Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

Article

Improving the generalization of Deep Learning Classification Models in Medical Imaging using Transfer Learning and Generative Adversarial Networks

Sagar Kora Venu *

Department of Analytics, Harrisburg University of Science and Technology, Harrisburg, PA 17101, USA
* Correspondence: SKora@my.HarrisburgU.edu

Abstract: Data sets for medical images are generally imbalanced and limited in sample size because of high data collection costs, time-consuming annotations, and patient privacy concerns. The training of deep neural network classification models on these data sets to improve the generalization ability does not produce the desired results for classifying the medical condition accurately and often overfit the data on the majority of class samples. To address the issue, we propose a framework for improving the classification performance metrics of deep neural network classification models using transfer learning: pre-trained models, such as Xception, InceptionResNet, DenseNet along with the Generative Adversarial Network (GAN) – based data augmentation. Then, we trained the network by combining traditional data augmentation techniques, such as randomly flipping the image left to right and GAN-based data augmentation, and then fine-tuned the hyper-parameters of the transfer learning models, such as the learning rate, batch size, and the number of epochs. With these configurations, the Xception model outperformed all other pre-trained models achieving a test accuracy of 98.7%, the precision of 99%, recall of 99.3%, f1-score of 99.1%, receiver operating characteristic (ROC) - area under the curve (AUC) of 98.2%.

Keywords: Generative Adversarial Networks; Transfer Learning; Medical Imaging; Deep Learning Classification; Chest X-ray's

1. Introduction

In general, medical image datasets, such as Chest X-ray images, are usually imbalanced and come with limited samples due to the high costs of obtaining the data and time-consuming annotations. Training a deep neural network model on such datasets to accurately classify the medical condition does not yield the desired results. Every so often over-fits the majority class samples' data. Usually, transfer learning and data augmentation are performed on the training data to improve the deep learning model's classification performance to address the issue.

First, for classification tasks with limited datasets, transfer learning is adopted. It improves learning in a new domain by transferring knowledge from a related domain, reducing the neural network's training time and generalization error. It is a common practice in the field of computer vision to use transfer learning for limited datasets via pre-trained models. The pre-trained models are those trained on large benchmark datasets, where the models have already learned to extract a wide variety of features, which can be used as a starting point to learn on a new task in a related domain. To enhance the models' performance, it is not uncommon to overlook the fine-tuning of the hyper-parameters of transfer learning models.

Second, to balance the datasets, there are a few traditional methods. Random over-sampling, which produces copies of minority class samples, and Synthetic Minority Over-sampling Technique (SMOTE) [1], which generates synthetic data from dataset

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

samples from nearest k-nearest neighbors. These methods of augmentation are not guaranteed to be advantageous and are only well suited to low-dimensional data. Deep generative models such as Generative Adversarial Networks (GANs) are known to augment high-dimensional image data effectively.

To address the issues of class imbalance and limited sample sizes for classification tasks, we propose a framework for improving the classification performance metrics of deep neural network classification models using transfer learning: pre-trained models, such as Xception, InceptionResNet, DenseNet, and along with the GAN – based data augmentation. We show the proposed framework in the Figure 1.

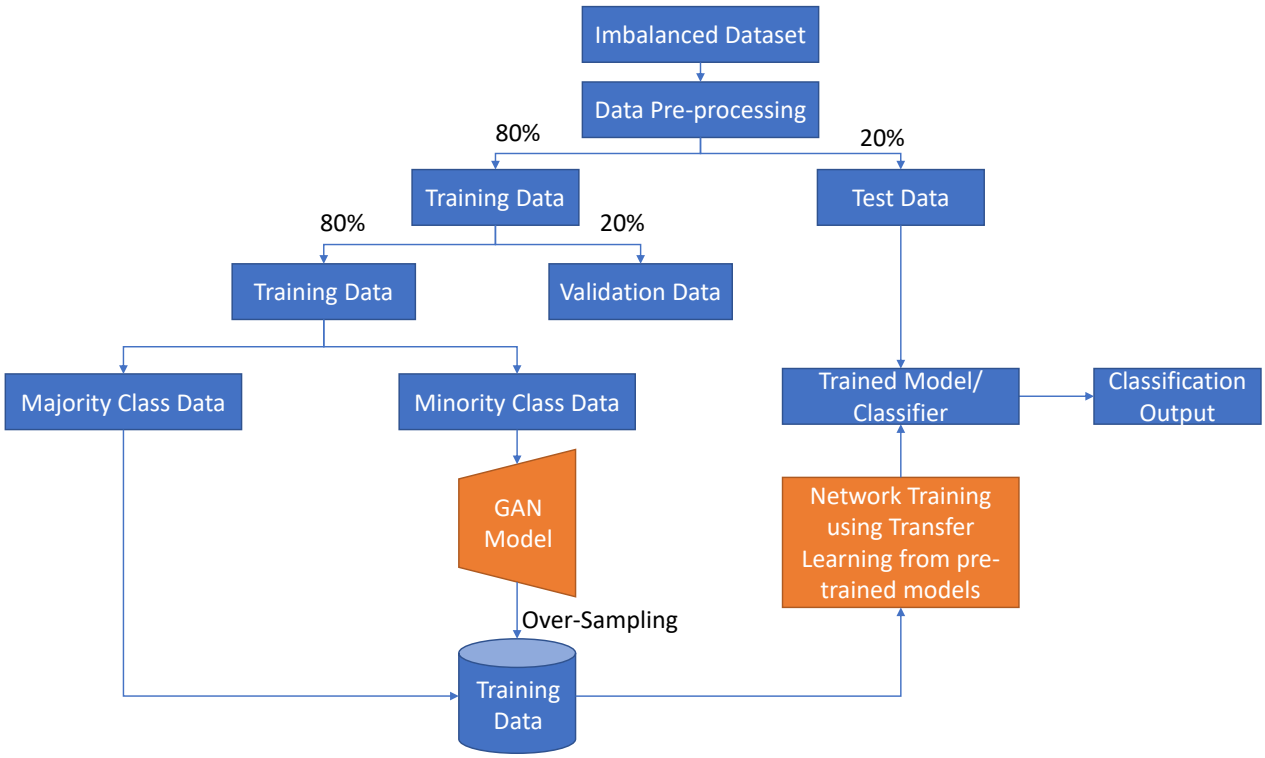


Figure 1. Proposed framework.

In one of our previous studies, we explored the GANs in creating artificial instances of chest X-ray images [2]. In another study, we investigated transfer learning to classify pneumonia from chest X-ray images [3]. In this study, we evaluated the combination of GAN - based data augmentation and transfer learning approaches.

The rest of the paper is structured as follows: In Section 2, we briefly present the related work from the literature. We introduce the materials and methods used in the study in Section 3. In Sections 4 and 5, we present the results of the proposed framework and discuss our findings. Finally, in Section 6, we conclude our study by providing directions to future work.

2. Related Work

The lack of availability of large, labeled datasets is one of the significant problems with deep learning in medical imaging. As mentioned in section 1, Medical images are not only annotated expensively but also time-consuming. By producing synthetic samples with real images' appearance, Generative Adversarial Networks (GANs) provide a novel way to create additional information from a dataset. In this section, we briefly review the literature on the use of transfer learning and GANs in the analysis of medical image data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

2.1. Transfer Learning in Medical Imaging

Rajpurkar et al. [4] developed an algorithm CheXNet, a 121-layer Dense convolutional neural network (DenseNet) that detects Pneumonia from chest X-ray images at a level exceeding the practicing radiologists. They trained their network on the ChestX-ray14 dataset released by Wang et al. [5] and assessed the performance with four practicing radiologists on the f1 score metric. CheXNet achieved an f1 score of 0.435, higher than the radiologist's average of 0.387. They later extended the CheXNet model to detect all 14 diseases in the ChestX-ray14 dataset and achieved state-of-the-art results on all 14 diseases. Similarly, Antin et al. [6] utilized transfer learning approach from a pre-trained model, DenseNet121 to classify Pneumonia from chest X-ray image dataset [5]. The only reported metric on the test data is the area under the curve (AUC), which they reported to be 60%.

Ayan and Ünver [7] proposed to use transfer learning approach from two state-of-the-art pre-trained architectures, Xception and VGG16, for classifying Pneumonia from chest X-ray images. They trained the two networks individually and reported that the Vgg16 network outperformed the Xception network by accuracy 87%, specificity 91%, precision 91%, and f1 score 90% with respect to pneumonia images. In contrast, the Xception network outperformed the Vgg16 network by sensitivity 85%, precision 86% for normal images, and recall 94% concerning pneumonia images.

The deep learning framework proposed by Liang et al. [8] incorporates transfer learning combined with residual thought and dilated convolution for the classification of pediatric pneumonia images. The deep neural network consisted of 49 convolutional layers combined with the ReLU activation, followed by a global average pooling layer and two dense layers. They used transfer learning by transferring the network weights from a pre-trained model on the large-scale dataset: ChestX-ray14 dataset [5] to accelerate neural network training and overcome the problem of insufficient data. The network's training is then carried out by introducing dilated convolutions and using the Adam as an optimizer to minimize the cross-entropy loss function. They achieved a test recall of 96.7%, and an f1 score of 92.7% in classifying Pneumonia from the chest X-ray image dataset [9].

Chouhan et al. [10] proposed a deep learning framework combined with the use of transfer learning to classify Pneumonia from chest X-ray images by adopting an ensemble-based approach to pre-trained architectures, such as AlexNet, InceptionV3, DenseNet121, ResNet18, and GoogLeNet. They trained the AlexNet for 200 epochs with an initial learning rate of 0.001 and then retrained with a learning rate of 0.00001. To prevent overfitting and improving generalization, they trained the DenseNet121 and InceptionV3 for 100 epochs and the GoogLeNet for 50 epochs. The network training is performed using Adam as an optimizer to minimize the cross-entropy loss function. Unfortunately, the authors did not provide details on the metrics/methods used to choose the learning rate and the number of epochs used to train the network. The final prediction was based on majority voting by combining the results of pre-trained neural networks in the Pneumonia classification. Their proposed ensemble model achieved an accuracy of 96.4% with a recall of 99.62% on unseen data from the Guangzhou Women and Children's Medical Center dataset [9].

Similarly, Hashmi et al. [11] proposed a weighted classifier that optimally combined the weighted predictions from the state-of-the-art deep learning models ResNet18, Xception, InceptionV3, DenseNet121, and MobileNetV3 for Pneumonia classification from chest X-ray images. They achieved a test accuracy of 98.43%, and an AUC score of 99.76% on the unseen data from the Guangzhou Women and Children's Medical Center pneumonia dataset [9]. In another study, Rahman et al. [12] used transfer learning from pre-trained networks such as AlexNet, ResNet18, DenseNet201, and SqueezeNet for classifying Pneumonia from chest X-ray images. The authors did not mention any details on the hyper-parameters used for the study. They showed that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

DenseNet201 outperformed the other three pre-trained networks, achieving a 98% accuracy on Pneumonia classification from the chest X-ray image dataset [9].

2.2. Generative Adversarial Networks in Medical Imaging

Bowles et al. [13] demonstrated GAN's feasibility to create synthetic data for two brain segmentation tasks. They used Progressive Growing of GANs [14], which improved the training stability at large image sizes to generate synthetic data. They reported that when synthetic data created using GAN's combined with the training data improved the Dice Similarity Coefficient anywhere from one and five percentage points. Similarly, Beers et al. [15] also used Progressive Growing of GANs for generating realistic medical images in two different domains. First, they could generate realistic fundus photographs exhibiting vascular pathology associated with retinopathy of prematurity (ROP). Second, they were able to generate synthetic multi-modal magnetic resonance images of glioma. In another research, Korkinof et al. [16] explored the use of progressively trained generative adversarial networks (GANs) for generating highly realistic, high-resolution synthetic Full Field Digital Mammograms (FFDM). They reported achieving the highest resolution of 1280x1024 pixels.

In another study, Sandfort et al. [17] evaluated the CycleGAN [18] for data augmentation in CT segmentation tasks by first transforming the contrast CT images to non-contrast CT images and then generated synthetic non-contrast CT images [19], [20], [21]. They trained the network by comparing the segmentation performance of a U-Net [22] trained on the original dataset compared to a U-Net trained on the combined dataset of original data and synthetic non-contrast images demonstrated substantial improvement in the segmenting performance of the CT images, with the Dice score increasing from 0.09 to 0.66.

In another research, Welander et al. [23], evaluated two unsupervised GAN models, such as CycleGAN [18] and UNIT [24] for image-to-image translation of T1- and T2-weighted MR images, by comparing generated synthetic MR images to ground truth images. They also evaluated two supervised models; a modification of CycleGAN (CycleGAN_s) and a pure generator model (Generator_s), and reported that all the GAN models would synthesize visually realistic MR images [25], [26]. Iqbal and Ali [27] proposed another Generative Adversarial Network for Medical Imaging, MI-GAN, for synthesizing Retinal images. They used the STARE [28], and DRIVE [29] datasets for evaluating the MI-GAN model and reported that they achieved a Dice coefficient of 0.837 on the STARE dataset and a Dice coefficient of 0.832 on the DRIVE dataset.

Dar et al. [30] demonstrated an end-to-end image synthesis approach for MRI that successfully estimated the image in the target contrast given the image in the source contrast, by utilizing conditional generative adversarial networks, cGANs [31], with pixel-wise and cycle-consistency loss functions. They trained the conditional GAN on three datasets, such as MIDAS dataset [32], the IXI dataset [33], and the BRATS dataset [34]. In order to generate realistic lung nodule samples, Chuquicusma et al. [35] proposed the use of unsupervised learning through the Deep Convolutional Generative Adversarial Networks (DCGANs). Likewise, Salehinejad et al. [36] showed an improvement in chest pathology classification performance by augmenting the original imbalanced data set with DCGAN. Similarly, in another study, the DCGAN was investigated by Madani et al. [37] to generate chest X-ray images for the augmentation of the original data and trained a convolutional neural network to classify cardiovascular abnormalities, showing a higher classification accuracy.

In another study, Qin et al. [38] investigated data sampling methods such as undersampling the majority class, in which the majority class in the training dataset was randomly dropped to achieve a 1:1 ratio between classes, and over-sampling/augmentation of the minority class, such as affine transformations and GAN-based data augmentation, to learn from the imbalanced and limited chest X-ray dataset. They reported achieving improved classification metrics with an accuracy of 89.9%, recall of 89.7%, the preci-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

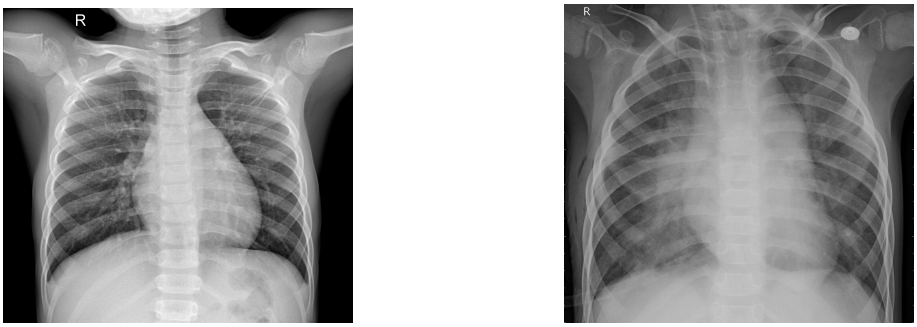
sion of 93.8%, F1score of 91.7%, and an AUC of 95.4% in detecting pneumonia with GAN-based data augmentation when trained with a deep convolutional neural network.

The combination of fine-tuning the hyper-parameters of transfer learning models and GAN-based data augmentation has received very little attention in the literature. This study addresses improving the classification performance metrics of an imbalanced and limited dataset by fine-tuning the hyper-parameters of the transfer learning models and utilizing GAN-based data augmentation.

3. Materials and Methods

3.1. Dataset Description and Pre-processing

We used a chest X-ray image dataset published by Kermany et al. [9] and chest X-ray images generated using GANs by Kora Venu et al. [2] for all the experiments conducted in this study. The dataset by Kermany et al. comprises 5,856 chest X-ray images in total, of which 1583 images labeled as Normal and 4273 images labeled as Pneumonia. The dataset was then shuffled and split into training and test sets, of which 4,684 images in the training set (Normal Images: 1,266 and Pneumonia images: 3,418), and 1,172 images in the test set (Normal images: 317, and Pneumonia images: 855) We further split the training dataset to have 80% as training data (3,748 images in total, of which 1013 are Normal images and 2735 are Pneumonia images) and 20% as validation data (936 images in total, of which 253 are Normal images and 683 are Pneumonia images). We show a Normal image and Pneumonia image sample in Figure 2.



(a) Normal Image. **(b)** Pneumonia Image.

Figure 2. Sample of Normal and Pneumonia Images.

To balance the training dataset, we have combined the Normal chest X-ray images generated with GANs [2] to the original training dataset to have an equal proportion of Normal and Pneumonia Images. We show a sample of Normal chest X-ray image generated using GANs in Figure 3.

Figure 3. Normal Chest X-ray image generated using GAN

The images are resized to have a shape of 224x224x3, and then we created TFRecords of the data to train on Tensor Processing Units (TPUs). Each pre-trained model expects

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

a specific kind of input pre-processing, and they all have the methods to pre-process the inputs before passing them to the model. For example, the Xception and Inception-ResNetV2 networks expects to have the input pixel values scaled between -1 and 1, the DenseNet network expects to have the input pixel values scaled between 0 and 1.

3.2. Transfer Learning Models

In the following sub-sections, we will discuss in detail the pre-trained models used in this study, such as Xception, DenseNet, and InceptionResNetV2.

3.2.1. Xception

Xception network, also known as extreme version of Inception is one of the state-of-the-art neural network architectures introduced by Chollet [39], which is based on depth-wise separable convolution layers. The detailed architecture of the Xception network is shown in the Figure 4.

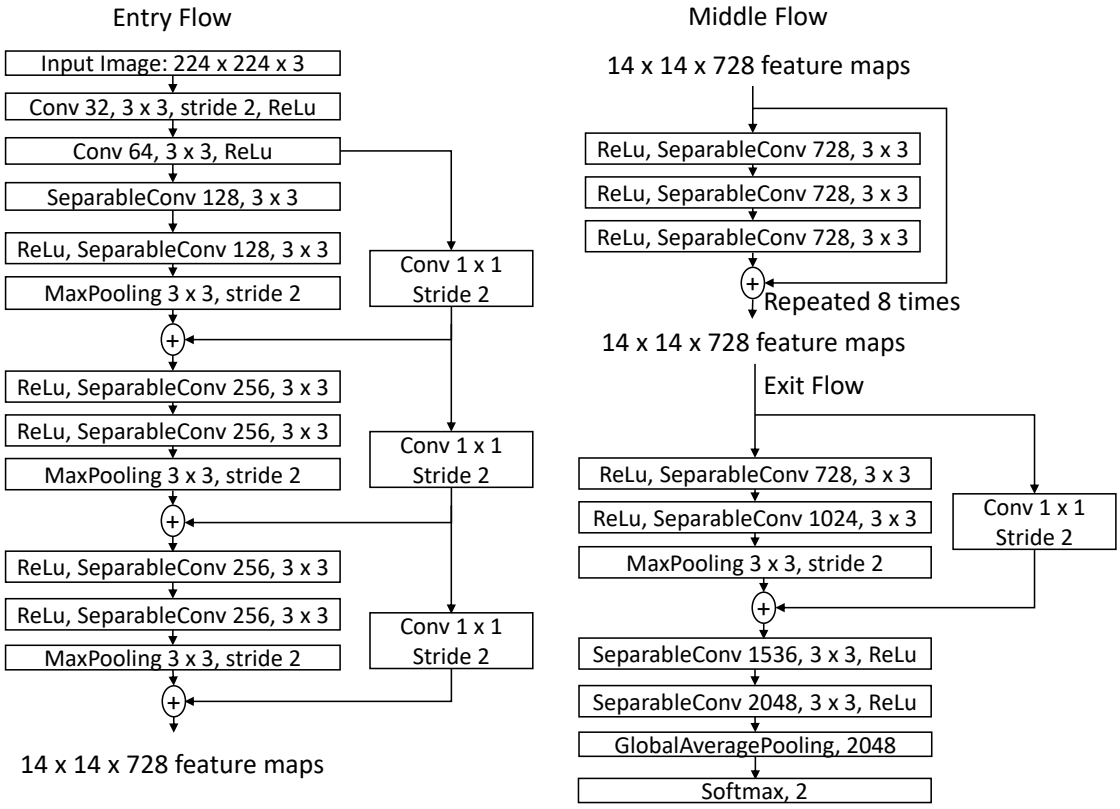


Figure 4. Xception Architecture.

The entire architecture consists of three flows, namely, the Entry flow, the Middle flow, and the Exit flow. The entry flow consists of four blocks, with traditional convolutional layers in the first block and depth-wise separable convolutional layers in the remaining three blocks. There is only one block of depth-wise separable convolution layers in the middle flow, repeated eight times. The exit flow has two blocks: the first block has the depth-wise separable convolutional layers, and the second block consists of depth-wise separable convolutional layers followed by a GlobalAveragePooling layer and a dense layer with softmax activation to output the probability of the input image being Normal or Pneumonia.

The Entry flow takes the pre-processed image of size 224x224x3 as an input, followed by two traditional convolutions. to output the representation of size 14x14x728. The Middle flow takes the representation of size 14x14x728 as input to output the repre-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

sentation of size 14x14x728. The Exit flow takes the representation of size 14x14x728 as input to output the probability of image as Normal or Pneumonia.

3.2.2. DenseNet

Huang et al. [40] introduced Densely Connected Convolutional Networks (DenseNet), which won the best paper award at the CVPR 2017 conference [41]. The DenseNet architecture is shown in Figure 5, which connects each layer of the network in a feed-forward manner to every other layer. In other words, all previous layers' feature maps are used as inputs into each layer, and its own feature maps are used as inputs into all subsequent layers [40]. With L layers, the DenseNet network has $L(L+1)/2$ direct connections compared to L connections of a traditional convolutional network, thereby significantly reducing the network's overall learnable parameters.

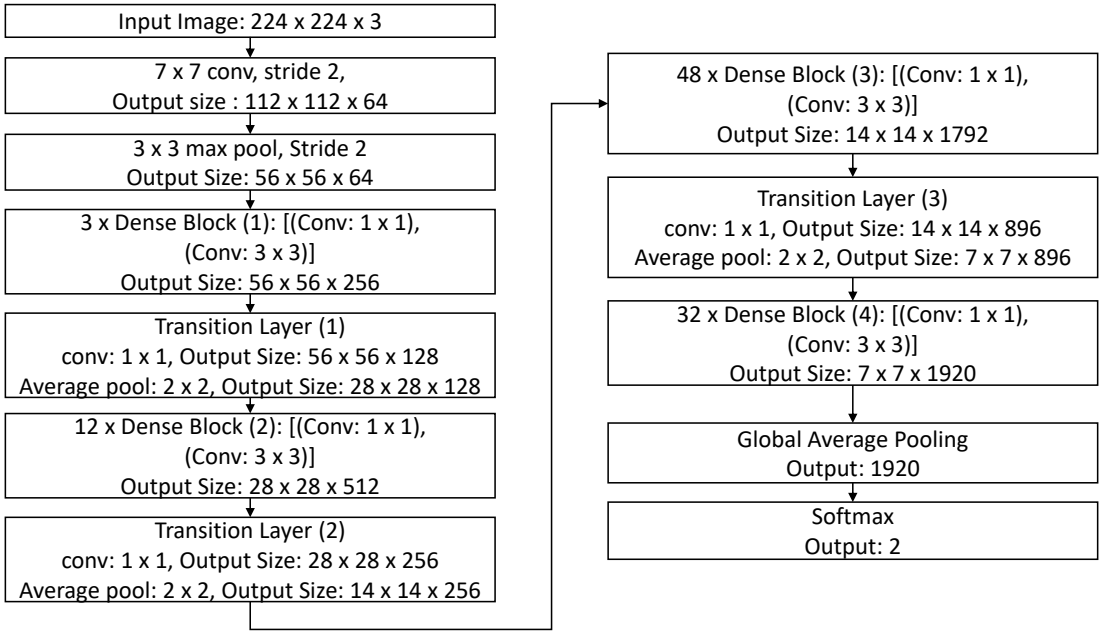


Figure 5. DenseNet201 Architecture.

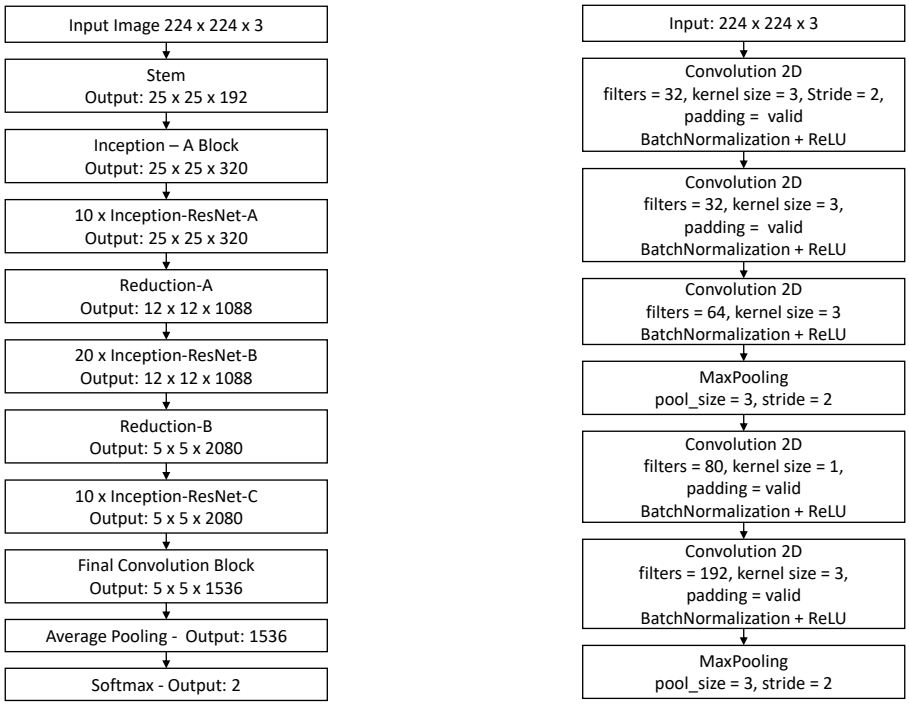
As shown in Figure 5, the DenseNet network takes an image of size 224x224x3 as input. The image then goes through an initial convolutional layer with a kernel size of 7x7 and a stride of 2 to output a representation of size 112x112x64, followed by a MaxPooling operation with a kernel size of 3x3 and a stride of 2, halving the representation size to 56x56x64. The 56x56x64 representation is subjected to a series of dense blocks and transition layers. Each dense block consists of a 1x1 convolution, followed by a 3x3 convolution, and each transition block consists of a 1x1 convolution followed by a 2x2 average pooling operation. The final dense block's output is a representation of size 7x7x1920, which is passed on to the global average pooling layer with a softmax activation to output the probability of the image as Normal or Pneumonia.

3.2.3. InceptionResNet

Szegedy et al. [42] introduced the InceptionResNet architecture based on the Inception Architectures family by replacing the Inception modules with the Inception-ResNet hybrid modules. The network training is significantly accelerated due to the presence of residual connections in the network. The InceptionResNet network architecture is shown in Figure 6a. The InceptionResNet network takes an image of size 224x224x3 as input, followed by a Stem module where the input image undergoes a series of convolutions as shown in Figure 6b. The output of final convolution in the stem module is followed by a Max-Pooling layer to output a representation of size 25x25x192. The output from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

the stem module is passed on to the Inception - A block as shown in Figure A1 to output a representation of size 25x25x320. The output from the Inception - A block is subjected to a series of hybrid Inception-ResNet modules and Reduction modules, such as Inception-ResNet-A (see Figure A2) followed by a Reduction-A module (see Figure A5), Inception-ResNet-B (see Figure A3) followed by a Reduction-B module (see Figure A6), and Inception-ResNet-C (see Figure A4). The detailed architectures of the hybrid Inception-ResNet modules and the corresponding Reduction modules are shown in the Appendix. The final hybrid Inception-ResNet module's output (Inception-ResNet-C) is fed to the average pooling layer, followed by a softmax layer to output the predictions.



(a) Large scale schema structure. **(b)** Schema for Stem module.

Figure 6. InceptionResNet Architecture and Stem Module.

3.3. Generative Adversarial Networks (GAN)

GANs are gaining traction as effective tools for dealing with data imbalance, which is quite common in the domain of medical imaging. The core principle behind GAN is to produce plausible synthetic data that is as realistic as to the original data from the training dataset. The architecture of the GAN is shown in the Figure 7. The generator network (G) and the discriminator network (D) are the two main building blocks of the GAN architecture, with the generator network learning to produce images as realistic as the original training data by taking random noise (Z) as input and the discriminator network randomly guessing at 50% probability that the image is from the generator distribution or the original training data. In this study, we use data generated by GAN from one of our previous studies [2], in which we discussed the individual architectures of the generator network and discriminator network, and the training process in detail.

4. Results

As per the proposed framework, we over-sampled the minority class samples using GAN's and trained the deep neural network by fine-tuning the hyper-parameters of the transfer learning models, such as the learning rate, batch size and the number of epochs to train the model. We trained the network using three pre-trained models, such as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

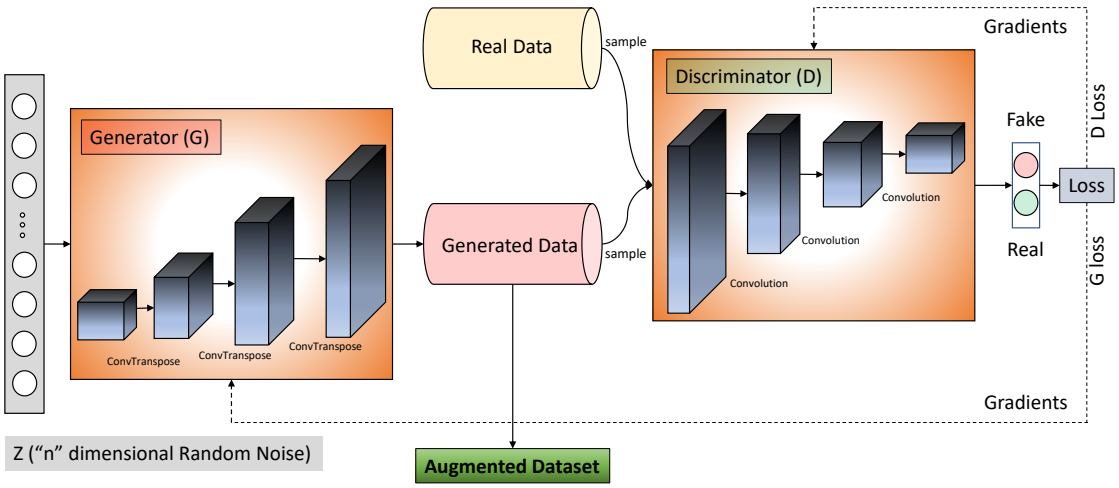


Figure 7. GAN Architecture.

Xception, DenseNet201, and InceptionResNetV2 individually. We show the classification metrics, such as accuracy, precision, recall, and F1-score in Table 1.

Model	Accuracy	Precision	Recall	F1-Score
Xception	98.7%	99%	99.3%	99.1%
DenseNet201	98.4%	98.5%	99.3%	98.9%
InceptionResNetV2	98.5%	98.7%	99.2%	99%

Table 1: Classification performance metrics.

We also show the confusion matrix and Receiver Operating Characteristics metrics in Tables 2 and 3.

Model	True Negative	False Positive	False Negative	True Positive
Xception	308	9	6	849
DenseNet201	304	13	6	849
InceptionResNetV2	306	11	7	848

Table 2: Confusion Matrix.

Model	False Positive Rate	True Positive Rate	Area Under the Curve
Xception	0.028	0.993	99.3%
DenseNet201	0.041	0.993	99.6%
InceptionResNetV2	0.035	0.992	99.5%

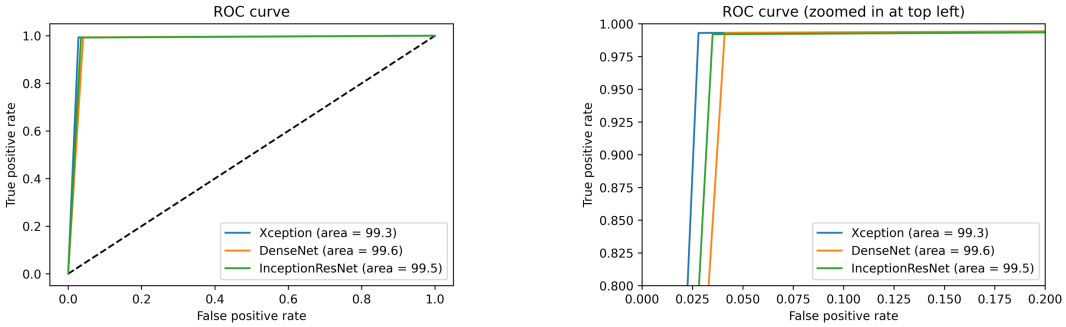
Table 3: Receiver Operating Characteristics.

The Xception model showed superior performance than the other two pre-trained models, acheiving an accuracy of 98.7%, precision of 99%, recall of 99.3%, F1-score of 99.1%, and AUC of 99.3%. To support our findings, we also plotted the ROC curve of the models as shown in Figure 8. The ROC curve graph confirms that the Xception model performed better than the DenseNet201 and InceptionResNetV2 models.

4.1. Comparison of results with other recent similar works

We compare the results of this study with other recent similar works in this section - see Table 4. The proposed model results, i.e., over-sampling of the minority class samples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1



(a) ROC Curve. **(b)** ROC Curve - Zoomed in at top left.

Figure 8. Receiver Operating Characteristics Curve.

using GAN's and training the neural network by fine-tuning the transfer learning models' hyper-parameters, outperformed all the previous studies in the majority of the classification metrics. As discussed in the Section 4, the Xception architecture achieved the best classification metrics with an accuracy of 98.7%, precision of 99%, recall of 99.3%, F1-score of 99.1% and ROC-AUC of 99.3%.

	Accuracy	Precision	Recall	F1 Score	AUC
Kermany et al. [9]	92.80	87.20	93.20	90.10	96.80
Nahid et al. [43]	97.92	98.38	97.47	97.97	-
Stephen et al. [44]	93.73	-	-	-	-
Qin et al.[38]	89.90	93.80	89.70	91.70	95.40
Chouhan et al. [10]	96.39	93.28	**99.62**	96.35	99.34
Rajaraman et al. [45]	96.20	97.00	99.50	-	99.00
Hashmi et al. [11]	98.43	98.26	99.00	98.63	**99.76**
Mittal et al. [46]	96.36	-	-	-	-
Rahman et al. [12]	98.00	97.00	99.00	98.10	98.00
Kora Venu [3]	98.46	98.38	99.53	98.96	99.60
Kora Venu et al. [2]	95.50	96.20	97.70	97.00	93.60
Current work	**98.70**	**99.00**	99.30	**99.10**	99.30

Table 4: Comparison of results with other recent similar works

5. Discussion

Medical image datasets generally come by limited samples and are often imbalanced on majority class samples. Transfer learning from pre-trained models, i.e., the models trained on a large-scale benchmark datasets like Imagenet, is commonly used for classification tasks when encountered with small datasets, which is more common in the medical imaging domain. Often over-sampling of minority class samples or under-sampling of majority class samples, or augmentation of minority class samples using traditional data augmentation techniques, such as position or color augmentation, is carried out to balance the dataset before training the model to This makes sure that the trained model's classification performance is not biased against the majority class samples. In this study, we proposed a framework where we over-sampled the minority class samples using Generative Adversarial Networks and then fine-tuned the hyper-parameters of transfer learning models to improve the classification performance of the trained model. Using Generative Adversarial Networks for data augmentation has two significant advantages, i.e., 1. Diversity - GAN's can generate more varied images than the sampling or traditional data augmentation techniques, and 2. Fidelity - GAN's improve the quality of generated images. Before training the models on TPU's, the models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

were fine-tuned and compiled. The results show that the combination of GAN-based data augmentation and fine-tuning of the transfer learning models' hyper-parameters demonstrates a significant improvement in classification metrics.

6. Conclusions and Future Work

The lack of availability of large, labeled datasets is one of the significant problems with deep learning classification tasks in the domain of medical imaging. We demonstrated the ability to generate synthetic samples of chest X-ray images using Generative Adversarial Networks in one of our previous studies [2], and we demonstrated that fine-tuning the hyper-parameters of the transfer learning models improves classification performance metrics in another study [3]. These studies gave confidence and motivation in conducting the present research by combining the GAN-based data augmentation for over-sampling the minority class samples to balance the dataset and fine-tuning the hyper-parameters of the transfer learning models. We later trained the deep neural network classification models on three pre-trained state-of-the-art transfer learning models, such as Xception, DenseNet201, and InceptionResNetV2. The Xception model outperformed the other two models achieving the test accuracy of 98.7%, the precision of 99%, recall of 99.3%, f1-score of 99.1%, receiver operating characteristic (ROC) - area under the curve (AUC) of 98.2%. Future work may include the investigation of other GAN methods for generating synthetic data, such as Wasserstein GAN with a gradient penalty, and the investigation of the generalization of the proposed framework to improve the classification performance metrics of other common medical conditions.

Funding: "This research received no external funding"

Acknowledgments: We thank Kermany et al. [9] for making the datasets publicly accessible and we also thank Harrisburg University of Science and Technology for their support.

Conflicts of Interest: "The author declares no conflict of interest."

Appendix A

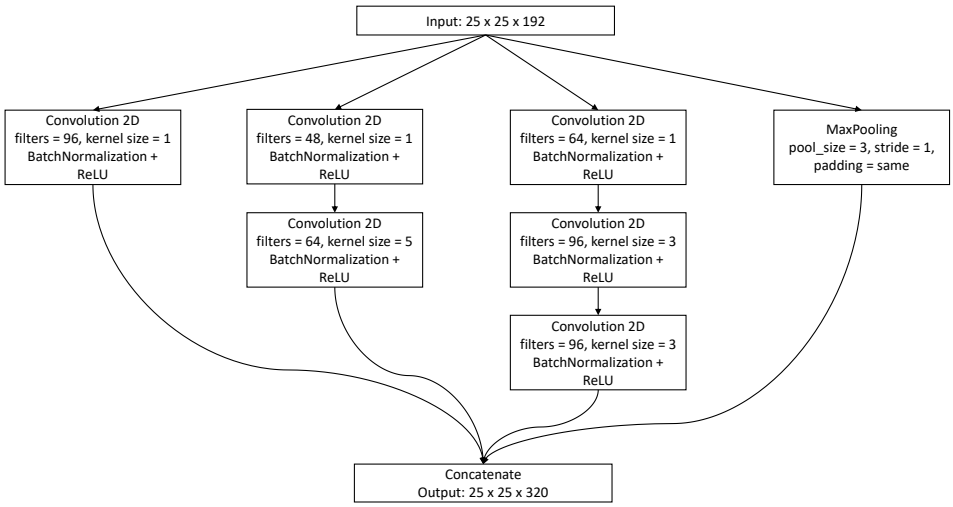


Figure A1. Inception - A Block.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

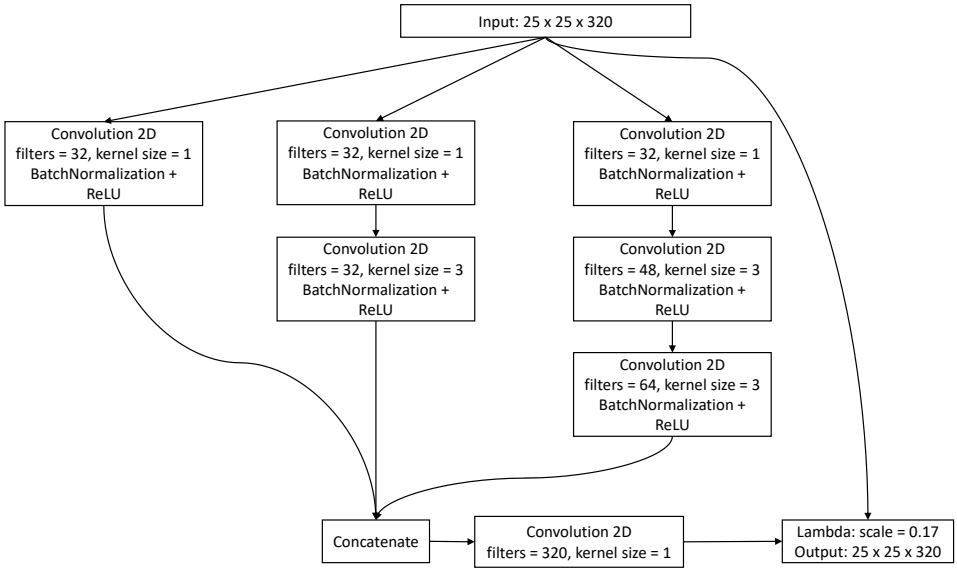


Figure A2. InceptionResNet - A Block.

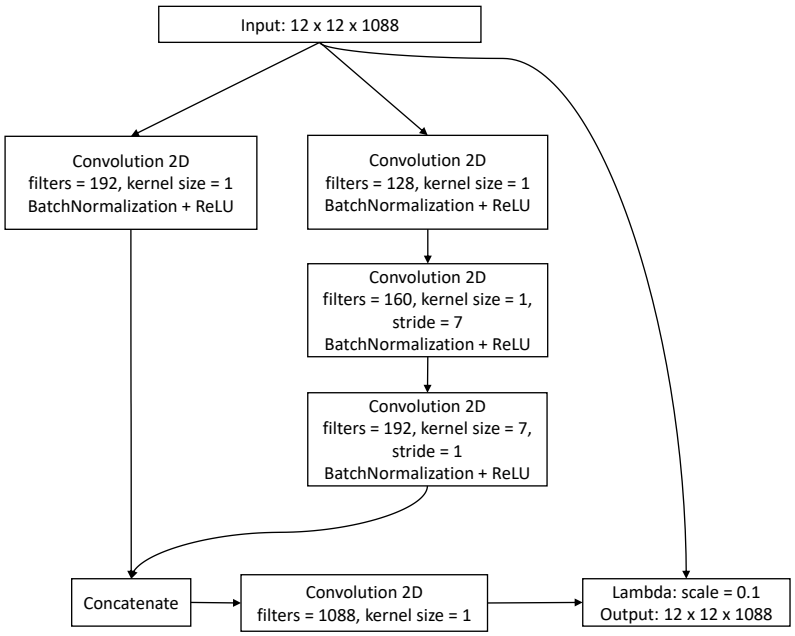


Figure A3. InceptionResNet - B Block.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

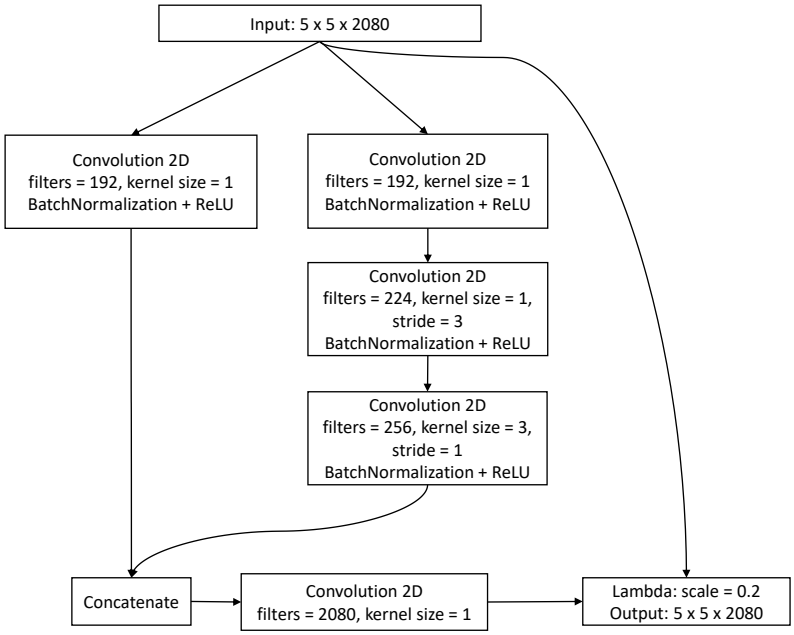


Figure A4. InceptionResNet - C Block.

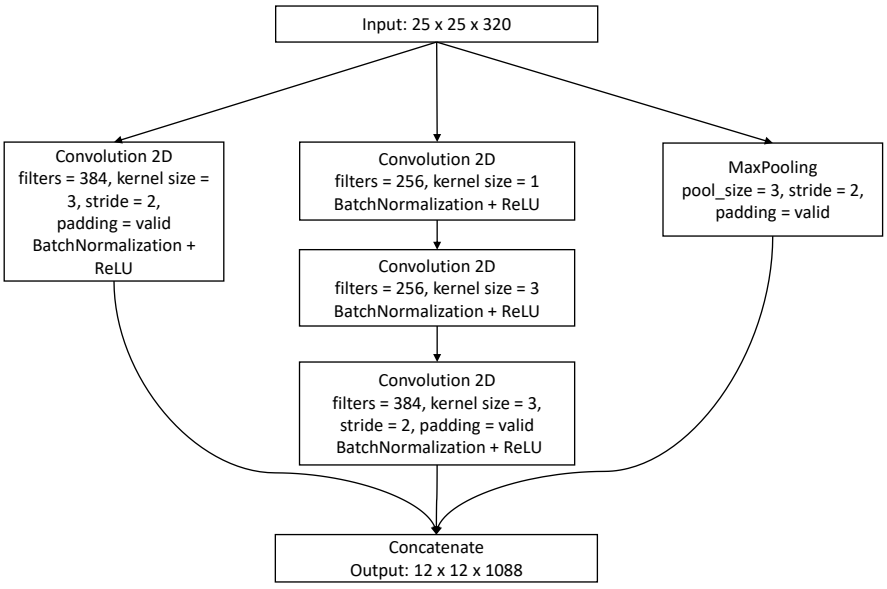


Figure A5. Reduction - A Block.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

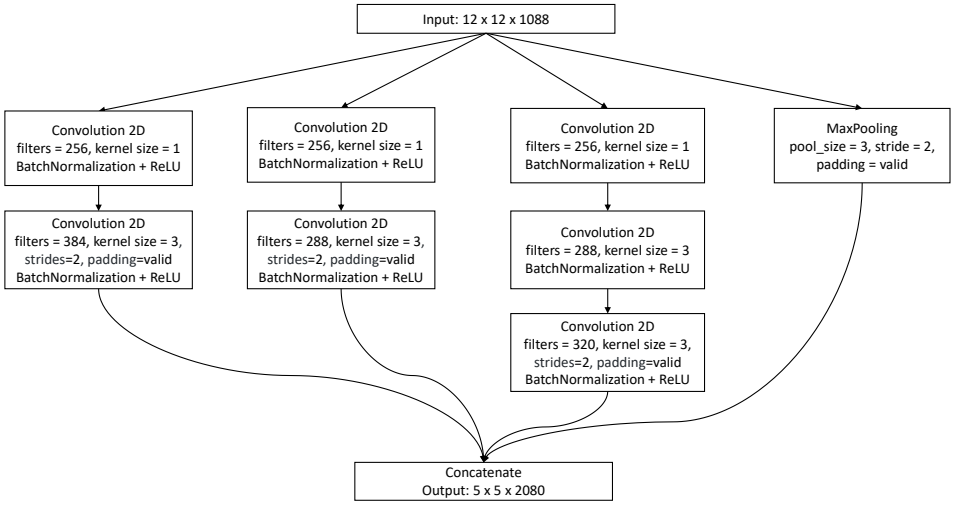


Figure A6. Reduction - B Block.

References

1. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **2002**, *16*, 321–357.
2. Kora Venu, S.; Ravula, S. Evaluation of Deep Convolutional Generative Adversarial Networks for Data Augmentation of Chest X-ray Images. *Future Internet* **2021**, *13*. doi:10.3390/fi13010008.
3. Kora Venu, S. An Ensemble-based Approach by Fine-Tuning the Deep Transfer Learning Models to Classify Pneumonia from Chest X-Ray Images. Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART,. INSTICC, SciTePress, 2021, pp. 390–401. doi:10.5220/0010377403900401.
4. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; others. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* **2017**.
5. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.
6. Antin, B.; Kravitz, J.; Martayan, E. Detecting pneumonia in chest X-Rays with supervised learning. *Semanticscholar. org* **2017**.
7. Ayan, E.; Ünver, H.M. Diagnosis of Pneumonia from Chest X-Ray Images Using Deep Learning. 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT). IEEE, 2019, pp. 1–5.
8. Liang, G.; Zheng, L. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer methods and programs in biomedicine* **2020**, *187*, 104964.
9. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; others. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131.
10. Chouhan, V.; Singh, S.K.; Khamparia, A.; Gupta, D.; Tiwari, P.; Moreira, C.; Damaševičius, R.; De Albuquerque, V.H.C. A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Applied Sciences* **2020**, *10*, 559.
11. Hashmi, M.F.; Katiyar, S.; Keskar, A.G.; Bokde, N.D.; Geem, Z.W. Efficient pneumonia detection in chest xray images using deep transfer learning. *Diagnostics* **2020**, *10*, 417.
12. Rahman, T.; Chowdhury, M.E.; Khandakar, A.; Islam, K.R.; Islam, K.F.; Mahbub, Z.B.; Kadir, M.A.; Kashem, S. Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection using Chest X-ray. *Applied Sciences* **2020**, *10*, 3233.
13. Bowles, C.; Chen, L.; Guerrero, R.; Bentley, P.; Gunn, R.; Hammers, A.; Dickie, D.A.; Hernández, M.V.; Wardlaw, J.; Rueckert, D. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863* **2018**.
14. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* **2017**.
15. Beers, A.; Brown, J.; Chang, K.; Campbell, J.P.; Ostmo, S.; Chiang, M.F.; Kalpathy-Cramer, J. High-resolution medical image synthesis using progressively grown generative adversarial networks. *arXiv preprint arXiv:1805.03144* **2018**.
16. Korkinof, D.; Rijken, T.; O'Neill, M.; Yearsley, J.; Harvey, H.; Glocker, B. High-resolution mammogram synthesis using progressive generative adversarial networks. *arXiv preprint arXiv:1807.03401* **2018**.
17. Sandfort, V.; Yan, K.; Pickhardt, P.J.; Summers, R.M. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific reports* **2019**, *9*, 1–9.
18. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 July 2021 doi:10.20944/preprints202107.0636.v1

19. Simpson, A.L.; Antonelli, M.; Bakas, S.; Bilello, M.; Farahani, K.; Van Ginneken, B.; Kopp-Schneider, A.; Landman, B.A.; Litjens, G.; Menze, B.; others. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* **2019**.
20. Pickhardt, P.J.; Lee, S.J.; Liu, J.; Yao, J.; Lay, N.; Graffy, P.M.; Summers, R.M. Population-based opportunistic osteoporosis screening: Validation of a fully automated CT tool for assessing longitudinal BMD changes. *The British journal of radiology* **2019**, *92*, 20180726.
21. Yan, K.; Wang, X.; Lu, L.; Summers, R.M. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging* **2018**, *5*, 036501.
22. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention. Springer, 2016, pp. 424–432.
23. Welander, P.; Karlsson, S.; Eklund, A. Generative adversarial networks for image-to-image translation on multi-contrast MR images-A comparison of CycleGAN and UNIT. *arXiv preprint arXiv:1806.07777* **2018**.
24. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. *arXiv preprint arXiv:1703.00848* **2017**.
25. Van Essen, D.C.; Smith, S.M.; Barch, D.M.; Behrens, T.E.; Yacoub, E.; Ugurbil, K.; Consortium, W.M.H.; others. The WU-Minn human connectome project: an overview. *Neuroimage* **2013**, *80*, 62–79.
26. Glasser, M.F.; Sotiropoulos, S.N.; Wilson, J.A.; Coalson, T.S.; Fischl, B.; Andersson, J.L.; Xu, J.; Jbabdi, S.; Webster, M.; Polimeni, J.R.; others. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **2013**, *80*, 105–124.
27. Iqbal, T.; Ali, H. Generative adversarial network for medical images (MI-GAN). *Journal of medical systems* **2018**, *42*, 1–11.
28. The STARE Project. https://cecas.clemson.edu/~ahoover/stare/. (Accessed on 01/27/2021).
29. Introduction - DRIVE - Grand Challenge. https://drive.grand-challenge.org/DRIVE/. (Accessed on 01/27/2021).
30. Dar, S.U.; Yurt, M.; Karacan, L.; Erdem, A.; Erdem, E.; Çukur, T. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. *IEEE transactions on medical imaging* **2019**, *38*, 2375–2388.
31. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* **2014**.
32. Bullitt, E.; Zeng, D.; Gerig, G.; Aylward, S.; Joshi, S.; Smith, J.K.; Lin, W.; Ewend, M.G. Vessel tortuosity and brain tumor malignancy: a blinded study1. *Academic radiology* **2005**, *12*, 1232–1240.
33. IXI Dataset – Brain Development. https://brain-development.org/ixi-dataset/. (Accessed on 01/27/2021).
34. BraTS 2015 - MICCAI BRATS 2017. https://sites.google.com/site/braintumorsegmentation/home/brats2015. (Accessed on 01/27/2021).
35. Chuquicusma, M.J.M.; Hussein, S.; Burt, J.; Bagci, U. How to Fool Radiologists with Generative Adversarial Networks? A Visual Turing Test for Lung Cancer Diagnosis, 2018, [arXiv:cs.CV/1710.09762].
36. Salehinejad, H.; Valaee, S.; Dowdell, T.; Colak, E.; Barfett, J. Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks. *CoRR* **2017**, *abs/1712.01636*, [1712.01636].
37. Madani, A.; Moradi, M.; Karargyris, A.; Syeda-Mahmood, T. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. Medical Imaging 2018: Image Processing; Angelini, E.D.; Landman, B.A., Eds. International Society for Optics and Photonics, SPIE, 2018, Vol. 10574, pp. 415 – 420. doi:10.1117/12.2293971.
38. Qin, X.; Bui, F.M.; Nguyen, H.H. Learning from an Imbalanced and Limited Dataset and an Application to Medical Imaging. 2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM). IEEE, 2019, pp. 1–6.
39. Chollet, F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
40. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
41. CVPR2017. https://cvpr2017.thecvf.com/program/main_conference#cvpr2017_awards. (Accessed on 10/31/2020).
42. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261* **2016**.
43. Nahid, A.A.; Sikder, N.; Bairagi, A.K.; Razzaque, M.; Masud, M.; Z Kouzani, A.; Mahmud, M.; others. A novel method to identify pneumonia through analyzing chest radiographs employing a multichannel convolutional neural network. *Sensors* **2020**, *20*, 3482.
44. Stephen, O.; Sain, M.; Maduh, U.J.; Jeong, D.U. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of healthcare engineering* **2019**, *2019*.
45. Rajaraman, S.; Candemir, S.; Kim, I.; Thoma, G.; Antani, S. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Applied Sciences* **2018**, *8*, 1715.
46. Mittal, A.; Kumar, D.; Mittal, M.; Saba, T.; Abunadi, I.; Rehman, A.; Roy, S. Detecting Pneumonia Using Convolutions and Dynamic Capsule Routing for Chest X-ray Images. *Sensors* **2020**, *20*, 1068.