Vision Transformers for Classification in Small-Sized Chest X-Ray Datasets

Kevin Jin
Department of Electrical and Computer Engineering, University of British Columbia
Vancouver, BC, Canada
khjin@ece.ubc.ca

Abstract

Chest X-ray imaging is a common medical imaging technique used to diagnose a variety of diseases. Many machine learning approaches have been developed to perform classification of chest X-ray images using convolution neural networks. Recently, vision transformers have been shown to have superior performance in image classification tasks compared to convolutional neural network techniques, when trained on large amounts of data. However, medical images are expensive to label, so large labelled datasets are typically not available. Lee et al. [1] proposed an approach using shifted patch tokenization and locality self-attention to train vision transformers on small-sized datasets. These techniques work by improving the local inductive biases of vision transformer models. Here, we evaluate the effectiveness of the shifted path tokenization and locality self-attention techniques for binary classification of chest X-ray images. We show that these techniques can be used to significantly improve the performance of vision transformers when applied to the domain of medical imaging, reducing the need for large datasets.

1 Introduction

Chest X-ray imaging is one of the most common medical imaging techniques in radiology, capable of diagnosing various diseases, including pneumonia, COVID-19, tuberculosis, lung cancer and more [2], [3]. Machine learning methods have gained popularity in the classification of chest X-ray images [4], with various groups demonstrating the effectiveness of deep learning techniques using convolutional neural networks (CNNs) [5].

Recently, transformers, which are based on a self-attention mechanism and were traditionally used for natural language processing (NLP) tasks, have been applied to image classification tasks, showing improved performance when compared to state-of-the-art CNN architectures when trained on large amounts of data [6]. Transformers are computationally efficient and scalable, being used extensively to process sequential data. In order to apply transformers to image data, patches of pixels can be extracted from images to reduce the computational cost of the transformer self-attention layers (as it would be too expensive computationally to apply self-attention to all the pixels in an image) [6]. However, transformers lack the inductive biases present in CNNs, so large amounts of data are typically needed to achieve optimal performance with transformer architectures [6]. This poses a challenge for medical imaging tasks, where labelled datasets tend to be scarce due to the high cost of having experts manually annotate and classify these images [3], [7]. For example, in the case of COVID-19, during the initial stages of the pandemic, there was limited data available to train machine learning models [8]. Therefore, groups have developed methods to improve

transformer performance on small-sized datasets [1], and we propose to evaluate the applicability of some of these methods for classification of chest X-ray images.

2 Related Work

Vision Transformers The original vision transformer (ViT) developed by Dosovitskiy et al. [6] was trained on large datasets, such as the JFT-300M dataset, which contains 18k classes and 303M images [6]. To the best of our knowledge, there is no dataset of medical images that comes anywhere near this size, so a similar training approach cannot be used for most medical image applications.

Small Datasets Lee et al. have proposed a method for applying ViTs to small-sized datasets using shifted patch tokenization (SPT) to embed more spatial information into tokens and locality self-attention (LSA) to attend locally using a softmax function with learnable parameters [1]. They show that these techniques improve the local inductive biases of ViT models [1]. By applying these two techniques, an average improvement of 2.96% in Tiny-ImageNet and an improvement of 4.08% in Swin Transformers was observed [1].

X-Ray Image Classification Chest X-ray image classification has traditionally been done with CNNs, achieving high accuracies, typically over 90% [9], [10]. More recently, ViTs have been applied for chest X-ray image classification as well [2], [3]. Usman et al. explore the transfer learning capabilities of transformers applied to chest X-ray image classification [3]. Okolo et al. present an improved version of the ViT [2], achieving improved performance over the "Base" and "Large" ViT variants in the original ViT model [6] when applied for classification of chest X-ray data. They iteratively add a representation of the original input layer to the output of each transformer encoder layer by using a CNN block in parallel with the ViT network to help the network "remember" the full input image after each transformer block output [2]. Their improved ViT model had a comparable performance to well-established CNN models on the datasets they examined [2].

Contribution Combining the previous approaches in literature, we apply the SPT and LSA approaches to ViTs for binary classification of chest X-ray images. We also evaluate the effect of training dataset size on the performance of ViT models. Hence, the main contributions of the proposed project are to: (1) apply ViTs with SPT and LSA to the medical imaging domain using chest x-ray images and (2) evaluate how ViTs with SPT and LSA scale with dataset size.

3 Method

Vision Transformer In the ViT architecture for image classification, 2D input image data is first split into patches to form a sequence of 1D vectors, which is then input into the transformer architecture, analogous to the sequence of tokens in NLP [6]. These patches are linearly embedded, 1D position encodings are added, and the result is input into the standard transformer encoder [1]. An extra learnable classification token is prepended to the sequence for classification purposes. The classification is implemented through a multi-layer perceptron (MLP), replacing the decoder portion of the original transformer, which is not needed for classification purposes [3]. Mathematically, given an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ where H, W, C represent the height, width, and channel dimensions of the input image respectively, we first divide the image into a sequence of non-overlapping patches and flatten the patches into a sequence of vectors

$$\mathcal{P}(\mathbf{x}) = [\mathbf{x}_p^1;\ \mathbf{x}_p^2; \ldots;\ \mathbf{x}_p^N]$$

where $\mathbf{x}_p^i \in \mathbb{R}^{P^2 C}$ is the i^{th} flattened vector, P is the patch size, and $N = \frac{HW}{P^2}$ is the number of patches [1]. Patch embeddings are learned through a linear projection $\boldsymbol{E} \in \mathbb{R}^{P^2 \times C \times d}$, where d is the hyperparameter for the dimension of the encoder [1]. This process generates tokens $\mathcal{T}(\mathbf{x}) = \mathcal{P}(\mathbf{x})\boldsymbol{E}$,

which are concatenated with a classification token $\mathbf{x}_{cls} \in \mathbb{R}^d$ and added to the positional embeddings $\boldsymbol{E}_{pos} \in \mathbb{R}^{(N+1)\times d}$ to be used as input to the transformer encoder [1]. The tokens are then passed to the encoder portion of the transformer consisting of multi-head self-attention, layer normalization, and feed-forward layers [3]. The transformer self-attention (SA) mechanism is applied with the learnable Query, Key, and Value matrices generated ($\boldsymbol{Q} = \mathbf{x}\boldsymbol{W}_Q, \boldsymbol{K} = \mathbf{x}\boldsymbol{W}_K, \boldsymbol{V} = \mathbf{x}\boldsymbol{W}_V$):

$$\mathrm{SA}(\mathbf{x}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}$$

where d_k is the dimension of the key. After passing through the encoder portion of the transformer, the output is fed to a MLP to perform the classification [3]. Our ViT model architecture follows this original ViT architecture, while also incorporating the method by Lee et al. for applying SPT and LSA to ViT models [1].

Shifted Patch Tokenization Applying SPT involves modifying the tokenization process by shifting each input image by half the patch size ($P/2$) in the 4 diagonal directions (i.e. left-up, left-down, right-up, right-down) [1]. The shifted features are cropped to the input image size and then concatenated with the original input [1]. The concatenated features are then split into non-overlapping patches and flattened, and layer normalization (LN) is applied [1]. The patch embeddings are constructed through tokenization with

$$\mathcal{T}(\mathbf{x}) = \mathrm{LN}\big(\mathcal{P}([\mathbf{x}\ \mathbf{s}^1\ \mathbf{s}^2\ \ldots\ \mathbf{s}^{N_s}])\big)\boldsymbol{E}$$

where $\boldsymbol{E} \in \mathbb{R}^{P^2\times C\times(N_s+1)\times d}$ is the learned linear projection for the tokens, d is the hyperparameter for the dimension of the encoder, $\boldsymbol{s}^i \in \mathbb{R}^{H\times W\times C}$ is the i^{th} shifted image, and $N_s = 4$ is the number of images shifted in the 4 diagonal directions [1]. Other shifting strategies are possible as well and are described in more detail in the original SPT paper [1]. Without SPT, the receptive field of a token would be the same as the patch size (P) of the ViT (which is 16 for the original ViT) [1], [6]. Meanwhile, CNN architectures like ResNet50 have much greater receptive fields (the receptive field of ResNet50 is 483) [11]. Applying SPT helps to increase the receptive field of the ViT to capture more spatial information and rival the inductive capabilities of CNNs [1]. Figure 1 shows a visualization of the SPT approach when applied to a chest X-ray image.

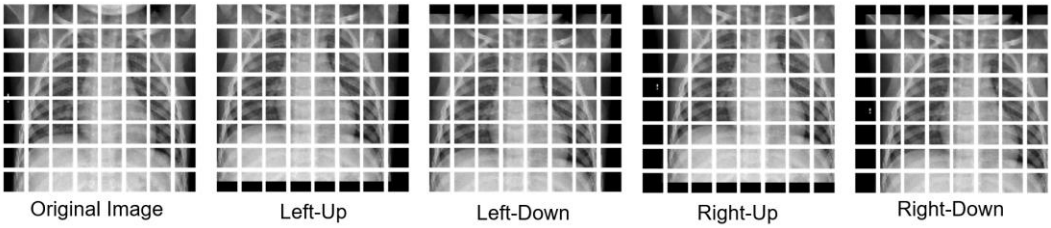


Figure 1: SPT applied to a chest X-ray image. The shifted images are concatenated with the original input and split into patches, with a patch size $P = 16$.

Locality Self-Attention Locality self-attention is applied by modifying the SA mechanism of the transformer encoder layer, in the argument of the softmax function. It utilizes two key ideas: diagonal masking and a learnable temperature scaling [1]. Diagonal masking involves emphasizing the inter-token relations by removing (masking) the self-token relations along the diagonals of the $\boldsymbol{R} = \boldsymbol{Q}\boldsymbol{K}^T$ matrix.

$$\tilde{R}_{i,j}(\mathbf{x}) = \begin{cases} R_{i,j}(\mathbf{x}) & i \neq j \\ -\infty & i = j \end{cases}$$

The learnable temperature scaling replaces the $\sqrt{d_k}$ term (in the denominator of the softmax argument) with a learnable temperature hyperparameter τ. This approach tends to result in a lower temperature (i.e. $\tau < \sqrt{d_k}$), which sharpens the score distribution and results in improved performance [1]. Put together, LSA involves the following modifications to the transformer SA mechanism:

$$\mathrm{SA}_{\mathrm{LSA}}(\mathbf{x}) = \mathrm{softmax}\left(\frac{\widetilde{\boldsymbol{R}}}{\tau}\right)\boldsymbol{V}$$

Model Overview Putting it all together, we arrive at our model architecture in Figure 2, which combines the SPT and LSA approaches by Lee et al. with [1] the ViT model by Dosovitskiy et al. [6].

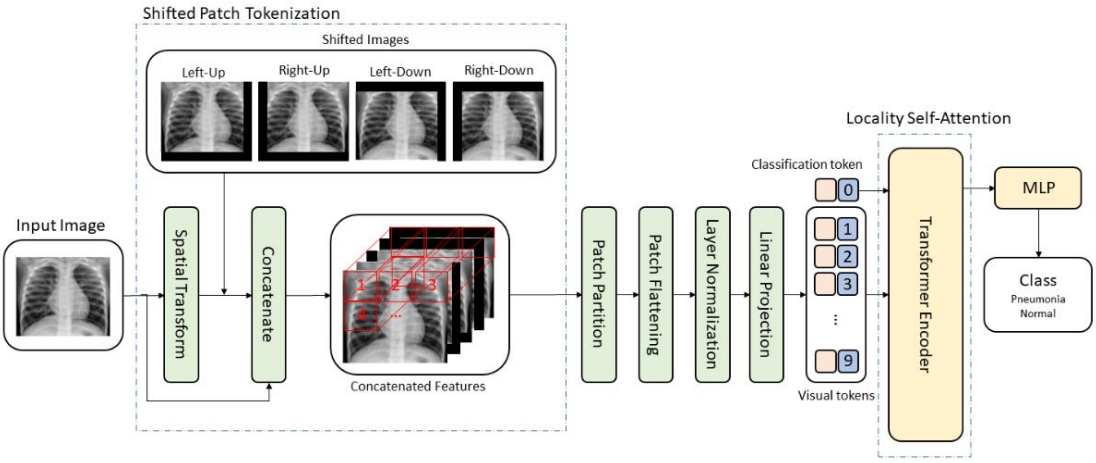


Figure 2: The general model architecture used in our experiments. SPT is applied prior to partitioning the input image into patches to be input into the ViT. For models without SPT, the SPT module is skipped, and the input image is directly partitioned into patches. LSA is applied within the transformer encoder layer's SA mechanism, if applicable.

4 Experiments

Dataset A chest X-ray image dataset of 5,856 frontal chest X-ray images from different patients was used for our experiments [12], [13]. The dataset consisted of 4,273 cases of pneumonia and 1,583 normal (healthy) lungs. Due to the class imbalance, class weights were computed to account for the skewed data distribution. A representative sample of the dataset is shown in Figure 3. An inexperienced viewer (like the author of this report) would have difficulty distinguishing the difference between the two images, making this task a somewhat challenging binary image classification task.

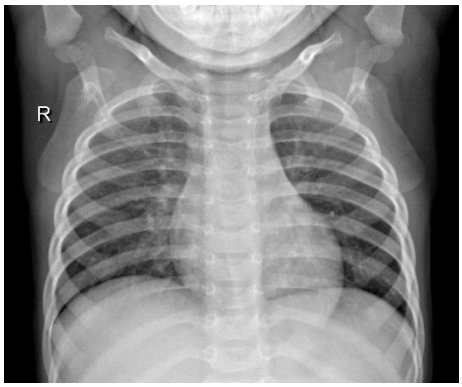
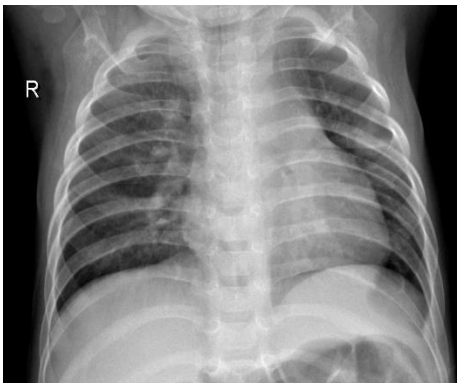
(a)  (b) 

Figure 3: Sample images from the dataset: (a) healthy lung (b) pneumonia.

Experiment Setup The images in the dataset were resized to 128×128 pixels and the default training, validation, and testing data split of the dataset [13] was used. The test dataset is used to report all results. The model hyperparameters were selected through trial and error and tweaked based on the validation dataset. The following four ViT models were developed:

(1) *Original*: an original ViT without any modifications

(2) *SPT only*: a ViT model that utilizes the SPT technique only

(3) *LSA only*: a ViT model that utilizes the LSA technique only

(4) *SPT and LSA*: a ViT model that combines both SPT and LSA techniques

For all four models, the transformer encoder architecture consisted of 8 encoder blocks and the multi-head attention layer consisted of 4 heads. A projection dimension $d = 64$ was used. For SPT, a patch size $P = 16$ was used, splitting the image into $N = 64$ patches. For classification, a MLP with 2048 and 1024 hidden units was used. A sigmoid activation function was applied at the output. An Adam optimizer with binary cross-entropy loss was used for training. The models were trained for 100 epochs, with a batch size of 512, learning rate 0.001. Dropout layers were used to improve model generalizability. Each experiment was performed 5 times for repeatability. All the development and model training for this work was done on the Google Colab platform.

Results To evaluate the effect of dataset size on the learning capabilities of ViTs with and without SPT and/or LSA, various fractions of the full training dataset were used to train the four models (10%, 30%, and 100%). The area under the receiver operating characteristic curve (AUC) for the test dataset was used to evaluate the model performance and the results are shown in Table 1. The AUC values for the training dataset were in the 0.95-0.99 range when trained on the full dataset, suggesting that the ViT models are powerful enough to fit to the training dataset. However, the model did not generalize as well to the validation and test datasets (i.e., they were overfit to the training data).

Model	DF = 10%	DF = 30%	DF = 100%
Original	0.47 ± 0.05	0.53 ± 0.04	0.58 ± 0.03
SPT only	0.49 ± 0.13	0.54 ± 0.05	0.68 ± 0.12
LSA only	0.49 ± 0.06	0.55 ± 0.09	0.68 ± 0.14
SPT and LSA	**0.55** ± 0.18	**0.55** ± 0.06	**0.73** ± 0.11

Table 1: Test AUC values of the 4 ViT models when trained with various dataset sizes (DF = dataset fraction). The full training dataset (DF = 100%) consisted of 5,216 images. Each experiment was repeated 5 times – the mean and standard deviation is reported in the table.

Generally, the model performance improved as the dataset size increased for most model variants, as expected (i.e., larger datasets improve transformer performance). In all cases, the model that utilized both SPT and LSA techniques showed the best performance for a given dataset, outperforming the original ViT without either of these techniques applied. It is interesting to note that the models with the SPT and LSA techniques applied generally resulted in larger variations in performance between runs, compared to the original ViT baseline.

Ablation Study When SPT and LSA are applied individually, small improvements in the model performance are observed when compared to the baseline original ViT. The baseline original ViT did not generalize well to the test dataset (the AUC values were in the 0.47-0.58 range, which is fairly close to random chance). These results suggest that applying both SPT and LSA can improve the inductive biases of ViT models, helping them learn relevant features and improving model generalization when applied to small chest X-ray datasets. However, their effect is greater when both techniques are applied together (Table 1).

Training Curves A comparison of representative training curves for the original ViT model and the ViT model with SPT and LSA applied are shown in Figure 4. It is interesting to note that it takes quite a few epochs before the ViT model starts learning patterns in the training data and the training loss starts to decrease further (while the validation loss increases), suggesting that the model starts to overfit to the training data. This point happens around epoch 55 for the original ViT model and epoch 35 for the ViT model with SPT and LSA. This suggests that SPT and LSA improves the model's ability to capture information about the dataset, requiring fewer epochs to achieve the same performance as a ViT model without these techniques applied.

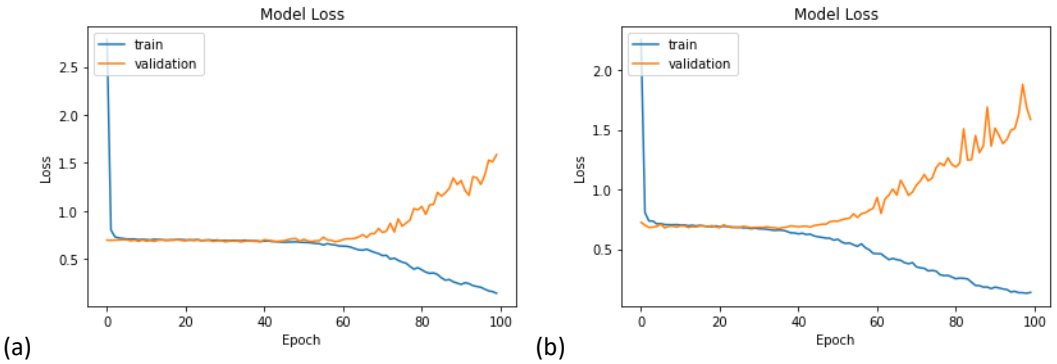


Figure 4: Training curve of (a) the original ViT and (b) the ViT with SPT and LSA.

Convolutional Neural Network Comparison As an interesting exercise, we also developed a basic CNN model (with 2 convolution layers and 3 fully-connected layers to perform the classification) to compare our ViT models to (Table 2). The CNN model outperformed the baseline ViT model but fell short of the ViT model with both SPT and LSA techniques applied when trained on the full training dataset. The CNN did not perform well on the small 10% and 30% datasets. Even the CNN was not able to achieve high test AUC values on this dataset, suggesting that perhaps the chest X-ray dataset is not an easy classification task. However, the CNN architecture and hyperparameters were not tuned as extensively, so the performance of the CNN model may still have some room for improvement.

DF = 10%	DF = 30%	DF = 100%
0.48 ± 0.10	0.47 ± 0.09	0.67 ± 0.11

Table 2: Test AUC values of the CNN model when trained with various dataset sizes (DF = dataset fraction). Each experiment was repeated 5 times – the mean and standard deviation is reported.

Discussion While there are notable differences between the ViT models compared, there is still much room for improvement in the model performance. For a binary classification task, an AUC of 0.50 would correspond to a random classifier. Hence, a few of the ViT models on the smallest dataset actually performed worse than random guessing. So, while SPT and LSA can help to improve the inductive capabilities of ViT models, ViTs still seem to require a relatively large dataset to achieve notable state-of-the-art performances. Furthermore, chest X-ray images may also be a challenge to classify: a single chest radiograph may not be sufficient to accurately diagnose whether a patient has pneumonia or not, and oftentimes, other diagnostic tests such as a clinical assessment or microbial test are used in conjunction to make an accurate diagnosis [14].

5 Conclusion

We have shown that the SPT and LSA techniques developed by Lee et al. [1] are applicable to small datasets in the domain of medical imaging, improving the local inductive biases of ViT models. Model performance is improved the most when both SPT and LSA are applied together. Without these techniques, ViT models struggle to learn patterns in small chest X-ray datasets, barely outperforming a random classifier. CNNs also struggle to learn relevant patterns with small chest X-ray datasets. Small datasets were used for this work primarily due to computational resource constraints. However, in the future, it may be helpful to explore larger datasets as well, to assess whether SPT and LSA can result in significant improvements for large datasets, where ViTs tend to perform best. A dataset where the classes are more easily distinguishable might be helpful as well. While AUC values were used as the primary evaluation metric in our work for simplicity of comparison across different models, other evaluation metrics should be explored as well for a more comprehensive comparison (e.g., accuracy, precision, recall, F1-score). Furthermore, instead of training ViT models from scratch, it may be beneficial to use a pre-trained ViT model and tweak it on the dataset instead. Lastly, many images in the medical domain are 3-dimensional, so while this work has been limited to 2-dimensional images for simplicity (and to reduce the computational cost), evaluating the generalizability of ViTs with SPT and LSA for 3D images would be an interesting area to explore in the future.

References

[1] S. H. Lee, S. Lee, and B. C. Song, “Vision Transformer for Small-Size Datasets.” arXiv, Dec. 26, 2021. Accessed: Oct. 13, 2022. [Online]. Available: http://arxiv.org/abs/2112.13492

[2] G. I. Okolo, S. Katsigiannis, and N. Ramzan, “IEViT: An enhanced vision transformer architecture for chest X-ray image classification,” *Comput. Methods Programs Biomed.*, vol. 226, p. 107141, Nov. 2022, doi: 10.1016/j.cmpb.2022.107141.

[3] M. Usman, T. Zia, and A. Tariq, “Analyzing Transfer Learning of Vision Transformers for Interpreting Chest Radiography,” *J. Digit. Imaging*, Jul. 2022, doi: 10.1007/s10278-022-00666-z.

[4] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, “Deep learning for healthcare applications based on physiological signals: A review,” *Comput. Methods Programs Biomed.*, vol. 161, pp. 1–13, Jul. 2018, doi: 10.1016/j.cmpb.2018.04.005.

[5] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights Imaging*, vol. 9, no. 4, Art. no. 4, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[6] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv, Jun. 03, 2021. Accessed: Sep. 30, 2022. [Online]. Available: http://arxiv.org/abs/2010.11929

[7] M. J. Willemink *et al.*, “Preparing Medical Imaging Data for Machine Learning,” *Radiology*, vol. 295, no. 1, pp. 4–15, Apr. 2020, doi: 10.1148/radiol.2020192224.

[8] F. Xu *et al.*, “An original deep learning model using limited data for COVID-19 discrimination: A multicenter study,” *Med. Phys.*, vol. 49, no. 6, pp. 3874–3885, Jun. 2022, doi: 10.1002/mp.15549.

[9] M. Sorić, D. Pongrac, and I. Inza, “Using Convolutional Neural Network for Chest X-ray Image classification,” in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, Sep. 2020, pp. 1771–1776. doi: 10.23919/MIPRO48935.2020.9245376.
[10] K. Almezhghwi, S. Serte, and F. Al-Turjman, “Convolutional neural networks for the classification of chest X-rays in the IoT era,” *Multimed. Tools Appl.*, vol. 80, no. 19, pp. 29051–29065, 2021, doi: 10.1007/s11042-021-10907-y.
[11] A. Araujo, W. Norris, and J. Sim, “Computing Receptive Fields of Convolutional Neural Networks,” *Distill*, vol. 4, no. 11, p. e21, Nov. 2019, doi: 10.23915/distill.00021.
[12] D. S. Kermany *et al.*, “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” *Cell*, vol. 172, no. 5, pp. 1122-1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.
[13] P. Mooney, “Chest X-Ray Images (Pneumonia),” *Kaggle*. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia (accessed Dec. 12, 2022).
[14] D. Wootton and C. Feldman, “The diagnosis of pneumonia requires a chest radiograph (x-ray)—yes, no or sometimes?,” *Pneumonia*, vol. 5, no. 1, Art. no. 1, Dec. 2014, doi: 10.15172/pneu.2014.5/464.