xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography

Arnab Kumar Mondal*, Arnab Bhattacharjee*, Parag Singla and Prathosh AP

***Abstract*— Since its outbreak, the rapid growth of COrona VIrus Disease 2019 (COVID-19) across the globe has pushed the health care system in many countries to the verge of collapse. Therefore, it is imperative to correctly identify COVID-19 positive patients and isolate them as soon as possible to contain the spread of the disease and reduce the ongoing burden on the healthcare system. The primary COVID-19 screening test, RT-PCR although accurate and reliable, has a long turn-around time. In the recent past, several researchers have demonstrated the use of Deep Learning (DL) methods on chest radiography (such as X-ray and CT) for COVID-19 detection. However, existing CNN based DL methods fail to capture the global context due to their inherent image-specific inductive bias. Motivated by this, in this work, we propose the use of vision transformers (instead of convolutional networks) for COVID-19 screening using the X-ray and CT images. We employ a multi-stage transfer learning technique to address the issue of data scarcity. Furthermore, we show that the features learned by our transformer networks are explainable. We demonstrate that our method not only quantitatively outperforms the recent benchmarks but also focuses on meaningful regions in the images for detection (as confirmed by Radiologists), aiding not only in accurate diagnosis of COVID-19 but also in localization of the infected area. The code for our implementation can be found here - `https://github.com/arnabkmondal/xViTCOS`.**

***Index Terms*— COVID-19 Detection, AI for COVID-19, Deep Learning, Vision Transformer, Chest Radiography, COVID-19 Detection Using CT Scan and CXR**

I. Introduction

A. Background

The novel COronaVIrus Disease 2019 (COVID-19) is a viral respiratory disease caused by Severe Acute Respiratory Syndrome COronaVirus 2 (SARS-CoV2). The World Health Organization (WHO) has declared COVID-19 a pandemic on 11 March 2020 [1]. This has pushed the health systems of several nations to the verge of collapse. It is, therefore, of utmost importance to screen the positive COVID-19 patients accurately for efficient utilization of limited resources. Two types of viral tests are currently popularly used to detect COVID-19 infection: Nucleic Acid Amplification Tests (NAATs) [2] and Antigen Tests [3]. NAATs can reliably detect SARS-CoV-2 and are unlikely to return a false-negative result of SARS-CoV-2. NAATs can use many different methods, among which Reverse Transcription Polymerase Chain Reaction (RT-PCR) is the most preferred test for COVID-19 due to its high specificity and sensitivity [4]. However, this test is expensive as it has an elaborate kit and time-consuming. An RT-PCR test uses nose or throat swabs to detect SARS-CoV-2 and requires trained professionals instructed for the RT-PCR kit to carry out the RT-PCR test. RT-PCR requires a complete set-up that includes the trained practitioners, laboratory, and RT-PCR machine for detection and analysis.

All the authors are with Indian Institute of Technology Delhi, New Delhi 110016, India. Email: anz188380@iitd.ac.in, arnab.bhattacharjee@uqidar.iitd.ac.in, paragsg@iitd.ac.in, prathoshap@iitd.ac.in., * indicates equal contribution.

B. Scope and Contributions

In the recent past, deep neural network models such as CheXNet [5] have been employed for detecting abnormalities such as Pneumonia from chest X-ray images. These networks achieved extraordinary results exceeding average radiologist performance [5], [6]. Motivated by such studies, several recent works have proposed the use of chest radiography images (X-ray and Computed Tomography, CT) as alternate modality to detect COVID-19 positive cases [7]–[13] (Elaborated in Sec. II). Unlike in the chest CT/X-ray of a healthy person, the lungs of COVID-19 affected patients show some visual marks like ground-glass opacity and/or mixed ground-glass opacity, and mixed consolidation [7].

While there has been a large body of literature on use of Deep Learning for Covid detection, most of them are based on Convolutional Neural Networks (CNNs) [13]–[16]. CNN, albeit powerful, lacks a global understanding of images because of its image-specific inductive biases. To capture long-range dependencies, CNNs require a large receptive field, which necessitates designing large kernels or immensely deep networks, leading to a complex model challenging to train. Recently, Vision transformers [17] have provided an alternative framework for learning tasks and overcome the issues associated with convolutional inductive bias as they can learn the most suitable inductive bias depending on the task at hand. Motivated by this, in this work, we propose to employ a vision transformer (ViT) based transfer learning method to detect COVID-19 infection from the chest radiography (X-ray and CT scan imaging). With this, we aim to develop an explainable model and employ a multi-stage transfer learning method to address the need for large-scale data. Specifically, the below are our contributions:

1) We propose a vision transformer based deep neural classifier, xViTCOS for screening of COVID-19 from chest radiography.

2) We provide explanability-driven, clinically interpretable visualizations where the critical patches responsible for the model's prediction are highlighted on the input image.
3) We employ a multi-stage transfer learning approach to address the problem of need for large-scale data.
4) We demonstrate the efficacy of the proposed framework in distinguishing COVID-19 positive cases from non-COVID-19 Pneumonia and Normal control using both chest CT scan and X-ray modality, through several experiments on benchmark datasets.

II. Related Work

The research community has proposed various novel deep neural networks for automated screening of COVID-19 cases using chest radiography. Most works in literature focus either employ an off-the-shelf CNN pre-trained on generic image datasets or prior feature extraction via different feature engineering and selection techniques for COVID-19 diagnosis from radiology data. While some of the existing methods are binary classifiers to distinguish between COVID-19 positive and negative cases [14], [18], [19]. Several other works [16], [20]–[25] propose a three-class-classifier (COVID-19, non-COVID-19 pneumonia and normal). Some work [13], [26] differentiate non-COVID-19 Pneumonia further, e.g., bacterial Pneumonia, viral Pneumonia. In the following two sections, we describe the prior literature for diagnosing COVID-19 using chest radiography.

A. COVID-19 Detection Using Chest CT

Chest Computed Tomography (CT) imaging has been proposed as an alternative screening tool for COVID-19 infection [7], [8]. It has been observed that chest CT exhibits higher sensitivity as compared to RT-PCR [9], [10].

In [27] multiple types of features, like Volume, Radiomics features, Infected lesion number, Histogram distribution and Surface area are extracted first from the CT images following which a deep forest algorithm, consisting of cascaded layers of multiple random forests, is used for discriminative feature selection and classification. The final label is obtained by aggregating the predictions of the last layer of random forests. However, the use of manual feature engineering often leads to sub-optimal performance of the classification module. This makes deep learning based methods more attractive.

The work in [14] performs a comparative study by exploiting transfer-learning to optimize 10 pre-trained CNN models viz AlexNet [28], VGG-16 [29], VGG-19 [29], SqueezeNet [30], GoogleNet [31], MobileNet-V2 [32], ResNet-18 [33], ResNet-50 [33], ResNet-101 [33], and Xception [34] on CT-scan images to differentiate between COVID-19 and non-COVID-19 cases. As per the results reported in [14], ResNet-101 and Xception achieve best performance. [25] segment out candidate infection regions from the pulmonary CT image set using a 3D CNN segmentation model and categorize these segments into the COVID-19, IAVP, and irrelevant to infection (ITI) groups, together with the corresponding confidence scores, using a location-attention classification model. COVNet [35] is a ResNet50 based CNN architecture that takes as input a series of CT slices and compute features from each slice of the CT series, which are combined by a max-pooling operation, and the resulting feature map is fed to a fully connected layer to generate a probability score for each class. [36] uses a pre-trained EfficientNet as the backbone and extracts features from each slice of CT data, and makes a binary prediction. Next, the slice level predictions are combined using a multi-layer perceptron (MLP) to make a final prediction at the patient level. COVIDNet-CT [16] is a architecture developed via machine-driven design exploration. Notable characteristics of COVIDNet-CT include high architectural diversity (heterogeneous composition of conventional spatial, point-wise, and depth-wise convolution layers), selective long-range connectivity, and lightweight design patterns (unstrided and strided projection-replication-projection-expansion). [37] proposes Contrastive COVIDNet which is built upon the COVIDNet [12] architecture by introducing domain specific batch normalization layers. The authors propose a joint learning algorithm where the model is trained to minimize a cross entropy classification loss and a contrastive loss that is meant to minimize the difference between same class but cross site image embeddings while maximizing the difference between different class embeddings. However, to apply this method prior information about the sources of the datasets is required making it difficult to apply it to heterogenous cases where the sources are unknown. In [38] a custom CNN model is built with two separate lines of forward pass and deep feature aggregation to classify COVID and non-COVID. The network is trained to work both on CT and X-ray data. It employs a deep feature aggregation strategy by aggregating layer outputs from varying depths following a classifier network. ResGNet-C [39] exploits Graph Convolution Network (GCN) [40] to perform binary classification task using the Resnet-101 [33] extracted features.

B. COVID-19 Detection Using Chest X-ray

Although chest-CT has more sensitivity as compared to RT-PCR [9], [10], associated cost and resource constraints makes routine CT screening for COVID-19 detection a less accessible solution to the third world's teeming millions. On the other hand, digital X-ray is an easily accessible modality.

ChestX-Ray8 [41] (later expanded to constitute ChestX-ray14 dataset), and CheXpert [42] are two large-scale datasets of chest X-rays (CXR) to facilitate the training of deep neural networks for automated interpretation of a wide variety of thoracic diseases. ChexNet [42] is DenseNet-121 [43] based deep neural network for Pneumonia detection using chest X-ray images and it achieved excellent results exceeding average radiologist performance. ChestNet [44] is another deep neural network designed to diagnose thoracic diseases using chest radiography images. The authors in [45] proposes to learn channel wise, element wise, and scale wise attention (triple attention) simultaneously to classify 14 thoracic diseases using chest radiography. Thorax-Net [46] is an attention regularized deep neural network for classification of thoracic diseases on chest radiography.

In [47] the authors propose a novel segmentation – classification pipeline defined in three stages for binary classification of chest X-ray images into COVID-19 and non COVID-19 routines. In the first stage, the significant lung region is cropped from the chest X-ray images using a bounding box segmentation. In the second stage, a GAN inspired class – inherent transformation network is used to generate two class inherent transformations $\boldsymbol{x}+$ and $\boldsymbol{x}-$ from each input image $\boldsymbol{x}$. These are then used to solve a four-class classification problem using a CNN with a Resnet-50 backbone and an aggregation strategy is developed to obtain the final class. However, as the number of classes increase, the number of generators to be trained in the second stage of this method will increase accordingly, making it difficult to scale for multi class classification. COVID-Net [12] leveraged a human-machine collaborative design strategy to produce a network architecture tailored for COVID-19 detection from chest X-ray images. CoroNet [13] uses Xception [34] backbone for extracting CXR image features which are classified using a multi-layer perceptron (MLP) classification head. CovidAID [26] finetunes a pretrained CheXNet [5]. [48] proposes a novel architecture with multiscale attention-based generation augmentation and guidance for training a CNN model for COVID-19 diagnosis. The multi-scale attention features are computed from the intermediate feature maps of a Resnet-50 [33] based feature extractor and are combined with the final feature map to obtain the predictions. The attention maps are also used for augmenting the input images, the predicted labels of which are then utilized to further regularize the training loss via a soft distance regularization technique. [49] proposes another attention based CNN model incorporating a teacher-student transfer learning framework for COVID-19 diagnosis from Chest X-ray and CT images.

Although numerous works in the computer vision literature [50]–[54] have highlighted the benefits of using self-attention along with convolutional neural networks for numerous vision tasks, it has been shown in a recent work [17] that the convolutional layers can simply be scraped and by using only stacked self attention layers, one can achieve SOTA results in image classification tasks. This is the motivation behind xViTCOS as will be described in section III. CHP-Net [22] consists of three networks: a bounding box regression network to extract bi-pulmonary region coordinates, a discriminator deep learning model to predict a differentiating probability distribution, and a localization deep network that represents all potential pulmonary locations. In [11] the authors propose using shape dependent Fibonacci p patterns to extract features from chest X-ray images and then apply conventional machine learning algorithms including SVM, KNN, Random Forest, AdaBoost, Gradient Tree Boosting, and Decision Trees are used for performing binary and ternary classification of X-ray images. [19] first extracts orthogonal moment features using Fractional Multichannel Exponent Moments (FrMEMs). Next, the most significant features are selected using a differential evolution based modified Manta-Ray Foraging Optimization (MRFO). Finally a KNN classifier is trained to distinguish COVID-19 positive cases from negative cases.

III. Proposed Method

As described in Sec. II, the existing state-of-the-art AI models for automated COVID-19 detection uses either pre-trained convolutional neural network (CNN) and fine tune on CXR dataset or design a novel network and train from scratch. Unlike the existing methods, we propose a vision transformer (ViT) [17] based model for automated COVID-19 screening and call it xViTCOS, illustrated in Figure 1. Since we use xViTCOS on two chest radiography modalities CT scan images and chest X-ray images, we refer to them as xViTCOS-CT and xViTCOS-CXR respectively in our further discussion when the two models are to be distinguished.

A. Vision Transformers

A Vision Transformer [17] is a deep neural model that adapts the attention-based transformer architecture [55] prevalent in the domain of natural language processing (NLP) to make it suitable for pattern recognition in visual image data. While the original transformer architecture comprises of an encoder and a decoder, vision transformer is an encoder-only architecture. The standard transformer was originally designed to handle sequence data and expects to receive 1D sequence of token embeddings. In case of non-sequential image analysis tasks, like image classification, the input image, $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$ is broken down into N image patches, $\boldsymbol{x}_p^{(i)} \in \mathbb{R}^{P\times P\times C}$, where $i \in \{1, \cdots N\}$, and each patch is of shape $P \times P$ in 2-D, C denotes the number of channels (e.g. $C = 3$ for RGB images) and $N = \frac{H\times W}{P\times P}$. These patches derived from the image is then effectively used as a sequence of input images for the Transformer. These input patches are first flattened and then mapped to a D dimensional latent vector through a trainable linear projection layer, leading to the generation of patch embeddings. Throughout its layers, the transformer maintains a constant latent vector size of D. Similar to the [class] token in BERT [56], a learnable embedding is embedded to the sequence of the patch embeddings ($\boldsymbol{Z}_0^0 = \boldsymbol{x}_{class}$). The final transformer layer state corresponding to this class token, $\boldsymbol{z}_L^0$, represents in a compact form the classification information that the model is able to extract from the image($\boldsymbol{y}$). The classification head is attached to $\boldsymbol{z}_L^0$ during both pre-training and fine-tuning. In order to retain crucial positional information, standard learnable 1D position embeddings are added to the patch embeddings. The final resulting sequence is provided as input to the encoder. During pre-training an MLP is used to represent the classification head and it is replaced by a single linear layer during the fine-tuning stage. As illustrated in the Figure 1, the transformer encoder of a vision transformer consists of alternating layers of multiheaded self-attention (MSA) and MLP blocks. Layernorm (LN) is applied before every block, and residual or skip connections after every block. The workings of the vision transformer can be mathematically described in Equations below:

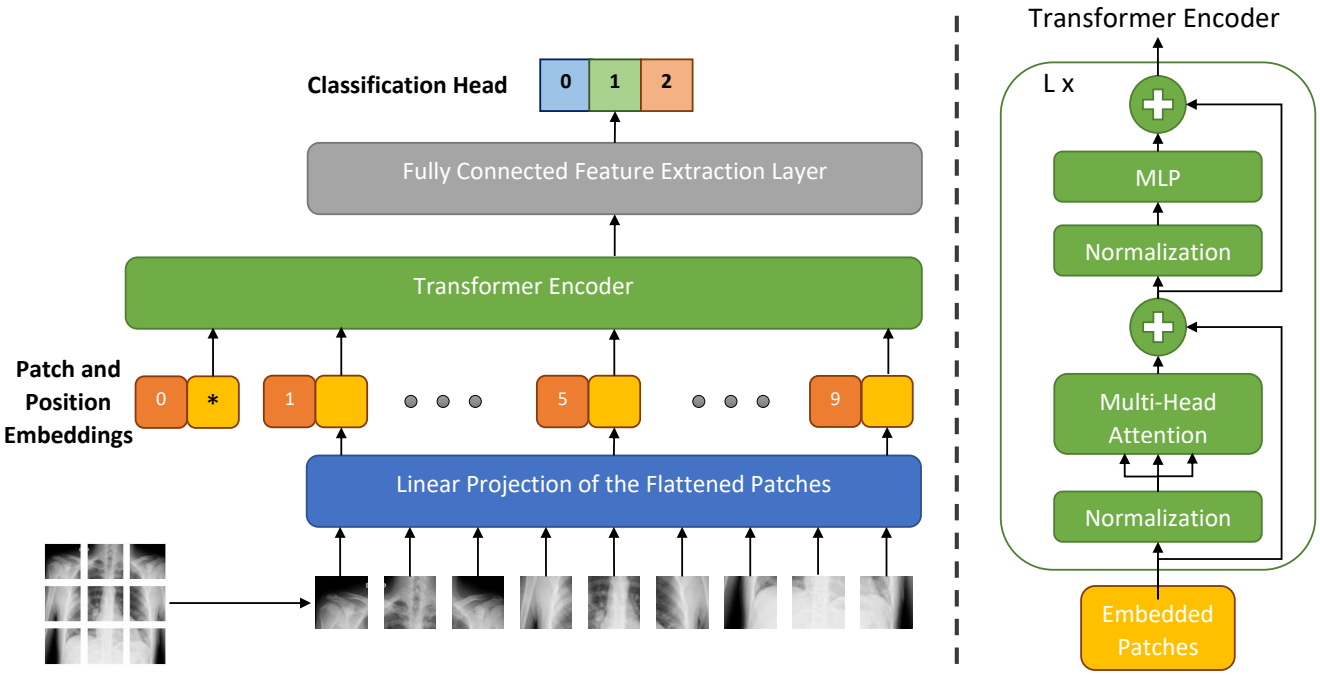


Fig. 1: xViTCOS: Illustration of our proposed network for COVID-19 detection using chest radiography (CT scan / CXR image). The input image is split into equal-sized patches and embedded using linear projection. Position embedding are added and the resulting sequence is fed to a Transformer encoder [55].

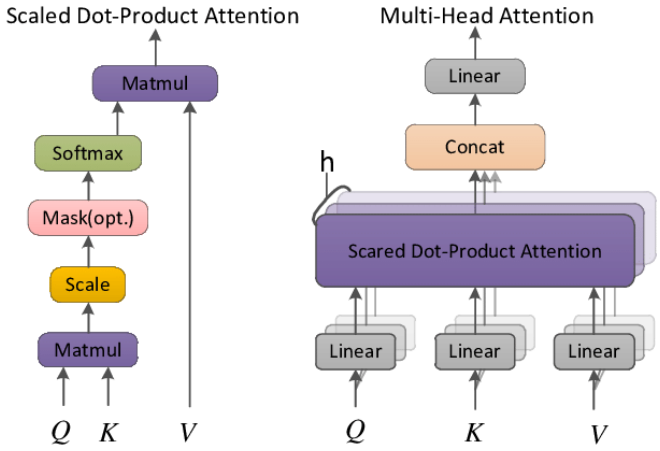


Fig. 2: xViTCOS: Block diagrams of Multi head Attention and Scaled Dot Product Attention Layers [57].

$$\boldsymbol{z}_0 = \left[\boldsymbol{x}_{class}; \boldsymbol{x}_p^1\boldsymbol{E}; \boldsymbol{x}_p^2\boldsymbol{E}; \cdots \boldsymbol{x}_p^N\boldsymbol{E}\right] + \boldsymbol{E}_{pos} \quad (1)$$

$$\boldsymbol{z}_l^{'} = \text{MSA}\left(\text{LN}\left(\boldsymbol{z}_{l-1}\right)\right) + \boldsymbol{z}_{l-1}, \forall l = 1 \cdots L \quad (2)$$

$$\boldsymbol{z}_l = \text{MLP}\left(\text{LN}\left(\boldsymbol{z}_l^{'}\right)\right) + \boldsymbol{z}_l^{'}, \quad \forall l = 1 \cdots L \quad (3)$$

$$\boldsymbol{y} = \text{LN}\left(\boldsymbol{z}_L^0\right) \quad (4)$$

where $\boldsymbol{E} \in \mathbb{R}^{(P^2C)\times D}$ and $\boldsymbol{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$

B. The Transformer Encoder

In general, a multi head attention block in a Transformer encoder is composed of horizontally stacked scaled dot product attention blocks. The scaled dot product attention block as shown in figure 2 takes as inputs three vectors: a query vector and a key vector each of dimensions d_p and a value vector of dimension d_v. The query vector is that particular image embedding with reference to which we want to calculate the attention values it receives from the other image embeddings in the sequence. Any of the other image embeddings about which we want to calculate the compatibility of our query vector can be the key vector. Suppose we want to quantify the compatibility or the influence that image embedding p_2 has on embedding p_1. In that case p_1 is our query and p_2 is the key. The dot product of the query vector with respect to all the key vectors are computed, scaled and then a softmax function is applied to get the weights on the value vectors. In practice the attention function is carried out on a set of query vectors stacked as a matrix Q. The keys and values are also stacked together to form matrices K and V respectively.

$$\text{Attention}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \text{Softmax}\left(\frac{\boldsymbol{Q} \times \boldsymbol{K}^{\text{T}}}{\sqrt{d_p}}\right)\boldsymbol{V} \quad (5)$$

Instead of using d_n dimensional vectors to calculate a single attention, it has been found beneficial to obtain h sets of d_p, d_p and d_v dimensional query, key and value vectors respectively through h different, trainable linear transformations. Attention is then performed in parallel on these h different sets of query, key and value vectors, the d_v dimensional outputs of which are concatenated and again projected linearly to the model dimension d_n to obtain the final results. Through multi head attention, it becomes possible to extract attention representation from a multitude of transformation spaces resulting in rich

representations.

$$\text{MultiHead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Concat}\left(\text{att}_1,\cdots\text{att}_h\right)\times\boldsymbol{W_L} \quad (6)$$

$$\text{att}_i=\text{Attention}\left(\boldsymbol{Q}\times\boldsymbol{W_{Qi}},\boldsymbol{K}\times\boldsymbol{W_{Ki}},\boldsymbol{V}\times\boldsymbol{W_{Vi}}\right) \quad (7)$$

where $\boldsymbol{W_{Qi}}\in\mathbb{R}^{d_n\times d_p}$, $\boldsymbol{W_{Ki}}\in\mathbb{R}^{d_n\times d_p}$, $\boldsymbol{W_{Vi}}\in\mathbb{R}^{d_n\times d_v}$ and $\boldsymbol{W_L}\in\mathbb{R}^{hd_v\times d_n}$ are the linear transformation matrices. Using attention mechanism, the transformer encoder architecture is able to generate representations with the capability of capturing a specific piece of information from a potentially infinitely-large context, the context being provided by the sequence of patch embeddings. In this scenario, context intuitively refers to the amount of information that one patch embedding or simply a patch of image has about another and it quantifies how closely related they are.

C. Inductive Bias in ViT

Unlike CNN based models that exploit the inherent bias associated with CNN such as translation invariance and a locally restricted receptive field, vision transformer (ViT) [17] has much less image specific inductive bias. This is because ViT treats an image as a sequence, hence loses any structural and neighborhood information a CNN can easily recognise. Although MLP layers are local and translationally equivariant, the self-attention layers are global. In fact in a ViT the spatial and two dimensional neighborhood relationships in an image needs to be learnt from scratch as the images are broken down into patches and fed as a sequence. The only mechanism that adds inductive bias and provides structural information about the image to the encoder are the position embeddings, that are concatenated with the patch embeddings. Without those, the Vision Encoder might find it difficult to make sense of the image patch sequence. Consequently, ViT does not generalize well when trained using insufficient amount of data. This might be a bit discouraging but the entire status quo changes as the size of the dataset increases. The large size of the training dataset overshadows the dependence of the model on inductive bias for generalization. As can be expected, using a ViT model pretrained on a large training dataset under a transfer learning framework on a smaller target dataset leads to improved performance. To combat this, we propose a multi-stage transfer learning strategy.

D. Multi-stage Transfer Learning

A domain and a task are the two main components of a typical learning problem. For the specific case of a supervised classification problem, the domain, $\mathcal{D}$ might be defined as the tuple of the feature space, $\mathcal{X}$, and the marginal feature distribution, $P(X)$, i.e. $\mathcal{D}=\langle\mathcal{X},P(X)\rangle$. The task, $\mathcal{T}$ is a tuple of label space, $\mathcal{Y}$, and the posterior of the labels conditioned on features, $P(Y|X)$, i.e. $\mathcal{T}=\langle\mathcal{Y},P(Y|X)\rangle$. Any change in either of the two components of a machine learning problem would cause severe degradation in the performance of the trained model and necessitates rebuilding the model from scratch. Transfer Learning is a way to combat this issue.

Given a source domain, $\mathcal{D}_s$ and a corresponding task, $\mathcal{T}_s$, and a target domain, $\mathcal{D}_t$ and a corresponding task, $\mathcal{T}_t$, the objective of transfer learning is to improve the performance of a machine learning model in $\mathcal{D}_t$ using the knowledge acquired in $\mathcal{D}_s$ and $\mathcal{T}_s$ [58]. Transfer learning has played a significant role in the facilitating the use of deep learning in numerous applications [59]–[64]. Along with deep convolutional neural networks (DCNNs), it has demonstrated tremendous success in medical image classification [14], [65] tasks where datasets are often sparse. In this work, we empirically demonstrate how knowledge transfer is equally effective for vision transformer based framework in medical image classification.

In our case, the target domain consists of chest radiography image data that the proposed model is ultimately supposed to explain, i.e., for xViTCOS-CXR, the target data is the COVID-19 CXR dataset and for the xViTCOS-CT model, the target data consists of the COVIDx-CT-2A dataset [66]. The target task to be learned is to classify the radiography images into three classes – COVID-19 Pneumonia, non-COVID-19 Pneumonia, and normal.

The first source domain $\mathcal{D}_{S_1}$ that our proposed ViT model is trained on consists of a large-scale dataset, ImageNet [67]. It is a widely used dataset for pre-training deep learning algorithms for a plethora of vision tasks. Since effective ViT training demands access to a sufficiently large number of data points, we choose a model which is pretrained on ImageNet-21k [67] $(\mathcal{T}_{S_1})$ in a self-supervised manner and later finetuned on ImageNet-2012 [68] $(\mathcal{T}_{S_2})$. This pre-training aims to ensure that the model learns to extract crucial but generic image representations to classify natural images.

In our case, the underlying distribution of clinical radiographic images is vastly different from an unconnected set of natural images like those in ImageNet, and distributional divergence is very high between the two domains. Hence in cases where the target dataset is of insufficient capacity, the pre-trained ViT model might find it highly difficult to bridge the domain shift between the learned source domain and the unseen target domain. However, with a sufficient number of training examples available from the target domain, the ViT model can overcome the gap between these two domains.

Keeping this in mind, an intermediate stage of knowledge transfer is used in this paper to train our proposed model depending on the size of the target domain training data. The primary goal of this stage of transfer learning is to help the ViT model, pre-trained on a generic image domains $\mathcal{D}_{S_1},\mathcal{D}_{S_2}$, to learn chest radiography specific representations to overcome the existing domain shift. In order to achieve this, we further finetune the pre-trained ViT model on a large collection of chest radiographic data $(\mathcal{D}_{S_3})$ after replacing its existing classification head with one suitable for the corresponding classification task $(\mathcal{T}_{S_3})$.

In our case, the COVIDx-CT-2A dataset [66] begin a moderate-sized dataset (refer to Table I), xViTCOS-CT model was able to overcome the domain shift and achieved state-of-the-art performance without the need for the intermediate finetuning stage. However, due to a limited number of COVID-19 CXR images (refer to Table II), an intermediate stage of knowledge transfer was employed to improve the performance of xViTCOS-CXR model. A publicly available large-scale CXR dataset, CheXpert [42] was used, and xViTCOS-CXR

TABLE I: Summarized description of COVIDx CT-2A dataset [66].

Split	Normal	Pneumonia	COVID-19	Total
Train	35996	25496	82286	143778
Validation	11842	7400	6244	25486
Test	12245	7395	6018	25658

TABLE II: Summarized description of CXR dataset.

Split	Normal	Pneumonia	COVID-19	Total
Train	1079	3106	1726	5911
Validation	270	777	432	1479
Test	234	390	200	824

was finetuned to classify five medical conditions (Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion) and the case of no finding on that dataset. Following this, the existing classification head of the ViT network was replaced by a new head suited for the particular target task, i.e., COVID-19 detection, and the model was further finetuned on the target domain.

Section IV-E presents an ablation study to understand the impact of multi-stage transfer.

E. Implementation Details

In all of our experiments, we have used ViT-B/16 network with the following configuration- Patch size: 16×16, Fraction of the units to drop for dense layers (Dropout rate): 0.1, Dimensions of the MLP output in the transformers: 3072, Number of transformer heads: 12, Number of transformer layers: 12, Hidden size: 768. The model parameters are initialized with the parameters of a model pretrained on ImageNet-21k [67] and fine-tuned on ImageNet-2012 [68].

While training xViTCOS-CXR, for the intermediate finetuning step using CheXpert [42], we use standard binary cross-entropy loss. This is because the classification task using CheXpert is a multi-label classification problem. Finally, while finetuning in the target COVID-19 CXR images, categorical cross-entropy loss is used to solve a multi-class classification problem. While training xViTCOS-CT, we utilize categorical cross-entropy. We use Keras [69] with Tensorflow [70] backend and vit-Keras[1] package for implementation of our code.

IV. Experiments and Results

A. Datasets

Some of the existing works validate their methods using private datasets [39], and several other works [13], [15], [16], [26], [71] combine data from different publicly available sources. While combining data from different public repository, researchers should be careful to avoid duplication as a contributor might upload the same image to many of the repositories. Another interesting way to mitigate the issue of data scarcity is through generative data augmentation where a neural generative framework [72]–[77] is trained to generate novel data samples. However in this work, we use the datasets described in the next section. We have rerun the codes of the baseline models using same dataset and same split to ensure a fair comparison.

1) CT Scan Dataset: To demonstrate the efficacy of xViTCOS-CT, we use COVIDx CT-2A dataset [66], derived from several public repositories [24], [78]–[84]. This dataset contains $194,922$ CT scans from 3,745 patients across the globe with clinically verified findings. Table I summarizes the important statistics of COVIDx CT-2A dataset.

2) Chest X-ray Dataset: To benchmark xViTCOS-CXR against other deep learning based methods for COVID-19 detection using CXR images, we construct a custom dataset consisting of three cases: Normal, Pneumonia, and COVID-19. Like in [13], [26], Normal and Pneumonia CXR images were obtained from the Kaggle repository 'Chest X-Ray Images (Pneumonia)' [85], which is derived from [86]. COVID-19 images were collected from the Kaggle repository 'COVIDx CXR-2' [87], which is a compilation of several public repositories [88]–[93].

COVIDx-CXR-2 [87] provides only Train-Test split of the data. To automatically select the best model based on validation-set performance, we split Training set in $80:20$ ratio as train and validation set. This would have caused huge class imbalance in the validation set as 'Chest X-Ray Images (Pneumonia)' [86] contains only 8 images per class in the validation set. Therefore, we combine the training and validation split and reconstruct the training and validation split in $80:20$ ratio. Table II summarizes split-wise image distribution. Note that, we have kept the test split intact in both the datasets to prevent patient-wise information leakage as multiple images for the same patient could be present in the dataset.

B. Data Preprocessing and Augmentation

1) CT Images: COVIDx CT-2A dataset [66] provides bounding box annotations for the body regions within the CT images. To standardize the field-of-view in the CT images, we crop the images to the body region using this additional information. Next each cropped image is resized to a fixed size of 224×224 pixels. To improve generalizability of the model, we augment the training data on the fly by applying random affine transformations such as rotation, scaling and translation, random horizontal flip and random shear.

2) CXR Images: In the compiled dataset, the chest X-ray images are of various sizes. To fix this issue, all the images were resized to a fixed size of 224×224 pixels. Again as in the case of CT images, to improve the generalizability of the model, we apply the same sets of autmentation techniques (refer to Section IV-B.1). In addition, we apply random zoom in and zoom out, and random channel shift.

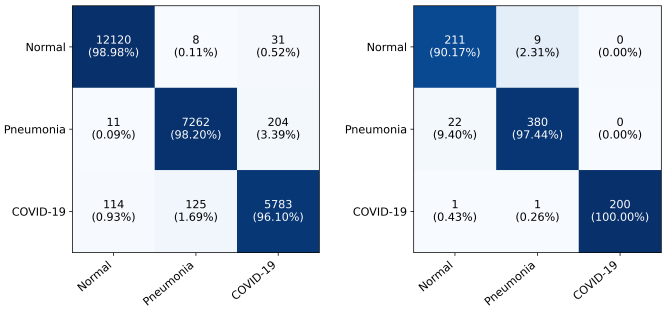
C. Quantitative Results

To quantify and benchmark the performance of xViTCOS, we compute and report Accuracy, Precision (Positive Prediction Value), Recall (Sensitivity), F1 score, Specificity, and

[1] https://github.com/faustomorales/vit-keras

TABLE III: Comparison of performance of xViTCOS-CT on CT scan dataset against state-of-the-art methods

Method	Class Label	Precision	Recall	F1-score	Specificity	NPV	Overall Accuracy
Resnet + Location Attention [25]	Normal	0.920	0.989	0.954	0.922	0.989	0.932
	Pneumonia	0.963	0.799	0.873	0.987	0.924	
	COVID-19	0.906	0.955	0.930	0.969	0.986	
COVIDNet-CT [16]	Normal	0.958	0.987	0.973	0.957	0.986	0.949
	Pneumonia	**0.981**	0.805	0.884	**0.989**	0.942	
	COVID-19	0.906	**0.988**	0.945	0.960	**0.995**	
Teacher-student Attention [49]	Normal	0.969	0.989	0.979	0.971	0.990	0.964
	Pneumonia	0.951	**0.982**	0.966	0.979	0.992	
	COVID-19	0.957	0.877	0.915	0.987	0.963	
ResGNet-C [39]	Normal	0.942	0.974	0.958	0.946	0.975	0.939
	Pneumonia	0.951	0.855	0.901	0.982	0.944	
	COVID-19	0.910	0.961	0.934	0.971	0.987	
xViTCOS-CT (Proposed)	Normal	**0.997**	**0.990**	**0.993**	**0.997**	**0.991**	**0.981**
	Pneumonia	0.971	**0.982**	**0.977**	0.988	**0.993**	
	COVID-19	**0.960**	0.961	**0.961**	**0.988**	0.988	



(a) xViTCOS-CT (b) xViTCOS-CXR

Fig. 3: Confusion Matrix: The horizontal axis consists of the ground true labels and the vertical axis corresponds to the predicted classes.

Negative Prediction Value (NPV). Let, for a particular class, TP, FP, TN, and FN denote the number of true positive (ground truth: positive, prediction: positive), false positive(ground truth: negative, prediction: positive), true negative (ground truth: negative, prediction: negative) and false negative (ground truth: positive, prediction: negative) predictions respectively. Next, we define the metrics considered in terms of TP, FP, TN, and FN.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{12}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{13}$$

1) xViTCOS-CT: Table III presents the overall accuracy of xViTCOS-CT on the test split of COVID-CT-2A dataset [66]. As can be observed, the proposed method achieves the best accuracy score of 98.1%, surpassing the current state of art methods. Next, we discuss the precision, recall, specificity, PPV, NPV, and F1-scores attained by the model on test COVID CT images and interpret their significance in determining the classification caliber of the model. From table III, it can be observed that xViTCOS-CT achieves a high value of recall or sensitivity at 96%, implying that a small proportion of pneumonia cases caused due to COVID-19 are incorrectly classified as having non-COVID-19 origin. This implies a significantly low number of false-negative cases, which is a highly sought-after characteristic in a medical data classifier as in such cases, a false negative situation may lead to denial or delay of treatment to a person genuinely infected by the disease. The proposed method also attains a high precision or positive predictive value of 96% for COVID-19 cases, implying a little chance of the model classifying a non-COVID case as having a COVID-19 origin. However, the usefulness of our proposed method lies in the fact that it achieves the highest F1 scores for all the classes, implying that in terms of both precision and recall, the proposed method is the most balanced amongst all the baseline models. Also, it is well able to differentiate between the normal and Pneumonia cases of patients as well. Similarly, we can see that the proposed model attains high specificity and NPV values of 98.8% for the COVID-19 case, implying that false positives are also very low. This is a useful characteristic in clinical scenarios since the model correctly rejects all the negative cases (patients who do not have COVID-19), facilitating efficient utilization of limited resources.

The prowess of the proposed model can be further understood From examining the confusion matrix (Figure 3a). The proposed model can distinguish the healthy patients from both covid and non-covid pneumonia cases very efficiently, with an accuracy of almost 99%. Particularly, out of a total of 12245 normal cases, 12120 have been classified correctly, while 11 (0.09%) and 114 (0.93%) cases have been wrongly classified as non-COVID pneumonia and COVID pneumonia classes, respectively. Another interesting point to note here is that while 114 normal cases have been misclassified as COVID-

TABLE IV: Comparison of performance of xViTCOS-CXR on chest X-ray dataset against state-of-the-art methods

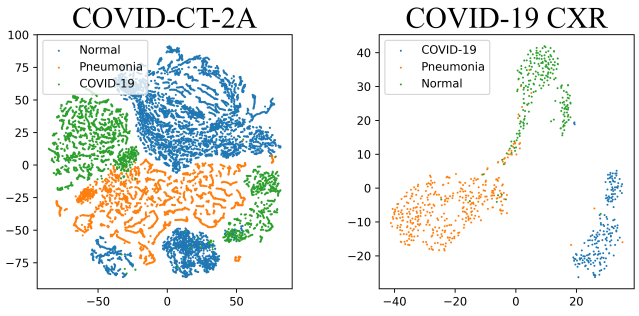
Method	Class Label	Precision	Recall	F1-score	Specificity	NPV	Overall Accuracy
InceptionV3 [94], [95]	Normal	0.932	0.876	0.903	0.974	0.952	0.946
	Pneumonia	0.933	0.964	0.948	0.937	0.967	
	COVID-19	0.990	0.995	0.992	0.997	0.998	
CoroNet [13]	Normal	0.812	**0.923**	0.864	0.915	**0.967**	0.917
	Pneumonia	**0.953**	0.941	0.947	**0.958**	0.947	
	COVID-19	**1.000**	0.865	0.927	**1.000**	0.958	
CovidNet [15]	Normal	0.826	0.918	0.870	0.923	0.966	0.919
	Pneumonia	0.950	0.882	0.915	**0.958**	0.900	
	COVID-19	0.985	0.995	0.99	0.995	0.998	
Teacher Student Attention [49]	Normal	0.913	0.902	0.908	0.966	0.961	0.932
	Pneumonia	0.918	0.974	0.945	0.922	0.976	
	COVID-19	0.989	0.885	0.934	0.997	0.964	
MAG-SD [48]	Normal	0.954	0.901	0.927	0.983	0.962	0.951
	Pneumonia	0.931	**0.974**	0.952	0.935	0.975	
	COVID-19	0.989	0.965	0.977	0.996	0.988	
xViTCOS-CXR (Proposed)	Normal	**0.959**	0.902	**0.929**	**0.985**	0.962	**0.960**
	Pneumonia	0.945	**0.974**	**0.959**	0.949	**0.976**	
	COVID-19	0.990	**1.000**	**0.995**	0.997	**1.000**	

19 and 204 COVID-19 cases have been assigned the non-COVID pneumonia label; the classifier has assigned only 31 COVID-19 originated pneumonia cases a normal class. This implies that the proposed method can distinguish the normal cases from the diseased cases, implying that genuine patients can be very quickly and efficiently segregated from healthy individuals. This makes xViTCOS-CT a genuinely valuable tool for COVID-19 diagnosis from chest CT images and can be used effectively along with the RT-PCR Testing.

2) xViTCOS-CXR: The observations regarding the performance of xViTCOS-CXR compared to its contemporaries are on the same lines as that of xViTCOS-CT, if not better. In terms of classification accuracy, xViTCOS-CXR achieves an accuracy of 96%, outperforming the baseline methods by a considerable margin as can be seen from Table IV. Further, it can be observed that xViTCOS-CXR achieves high recall (100%) and precision values (99%) on the COVID-19 cases, implying that the number of occasions on which the proposed model classified a COVID-19 model as a non-COVID-19 model or vice-versa is extremely low. Interestingly, as before, the F1-score of the proposed method for each of the classes is the highest among its contemporary methods on the test dataset. Examining the entries of Table IV, one can observe that the proposed method is the most balanced in terms of precision-recall when compared with the SOTA baselines. Similarly, we can see that the proposed model attains high specificity and NPV values of almost 100% for the COVID-19 case implying that the number of false positives is almost negligible. This is a valuable characteristic in clinical scenarios since it allows for rapid identification of patients who do not have COVID-19.

Analysing figure 3b, it can be seen that the class-wise accuracy of COVID-19 is 100%, i.e., all the ground truth COVID-19 cases have been classified as COVID-19, implying that the number of false negatives is zero. This confirms the efficacy of the proposed model in distinguishing between COVID and non-COVID cases, which is an essential clinical trait to have.

D. Qualitative Results



(a) (b)

Fig. 4: Visualization of penultimate feature space of xViTCOS using t-SNE plots.

1) Visualization of Feature Space: To visually analyze how clustered the feature space is, we perform a t-SNE visualization of the penultimate features for both the models using the test splits. As can be seen from Figure 4, the features in the penultimate layer clusters nicely for the three different classes.

2) Explainability: For qualitative evaluation of xViTCOS we present samples of CXR images and CT scans along with their ground truth labels and corresponding saliency maps along with the prediction in Figure 5. We have leveraged explanability driven approach outlined in [96], to better understand the diagnostic relevance of the visual factors leading to the predicted outcome of xViTCOS. Figure 5a, 5b and 5c presents CT scans of normal, Pneumonia and COVID-19 cases respectively; Figure 5d, 5e and 5f presents CXR images of normal, Pneumonia and COVID-19 cases respectively.

Report corresponding to Figure 5b as interpreted by a practicing radiologist: ground glass opacities, consolidation and secondary interlobar septal thickening, in bilateral lung, more extensive in right. xViTCOS-CT correctly highlighted these suspected regions. In Figure 5c xViTCOS-CT localized suspicious lesion regions exhibiting ground glass opacities, consolidation, reticulations in bilateral postero basal lung with subpleural predominance. In Figure 5e Patchy air space opacities noted in right upper and midzone matches the regions

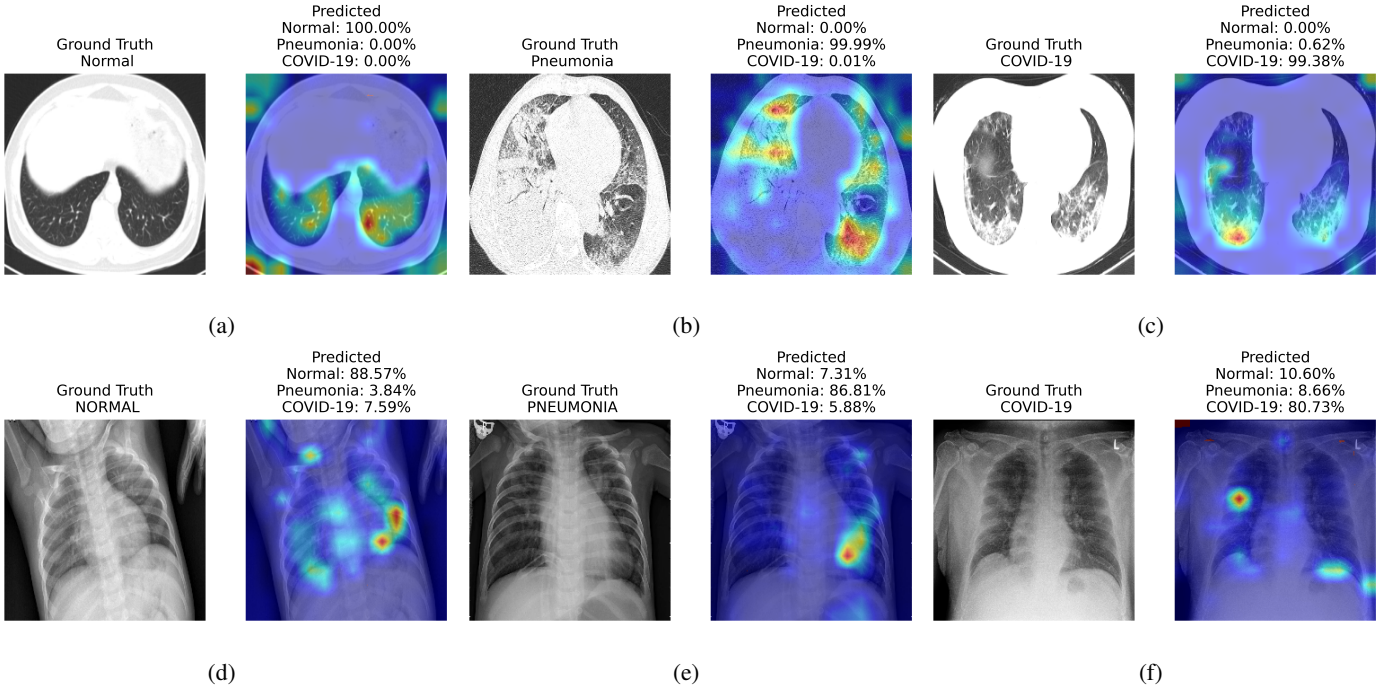


Fig. 5: Visualization of different cases (normal, Pneumonia, COVID-19) considered in this study and their associated critical factors in decision making by xViTCOS as identified using the explanability method laid out in [96] for vision transformers [17]. In each subfigure, the left figure presents the input to xViTCOS and its ground truth label; the right figure presents the predicted probabilities for each class and highlight the factors critical corresponding to the top predicted class. We have used jet colormap to colorize heatmap. Figure 5a, 5b and 5c corresponds to CT scan and Figure 5d, 5e and 5f corresponds to CXR images.

TABLE V: Ablation Studies for xViTCOS-CXR: Impact of multi-stage transfer

Method	Class Label	Precision	Recall	F1-score	Specificity	NPV	Overall Accuracy
Training ViT from scratch on COVID-19 CXR data	Normal	0.754	0.444	0.559	0.942	0.811	0.710
	Pneumonia	0.688	0.897	0.779	0.634	0.873	
	COVID-19	0.740	0.655	0.694	0.926	0.893	
Training ViT from scratch on CheXpert and finetuning on COVID-19 CXR data	Normal	0.777	0.641	0.702	0.927	0.866	0.821
	Pneumonia	0.819	0.882	0.849	0.824	0.886	
	COVID-19	0.867	0.915	0.891	0.955	0.972	
No intermediate finetuning on CheXpert	Normal	0.894	0.906	0.900	0.957	**0.962**	0.943
	Pneumonia	0.944	0.946	0.945	**0.949**	0.952	
	COVID-19	**1.000**	0.980	0.990	**1.000**	0.994	
xViTCOS-CXR (multi-stage transfer)	Normal	**0.959**	0.902	**0.929**	**0.985**	**0.962**	**0.960**
	Pneumonia	**0.945**	**0.974**	**0.959**	**0.949**	**0.976**	
	COVID-19	0.990	**1.000**	**0.995**	0.997	**1.000**	

TABLE VI: Ablation Studies for xViTCOS-CXR: Impact of freezing layers.

Method	Class Label	Precision	Recall	F1-score	Specificity	NPV	Overall Accuracy
Only the final classification head was trained	Normal	0.904	0.846	0.874	0.964	0.940	0.921
	Pneumonia	0.913	0.936	0.924	0.919	0.941	
	COVID-19	0.956	0.980	0.968	0.985	0.993	
First nine encoders of ViT were frozen	Normal	0.908	0.885	0.896	0.964	0.954	0.938
	Pneumonia	**0.949**	0.951	0.950	0.954	0.956	
	COVID-19	0.951	0.975	0.963	0.984	0.992	
First six encoders of ViT were frozen	Normal	**0.961**	0.846	0.900	**0.986**	0.942	0.945
	Pneumonia	0.920	**0.979**	0.949	0.924	**0.980**	
	COVID-19	0.980	0.995	0.987	0.993	0.998	
First three encoders of ViT were frozen	Normal	0.919	**0.927**	0.923	0.968	**0.971**	0.953
	Pneumonia	0.958	0.954	0.956	**0.963**	0.959	
	COVID-19	0.980	0.980	0.980	0.993	0.993	
xViTCOS-CXR (All the layers were finetuned)	Normal	0.959	0.902	**0.929**	0.985	0.962	**0.960**
	Pneumonia	0.945	0.974	**0.959**	0.949	0.976	
	COVID-19	**0.990**	**1.000**	**0.995**	**0.997**	**1.000**	

highlighted by xViTCOS-CXR. In Figure 5f, radiologist's interpretation is: thick walled cavity in right middle zone with surrounding consolidation. xViTCOS-CXR is able to correctly identify it. For the cases, where no abnormality is detected (Figure 5a and 5d), xViTCOS focuses on the entire lungs and chest respectively to make a final decision.

E. Ablation Studies

In this section we analyse the results of an ablation study performed on the xViTCOS-CXR model proposed in this paper. The ablation study can be loosely grouped in three sections depending on the specific aspect of training or model architecture targeted.

To understand the contributions made by each of the proposed training (finetuning) steps in training xViTCOS-CXR, we conduct several ablation experiments in this section. Table V presents the results. When ViT is trained on COVID-19 CXR data, its performance is worst as the dataset has very less training samples. CheXpert [42] consists of 224,316 chest radiographs of 65,240 patients. However, these many images are not sufficient for training ViT from scratch. Therefore, although the performance of the model improves, it is not comparable to the SOTA results. When the ViT model pretrained on imagenet is directly used for finetuning on CXR dataset, we see a huge boost in the performance. However, the best performance is achieved when the training procedure involves an intermediate finetuning step using CheXpert [42]. We can safely conclude, the intermediate finetuning helps the model learn useful features related to chest X-ray.

In order to analyse the effects of freezing a subset of layers on the classification performance of the proposed model, we conduct three experiments by subsequently freezing the first three, six and nine encoder layers of the model finetuned on the CheXpert data. These models are then trained on the COVID-19 CXR dataset. A fourth experiment is conducted where only the classification head of the model is allowed to train on the Covid CXR images and all the remaining layers are frozen, following the intermediate finetuning on CheXpert. The results are shown in Table VI. As expected freezing more layers during training on the CXR dataset leads to decreasing accuracy of classification, with the model whose first three encoder layers were frozen performing the best amongst the lot and the model where only the classification head is trained performs the worst. This implies that the more the number of trainable layers, the more is the capacity leading to a better performance of xViTCOS-CXR.

V. Conclusion

In this study, we introduce a novel vision transformer based method, xViTCOS for COVID-19 screening using chest radiography. We have empirically demonstrated the efficacy of the proposed method over CNN based SOTA methods as measured by various metrics such as precision, recall, F1 score. Additionally, we examine the predictive performance of xViTCOS utilizing explanability-driven heatmap plot to highlight the important factors for the predictive decision it makes. These interpretable visual cues are not only a step towards explainable AI, also might aid practicing radiologists in diagnosis. We also analzed the failure cases of our method. Thus, to enhance the effectiveness of diagnosis we suggest that xViTCOS be used to complement RT-PCR testing. In the next phase of this project, we aim to extend this work to automate the analysis of the severity of infection using vision transformers.

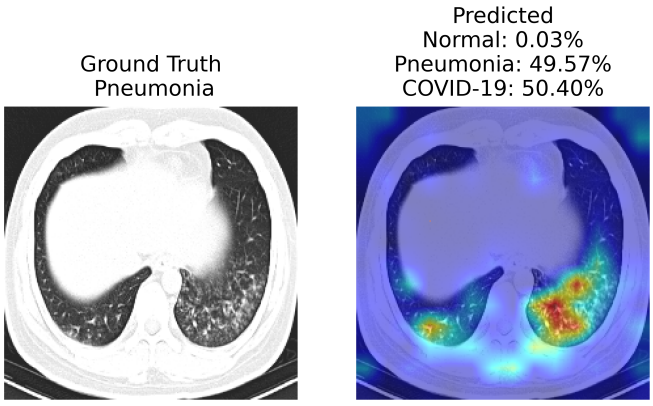


Fig. 6: A case of failure. xViTCOS-CT fails to predict the ground truth non-COVID-19 Pneumonia with confidence as it predicts non-COVID-19 Pneumonia with $\approx 50\%$ probability and COVID-19 with $\approx 50\%$ probability. This might happen as the findings on chest imaging in COVID-19 are not exclusive and overlap with many other type of infections [97]. In such cases, human expert intervention is necessary. For a detailed discussion refer to Section V.

Acknowledgement

We thank Dr. Arindam Mukherjee and Dr. Sabyasachi Mandal for their help in interpreting the qualitative results of chest radiography. We thank IIT Delhi HPC facility[2] for computational resources. Parag Singla is supported by the DARPA Explainable Artificial Intelligence (XAI) Program with number N66001-17-2-4032, the Visvesvaraya Young Faculty Fellowships by Govt. of India and IBM SUR awards. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of the funding agencies.

References

[1] "WHO director-general's opening remarks at the media briefing on covid-19 - 11 march 2020." https://www.who.int/director-general/speeches/detail.

[2] M. M. Hellou, A. Górska, F. Mazzaferri, E. Cremonini, E. Gentilotti, P. De Nardo, I. Poran, M. Leeflang, E. Tacconelli, and M. Paul, "Nucleic-acid-amplification tests from respiratory samples for the diagnosis of coronavirus infections: systematic review and meta-analysis," *Clinical Microbiology and Infection*, 2020.

[2] http://supercomputing.iitd.ac.in

[3] F. Colavita, F. Vairo, S. Meschi, M. B. Valli, E. Lalle, C. Castilletti, D. Fusco, G. Spiga, P. Bartoletti, S. Ursino, M. Sanguinetti, A. Di Caro, F. Vaia, G. Ippolito, and M. R. Capobianchi, "Covid-19 rapid antigen test as screening strategy at points of entry: Experience in lazio region, central italy, august–october 2020," *Biomolecules*, vol. 11, no. 3, 2021.

[4] K. Munne, V. Bhanothu, V. Bhor, V. Patel, S. D. Mahale, and S. Pande, "Detection of sars-cov-2 infection by rt-pcr test: factors influencing interpretation of results," *VirusDisease*, 2021.

[5] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. Langlotz, K. S. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *CoRR*, vol. abs/1711.05225, 2017.

[6] I. Pan, S. Agarwal, and D. Merck, "Generalizable inter-institutional classification of abnormal chest radiographs using efficient convolutional neural networks," *Journal of digital imaging*, vol. 32, no. 5, pp. 888–896, 2019.

[7] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu, "Chest ct for typical coronavirus disease 2019 (covid-19) pneumonia: relationship to negative rt-pcr testing," *Radiology*, vol. 296, no. 2, pp. E41–E45, 2020.

[8] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, *et al.*, "Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection," *Radiology*, p. 200463, 2020.

[9] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, and W. Ji, "Sensitivity of chest ct for covid-19: comparison to rt-pcr," *Radiology*, vol. 296, no. 2, pp. E115–E117, 2020.

[10] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of chest ct and rt-pcr testing for coronavirus disease 2019 (covid-19) in china: a report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. E32–E40, 2020.

[11] K. Panetta, F. Sanghavi, S. Agaian, and N. Madan, "Automated detection of covid-19 cases on radiographs using shape-dependent fibonacci-p patterns," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 1852–1863, 2021.

[12] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[13] A. I. Khan, J. L. Shah, and M. M. Bhat, "Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105581, 2020.

[14] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi, "Application of deep learning technique to manage covid-19 in routine clinical practice using ct images: Results of 10 convolutional neural networks," *Computers in Biology and Medicine*, vol. 121, p. 103795, 2020.

[15] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, p. 19549, 2020.

[16] H. Gunraj, L. Wang, and A. Wong, "Covidnet-ct: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images," *Frontiers in medicine*, vol. 7, 2020.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. of ICLR*, 2021.

[18] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P. M. Robson, M. Chung, *et al.*, "Artificial intelligence–enabled rapid diagnosis of patients with covid-19," *Nature medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.

[19] M. A. Elaziz, K. M. Hosny, A. Salah, M. M. Darwish, S. Lu, and A. T. Sahlol, "New machine learning method for image-based diagnosis of covid-19," *Plos one*, vol. 15, no. 6, p. e0235187, 2020.

[20] E. Tartaglione, C. A. Barbano, C. Berzovini, M. Calandri, and M. Grangetto, "Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data," *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, p. 6933, 2020.

[21] N. Tsiknakis, E. Trivizakis, E. E. Vassalou, G. Z. Papadakis, D. A. Spandidos, A. Tsatsakis, J. Sánchez-García, R. López-González, N. Papanikolaou, A. H. Karantanas, *et al.*, "Interpretable artificial intelligence framework for covid-19 screening on chest x-rays," *Experimental and Therapeutic Medicine*, vol. 20, no. 2, pp. 727–735, 2020.

[22] Z. Wang, Y. Xiao, Y. Li, J. Zhang, F. Lu, M. Hou, and X. Liu, "Automatically discriminating and localizing covid-19 from community-acquired pneumonia on chest x-rays," *Pattern Recognition*, vol. 110, p. 107613, 2021.

[23] D. Ezzat, A. E. Hassanien, and H. A. Ella, "An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization," *Applied Soft Computing*, p. 106742, 2020.

[24] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, L. Ye, M. Gao, Z. Zhou, L. Li, J. Wang, Z. Yang, H. Cai, J. Xu, L. Yang, W. Cai, W. Xu, S. Wu, W. Zhang, S. Jiang, L. Zheng, X. Zhang, L. Wang, L. Lu, J. Li, H. Yin, W. Wang, O. Li, C. Zhang, L. Liang, T. Wu, R. Deng, K. Wei, Y. Zhou, T. Chen, J. Y.-N. Lau, M. Fok, J. He, T. Lin, W. Li, and G. Wang, "Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433.e11, 2020.

[25] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Q. Ni, Y. Chen, J. Su, *et al.*, "A deep learning system to screen novel coronavirus disease 2019 pneumonia," *Engineering*, vol. 6, no. 10, pp. 1122–1129, 2020.

[26] A. Mangal, S. Kalia, H. Rajgopal, K. Rangarajan, V. Namboodiri, S. Banerjee, and C. Arora, "Covidaid: Covid-19 detection using chestx-ray," *arXiv 2004.09803*, 2020.

[27] L. Sun, Z. Mo, F. Yan, L. Xia, F. Shan, Z. Ding, B. Song, W. Gao, W. Shao, F. Shi, H. Yuan, H. Jiang, D. Wu, Y. Wei, Y. Gao, H. Sui, D. Zhang, and D. Shen, "Adaptive feature selection guided deep forest for covid-19 classification with chest ct," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2798–2805, 2020.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of NeurIPS* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), Curran Associates, Inc., 2012.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of ICLR*, 2015.

[30] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and $<$ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of CVPR*, 2015.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of CVPR*, 2018.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016.

[34] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of CVPR*, 2017.

[35] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, *et al.*, "Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: evaluation of the diagnostic accuracy," *Radiology*, vol. 296, no. 2, pp. E65–E71, 2020.

[36] H. X. Bai, R. Wang, Z. Xiong, B. Hsieh, K. Chang, K. Halsey, T. M. L. Tran, J. W. Choi, D.-C. Wang, L.-B. Shi, *et al.*, "Artificial intelligence augmentation of radiologist performance in distinguishing covid-19 from pneumonia of other origin at chest ct," *Radiology*, vol. 296, no. 3, pp. E156–E165, 2020.

[37] Z. Wang, Q. Liu, and Q. Dou, "Contrastive cross-site learning with redesigned net for covid-19 ct classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2806–2813, 2020.

[38] M. Owais, Y. W. Lee, T. Mahmood, A. Haider, H. Sultan, and K. R. Park, "Multilevel deep-aggregated boosted network to recognize covid-19 infection from large-scale heterogeneous radiographic data," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 1881–1891, 2021.

[39] X. Yu, S. Lu, L. Guo, S.-H. Wang, and Y.-D. Zhang, "ResGnet-C: A graph convolutional neural network for detection of covid-19," *Neurocomputing*, vol. 452, pp. 592–605, 2021.

[40] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. of ICLR*, 2017.

[41] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. of CVPR*, 2017.

[42] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. of AAAI*, 2019.

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. of CVPR*, 2017.

[44] H. Wang and Y. Xia, "Chestnet: A deep neural network for classification of thoracic diseases on chest radiography," *arXiv preprint arXiv:1807.03058*, 2018.

[45] H. Wang, S. Wang, Z. Qin, Y. Zhang, R. Li, and Y. Xia, "Triple attention learning for classification of 14 thoracic diseases using chest radiography," *Medical Image Analysis*, vol. 67, p. 101846, 2021.

[46] H. Wang, H. Jia, L. Lu, and Y. Xia, "Thorax-net: An attention regularized deep neural network for classification of thoracic diseases on chest radiography," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 475–485, 2020.

[47] 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. A. Valero-González, P. García-Villanova, E. Olmedo-Sánchez, and F. Herrera, "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.

[48] J. Li, Y. Wang, S. Wang, J. Wang, J. Liu, Q. Jin, and L. Sun, "Multiscale attention guided network for covid-19 diagnosis using chest x-ray images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1336–1346, 2021.

[49] W. Shi, L. Tong, Y. Zhu, and M. D. Wang, "Covid-19 automatic diagnosis with radiographic imaging: Explainable attentiontransfer deep neural networks," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2021.

[50] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," 2020.

[51] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," 2018.

[52] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," in *International Conference on Learning Representations*, 2020.

[53] N. Parmar, A. Vaswani, J. Uszkoreit, Łukasz Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," 2018.

[54] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020.

[55] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NeurIPS*, 2017.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, June 2019.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[58] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[59] M. Volpp, L. P. Fröhlich, K. Fischer, A. Doerr, S. Falkner, F. Hutter, and C. Daniel, "Meta-learning acquisition functions for transfer learning in bayesian optimization," in *International Conference on Learning Representations*, 2020.

[60] A. Shafahi, P. Saadatpanah, C. Zhu, A. Ghiasi, C. Studer, D. Jacobs, and T. Goldstein, "Adversarially robust transfer learning," in *International Conference on Learning Representations*, 2020.

[61] A. Bhattacharjee, A. Verma, S. Mishra, and T. K. Saha, "Estimating state of charge for xev batteries using 1d convolutional neural networks and transfer learning," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3123–3135, 2021.

[62] K. Wang, X. Gao, Y. Zhao, X. Li, D. Dou, and C.-Z. Xu, "Pay attention to features, transfer learn faster cnns," in *International Conference on Learning Representations*, 2020.

[63] J. Lin, L. Zhao, Q. Wang, R. Ward, and Z. J. Wang, "Dt-let: Deep transfer learning by exploring where to transfer," *Neurocomputing*, vol. 390, pp. 99–107, 2020.

[64] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," 2018.

[65] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[66] H. Gunraj, "COVIDx CT-2A: A large-scale chest CT dataset for covid-19 detection." `https://www.kaggle.com/hgunraj/covidxct`, 2021.

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. of CVPR*, pp. 248–255, 2009.

[68] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[69] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[70] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[71] J. P. Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," *arXiv preprint arXiv:2003.11597*, 2020.

[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. of NeuRIPS*, 2014.

[73] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. of ICML*, 2017.

[74] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Scholkopf, "Wasserstein auto-encoders," in *Proc. of ICLR*, 2018.

[75] A. K. Mondal, S. P. Chowdhury, A. Jayendran, P. Singla, H. Asnani, and A. Prathosh, "MaskAAE: Latent space optimization for adversarial auto-encoders," in *Proc. of UAI*, 2020.

[76] A. K. Mondal, H. Asnani, P. Singla, and A. Prathosh, "FlexAE: Flexibly learning latent priors for wasserstein auto-encoders," in *Proc. of UAI*, 2021.

[77] E. Acar, E. Şahin, and İ. Yilmaz, "Improving effectiveness of different deep learning-based models for detecting covid-19 from computed tomography (ct) images," *medRxiv*, 2020.

[78] A. P, X. S, H. SA, T. EB, S. TH, A. A, K. M, V. N, B. M, A. V, P. F, C. G, T. BT, and W. BJ, "Ct images in covid-19 [data set]," 2020.

[79] M. Rahimzadeh, A. Attar, and S. M. Sakhaei, "A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset," *Biomedical Signal Processing and Control*, vol. 68, p. 102588, 2021.

[80] W. Ning, S. Lei, J. Yang, Y. Cao, P. Jiang, Q. Yang, J. Zhang, X. Wang, F. Chen, Z. Geng, *et al.*, "Open resource of clinical data from patients with pneumonia for the prediction of covid-19 outcomes via deep learning," *Nature biomedical engineering*, vol. 4, no. 12, pp. 1197–1207, 2020.

[81] J. Ma, Y. Wang, X. An, C. Ge, Z. Yu, J. Chen, Q. Zhu, G. Dong, J. He, Z. He, *et al.*, "Towards efficient COVID-19 ct annotation: A benchmark for lung and infection segmentation," *arXiv preprint arXiv:2004.12537*, 2020.

[82] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. Van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke, "Data from lidc-idri [data set]," 2015.

[83] Radiopaedia, "COVID-19." `https://radiopaedia.org/articles/covid-19-4`.

[84] S. Morozov, A. Andreychenko, N. Pavlov, A. Vladzymyrskyy, N. Ledikhova, V. Gombolevskiy, I. A. Blokhin, P. Gelezhe, A. Gonchar, and V. Y. Chernina, "Mosmeddata: Chest ct scans with covid-19 related findings dataset," *arXiv preprint arXiv:2005.06465*, 2020.

[85] P. Mooney, "Chest x-ray images (pneumonia)." `https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia`, 2018.

[86] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," 2018.

[87] A. Zhao, H. Aboutalebi, A. Wong, H. Gunraj, N. Terhljan, *et al.*, "COVIDx CXR-2: Chest x-ray images for the detection of covid-19." https://www.kaggle.com/andyczhao/covidx-cxr2, 2021.

[88] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv 2006.11988*, 2020.

[89] L. Wang, A. Wong, Z. Q. Lin, P. McInnis, A. Chung, H. Gunraj, J. Lee, M. Ross, B. VanBerlo, A. Ebadi, K.-A. Git, and A. Al-Haimi, "Figure 1 covid-19 chest x-ray dataset initiative." https://github.com/agchung/Figure1-COVID-chestxray-dataset, 2020.

[90] L. Wang, A. Wong, Z. Q. Lin, P. McInnis, A. Chung, H. Gunraj, J. Lee, M. Ross, B. VanBerlo, A. Ebadi, K.-A. Git, and A. Al-Haimi, "Actualmed covid-19 chest x-ray dataset initiative." https://github.com/agchung/Actualmed-COVID-chestxray-dataset, 2020.

[91] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam, "Can AI help in screening viral and COVID-19 Pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[92] R. S. of North America, "RSNA Pneumonia Detection Challenge: Can you build an algorithm that automatically detects potential pneumonia cases?." https://www.kaggle.com/c/rsna-pneumonia-detection-challenge, 2018.

[93] E. B. Tsai, S. Simpson, M. Lungren, M. Hershman, L. Roshkovan, E. Colak, B. J. Erickson, G. Shih, A. Stein, J. Kalpathy-Cramer, *et al.*, "Data from medical imaging data resource center (midrc) - rsna international covid radiology database (RICORD) release 1c - chest x-ray, covid+ (midrc-ricord-1c)," 2021.

[94] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. of CVPR*, pp. 2818–2826, 2016.

[95] N. Tsiknakis, E. Trivizakis, E. Vassalou, Evangelia, Z. Papadakis, Georgios, A. Spandidos, Demetrios, A. Tsatsakis, J. Sánchez-García, R. López-González, N. Papanikolaou, H. Karantanas, Apostolos, and K. Marias, "Interpretable artificial intelligence framework for covid-19 screening on chest x-rays," *Exp Ther Med*, vol. 20, pp. 727–735, Aug 2020.

[96] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. of CVPR*, 2021.

[97] A. C. of Radiology, "ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection," 2020.