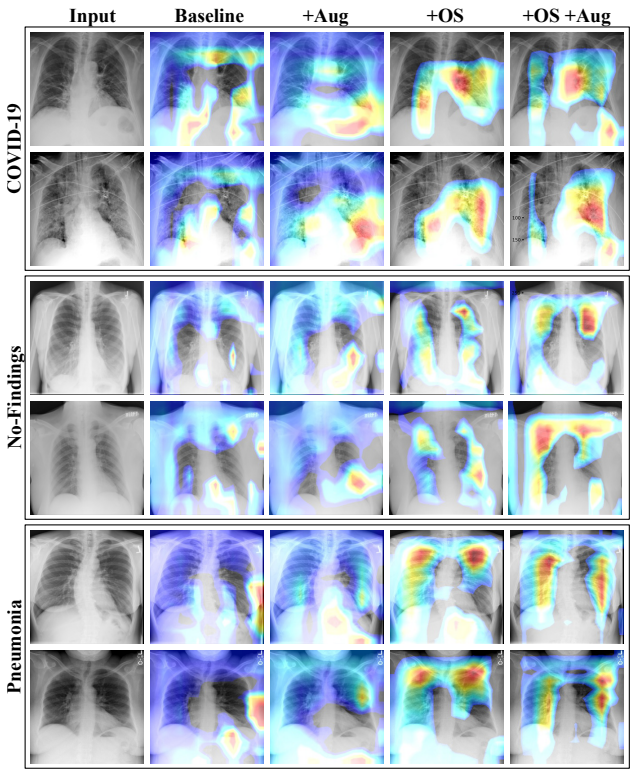


Figure 7. Comparison of Grad-CAM visualization obtained from DarkCovidNet (baseline) and OS-constrained models with and without data augmentation. The OS constraint uses $k = 3$.

represents the number of partitions of the flatten layer. The performance of the OS-constrained model with different k values was evaluated for two- and three-class classification tasks. Figure 4 and 3 show performance for 2 class and 3 class classification respectively. In case of 2 classes, OS-

[h!]

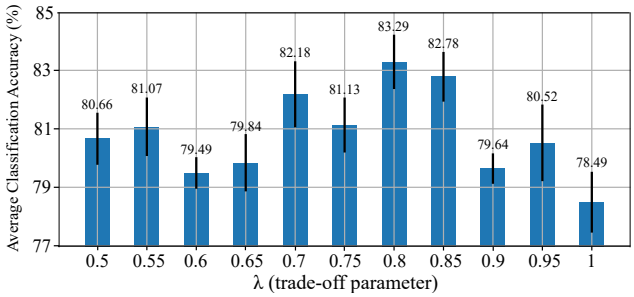


Figure 8. Average classification accuracy over 5 folds obtained for different values of λ. $\lambda = 1$ indicates the results obtained from the baseline model trained only with cross-entropy loss.

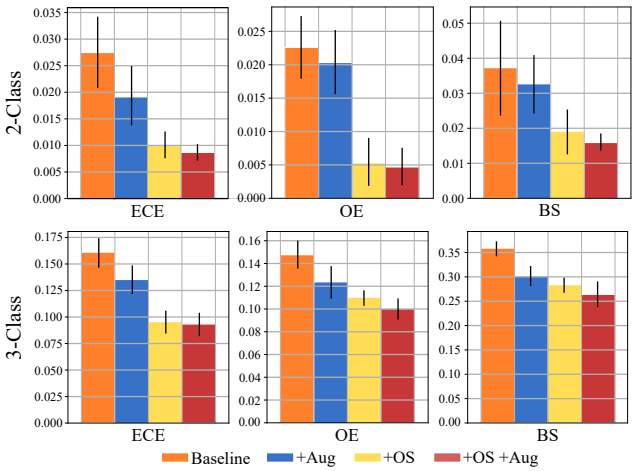


Figure 9. Comparison of average calibration metric scores across 5-folds for 2 (top row) and 3 (bottom row) class problem across different methods.

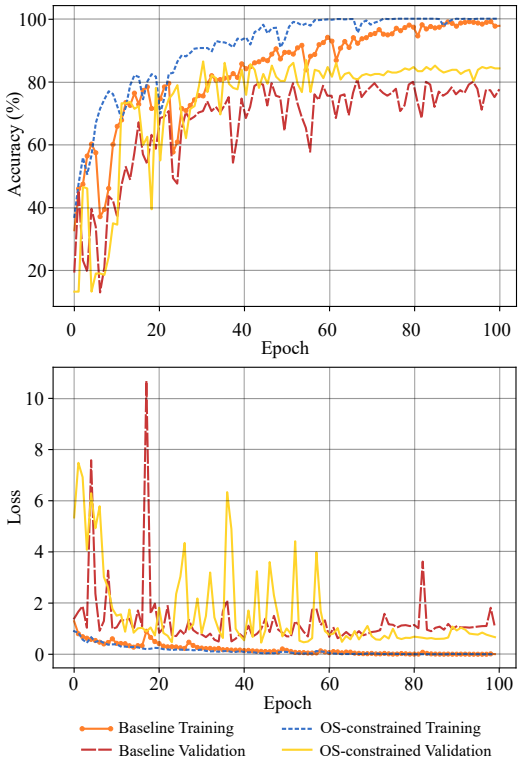


Figure 10. Training and validation accuracy/loss curves for the baseline and OS-constrained (baseline +OS) models.

constrained model (+OS) and OS-constrained model with data augmentation (+OS+Aug), achieve marginally higher performance for $k = 6$. On the other hand, Figure 3 reports classification results of the OS-constrained model for three classes. With three classes, we find that the average accuracy is highest for $k = 3$. These respective k values are used in all other experiments.

6.2. Three-Class Classification

As shown in Figure 3, the OS-constrained model performs slightly better than the baseline model for all values of k. Taking the optimal value, *i.e.,* $k = 3$, Table 2 shows the average accuracy, precision, recall, and F1-score across 5 folds for the OS-constrained and baseline models. The DarkCovidNet model obtains an average classification accuracy of 78.49% and 81.69% without and with data augmentation respectively. In comparison, the OS-constrained models obtains an average accuracy of 83.29% and 83.27% in the same scenario, with approximately 3-5 % improvement over the baseline models. It can be noted that data augmentation marginally improves classification performance for both models. Additionally we also computed the confusion matrices are shown in Figure 5 for more detailed analysis of the three class problem. As pointed in [17], the deep learning model is better at classifying COVID-19 than pneumonia and no-findings classes. These improvements in classification performance in the OS-constrained model can be attributed to more diverse feature representations, reduced model calibration error, and improved robustness by the OS regularizer.

6.3. Two-Class Classification

Next we evaluate the performance of our OS-regularized model for the two-class classification task, involving only the COVID-19 and No-Findings classes. Figure 4 displays the average accuracy obtained from the OS and baseline models for various k values. Again, we find that the classification performance of the OS-constrained model is consistently higher than the DarkCovidNet model by a slight margin. For 2-class classification, k is optimized at 6, and Table 3 details specific performance metrics across 5 folds. The average accuracy of the OS-constrained model was 99.04% compared to 97.44% by the baseline model and 99.68% compared to 97.92% for the models with data augmentation, reflecting a 1-2 percentage point difference. It can be noted that the performance of both OS-constrained models surpassed the 98.08% accuracy reported by T. Ozturk et al.[17] for the DarkCovidNet model.

We have also included in Figure 6 the overlapped confusion matrices obtained over 5 folds, where we find that our OS-constrained model achieved slightly higher performance overall.

6.4. Grad-CAM Visualizations

We obtained Grad-CAM[18] heat maps to visually depict decisions made by the deep learning model. The heatmap reveals regions of the X-ray image which contributed most to the model's classification. The images in Figure 7 represent Grad-CAM visualizations of 6 test images from the chest X-ray dataset, with 2 images per class, obtained from four experimental models for 3-class classification. Similar to the findings of T. Ozturk et al.[17], the baseline DarkCovidNet model highlights more scattered areas outside the lungs, such as the chest bone, shoulders, and diaphragm, which are generally irrelevant to diagnosis and may hinder post-hoc interpretability. Although applying data augmentation to the baseline model seems to consolidate some regions, overall these areas are not helpful in understanding model decisions. Instead, the OS regularizer captures more exact and localized areas within the lobes of the lungs, suggesting improved semantic interpretation as regions of interest are better preserved. Similar to the baseline model, applying data augmentation to the OS-constrained model helped identify more relevant areas in the image. It can be noted that the OS-constrained model seems to focus more on the right side of the lung when classifying COVID-19, but emphasizes both sides of the lung for the No-Findings and Pneumonia classes. We observe that

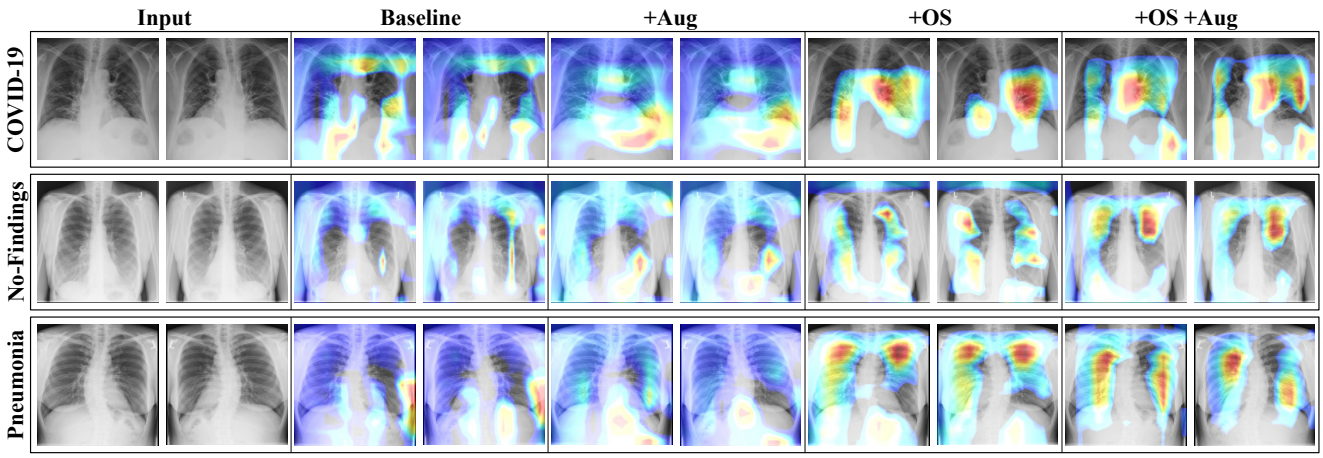


Figure 11. Grad-CAM visualization for baseline and OS-constrained models with and without data augmentation methods. The OS constraint uses k = 3. For each pair of two columns, the first column displays the visualizations obtained from the original images, and the second column represents visualizations obtained from the horizontally flipped images.

the Grad-CAM heatmaps obtained from the OS-constrained model highlight very specific lung regions which may help radiologists identify diagnostic features such as ground-glass opacities and consolidation[12].

7. Ablation Study

7.1. Effect of λ parameter

The λ parameter in Eq. 2 governs the contribution of OS constraint during network training. We analyzed the behaviour of network performance for different value of λ in the range $[0, 1]$ Figure 8 shows the average accuracy obtained for different value of λ parameter for 3-class classification performance with $k = 2$. We observe that the optimal performance of the model is achieved for λ=0.8. This value of λ used for all other experiments.

7.2. Model Calibration

We also evaluate how well models were calibrated using the OS regularizer. Calibration metrics allow us to determine whether the predicted softmax scores obtained from the model are good indicators of the actual probability of the correct predictions. Our models are assessed using the Expected Calibration Error (ECE), Overconfidence Error (OE), and Brier Score (BS) [8, 7]. These calibration metrics is defined as:

- ECE = $\sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|$
- OE = $\sum_{m=1}^{M} \frac{|B_m|}{N} [\text{conf}(B_m) \times \max(\text{conf}(B_m) - \text{acc}(B_m), 0)]$
- BS = $\frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} [p_\theta(\hat{y}_n = k|x_n) - 1(y_n = k)]^2$

Here, B_m represents the number of predictions falling in bin m and K represents the number of classes. acc(B_m) denotes the accuracy of the model and conf(B_m) denotes the model's average confidence. Figure 9 shows the calibration metric scores obtained from our baseline and OS-constrained models. Note, models with lower calibration scores are better. We find that lower calibration scores are obtained when we implement the OS regularizer with the baseline model, and data augmentation has slightly reduces calibration scores. These findings are especially significant for the 2-class classification task.

7.3. Effect on Network Training

Figure 10 shows the validation and training accuracy and loss curves for the baseline and OS-constrained models. The accuracy curves reveal that the OS-constrained model tends to achieve higher training and validation accuracy compared to the baseline model throughout the training period. The loss curves for both models are relatively similar, although the validation loss of the OS-constrained model shows slightly more volatility than the baseline model.

7.4. Horizontal Flipping

In this subsection we study the effect of horizontally flipping input images on GRAD-Cam visualizations. Input images were mirrored across the vertical axis for testing. Using the OS-regularized and baseline models for 3-class classification task to obtain predictions, we evaluate the Grad-CAM heatmaps resulting from these modified images. Despite flipping the images, the heatmaps shown in Figure 11 stayed relatively consistent as those obtained from our previous experiments for all models, with highlighted regions only exhibiting slight shifts. For example, the regions

emphasized by the OS-constrained models with data augmentation remained concentrated on the right side of the lung in the COVID-19 class. Since these highlighted regions were not mirrored after horizontally flipping the input images, these results suggest that despite improved performance achieved by the OS regularizer, our model still lacks robustness to transformed data. In future research, other techniques may be further explored in conjunction with OS-constraints to improve the robustness of deep learning models.

Acknowledgment

This work was supported in part by NSF RAPID grant 2029044.

8. Conclusion

In this work, we studied orthogonality constraint imposed on a deep learning model to classify COVID-19 cases from chest X-ray images. The proposed OS regularization yields improved performance compared to the baseline DarkCovidNet model, obtaining a classification accuracy of 83.29% over 78.49% for three classes, and 99.04% over 97.44% accuracy for two classes without augmentation. Our OS-constrained model generates more localized and interpretable activation maps that can assist radiologists in understanding classification decisions and improving acceptance of deep learning models in the clinical settings. In future work, it is promising to explore applications of orthogonality constraints in other medical imaging tasks such as the diagnosis of chest-related diseases including pneumonia or tuberculosis.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019. 3

[2] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in china: A report of 1014 cases. *Radiology*, 296(2):E32–E40, 2020. 2

[3] Adam Bernheim, Xueyan Mei, Mingqian Huang, Yang Yang, Zahi A. Fayad, Ning Zhang, Kaiyue Diao, Bin Lin, Xiqi Zhu, Kunwei Li, Shaolin Li, Hong Shan, Adam Jacobi, and Michael Chung. Chest ct findings in coronavirus disease 2019 (covid-19): Relationship to duration of infection. *Radiology*, 295:685–691, 2020. 2

[4] Hongjun Choi, Anirudh Som, and Pavan Turaga. Role of orthogonality constraints in improving properties of deep networks for image classification, 2020. 2, 4

[5] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection, 2020. 2, 4

[6] Jia Deng, Wei Dong, Richard Sochard, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 4

[7] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007. 8

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, page 1321–1330, 2017. 8

[9] Ezz El-Din Hemdan, Marwa A. Shouman, and Mohamed Esmail Karar. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images, 2020. 1, 2, 3

[10] Jeffrey Kanne, Brent Little, Jonathan Chung, Brett Elicker, and Loren Ketai. Essentials for radiologists on covid-19: An update- radiology scientific expert panel. *Radiology*, 296(2):E113–E114, 2020. 2

[11] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 4

[12] Yan Li and Liming Xia. Coronavirus disease 2019 (covid-19): Role of chest ct in diagnosis and management. *American Journal of Roentgenology*, 214(6):1280–1286, 2020. 2, 8

[13] Medhani Menikdiwela, Chuong Nguyen, and Marnie Shaw. Deep learning on brain cortical thickness data for disease classification. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–5, 2018. 1

[14] Shervin Minaee, Rahele Kafieh, Milan Sonka, Shakib Yazdani, and Ghazaleh Jamalipour Soufi. Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning. *Medical Image Analysis*, 65:101794, 2020. 1, 2

[15] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks, 2020. 3

[16] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. In *Neural Information Processing Systems*, 2011. 4

[17] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U. Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in Biology and Medicine*, 121:103792, 2020. 1, 2, 3, 4, 7

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2, 7

[19] Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*, 9(1), Aug 2019. 1

[20] Ankita Shukla, Sarthak Bhagat, Shagun Uppal, Saket Anand, and Pavan Turaga. Product of orthogonal spheres parameterization for disentangled representation learning, 2019. 2, 3

[21] Buddhisha Udugama, Pranav Kadhiresan, Hannah N. Kozlowski, Ayden Malekjahani, Matthew Osborne, Vanessa Y. C. Li, Hongmin Chen, Samira Mubareka, Jonathan B. Gubbay, and Warren C. W. Chan. Diagnosing covid-19: The disease and tools for detection. *ACS Nano*, 14(4):3822–3835, 2020. 2

[22] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549, 2020. 1, 2, 3

[23] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 2, 4

[24] Qian Wu and Abbas Cheddad. Segmentation-based deep learning fundus image analysis. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–5, 2019. 1

[25] Xingzhi Xie, Zheng Zhong, Wei Zhao, Chao Zheng, Fei Wang, and Jun Liu. Chest ct for typical coronavirus disease 2019 (covid-19) pneumonia: Relationship to negative rt-pcr testing. *Radiology*, 296(2):E41–E45, 2020. 2

[26] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network, 2015. 3

[27] Zi Yue Zu, Meng Di Jiang, Peng Peng Xu, Wen Chen, Qian Qian Ni, Guang Ming Lu, and Long Jiang Zhang. Coronavirus disease 2019 (covid-19): A perspective from china. *Radiology*, 296(2):E15–E25, 2020. 3