AI6102- Machine Learning Applications and Methodologies

Stacking Ensemble Approach to Pediatric Pneumonia Diagnosis

Research Paper

Chithra Ramesh Asswin
G2302832A
ASSWIN001@e.ntu.edu.sg

Aradhya Dhruv
G2303518F
AR0001UV@e.ntu.edu.sg

Maheswaran Rohin Kumar
G2303513K
rohinkum001@e.ntu.edu.sg

Abstract

Pediatric pneumonia poses a significant health risk as it involves the accumulation of fluid in the air sacs of the lungs. Although, X-rays do help to some extent, the challenge lies in accurately identifying the condition in children due to their sensitivity to X-ray radiation which leads to diagnostic errors.To overcome this obstacle, there has been a growing adoption of deep learning models, particularly Convolutional Neural Networks (CNNs), for computer-aided diagnosis using chest X-ray images.

***We propose a novel integration** of an Efficient Channel Attention (ECA) module along with pre-trained ResNet50 and DenseNet121, VGG19 models. Additionally, we present a **stacking ensemble** method based on the performance of our proposed model. Our approach demonstrates remarkable results, achieving an accuracy of 95.29% and an AUC curve of 92.39% on the pediatric pneumonia dataset.*

In summary, our proposed architecture exhibits great potential in facilitating real-time diagnosis of pediatric pneumonia, with the prospect of enhancing patient outcomes.

1. Introduction

There has been growing concern about the increase in diseases in recent years. This surge in illnesses has resulted in mortality rates and long term economic losses in various countries [1]. Pneumonia is one disease that causes respiratory infections disrupting the normal functioning of the human body. Viruses and bacteria are known to be some of the pathogens for causing pneumonia. One significant factor contributing to the transmission of these viruses and bacteria is the degradation of air quality [2]. Within our lungs there are sacs called alveoli, which play a crucial role in exchanging essential gases like oxygen and carbon dioxide. When someone contracts pneumonia these sacs become filled with pus and fluid impairing the exchange of gases between the blood and lungs. Symptoms of pneumonia include difficulty breathing, well as complications such, as chest pain, coughing, vomiting, diarrhea and fatigue. Various diagnostic methods, including pulse oximetry, Complete Blood Count (CBC), and sputum tests, are employed for pediatric pneumonia diagnosis. However, these indicators can be nonspecific, and chest X-rays, while affordable, suffer from manual examination challenges, leading to human errors. This necessitates a precise Computer-Aided Diagnostic (CAD) model for prompt and accurate predictions.

Our **proposed novel solution** for pediatric pneumonia diagnosis employs an efficient channel attention module with a stacking ensemble, offering simplicity and applicability Fig. 2.

2. Litrature Review

In our exploration for medical image diagnosis, several noteworthy studies have contributed valuable insights, each with its unique strengths and limitations. Transfer learning methodologies have shown promise in leveraging pre-trained models for medical image analysis. Liang and Zheng [3] addressed the challenge of preserving spatial information in the context of increasingly complex convolutional neural networks (CNNs). Their approach incorporated residual connections within the transfer learning framework, effectively mitigating the risk of losing fine-grained details crucial for accurate diagnosis.

Another significant contribution comes from Kanakaprabha and Radha [4], who employed CNNs for the differentiation of COVID-19 and pneumonia subtypes in chest X-rays. Their work exemplifies the potential of CNNs in enabling rapid and accurate diagnosis, particularly in the context of distinguishing between related medical conditions.

Traditional CNNs, while powerful, exhibit limitations in selectively attending to the most informative regions of input, impacting performance when specific parts of the input carry more relevance. Addressing this, attention-based CNN models have emerged as a solution by allowing models to focus on crucial regions while ignoring less informative ones. Hu et al. [5] introduced Squeeze-and-Excitation Networks (SENet), a notable example of such models. SENet performs channel-wise feature recalibration by incorporating global information, enhancing the model's ability to focus on salient features.

2.1. Limitations of Transfer Learning Approaches

While transfer learning proves advantageous, it is not without its challenges:

- **Loss of Spatial Information:** As CNN complexity increases, there's a risk of losing spatial information crucial for fine-grained details, potentially impacting diagnostic accuracy.
- **Sensitivity to Hyperparameters:** The performance of transfer learning models can be sensitive to hyperparameter choices, such as learning rates and training epochs, making model tuning challenging.
- **Lack of Generalizability:** Transfer learning models may lack generalizability to diverse datasets, limiting their applicability to real-world scenarios with varying image characteristics.

2.2. Limitations of Attention-based CNN Models

Attention-based CNN models, while offering improved focus, come with their set of limitations:

- **Computationally Expensive:** The attention mechanisms used in these models can be computationally expensive, posing challenges, especially in large-scale medical image analysis and hindering deployment in real-world clinical settings.
- **Sensitivity to Attention Mechanisms:** The performance of attention-based CNN models can be sensitive to the choice of attention mechanisms, with varying strengths and weaknesses that might depend on the specific task and dataset.

This comprehensive review lays the groundwork for our research, guiding our approach to overcome these limitations and contribute to the advancement of pediatric pneumonia diagnosis using deep learning techniques.

3. Dataset Description

The experiments, in this study utilized the Kermany et al. [6] dataset obtained from the Guangzhou Women and Childrens Medical Center. The dataset consisted of 5856 chest X rays taken from children aged 1 to 5 which were categorized into two classes; Normal (1583 X rays) and Pneumonia (4273 X rays). Tab. 1 provides an overview of the distribution. To address the imbalance in the dataset we created a dataset by applying various geometrical transformations. These transformations included a rotation range of 20 degrees, a zoom range between 0.8 and 1.2 height and width shifts of 0.2 and flipping. Randomly selected augmented images, from the training set were employed for validation purposes. Tab. 2 displays the distribution of data in this dataset. Fig. 1 gives an visulalization distribution of Normal and Pneumonia classes in Training Dataset.

Class	Train	Test
Normal	1349	234
Pneumonia	3883	390
Total	5232	624

Table 1. Distribution of original Kermany dataset

Class	Train	Test	Validation
Normal	3884	234	970
Pneumonia	3883	390	970
Total	7767	624	1940

Table 2. Distribution of augmented Kermany dataset

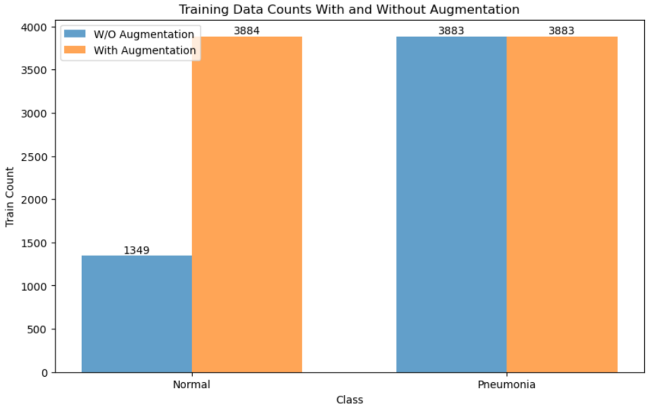


Figure 1. Class distribution before and after data augmentation

4. Methodology

In this study,**we propose a novel integration of an Efficient Channel Attention (ECA) module** along with pre-trained ResNet50 and DenseNet121, and VGG19 models for pediatric pneumonia diagnosis. Fig. 2

4.1. Efficient Channel Attention

Working Principle

The ECA module is a **lightweight attention mechanism** that can be plugged into any CNN architecture to improve

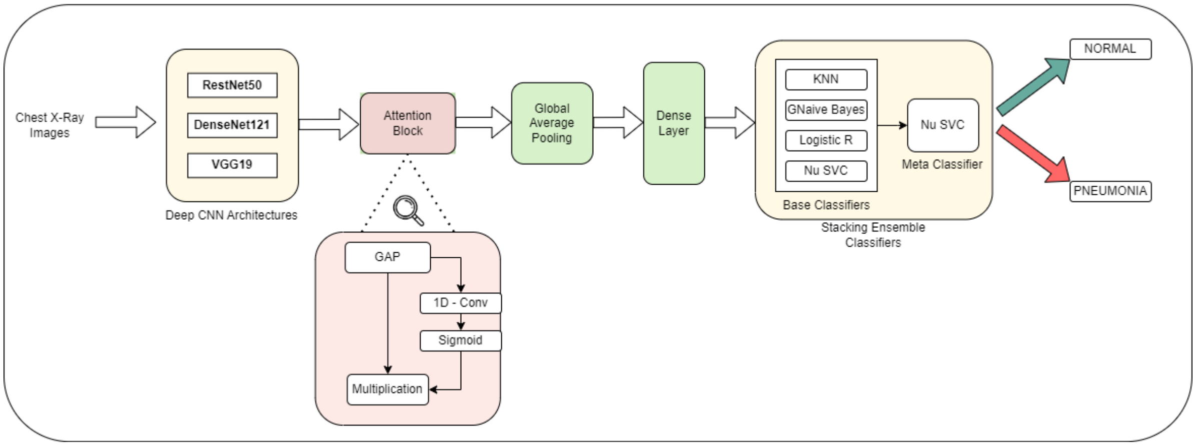


Figure 2. Proposed architecture pipleline diagram

its performance. It works by adaptively learning the importance of each channel in the feature maps of a CNN [7]. This is done by applying a 1D convolution to each feature map, where the kernel size of the convolution is determined by the number of channels in the feature map. (Please refer Fig. 3 for details)

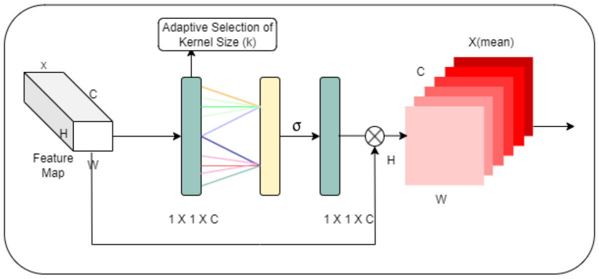


Figure 3. Efficient Net Architecture

In our proposed architecture from Fig. 3, the ECA attention block extracts the features from ResNet50, DenseNet121, and VGG19. The resultant **attention-aware** features are fed into a **global average pooling layer**. Between the final dense layers, dropout layers have been added to avoid overfitting of the model. Finally, the individual deep CNN architecture predictions will be sent to the stacking ensemble for final classification.

4.2. Stacking Ensemble

Ensemble learning is a highly effective technique for improving the performance and reliability of machine learning models. Among various strategies, stacking stands out as an especially powerful approach. Unlike other ensemble methods like bagging and boosting that rely on using the same algorithms, stacking combines diverse machine learning algorithms. Stacking operates through a two tiered structure; the **base classifiers** comprises multiple models that are trained simultaneously and their predictions serve as input features for the second layer. In turn, the second layer, known as the **meta classifier,** utilizes these inputs to make the ultimate prediction by harnessing the strengths of each individual model in a cohesive manner.

4.3. Training Steps:

In addition to the ECA module, we also propose a stacking ensemble method to further improve the performance of our model. Stacking ensemble involves combining the predictions of **multiple machine learning models** to produce a more accurate prediction. In our case, we combine the predictions of the 3 models specified earlier with the ECA module.

The stacking ensemble method works as follows:

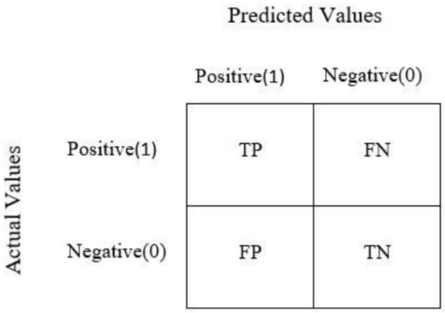
- First, we train each of the base models (ResNet50, DenseNet121, VGG19) on the training data
- For each data point in the test set, obtain the predictions from each of the base models
- Train a meta-learner (another machine learning model) to predict the true label (normal or pneumonia) based on the predictions of the base classifiers
- Use the meta-learner to predict the labels for the entire test set

By combining the predictions of the base models using the meta-learner, we can achieve a more accurate prediction than any of the base models alone. This is because the meta-learner has access to the predictions of all three models, which allows it to capture more complex patterns in the data.

We evaluated our proposed model on a public dataset of pediatric pneumonia chest X-rays. **Our model achieved an accuracy of 95.28% and an AUC curve of 92.39%**

5. Performance Metrics

In this analysis we evaluate the performance of classification algorithms by considering indicators such as accuracy, precision, recall, F1 score and the area under the receiver operating characteristic (ROC) curve known as the AUC value. The models predictions are visually presented in a confusion matrix (Fig. 4) where the horizontal axis represents predicted classes and the vertical axis represents actual classes.



Actual Values	Predicted Values: Positive(1)	Predicted Values: Negative(0)
Positive(1)	TP	FN
Negative(0)	FP	TN

Figure 4. Confusion Matrix

The accuracy of the model is determined by calculating the proportion of predictions, out of all predictions made (equation 1). Precision (equation 2) measures how positive labels are predicted by considering true positives and false positives. Recall (equation 3) assesses the models ability to identify all instances of a class by comparing true positives to the sum of true positives and false negatives.The F1 score as described in equation 4 represents a combination of precision and recall. It provides a comprehensive view of the models reliability by considering both aspects and weighing them accordingly. Additionally the AUC score measures the models ability to distinguish between classes without relying on a specific threshold. **For instance in this scenario it assesses how well the model can differentiate children with pneumonia from those without it**. If the AUC score is 1 it means that the model has discriminatory capabilities.

These metrics together form a framework for evaluating the performance of the model. They go beyond accuracy measurements and allow for a thorough assessment especially when dealing with datasets that have imbalanced class distributions. This knowledge is particularly crucial, in diagnostic tools where **false negatives** carry significant consequences and costs.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$F1\text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

6. Results & Discussion

In this study a novel architecture is proposed that combines ResNet50 Attention, DenseNet121 Attention and VGG19 Attention mechanisms. These combined mechanisms show performance improvements compared to established CNN models. The evaluation was conducted on the recognized Kermany dataset. The results are summarized in Tab. 3.

Table 3. Performance of Baseline Deep CNN Architectures(%)

Model	Accuracy	Precision	Recall	F1	AUC
Resnet50	**88.62**	**84.90**	**99.49**	**91.62**	**85.00**
Resnet101	84.78	80.41	100.0	89.14	79.70
Resnet152	86.86	82.63	100.0	90.49	82.48
Densenet121	**89.42**	**85.68**	**99.74**	**92.18**	**85.98**
Densenet201	80.76	76.47	100.0	86.66	74.35
VGG19	**91.63**	**90.00**	**99.23**	**94.39**	**90.43**
Xception	83.97	79.59	100.0	88.64	78.63

Tab. 3 presents a comparison of deep CNN architectures highlighting their performance across important metrics such, as accuracy, precision, recall, F1 score and AUC. It indicates that Resnet50, Densenet121, VGG 19 achieve accuracy levels. Specifically VGG19 stands out in recall and F1 score metrics showing its strength in both predicting results and consistently identifying positive classes. The AUC values further support the effectiveness of the proposed model by demonstrating its ability to differentiate between classes with error. Overall while the benchmarked architectures perform overall this new pipeline sets a standard, for accuracy and reliability within the context of the Kermany dataset.

The Proposed Pipeline, an ensemble of the 3 best performing attention-augmented models , exhibits superior performance with an accuracy of 95.28% and an AUC of 92.93% as observed from Tab. 4 , suggesting a robust capability in identifying relevant features for accurate classification. From Fig. 5, the loss curves for this pipeline indicate a consistent reduction in training loss and stable low validation loss, implying effective learning and **generalization** with minimal overfitting. The confusion matrix from Fig. 6 corroborates the model's efficacy, with a substantial number of true positives (384 for pneumonia) and true negatives

Table 4. Performance comparison between proposed attention model and baseline deep CNN model(%)

Model	Accuracy	Precision	Recall	F1	AUC
ResNet50	88.62	84.90	99.49	91.62	85.00
ResNet50-Attn	**94.07**	**94.46**	**96.15**	**95.30**	**93.38**
DenseNet121	89.42	85.68	99.74	92.18	85.98
DenseNet121-Attn	**94.07**	**94.68**	**95.90**	**95.29**	**93.46**
VGG19	83.97	79.59	100	88.64	78.63
VGG19-Attn	**94.55**	**93.41**	**98.21**	**95.75**	**93.33**
Proposed Model	**95.28**	**92.30**	**98.46**	**93.90**	**92.93**

(213 for normal), but also reveals room for improvement given the 21 false positives and 6 false negatives. These results collectively highlight the model's potential in medical diagnostic applications, while also emphasizing the importance of precision in reducing misclassifications for better clinical outcomes.

Table 5. Individual Machine Learning Classifier Performance Metrics(%)

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	51.60	62.79	55.38	58.86	50.34
SVC	60.42	62.93	89.23	73.81	50.81
XGB Classifier	54.65	65.51	57.95	61.50	53.55
Decision Tree	50.48	61.81	54.36	57.84	49.19
Naive	**95.19**	**95.23**	**97.18**	**96.19**	**94.53**
KNN	52.40	63.25	56.92	59.92	50.90
Nu-SVC	**88.94**	**93.73**	**88.21**	**90.89**	**89.22**
Logistic Regression	**95.35**	**94.79**	**97.95**	**96.34**	**94.49**
MLP Classifier	63.46	63.32	98.72	77.15	51.71
Adaboost	60.74	76.56	53.59	63.05	63.12
Bagging	48.40	61.97	45.13	52.23	49.49
Extra Tree	52.24	63.94	54.10	58.61	51.62
Proposed Model	**95.28**	**92.30**	**98.46**	**93.90**	**92.93**

The extracted features from each of the 3 best performing post ECA deep CNN architecures are concatenated column-wise and are trained and tested on a stacking classifier with KNeighborsClassifier, LogisticRegression, Gaussian Naive Bayes and Nu-Support Vector Classifier as the first stage classifiers. The hyperparameters for each best performing machine learning classifier employed in the first stage of the stacking classifier found using optimization strategy in the Optuna package. Tab. 5 illustrates that by utilizing all the classification strength of individual classifiers, the proposed stacking ensemble learning surpasses all existing machine learning classifiers.

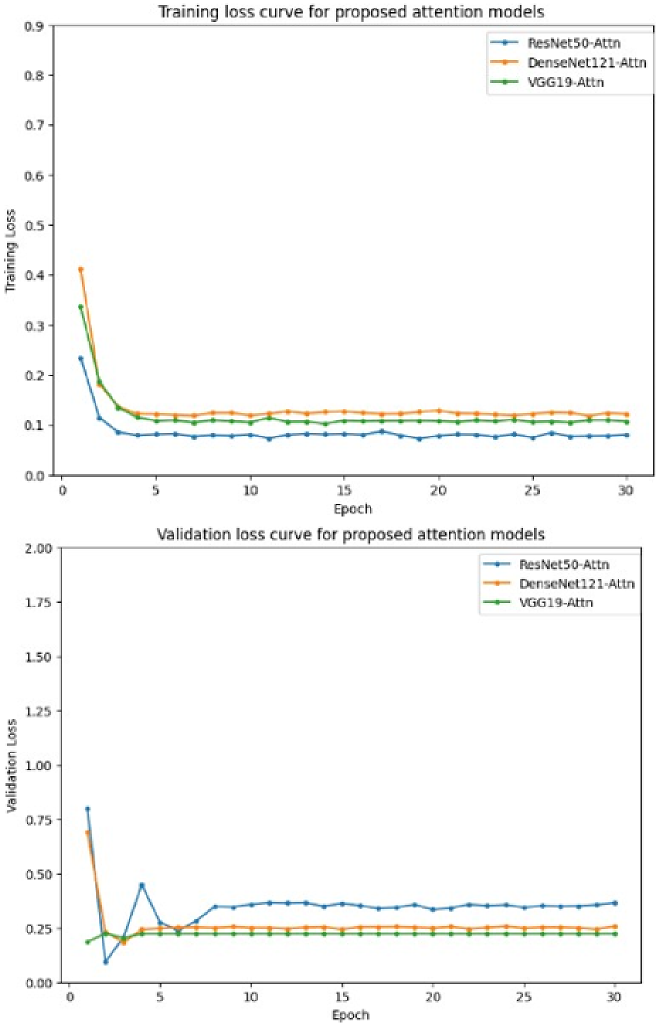


Figure 5. Training and Validation curves of the proposed model

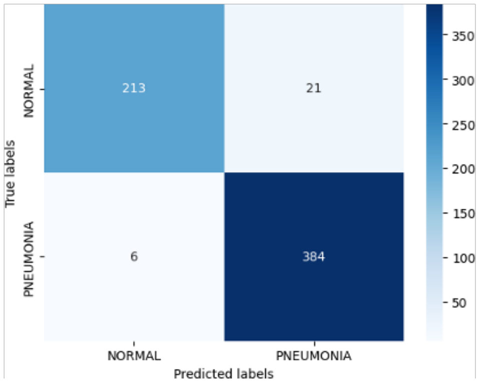


Figure 6. Confusion matrix for proposed pipeline on the test data

In Tab. 6 we can see a comparison, between the model

Table 6. Performance of other recent works on the Kermany dataset(%)

Author	Accuracy	Precision	Recall	F1-Score	AUC
Kermany et al. [6]	92.8	90.1	93.2	-	-
Stephen et al. [8]	93.73	-	-	-	-
Rajpurkar et al [9]	88.78	-	-	-	-
Siddiqi et al. [10]	94.39	92.0	99.0	-	-
Proposed Model	**95.28**	**92.30**	**98.46**	**93.90**	**92.93**

we propose and recent studies conducted on the Kermany et al. [6] dataset. Our proposed model shows performance with accuracy, precision, recall, F1 score and an AUC curve of 95.28%, 92.30%, 98.46%, 93.90% and 92.93% respectively. These results clearly demonstrate how effective our approach is, in detecting pneumonia from chest X ray images. To comprehend the region of interest offered by these deep CNN architectures (ResNet50, DenseNet121, VGG19), **Class Activation Maps (CAM)** are used.Fig. 7

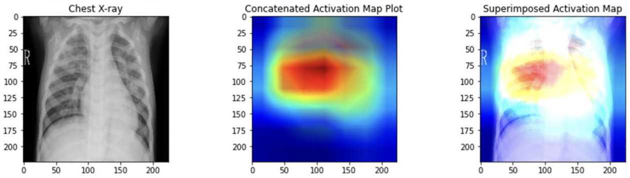


Figure 7. Class Activation Maps of our Deep CNN Architectures

7. Conclusion & Future Works

In this project, we proposed a novel model for pediatric pneumonia diagnosis using chest X-ray images, which integrates an Efficient Channel Attention (ECA) module with pre-trained ResNet50, DenseNet121, and VGG19 models, and combines their predictions using a stacking ensemble method. Our model achieved an **accuracy of 95.28% and an AUC of 92.39%** on the pediatric pneumonia dataset, outperforming existing transfer learning and attention-based CNN models. However, our model still has room for improvement, as it produced some false positives and false negatives, which could have serious implications in clinical settings.

For future work, we plan to extend our model to other medical image analysis tasks, such as COVID-19 detection, lung cancer screening, and tuberculosis diagnosis. We also aim to explore other attention mechanisms and ensemble methods that could further improve the performance and robustness of our model. Moreover, we intend to conduct a comprehensive evaluation of our model on larger and more diverse datasets, as well as compare it with state-of-the-art models in the field.

References

[1] M. Ramezani, S. Z. Aemmi, and Z. E. Moghadam. Factors affecting the rate of pediatric pneumonia in developing countries: a review and literature study. *Int J Pediatr*, 3(6):1173–1181, 2015. 2

[2] B. Neupane, M. Jerrett, R. T. Burnett, T. Marrie, A. Arain, and M. Loeb. Long-term exposure to ambient air pollution and risk of hospitalization with community-acquired pneumonia in older adults. *Am J Respir Crit Care Med*, 181(1):47–53, 2010. 2

[3] Gaobo Liang and Lixin Zheng. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer methods and programs in biomedicine*, 187:104964, 2020. 2

[4] S. Kanakaprabha and D. Radha. Analysis of covid-19 and pneumonia detection in chest x-ray images using deep learning. In *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*, 2021. 2

[5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3

[6] Daniel S. Kermany et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 3, 7

[7] Qilong Wang et al. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 4

[8] Okeke Stephen et al. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of Healthcare Engineering*, 2019, 2019. 7

[9] Pranav Rajpurkar et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 7

[10] Raheel Siddiqi. Automated pneumonia diagnosis using a customized sequential convolutional neural network. In *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*, 2019. 7