Deep Learning-Based Pneumonia Detection Using X-Ray Images

Dennis Marinissen Malin Rekdal Siri Westgård Rusten Mina Mangseth Svorkmo

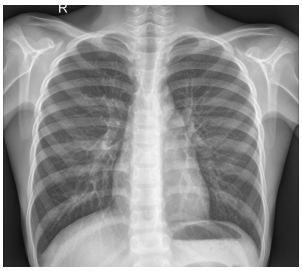
***Abstract*—This paper addresses the problem of the identification of pneumonia in human lungs through the use of chest X-ray images, followed by a proposed solution. Different existing classification methods for lung disease are discussed. We propose to generate synthetic data to compensate for an unbalanced dataset. Furthermore, we fine-tune a pre-trained DenseNet-based architecture for binary classification task of pneumonia detection. The goal of this method is to assist medical experts in diagnosing patients.**

***Index Terms*—Multimedia, Medical Imaging, X-ray, Pneumonia, Lung Disease Detection, Classification, CNN, Data Augmentation, Transfer Learning, ResNet, DenseNet, unbalanced dataset.**

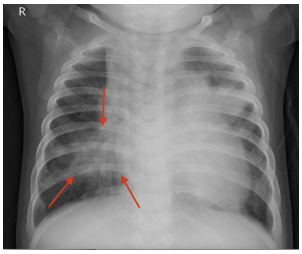
I. Introduction

Figure 1 shows an X-ray image of a healthy chest at the top, bacterial pneumonia in the middle and viral pneumonia at the bottom. A healthy chest X-ray has clear lungs without any abnormal areas. Bacterial pneumonia typically exhibits a focal lobar consolidation. That is when we have a pattern of lung abnormality that we can see in specific spots in the images because of small air sacs in the lungs that are filled with fluid or other substances. In this case we can see it in the right lung marked by the red arrows. Viral pneumonia gives a diffuse pattern of inflammation that is more widespread and affects the spaces between the air sacs in both lungs.

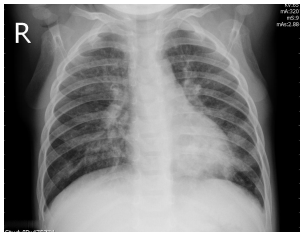
The problem at hand is to detect whether an X-ray image displays a healthy chest or one affected by pneumonia. This is done by using X-ray images of the chest with accompanying class labels. In short, the problem is to solve a supervised binary classification problem on chest X-ray images for pneumonia detection.



(a) Healthy lung.



(b) Bacterial pneumonia.



(c) Viral pneumonia.

Fig. 1: Chest X-ray images of healthy and pneumonia cases.

II. Why is this problem important?

Pneumonia is an inflammation of the lungs, caused by infection with viruses or bacteria, and is a disease that spreads rapidly. For children, old people, and people with heart or lung conditions, this disease can be a serious risk to their health and they may need treatment in hospital. The diagnosis of pneumonia today requires a qualified specialist. For pneumonia, especially viral pneumonia, X-ray images of the chest are used in the diagnosis. In addition to X-ray images of the chest, the specialist also need laboratory tests, vital signs and clinical history from the patient. This makes the diagnosis complex and challenging. As there is a lack of available specialists, the task is getting more demanding. Therefore, it would be valuable to investigate the possibilities of using learning-based models to detect pneumonia through X-ray images only, with the goal of assisting medical experts in the process of diagnosing patients.

III. How is the problem addressed?

A current method to identify pneumonia in the lungs is through the manual examination of X-ray images. We propose to use these X-ray images to train a model for detecting pneumonia. This method requires images that are labeled, with the label indicating either a healthy lung or a lung with pneumonia. By feeding labeled data to a machine learning-based model, we can train a model to classify images, distinguishing between images containing a lung with or without pneumonia. The target is to provide the model with an adequate data

upon training, so that after training the model can perform this classification on unseen images. Performance is therefore evaluated with unseen testing data.

The implications of the model outputs may affect modeling design decisions. False negatives may lead to inadequate treatment for a patient. False positives may lead to unnecessary stress in patients or doctors starting wrong treatments. Since the goal is to assist (i.e. advice) medical experts and not replace them, we aim to maximize model accuracy, therefore punishing false positives and false negatives equally.

The target of a model is to learn a classification function $f : X \rightarrow Y$ by training on the training data D_{train}. We aim to maximize the number of correctly mapped input images x to a binary class label $y \in \{\textit{healthy, pneumonia}\}$. A predicted class label is denoted by $\hat{y}$, where $\hat{y}_i = f(x_i)$. Thus, the goal is to maximize the number of cases where holds that $\hat{y}_i = y_i$. This is optimized during training using a loss function $\mathcal{L}$, which is a measure on the error term between the predicted label $\hat{y}$ and the true label y.

A. Already existing solutions

Using a Convolutional Neural Network (CNN) has shown to be an effective approach for image classification problems in recent literature, more specifically, this has shown to be the case for (deep) learning-based pneumonia detection [8]. Methods using CNN-based architectures offer the benefit of automatic feature learning for more complex problems, showing classification accuracy of over 90% [15, 14]. Using data augmentation to compensate for small data availability and increase model robustness shows to be beneficial [14]. Many existing solutions in literature apply transfer learning for pneumonia detection problems. Existing pre-trained architectures have been proposed for this problem, such as ResNet [1, 9], DenseNet [11, 12], AlexNet [5] and VGG-16 [13]. The advantage of using transfer learning is that pre-trained weights of complex models can be used, saving high computational costs. The disadvantage is that, without further customization or optimization, these approaches using transfer learning show inadequate accuracy performance for our application. Many publications aim to provide a high-complexity solution as a one-size-fits-all solution for multiple classification problems. Only when further optimizing and re-training for our specific pneumonia detection problem, these models show to be promising [8]. Furthermore, one-shot learning has been explored in existing research to compensate for data scarcity [4]. Lastly, ensemble learning has been proposed in literature as a fusion of architectures, to benefit from the properties of multiple base learners [10, 8].

In [11] it is suggested, and compared, multiple backbone architectures for pneumonia detection, and the conclusion is that DenseNet201 is most promising. Since it is shown that DenseNet201 is most promising for a pneumonia detection problem, we propose to use this network and fine-tune it with further improvements. In [2] a proposed solution using ResNet is used. This is used as a starting point for this project. Later we verify if we can outperform existing implementations.

TABLE I: Description of the original dataset.

	Healthy	**Pneumonia**	**Total**
Train set	1349	3883	**5232**
Test set	234	390	**624**
Total	**1583**	**4273**	**5856**

B. Improvements to the solution

Existing solutions in literature show promising results. In the work of [2] high accuracy is achieved. The recall value on the 'healthy' class of 0.82, seems to identify a challenging part of the task at hand. This is because of the large amount (42) of false positives. This can clearly be explained by the class imbalance in the data. Furthermore, overall accuracy of 0.92 leaves potential for improvement.

We will explore already existing solutions as a starting point and look at different methods to improve the model performance. Special focus should be on improving the problem of false positives for healthy lungs. We will generate data to validate if it can improve classification results. Lastly we will do data augmentation to add artificial variation and increase the training set to improve model performance and avoid overfitting.

IV. System architecture and main modules

A. Data availability and description

First of all, we consider data availability. The available datasets are too similar to have any advantage from combining multiple datasets. It will therefore be focused on using one dataset and solve the class imbalance issue. We will improve the reliability of the neural network using data augmentation methods to increase model robustness by introducing more variations in the data, further explained in section IV-C. We also modify the backbone tailored to our application.

The dataset used is from patients of one to five years olds from Guangzhou Women and Children's Medical Center [6]. All chest X-ray imaging was performed as part of patients' routine clinical care. The distribution of the images among training and test data is shown in table I. The dataset consists of 5856 images where 1583 images is of healthy lungs, and 4273 images of lungs with pneumonia.

A class imbalance exists in the dataset, with 2.8 times more pneumonia than healthy instances in the training data. This class imbalance may lead to biased training, in terms of overfitting on one class. This is probably the reason for why existing methods [2] get many false positives. We want to improve this by using data augmentation on the healthy images. This will be further explained in section IV-C.

B. Overall structure

Based on the already existing solutions mentioned in section III-A, we propose to use a CNN-based architecture with transfer learning. We will be fine-tuning a pre-trained DenseNet architecture as explained more in detail in section IV-D. We will use our training data to fine-tune the model

TABLE II: Description of our dataset after data augmentation.

	Healthy	Pneumonia	Total
Train set	3849	3883	**7732**
Test set	234	390	**624**
Total	**4083**	**4273**	**8356**

for pneumonia detection. We will test final performance on unseen test data. Our pipeline for the training of our model will therefore be composed of the elements shown in figure 2. As mentioned in section IV-A, we take the original data and perform data augmentation on the healthy data, generating extra instances to compensate for a class imbalance. Upon model fitting of the pre-trained DenseNet201, we also use data augmentation as a training strategy for all classes to prevent it from over-fitting on the training data. This is to improve performance for detection on unseen cases.

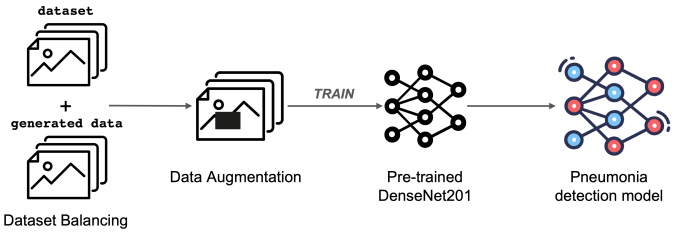


Fig. 2: Main modules in our approach.

C. Data augmentation and validation split

As argued in section IV-A we want to do two types of data augmentation. Firstly, we want to balance out the class imbalance to prevent over-fitting on the class that has pneumonia. To do this we generate extra "healthy"-only samples and store them in the data directory we have. The transformations done are subtle skewing, distortion, zoom and rotation.

The new distribution of the data after data augmentation is shown in table II. When we do the augmentation we get 7732 images in total for training, where 3849 of them is for healthy lungs, and 3883 of them is for data with pneumonia. This is around the same amount for both cases, and we can now expect to not get an over-fitting on one of the classes, therefore preventing a biased model.

We also want to have a part of the data for validation to choose the models hyperparameters to know when to stop training, to avoid either under- or overfitting the model. We therefore split the training data into both training and validation data. We use 80% of the training data for training, and 20% for validation.

Further, we want to increase the model robustness by introducing more variations in the data, and this we do when passing the samples to the model from the data generator flow. The data generator flow is an efficient way of passing data to the model. This data augmentation is preventing the model to just "memorize" the training samples. In other words we want to prevent the model from over-fitting on the training set. Here we introduce slight variations in brightness, rotation, zoom and shift. Note, that in both cases of augmentation we do not flip or mirror any data, since X-ray imaging is standardized in terms of view.

D. DenseNet structure

From section III-A, we know that the DenseNet structure is most promising for the image classification tasks of pneumonia detection. Therefore our structure is based on this deep neural network architecture.

More specifically, DenseNet201 is a type of DenseNet network with 201 layers. Each of these 201 layers has batch normalization, a ReLU activation function and a convolutional layer. After blocks, pooling is applied to down-sample the layers, changing the size of feature-maps. The number of filters and kernel sizes varies depending on the layers position in the network.

Three blocks in DenseNet201 are shown in figure 3. In DenseNet201 we have dense connections which means that all the layers are connected to all the other layers through feeding forward fashion. Besides other beneficial reasons, this will in our case give a desired flow of information through the network. The dense layers are combined into multiple dense blocks.

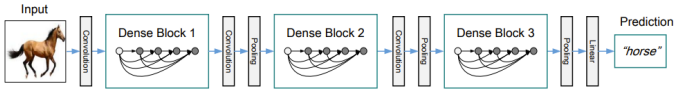


Fig. 3: DenseNet structure with 3 blocks.

E. Approach description

In the original DenseNet201 structure, described in section IV-D, the ReLU activation function is used for its layers. In order to use the pre-trained DenseNet, we need to adjust it for our application. Therefore, we replace the output layers of the model. We replace the "head" of the original architecture by a fully connected layer, together with a pooling and dropout layer. These are followed by the final output layer. For the final output layer we use a single node with sigmoid activation function, responsible for the final classification. In our application it is beneficial to use the sigmoid function because sigmoid is naturally suited for binary classification tasks. This is because it models probabilities of binary events. The output can be directly interpreted as the estimated probability of belonging to either one of the binary classes, as represented by 0 and 1. This enables higher computational efficiency, since a single node can be used in the output layer in our case, whereas other activation functions (e.g. softmax) requires parameter calculations on two nodes.

The DenseNet201 has been pre-trained on the widely used ImageNet data [3]. In order to fine-tune the model for classifying pneumonia, we want to re-train a part of those weights. We freeze the weights of the pre-trained model, except for the last 30 layers of the backbone. This means that during training

TABLE III: Table of precision, recall and F1-score for our model.

	Precision	**Recall**	**F1-score**	**Support**
Healthy	0.98	0.91	0.94	234
Pneumonia	0.95	0.99	0.97	390

we, fit the parameters of the last 30 layers of the DenseNet model, as well as the head and output layers.

Validation data is used to improve the fitting of our model, allowing the use of callbacks during training. Callbacks are a way of mediating the training process, in our case with the goal of avoiding over-fitting. After each epoch during training, the validation loss is monitored. In case the validation loss reaches a plateau, the learning rate is reduced with patience 2. Also, we apply early stopping with patience 5, which assists in stopping the training process before it starts over-fitting. Additional parameters are presented in accompanying code.

The final results were achieved after 70 epochs (stopped by early stopping), initialized with Adam optimizer [7] and learning rate of 0.000002.

V. Results

The model's performance on the test data are shown in the confusion matrix in figure 4. There we for example can see that there are more false positives than false negatives. The testing also gives the accuracy 0.96, and the values of precision, recall and f1-score are presented in table III. For further discussion we will compare with the results from [2].

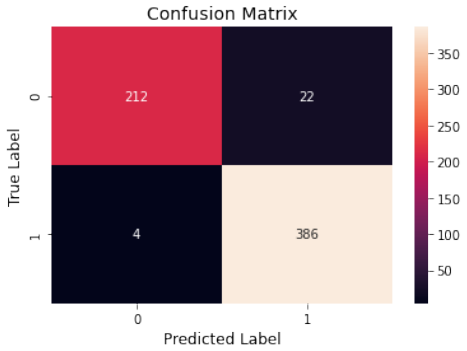


Fig. 4: Confusion matrix for the model on the testing data. Label 0 is for healthy and label 1 for pneumonia.

VI. Discussion

Firstly, when comparing, we observed that DenseNet with our data augmentation strategies outperformed ResNet in terms of accuracy and overall performance. The most impressive improvement is that the recall on the healthy data went from 0.82 to 0.91 by using DenseNet and balancing the datasets using data augmentation. In addition to overall improvements in recall for pneumonia and precision, this resulted to a significantly increased accuracy from 0.92 to 0.96. It is difficult to know whether the performance improvement is caused by the DenseNet or the data augmentation since we did not try them separately, but it is likely to assume that both have contributed to improving the results.

Moreover, we used data augmentation techniques on all the training data to increase the diversity and size of the dataset, in addition to balancing the dataset. The inclusion of data augmentation on all of the data afterwards resulted in notable improvements in the model's performance. The improved performance after data augmentation suggests that it effectively mitigates over-fitting and helps the model learn more robust and generalized representations.

Nevertheless, there is more false positives than false negatives as shown in figure 4. This means that more healthy images is classified as pneumonia than the opposite. This can come from noise in the healthy images that can have the appearance as either the abnormal areas in bacterial pneumonia or the widespread inflammation in viral pneumonia. This can also still be the result of the original dataset being unbalanced, as data augmentation might not be able to fully compensate for that.

A. Further work

For further work it would be interesting to look at what happens if we change from ResNet to DenseNet and do the data augmentation on the healthy data separately instead of together. Because of computational limitations, investigating this was considered out of scope. However, it would be interesting to get information on which of those improvements that contributed the most to the results we got. Furthermore, in the future, newly developed architectures could also be explored over time for their applicability.

It would also be interesting to extend the solution to a multiple class model which can distinguish between bacterial and viral pneumonia. We do not know if the miss-classified images are bacterial or viral pneumonia. Furthermore, we can not say how robust our model really is to any data besides the data we have at hand. Also, it would be a good improvement to combine the data with data on other lung diseases to develop a solution that can classify multiple lung diseases. It would be interesting to look at how super resolution can enhance the detection performance as well as focus more on the explainable and inseparability of our solution, e.g. by quantifying model prediction confidence.

VII. Conclusion

An accuracy of 0.96 with F1-scores of 0.94 and 0.97 makes it a strong proposition that it might e possible to assist medical professionals in detecting pneumonia with help from CNNs. We clearly see the importance of having a good dataset with approximately the same amount of data in all classes when doing classification to avoid biasing. In conclusion, there is a lot of factors that affect the result of automated classification. It is still a far stretch before Neural Networks can replace medical professionals. However, considering the high performance in our results, we see a promising case of assisting experts in the field using learning-based models.

References

[1] Saleh Albahli et al. "AI-driven deep CNN approach for multilabel pathology classification using chest X-Rays". In: *PeerJ Computer Science* 7 (2021), pp. 1–17. ISSN: 23765992. DOI: 10.7717/peerj-cs.495.

[2] Jonas Paluci Barbosa. *Kaggle: Chest X-Ray (Pneumonia) - CNN & Transfer Learning*. 2022. URL: https://www.kaggle.com/code/jonaspalucibarbosa/chest-x-ray-pneumonia-cnn-transfer-learning.

[3] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 248–255. DOI: 10.1109/cvpr.2009.5206848.

[4] Tatiana Gabruseva, Dmytro Poplavskiy, and Alexandr Kalinin. "Deep learning for automatic pneumonia detection". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 2020-June (2020), pp. 1436–1443. ISSN: 21607516. DOI: 10.1109/CVPRW50498.2020.00183. arXiv: 2005.13899.

[5] Abdullahi Umar Ibrahim et al. "Pneumonia Classification Using Deep Learning from Chest X-ray Images During COVID-19". In: *Cognitive Computation* 0123456789 (2021). ISSN: 18669964. DOI: 10.1007/s12559-020-09787-5. URL: https://doi.org/10.1007/s12559-020-09787-5.

[6] Daniel S. Kermany et al. "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning". In: *Cell* 172.5 (2018), 1122–1131.e9. ISSN: 10974172. DOI: 10.1016/j.cell.2018.02.010. URL: https://doi.org/10.1016/j.cell.2018.02.010.

[7] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (Dec. 2014), pp. 1–15. arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

[8] Rohit Kundu et al. "Pneumonia detection in chest X-ray images using an ensemble of deep learning models". In: *PLoS ONE* 16.9 September (2021). ISSN: 19326203. DOI: 10.1371/journal.pone.0256630. URL: http://dx.doi.org/10.1371/journal.pone.0256630.

[9] Gaobo Liang and Lixin Zheng. "A transfer learning method with deep residual network for pediatric pneumonia diagnosis". In: *Computer Methods and Programs in Biomedicine* 187 (2020), p. 104964. ISSN: 18727565. DOI: 10.1016/j.cmpb.2019.06.023. URL: https://doi.org/10.1016/j.cmpb.2019.06.023.

[10] Ian Pan, Alexandre Cadrin-Chênevert, and Phillip M. Cheng. "Tackling the radiological society of North America pneumonia detection challenge". In: *American Journal of Roentgenology* 213.3 (2019), pp. 568–574. ISSN: 15463141. DOI: 10.2214/AJR.19.21512.

[11] Tawsifur Rahman, Muhammad E H Chowdhury, and Amith Khandakar. "applied sciences Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using". In: *MDPI, J. app Sci.* 3233 (2020), pp. 1–17.

[12] Pranav Rajpurkar et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". In: (2017), pp. 3–9. arXiv: 1711.05225. URL: http://arxiv.org/abs/1711.05225.

[13] Uzair Shah et al. "An efficient method to predict pneumonia from chest X-rays using deep learning approach". In: *Studies in Health Technology and Informatics* 272 (2020), pp. 457–460. ISSN: 18798365. DOI: 10.3233/SHTI200594.

[14] Harsh Sharma et al. "Feature extraction and classification of chest X-ray images using CNN to detect pneumonia". In: *Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering* (2020), pp. 227–231. DOI: 10.1109/Confluence47617.2020.9057809.

[15] Okeke Stephen et al. "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare". In: *Journal of Healthcare Engineering* 2019 (2019). ISSN: 20402309. DOI: 10.1155/2019/4180949.