A Biologically Inspired Attention-guided Machine Learning Approach for Pneumonia Diagnosis

PHUC LE, Cypress Woods High School, USA

Pneumonia is a common respiratory infection characterized by the inflammation of air sacs in the lungs. A common method for diagnosis of pneumonia is through chest X-rays, which can be used by computer vision systems to create models for machine learning-based diagnosis. In this paper, we propose the implementation of attention mechanisms, which learns to "guide" the model to focus on specific regions of interest, into a traditional convolutional neural network (CNN) to improve classification performance and interpretability of model results. We evaluate and compare our proposed model against a baseline CNN and find that the implementation of attention mechanisms can lead to significant improvements, with our model achieving an accuracy of 95.48% and precision 0.87 compared to the baseline's accuracy of 90.27% and precision of 0.74. We also compared Grad-CAM heatmaps generated by both models to find that our proposed approach mostly succeeds in ensuring that the model focuses on the lungs, which is the main region of interest for pneumonia diagnosis.

1 INTRODUCTION

Pneumonia is described as a mild to severe infection of the lungs, usually characterized by inflamed air sacs filled with fluid or pus. If left unchecked, pneumonia can be life-threatening and require extensive medical treatment. It is also a very common disease, affecting over three million American adults and causing the death of over 50,000 every year [2]. A large reason for this high number is due to unclear symptoms. In the early stages, the symptoms of pneumonia can appear mild, resembling that of a common cold or flu. This is further compounded by the fact that lab tests or imaging are generally required to give a proper medical diagnosis of pneumonia.

Currently, chest X-rays are one of the best readily available methods for the diagnosis of pneumonia, with the World Health Organization stating that a readable X-ray is crucial for the diagnosis of radiological pneumonia [1]. However, due to the textural nuance and noise contained within chest X-ray imaging, detecting pneumonia through chest X-rays is a challenging task for radiologists, even for those who have had extensive training and professional guidance [22]. Hence, it may be critical to develop computational support systems for radiologists to better diagnose pneumonia and other thoracic diseases to ensure that patients receive timely treatment.

Recent advances in machine learning algorithms have allowed for their gradual implementation into the medical field. Of these applications, there exist methods for prediction, diagnosis, segmentation, natural language processing, and more [25]. By leveraging the large-scale knowledge of patterns and rules inherent in a machine learning algorithm, they have been a useful computational aid for medical professionals and researchers alike.

However, a large obstacle for the use of machine learning in the medical domain, specifically the use of deep learning in image classification, is the lack of a guarantee that models will classify based on features that are actually relevant. When looking at a large database of chest X-rays for the purpose of pneumonia detection, a machine learning model may instead focus on a graphical defect in the spine that is specific to a single dataset for its classification, thereby failing to produce generalizable results. To address this problem, we propose the use of attention mechanisms [12] as an augmentation tool for the task of chest X-ray interpretation. Through the use of attention mechanisms, we can ensure that our machine learning model places particular emphasis on the region of interest—the lungs. For classification, we propose the use of transfer learning for two-dimensional convolutional neural networks (CNN). With this, we believe

Author's address: Phuc Le, Cypress Woods High School, 13550 Woods-Spillane Blvd, Cypress, Texas, USA, 77429, JackLe61333@gmail.com.

we can improve not only the classification ability of our model but also the interpretability of the output to allow it to be used as a computational support for radiologists.

In this paper, we have successfully implemented the Convolutional Block Attention Module [31] as an augmentation tool for CNN-based pneumonia detection from chest X-rays. Through validation, we have found improvements in both classification ability and final interpretability of the results, indicating that the use of a biologically-inspired attention mechanism can be useful for this task. No longer does the model classify the chest X-ray scans based on confounding features; it now classifies them by looking specifically at regions of interest, which in our case is the lungs.

This paper will be organized as follows. In Section 2, we will discuss existing work regarding the use of machine learning in medical domains, with a specific emphasis on pneumonia diagnosis using chest X-rays. In Section 3, we will describe the methodology of our study, focusing on the network architecture. In Section 4, we will discuss the materials and methods of our experiments, such as the dataset used, hyperparameter optimization, and the experimental procedure. In Section 5, we will discuss the results of our experiment. We will conclude the paper with Section 6, where we will provide a brief overview of the paper and propose future areas of study in this subject. Section 6 will also contain information regarding the availability of the source code for this study.

2 RELATED WORK

2.1 Deep Learning on Medical Imaging

The analysis of medical images via deep learning has been extensively studied in recent years [15]. These studies have covered numerous aspects and modalities of medical imaging. Among these studies, there were two that we found particular interest in. Folego *et al.* utilized three-dimensional convolutional neural networks (CNN) for biomarker identification in magnetic resonance imaging (MRI) to help with the diagnosis of Alzheimer's Disease [8]. Kuo *et al.* implemented a deep residual network (ResNet) to predict the onset of chronic kidney disease from ultrasound imaging [17].

2.2 Deep Learning on Chest X-rays

There has been previous work on the application of deep learning on chest X-rays, with many in the past few years. Rajpurkar *et al.* created CheXNet, a 121-layer CNN that was able to outperform four practicing radiologists in the task of pneumonia detection [22]. Cohen *et al.* trained a densely-connected CNN to predict the severity of COVID-19 pneumonia for use in monitoring treatment efficacy [7]. Ayan *et al.* evaluated two well-known CNN models for the diagnosis of pneumonia from chest X-ray images and found that each individual model has its own specific capabilities that allow it to perform well on specific tasks within the problem [5].

2.3 Attention Mechanisms

There has been a lot of research on attention mechanisms in the past few years. In 2018, Woo *et al.* presented the Convolutional Block Attention Module (CBAM), a light and general attention module that can be implemented into a CNN architecture. The module adaptively refines features by inferring attention maps along a channel and spatial dimension, then multiplying it to the input feature map to produce the final feature output. Their work has shown consistent improvements across multiple tasks and has demonstrated wide applicability [31]. In 2020, Wang *et al.* introduced the Efficient Channel Attention (ECA) module with the hopes of overcoming the performance and model complexity trade-off. Their results have shown that despite having few parameters, their module is able to generate clear performance gains over traditional methods [29].

2.4 Attention Mechanisms in Medical Imaging

There have also been previous work in the use of attention mechanisms for medical imaging. Sinha *et al.* utilize self-guided attention mechanisms to overcome two highlighted limitations associated with standard CNN models: the lack of efficient modeling for long-range feature dependencies, and redundant use of information during multi-scale approaches such as encoder-decoder networks. They evaluate their proposed model on the task of semantic segmentation using three different medical imaging datasets, finding that their approach increases the accuracy of predictions while also reducing standard deviation. Their research demonstrates the efficiency of self-guided attention in CNNs in generating segmentations of medical images [28]. More close to our work, Guan *et al.* proposed a multi-branch attention-guided CNN that they found is able to learn from disease-specific regions of interest to reduce noise and improve alignment. They conducted evaluations on the ChestX-ray14 dataset and found meaningful improvements over state-of-the-art CNN architectures [9]. An *et al.* introduced the visual attention mechanism into a deep learning model for the use of medical image classification. Through experiments, they found that the implementation of visual attention was able to increase both the accuracy and the interpretability of the model. They evaluated their method using a lung nodule and a breast cancer dataset, and again found improved accuracy with good stability and robustness [4].

2.5 Discussion

Through our literature review, we have found that a majority of existing work on pneumonia detection from chest X-ray scans utilize the full scan, including areas of potential noise or bias. We believe this to be an issue in the generalizability of existing work, where models will learn based on graphical noise caused by confounding features such as differences in images created by different X-ray scanners. This paper attempts to rectify this issue by introducing a biologically-inspired attention mechanism for the augmentation of CNNs in training.

3 METHODOLOGY

In this section, we will describe our proposed machine learning model and provide background information to give a more intuitive understanding of our approach. We will then discuss our baseline architecture and the construction of our final network architecture.

3.1 Convolutional Neural Networks

A convolutional neural network is a class of neural networks that has seen widespread use in computer vision tasks. A main advantage of CNNs over previous methods is that they can take into account spatial information in its training process, which can be essential in differentiating one image from another. Current state-of-the-art CNNs, such as DenseNet or ResNet, are extremely complex and highly optimized, with numerous different types of layers that allow them to function. However, for the purpose of this paper, we will only be providing surface level and intuitive explanation of how basic CNNs function.

A basic CNN is typically built using three main types of layers: convolution layers, pooling layers, and fully connected layers. As seen in Figure 1, a combination of multiple convolutional and pooling layers typically makes up the hidden layers within a basic CNN. These hidden layers are where the main feature engineering of the model is done.

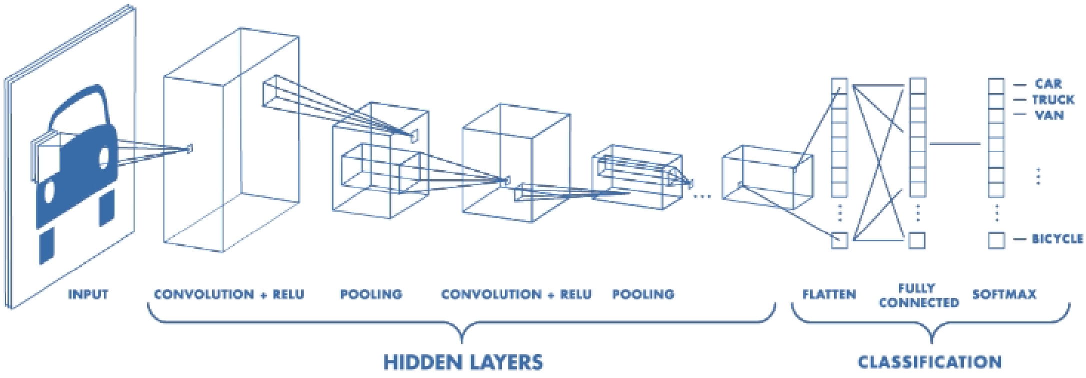


Fig. 1. The architecture of a basic CNN. Figure from [21].

The convolution layer is the main computational section of the CNN. It generates a new, more in-depth representation of the input image by performing a dot product between a matrix of learnable parameters (known as a kernel) and every “chunk” within an image. It does so by sliding the kernel matrix across the image in chunks and producing a mathematical representation of that chunk. Following this process, we would result in a smaller activation map that captures both the information contained in the pixel of the image as well as the spatial relationships between each pixel.

Following the convolution layer is the pooling layer. Pooling is a form of dimensionality reduction that is extremely useful for CNNs. Pooling works by separating the input slice into “neighborhoods”, where values close to each other are grouped together. There are two main types of pooling that are used: max pooling and average pooling. During max pooling, each neighborhood would be represented by the maximum value within that neighborhood, as seen in Figure 2. During average pooling, each neighborhood would be represented by the average of all values within that neighborhood. Pooling has two important uses in CNNs: first, it reduces the dimensions of the data as it goes further into the model, allowing for more complex networks to be run on a larger variety of machines. Second, it introduces translational invariance, allowing for objects to be recognized regardless of where they appear in the image.

Finally, the fully connected layers, as their name suggests, has full connectivity between every neuron within the layers. This enables them to map the representation between the input image and the expected output, thereby allowing for the model to create a prediction output based on its data input.

3.2 Baseline

To properly evaluate the improvement in performance and interpretability of the attention-guided pneumonia detection model, we will set up a baseline using a deep convolutional neural network that classifies on the full chest X-ray scan. To ensure a valid comparison, the baseline network will employ the exact same training and testing splits, image preprocessing methods, and random seeds. For the model architecture, we have chosen to replicate one of the top-performing submissions [19] for our particular dataset on Kaggle, an online machine learning platform that hosts competitions and stores datasets for open-access. The model uses a series of convolutional blocks, each containing

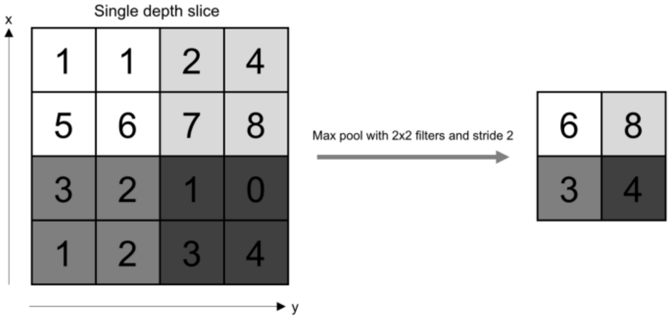


Fig. 2. A visual explanation of max-pooling. Figure from [20].

a convolutional layer, a batch normalization layer, and a max-pooling layer. On alternating convolutional blocks, a dropout layer is added to reduce overfitting. Following the convolutional blocks, a global average pooling layer is used, followed by fully-connected layers for classification. The full model architecture can be found on [19].

3.3 Proposed Attention-CNN Model

We build upon the baseline model mentioned in the previous subsection to construct our attention-guided CNN model. Attention mechanisms are a relatively new tool used to improve the performance of traditional convolutional neural networks. They can be considered biologically inspired since their functionality is similar to the perception system of human and animal vision, where certain objects are "in-focus" while others, such as those in the peripheral vision, are blurred and "out-of-focus". As said before, attention mechanisms are generally used to make CNNs place more emphasis on important information in the learning process , rather than learning non-useful or confounding background information. For the construction of the attention modules themselves, we have utilized the Convolutional Block Attention Module (CBAM) from [31].

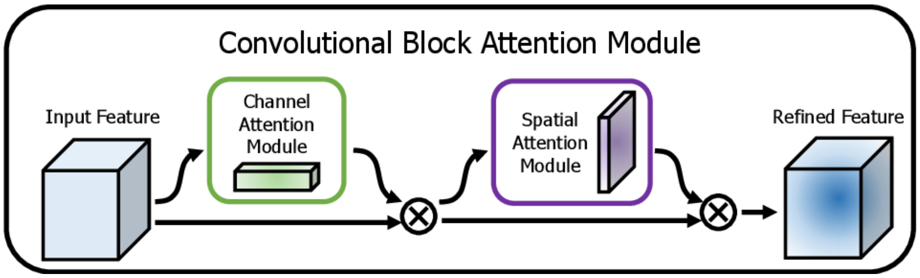


Fig. 3. An overview of the Convolutional Block Attention Module structure. Figure from [31].

CBAM works by applying attention to both the channel and spatial axes of an image (i.e., a color image normally has three channels; all images contain spatial information, which can be considered the spatial dimension). Through this, the modules can learn and inform the model on "what" parts of the image and "where" in the image it should be paying attention to. The computations on these two axes work sequentially. A more in-depth description of how CBAM works can be found in [31]. In this paper, we will give a short summary of their explanation for what each section of CBAM does. This process is defined as follows:

Given a feature map F, which represents the output following a previous convolution layer, CBAM generates a channel attention map M_c and a spatial attention map M_s.

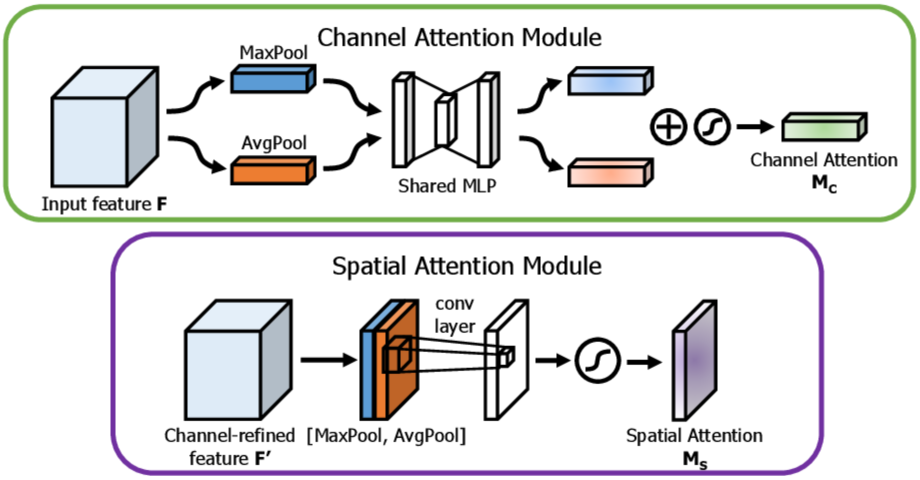


Fig. 4. A diagram modeling each of the two sub-modules in CBAM. Figure from [31].

The channel attention map is inferred by first gathering the results of average pooling F^c_{avg} and max-pooling F^c_{max} on the same feature map. A shared multi-layer perceptron network is then applied to each of these features. Finally, the results are aggregated, resulting in M_c.

The spatial attention map is inferred by utilizing the inter-spatial relationship of features. This is done by applying a convolution layer onto the feature map F, which creates a spatial attention map M_s that tells the model which areas to suppress and which to emphasize.

The final feature map is then calculated as follows:

$$F' = M_c(F) \oplus F$$

$$F'' = M_s(F') \oplus F'$$

where $\oplus$ indicates element-wise multiplication and F'' represents the final feature output of the CBAM module [31].

For our model architecture, we have placed a CBAM module following every convolution block of the baseline model. Although it is possible to run the two CBAM sub-modules in parallel, we have opted to layer then sequentially due to the suggestions and test results of [31].

4 EXPERIMENTS

In this section, we will describe the dataset used. Then, we will discuss our preprocessing and image augmentation techniques, as well as our reasoning for them. Next, we will explain our selection process for the hyperparameters of the model. Finally, we describe our experimental setup, training procedure, and evaluation procedure.

4.1 Data

4.1.1 Chest X-ray Datasets. There exists a large amount of chest X-ray datasets available for use in research and development, with some of the largest and most commonly used being ChestX-ray14, containing over 100,000 images [30], MIMIC-CXR, containing over 300,000 images [13], and CheXpert, containing over 200,000 images [11]. There is also a smaller dataset collected collected from Guangzhou Women and Children's Medical Center, which contains 5,856 chest X-rays separated into two categories: normal and pneumonia [14]. For our paper, we will be using the smaller dataset from [14] due to storage and computational limitations. We believe this will be enough to assess whether the implementation of attention mechanisms into a CNN-based detection of pneumonia can improve the interpretability of results. However, we hope to explore and use these larger datasets in future work.

4.1.2 Description of Dataset. This study uses the chest X-ray dataset presented in [14]. The dataset is a collection of 5,856 anterior-posterior chest X-ray images selected from a retrospective cohort of pediatric patients ranging from one to five years old from the Guangzhou Women and Children's Medical Center. The dataset is HIPAA-compliant, adherent to the tenets of the Declaration of Helsinki, and approved by an Institutional Review Board/Ethics Committee [14]. The dataset is publicly available and distributed through Kaggle.

Initial processing has been conducted by the authors of the dataset. All chest X-ray images that were low-quality or unreadable have been removed. Each scan was then diagnosed by two expert physicians. A subset of the images was then checked by a third expert to account for possible diagnosis errors.

4.2 Data splitting

We have separated the dataset into a training set, validation set, and testing set. The training set will be used to train and tune the model. The validation set will be used for evaluation and further tuning during the training process. Following the completion of training, the final model will be evaluated against a testing set that has not been seen by the model in any of the prior stages.

Our data splits are as follows: 75% of the dataset, totaling 4,392 scans, are used for training. 5% of the dataset, totaling 292 scans, is used for validation. The remaining 20% of the dataset, totaling 1,172 scans, is used for testing. The number of patients in each class within each split is shown in Table 1.

Table 1. Description of the training, validation, and testing splits from the dataset.

Diagnosis	**Training**	**Validation**	**Testing**	**Total**
Normal	1201	66	316	1583
Pneumonia	3191	226	856	4273

4.3 Preprocessing

In computer vision, it is common to perform image augmentation prior to training in order to increase the generalizability and classification accuracy of the model. It can also help with the problem of overfitting, which is likely to occur given the complexity of medical images and the number of parameters that are within state-of-the-art machine learning models. In simple terms, overfitting occurs when, due to the high complexity of network architectures or some other factor, the model essentially memorizes the dataset given to it instead of actually learning patterns from the data.

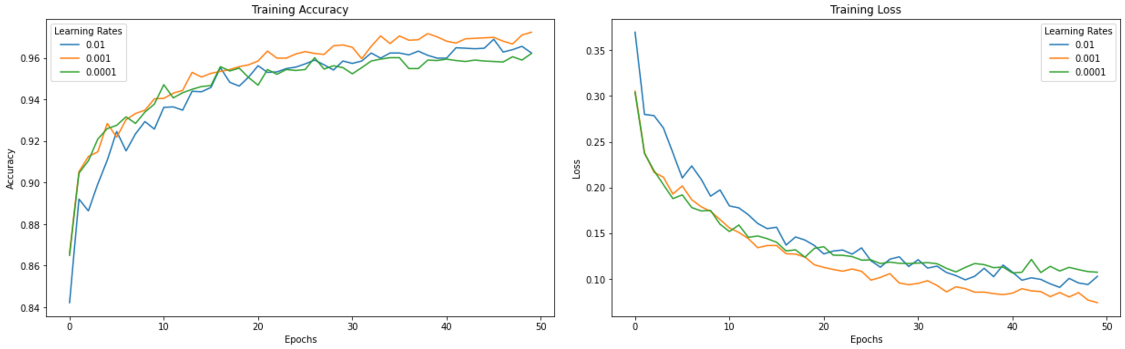
For our preprocessing, we have chosen a set of common augmentation steps that are computationally simple with proven performance improvements. We first normalize our images into a [0, 1] scale. Past works have shown that normalizing can improve the generalization and accelerate the training procedure of neural networks, while also helping with convergence [10]. We then ensure that our images are of the same dimensions by cropping and resizing them into a common size of 150 x 150 pixels. To further avoid the problem of overfitting, we apply various methods of image data augmentation in order to introduce variability as well as artificially expand the size of our dataset. These methods include: randomly rotating images, randomly zooming into images, randomly shifting images in one or more directions, and randomly flipping images in the horizontal direction. These have been chosen for their relative computational simplicity and reported performance improvements in previous works [27].

4.4 Hyperparameter Optimization

Hyperparameter tuning is an important process in ensuring that our model will perform well in our specific task. For our study, we will be tuning the learning rate of the optimizer and the batch size of the data given to the model during the training process.

4.4.1 Learning Rate. Our model contains a learning rate scheduler that will multiply the learning rate of the model by a factor of 0.8 after a plateau in validation accuracy is detected for 2 or more epochs. Therefore, the main parameter we will be tuning here is the starting learning rate, or the learning rate at the 0th epoch. To select the best learning rate for our model, we have selected and tested the learning rates used by previous studies on the use of CNNs on medical imaging. The learning rates we will be testing are 0.01 [9], 0.001 [22], and 0.0001 [6].

We will evaluate all learning rates using 50 epochs with a batch size of 16. The final evaluation will be done through graphs of the training accuracy and loss, as well as the trained model's accuracy on the testing set.



(a) Training accuracy across 50 epochs.

(b) Training loss across 50 epochs.

Fig. 5. Training curves comparing model performance for different learning rates.

Looking at the training curves from Figure 5, we see that a learning rate of 0.001, as used in [22], has the best performance during training. As seen in Figure 5a, we see that it consistently achieves the highest training accuracy throughout the epochs and results in the highest final training accuracy. As shown in Figure 5b, we also see that a learning rate of 0.001 results in quicker model convergence and a lower final training loss.

Table 2. Testing set accuracy of each learning rate.

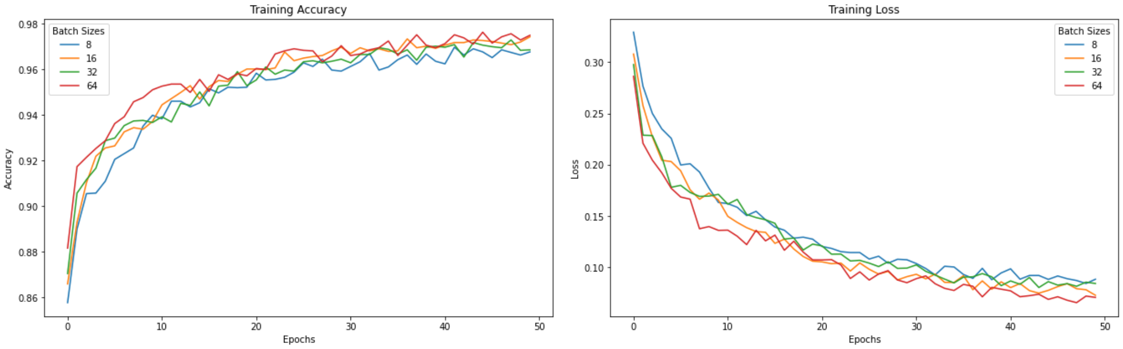
Learning Rate	Testing Set Accuracy
0.01	93.67%
0.001	96.25%
0.0001	93.17%

Our findings are further validated by Table 2, where we find that a learning rate of 0.001 significantly outperforms the other two learning rates when evaluating on the testing set.

4.4.2 Batch Size. Another important hyperparameter to optimize during the training process is the batch size, or the number of training samples seen by the model during one iteration, or step. With larger batch sizes, fewer steps will be done per epoch, and vice versa.

For our experiments, we have again selected batch sizes that have achieved high performance in previous studies. The batch sizes we will be testing are 8 [6], 16 [22], 32 [9], and 64 [9].

Following the learning rate experiments, we will be testing all batch sizes using a starting learning rate of 0.001. We will evaluate all batch sizes using 50 epochs, and compare graphs of the training accuracy and loss, as well as the trained model's accuracy on the testing set.



(a) Training accuracy across 50 epochs.

(b) Training loss across 50 epochs.

Fig. 6. Training curves comparing model performance for different batch sizes.

Looking at the training curves in Figure 6, we see that batch sizes of 16 and 64 perform similarly well, while batch sizes of 8 and 32 perform consistently worse. In Figure 6a, we see that a batch size of 64 seems to reach higher training accuracies quicker, yet results in a near-identical final training accuracy to a batch size of 16. This trend is also seen in Figure 6b, with the batch size of 64 converging quicker yet ending in a similar final training loss. Thus, we can conclude from these graphs that the best batch size for our particular situation will either be 16 or 64.

The results from Table 3 are more conclusive. We see here that a batch size of 16 has a significantly higher testing set accuracy than a batch size of 64. We believe the reason why a batch size of 64 achieved such high training set accuracy yet fail to repeat with the testing set accuracy is likely overfitting. One of the main trade-offs with larger batch sizes is that, while it may lead to computational speedups, it sacrifices generalization in the process [26]. Hence, to create a

Table 3. Testing set accuracy of each batch size.

Batch Size	Testing Set Accuracy
8	96.25%
16	96.84%
32	92.58%
64	93.77%

model that can perform well in both the training and testing sets, it may be beneficial to use smaller relative batch sizes. Thus, we believe that utilizing a batch size of 16 will be a safe balance between accuracy and generalizability.

4.5 Training Process

For the training of our proposed attention-guided model, we use a batch size of 16 and starting learning rate of 0.001. We trained the model for 50 epochs. The model will be trained using the Adam optimizer with standard parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) [16].

We have also implemented a plateau-based learning rate scheduler, where the learning rate will be reduced whenever it detects a plateau, or a period of low improvement in the validation accuracy, during the training process. For the parameters of the scheduler, we have utilized a patience of 2 (i.e., it will only reduce the learning rate when encountering plateaus lasting two or more epochs), and a reduction factor of 0.8. This means that every time it detects a valid plateau, it will multiply the current learning rate by 0.8.

For the baseline training, we did not change any significant parameters beyond what is presented in [19]. The baseline was trained using the RMSprop optimizer with standard parameters ($\rho = 0.9, momentum = 0.0$) and a starting learning rate of 0.001. The baseline also implemented a learning rate scheduler with a patience of 2 and a reduction factor of 0.3, up to a minimum learning rate of $1 * 10^{-6}$. The batch size used in the baseline is 32.

The authors of [19] only trained the baseline model for 12 epochs. However, to stay consistent with the rest of our training process, we will instead be training the model for 50 epochs. Furthermore, to allow for proper comparison between the two models, we will be using the same data splits used for our attention-guided model. This would ensure that there would no differences in data seen by the two models, thereby removing data splits as a possible confounding variable in the final evaluation process.

4.6 Evaluation Procedure

Following training, we will test both models using a separate testing set, which neither model has seen before. We will calculate the testing-set accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC).

To evaluate whether our proposed attention-guided model is improving the interpretability of diagnoses by “focusing” on specific regions of interest, we will also be generating Gradient-weighted Class Activation Mapping (Grad-CAM) [24] on a random set of scans. Grad-CAM will give us visual explanations of regions in the image considered important by the CNN. We will be comparing these Grad-CAMs between the baseline and our proposed approach to evaluate whether our implementation of attention mechanisms is actually working the way it is intended to.

4.7 Experimental Setup

For our experiments, we have utilized Python 3.8.11, running on Jupyter Notebooks in Manjaro Linux 21.1.1. For metrics, we used SciKit-Learn 1.0. We used TensorFlow 2.6.0 to create and train the model. In TensorFlow, both the central processing unit (CPU) and graphical processing unit (GPU) are used. To gather image batches during the training process, the CPU is used. For model convergence and training, a combination of both the CPU and GPU is used. For our setup, our CPU is an AMD Ryzen 7 3700X and our GPU is a single NVIDIA RTX 2070 Super.

5 RESULTS AND DISCUSSION

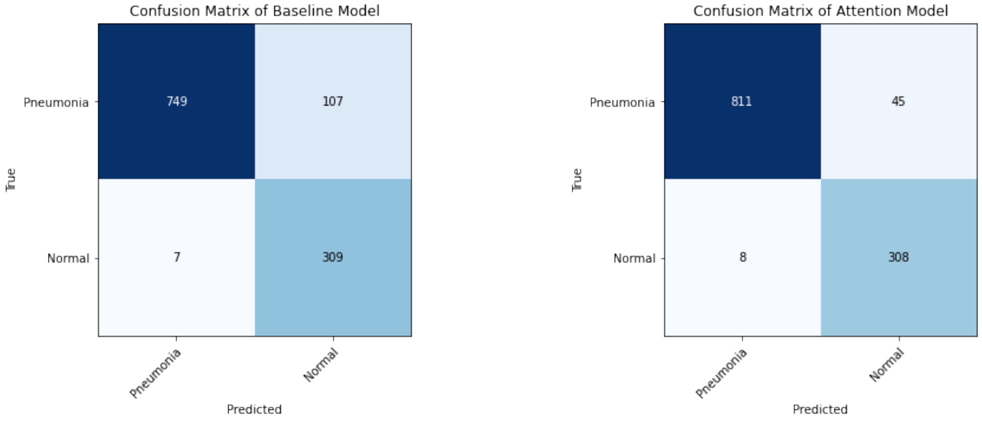
5.1 Classification Metrics

For our dataset, all images were graded and labeled by two expert physicians before being cleared for the dataset. We will be comparing these labels with the predictions by our model to calculate our metrics, which are the testing-set accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). These performance metrics are shown in Table 4.

Table 4. Classification metrics of model performance

Model	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
Baseline	90.27%	0.74	0.98	0.84	0.98	0.95
Attention-guided	95.48%	0.87	0.97	0.92	0.99	0.98

Comparing the baseline model and our proposed attention-guided model, we see that, although the accuracies are already remarkably high for both, the implementation of attention mechanisms into a traditional CNN can still bring significant performance improvements. The baseline model had an accuracy of 90.27%, whereas our attention-guided model had an accuracy of 95.48%.



(a) Confusion matrix of the Baseline model.

(b) Confusion matrix of the Attention-guided model.

Fig. 7. Confusion matrices of both models, evaluating on the testing set.

Both models show high recall with relatively low precision. Our approach has a higher precision of 0.87 compared to the baseline with 0.74. However, the baseline has a slightly higher recall than ours, at 0.98 compared to our 0.97. Figure 8 shows a sample of scans from patients affected by pneumonia that were incorrectly classified by the baseline model yet correctly classified by our proposed approach. There were 73 scans, or approximately 6.23% of the test set, where the two models differed in their prediction. Of these differences, 70 scans were of patients affected by pneumonia and 3 were scans of normal patients. From this, 67 of the 73 scans where the two models differed were correctly classified by our attention-guided model but incorrectly classified by the baseline model. Only 6 scans of 73 scans where the two models differed were correctly classified by the baseline model yet incorrectly classified by our attention-guided model. This difference in classification ability is the reason why we see the higher precision mentioned previously.

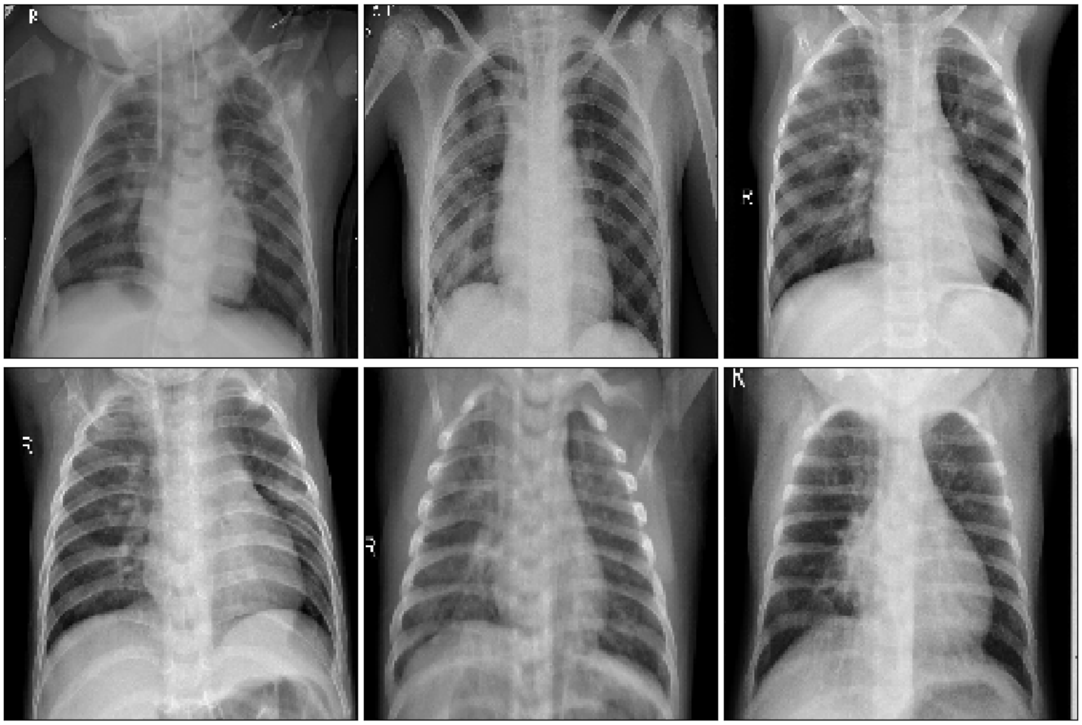


Fig. 8. Examples of pneumonia cases incorrectly classified by the baseline model but correct classified by our attention-guided model.

Further improvements are seen in the F1 score, area under the receiver operating characteristic (AUROC), and area under the precision-recall curve (AUPRC). While some differences in classification metrics are significant, others, such as the AUROC and the recall, are marginal. Thus, further testing using a larger and more comprehensive dataset such as CheXpert [11] may be necessary to confirm generalized performance improvements beyond the dataset used in this study.

5.2 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [24] is commonly used to visualize the "thinking" process of a CNN, thereby allowing users to better understand a model's prediction. It examines the gradient information flowing into the final convolutional layer in a given network to create a heatmap visualization that highlights areas of "focus". For example, when a CNN separates a picture of a dog from that of a cat, the Grad-CAM image may show "focus" on the facial features of each animal, since that is where the CNN must look to differentiate between the two. In our case, the Grad-CAM should show emphasis on the lungs, since that is where the main indicators of pneumonia can be found.

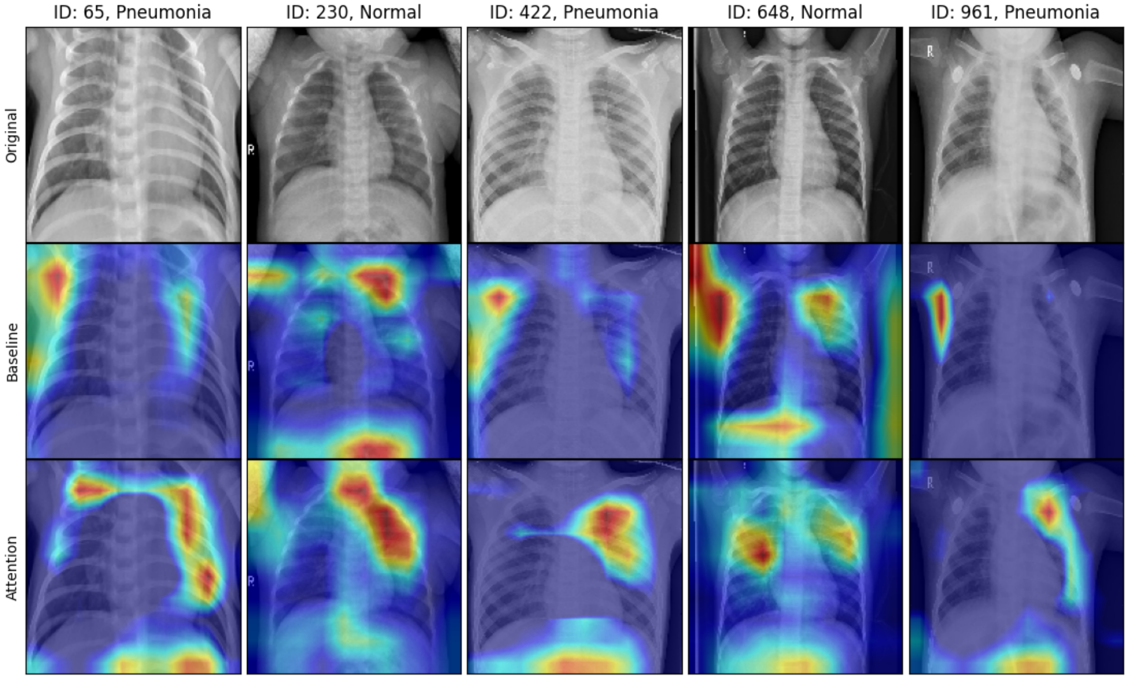


Fig. 9. Comparison images of the Grad-CAM of each model. The first row is the original image, the second row is the Grad-CAM from the baseline model, and the third row is the Grad-CAM image from our attention-guided model.

Looking at Figure 9, we clearly see that the implementation of attention into the CNN changes where the model looks at when making its prediction. The images in the second and third row show the Grad-CAM heatmap of the baseline model and our attention-guided model, respectively, superimposed onto the original image, where areas with warmer colors are considered "more important" by the model during its prediction procedure.

By comparing these images, we can see the aforementioned improved interpretability of our model. While the baseline model still places emphasis on some parts of the lung, it mostly focuses on other confounding features within the image. Looking at the images of patient ID. 65, 648, and 961, we see that the baseline model seems to focus on the axilla and upper arm instead of the lungs. In contrast, the attention-guided model correctly focuses on the lungs, which is our main region of interest. This is especially apparent in the image of patient ID. 65 and 961, where the activation mapping "wraps" around the heart and outer ribs, focusing only on the parts of the lungs that are shown and not anything else.

However, our attention-guided model does not always place emphasis on the correct parts. Looking at the images of patient ID. 230, we see that the baseline model incorrectly places large emphasis on the clavicle and diaphragm. With our model, while the emphasis on the diaphragm is gone and emphasis on the lungs is present, it seems to now focus on both the clavicle, upper sternum, and upper spine. This is also seen with patient ID. 422. Here, we see the attention model correctly focusing on the lungs. However, we also see that it places emphasis on the diaphragm, which the baseline model does not do.

In fact, this trend is seen in many of the scans we sampled. The attention-guided approach often correctly focuses on the lungs of a patient, yet also places emphasis on the diaphragm and lower regions of the image. While we believe this to still be an improvement over the baseline, which often focuses on non-informative or incorrect parts of the image, this is a finding that we believe warrants further research and confirmation.

5.3 Limitations

We identify and discuss various limitations with our study below.

First, previous works have shown that deep learning, such as deep convolutional neural networks, requires a very large amount of data to perform efficiently [3]. Due to our computational limits, we were only able to do our investigation using a smaller dataset. Although we were still able to achieve good results with just 5,856 images, we believe that our model will likely struggle when applied to more general and unfiltered images, such as those of a real-world clinical setting. To further investigate and confirm our findings, we hope to repeat this study in the future, when we have access to more computational resources, using a larger and more comprehensive dataset, such as ChestX-ray14 [30], MIMIC-CXR [13], or CheXpert [11].

Furthermore, our dataset, as well as most other chest X-ray datasets available, only includes frontal radiographs. It has been shown that a lateral radiograph is required for up to 15% of accurate disease diagnoses from chest X-rays [23]. Rajpurkar *et al.* has concluded that with this limitation, a setup of this nature only provides a conservative estimate of model performance, and we concur.

The dataset used in this study [14] contained images from retrospective cohorts of pediatric patients from the Guangzhou Women and Children's Medical Center, with all images being taken as part of a patient's routine clinical care. Therefore, due to patient privacy, the publicly available dataset did not contain patient history or results of other pathological studies, which may allow for more meaningful insights into a patient's condition beyond chest radiographs. As such, the scope of our study was limited to binary image classification between normal patients and those affected by pneumonia. With access to pathologies and patient history, the model could be further refined and improved, allowing for deployment in real-world use. Therefore, we believe our early research to be a valuable step towards creating a computational support for radiologists in clinical settings, such as a pre-screening tool for chest radiographs.

Finally, the images in our dataset were manually graded and labeled by two expert physicians before being confirmed by a third expert. Due to radiological interpretation errors during this labeling process, it is likely that there exist incorrect labels, which may lead to decreases in model performance in more general scenarios. We believe this limitation can be rectified using a larger dataset, where errors may not have as much weight.

6 CONCLUSION

In this paper, we have presented an attention-guided convolutional neural network for the task of pneumonia diagnosis from chest X-rays. We have shown that the implementation of attention mechanisms into a convolutional neural network for the task of medical image diagnosis can lead to significant performance improvements, with our model achieving an accuracy of 95.48% compared to the baseline model's 90.27%. Our attention-guided model has also shown higher precision than the baseline, indicating its possibility for use as a pre-screening tool and/or computational support for pneumonia diagnosis in clinical settings. We have also investigated improvements in the interpretability of our model by comparing the Grad-CAM heatmaps of both models confirming that attention mechanisms indeed work in ensuring the model places emphasis on correct regions of interest. Various limitations with our study have been identified, with most dealing with the limitations in our dataset.

For future works, we would like to address these limitations by using a larger and more comprehensive dataset of chest X-ray images to further investigate and confirm our findings. We would also like to investigate the use of deep transfer learning to overcome the limited availability of data. One of the current state-of-the-art approaches to chest X-ray classification is CheXnet [22], which implements a 121-layer Dense Convolutional Network (DenseNet). We believe it may interesting to investigate whether the implementation of attention mechanisms into this more complex model can still lead to further improvements in classification ability.

Finally, for the transparency and reproducibility of our academic work, we have made the source code of this paper publicly available at [18].

ACKNOWLEDGMENTS

This research was performed as part of the Summer 2021 term at Pioneer Academics, a fully accredited online research program for high school students (https://pioneeracademics.com/). We would like to thank Professor Avani Wildani of Emory University for her invaluable help and guidance through this process.

REFERENCES

[1] Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children, Jan 2001.

[2] Pneumonia can be prevented-vaccines can help, Oct 2020.

[3] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, Mar 2021.

[4] F. An, X. Li, and X. Ma. Medical image classification algorithm based on visual attention mechanism-mcnn. *Oxidative Medicine and Cellular Longevity*, 2021:6280690, Feb 2021.

[5] E. Ayan and H. M. Ünver. Diagnosis of pneumonia from chest x-ray images using deep learning. In *2019 Scientific Meeting on Electrical-Electronics Biomedical Engineering and Computer Science (EBBT)*, pages 1–5, 2019.

[6] G. Chada. Machine learning models for abnormality detection in musculoskeletal radiographs. *Reports*, 2(4), 2019.

[7] J. P. Cohen, L. Dao, P. Morrison, K. Roth, Y. Bengio, B. Shen, A. Abbasi, M. Hoshmand-Kochi, M. Ghassemi, H. Li, and T. Q. Duong. Predicting covid-19 pneumonia severity on chest x-ray with deep learning, 2020.

[8] G. Folego, M. Weiler, R. F. Casseb, R. Pires, and A. Rocha. Alzheimer's disease detection through whole-brain 3d-cnn mri. *Frontiers in Bioengineering and Biotechnology*, 8:1193, 2020.

[9] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification, 2018.

[10] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao. Normalization techniques in training dnns: Methodology, analysis and application, 2020.

[11] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. 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, 2019.

[12] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr. Learn to pay attention, 2018.

[13] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, Dec 2019.

[14] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, Feb 2018.

[15] M. Kim, J. Yun, Y. Cho, K. Shin, R. Jang, H.-J. Bae, and N. Kim. Deep learning in medical imaging. *Neurospine*, 16(4):657–668, Dec 2019. 31905454[pmid].

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[17] C.-C. Kuo, C.-M. Chang, K.-T. Liu, W.-K. Lin, H.-Y. Chiang, C.-W. Chung, M.-R. Ho, P.-R. Sun, R.-L. Yang, and K.-T. Chen. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. *npj Digital Medicine*, 2(1):29, Apr 2019.

[18] P. Le. A biologically inspired attention-guided machine learning approach for pneumonia diagnosis, 2021. Available at https://github.com/Jakl3/Pneumonia-Diagnosis-with-Attention.

[19] M. Mathur. Pneumonia detection using cnn, Mar 2020.

[20] M. Mishra. Convolutional neural networks, explained, Aug 2020.

[21] S. Patel and J. Pingel. Introduction to deep learning: What are convolutional neural networks?, Mar 2017.

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

[23] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugulpati, and E. C. Rosenow. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, Feb 2012.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019.

[25] K. Shailaja, B. Seetharamulu, and M. A. Jabbar. Machine learning in healthcare: A review. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 910–914, 2018.

[26] K. Shen. Effect of batch size on training dynamics, Jun 2018.

[27] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, Jul 2019.

[28] A. Sinha and J. Dolz. Multi-scale self-guided attention for medical image segmentation, 2020.

[29] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. Eca-net: Efficient channel attention for deep convolutional neural networks, 2020.

[30] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module, 2018.