Simon Fonck*, Sebastian Fritsch, Alina Nguyen, Stefan Kowalewski, and André Stollenwerk

Robustness of a DenseNet-121 for the Classification of ARDS in Chest X-Rays

https://doi.org/10.1515/cdbme-2024-2059

Abstract: Research in the field of artificial intelligence (AI) in medicine is increasingly relying on algorithms based on deep learning (DL), especially for radiology. Despite producing promising results, DL models have a major drawback: their reliance on large training datasets. Especially in medicine, large, annotated datasets are hard to obtain, leading to low robustness and a performance loss when exposed to unseen, new data. To address this problem, our research evaluates how well data augmentation is able to expand the used dataset and thus improve a DL model. We employ 17 different augmentation methods to test the robustness of a DenseNet-121 trained to classify Acute Respiratory Distress Syndrome (ARDS) in chest X-rays. Our experiments show that while the model has low robustness for augmented test data when trained on unaugmented data, the general performance for ARDS classification can be improved by augmenting the training data. Overall, this demonstrates that data augmentation is beneficial in training AI models for ARDS classification in order to create more robust and generalizable models.

Keywords: Deep Learning, Data Augmentation, Medical Application, Robustness, Artificial Intelligence

1 Introduction

Acute Respiratory Distress Syndrome (ARDS) is a lung injury characterized by a sudden and severe inflammatory response in the alveoli leading to severe impairment of alveolar gas exchange and consequently hypoxemia. It was first described by Ashbaugh et al. in 1967 and has since been a frequent topic in research [1, 2]. In 2012, the Berlin Definition (BD) has been established as a standard for the diagnosis of ARDS. The BD specifies four criteria that must be met for ARDS [3]. In 2016, Bellani et al. discovered in their LUNG SAFE study, that around 10.4 % of all mechanically ventilated ICU patients suffer from ARDS, with a hospital mortality of 40 % [4]. One of the reasons for the high mortality rate is the often too late or missed diagnosis [5]. To support physicians in diagnosing ARDS, artificial intelligence (AI) methods have been published in recent years [6]. They can support physicians in determining the BD criteria and therefore improve the diagnosis. In particular, deep learning (DL) models, like convolutional neural networks (CNN), are used to support the ARDS classification in image data, which is a time-consuming task typically done by a radiologist [7, 8]. But as useful as they are, DL models have one drawback, which is “their lack of generalisability and tendency to overfit when presented with small training sets” [9]. Since there are only a few publicly available datasets in the medical field, due to the sensitivity of medical data and protection of the patient’s privacy, the training dataset is often small. Thus, resulting models lack robustness and transferability [10]. To overcome this problem, data augmentation can be used to expand and generalize the dataset and thus prevent overfitting [11]. In their review, Chlap et al. present data augmentation techniques ranging from basic methods (such as rotation) to deformation and DL-based methods (such as generative adversarial networks) [11]. In our research we incorporated basic data augmentation techniques to evaluate the robustness of a DenseNet-121 for the classification of ARDS in chest X-rays [8].

2 Design & Concept

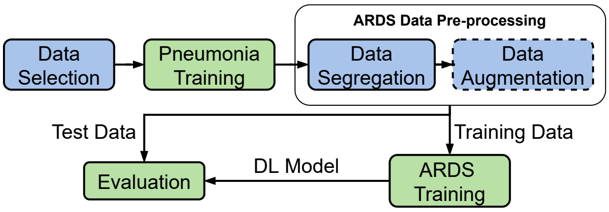

Fig. 1: Workflow for the training and evaluating of the robustness of a DL model.

In our experiments, we evaluate a DenseNet-121 using various augmentation techniques, that are suited for medical image data [11]. The model is trained using a transfer-learning process from pneumonia to ARDS, analog to the one presented in Fonck et al. [8]. First, we selected datasets for the training on pneumonia and a dataset for the fine-tuning on ARDS. Afterwards, the ARDS data is segregated in training and test data,

Simon Fonck*, Alina Nguyen, Stefan Kowalewski, André Stollenwerk, Embedded Software (Informatik 11), RWTH Aachen University, 52074 Aachen, Germany, ***Corresponding Author:** fonck@embedded.rwth-aachen.de
Simon Fonck*, Sebastian Fritsch, André Stollenwerk, Center for Advanced Simulation and Analytics (CASA), Forschungszentrum Jülich, 52428 Jülich, Germany
Sebastian Fritsch, Department of Intensive Care Medicine, University Hospital RWTH Aachen, 52074 Aachen, Germany
Sebastian Fritsch, Jülich Supercomputing Centre, Forschungszentrum Jülich, 52428 Jülich, Germany

which can then be augmented. With the according data the model is fine-tuned and finally evaluated using performance metrics. The process is depicted in Figure 1.

2.1 Data Selection

For our study, we used three publicly available databases: CheXpert [12] and a dataset by Kermany et al. [13] for the pretraining on pneumonia and MIMIC-CXR for the training on ARDS [14]. CheXpert contains 224,316 chest radiographs of 65,240 patients labelled for 14 different common chest radiographic observations (4,684 labeled for pneumonia and 17,000 with no findings) [12]. The dataset published by Kermany et al. contains 5,233 images, where 3,883 are labeled as pneumonia and 1,350 as normal [13]. The radiographs were taken of pediatric patients of one to five years old. Here, it is important to take the physical differences of adults and children into account. However, this fact may contribute to a more robust and generalizable DL model. Lastly, the MIMIC-CXR is a dataset containing 337,110 radiographs from 227,827 patients, which are also labeled for the 14 observations [14]. The database is published by the PhysioNet [15]. None of the three databases contain labels for the classification of ARDS. Therefore, we extracted 533 images from the MIMIC-CXR dataset based on the noteevents in the MIMIC-IV database, that was manually screened and annotated by a radiologist of the University Hospital RWTH Aachen for the presence of ARDS. In addition, the radiologist indicated (un)certainty for specific images. According to his (un)certainty level we weight the dataset by including ARDS images that the radiologist was (very) certain about up to four times in the training dataset. For the pretraining of the DL model we used two sets of the pneumonia labeled radiographs: a subset of 1,000 images (687 normal and 313 pneumonia) for initial testing and the combined dataset for further evaluation. For the fine-tuning on ARDS, we used the weighted dataset (see Table 1).

Tab. 1: Datasets that are used for the pneumonia training and fine-tuning for ARDS.

	Pneumonia		**No Findings**	
	CheXpert	**Kermany**	**CheXpert**	**Kermany**
Combined	**4,684**	**3,883**	**17,000**	**1,350**
Subset	**313**		**687**	
	ARDS		**No ARDS**	
Unweighted	**163**		**370**	
Weighted	**446**		**370**	

2.2 Data Segregation & Augmentation

After the training of the DL model on pneumonia, the weighted ARDS dataset is divided into a training (77 %) and a test dataset (23 %). The training data is used to fine-tune the DL model to ARDS. The test dataset is then used to evaluate the models. It is important that the data augmentation takes place after the segregation, as otherwise the same image, albeit augmented, is in the training and test dataset and a sound evaluation of the model is no longer possible. In the data augmentation step, we focused on methods, that according to a radiologist of the University Hospital Aachen may occur when using different X-ray settings or machines, like geometric shifts or changes in the intensity [11]. Overall, we tested 17 different data augmentation techniques. The overview of all used augmentation techniques can be seen in Figure 2.

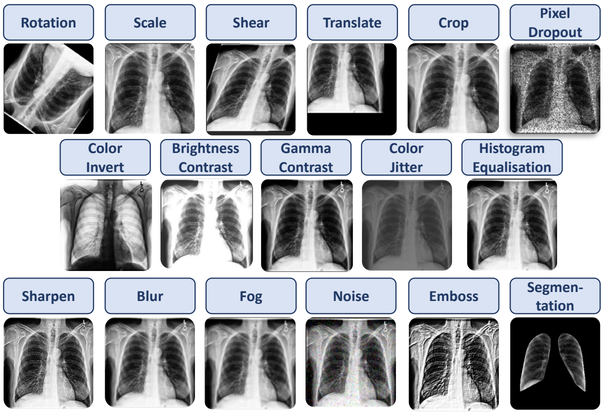

Fig. 2: Different data augmentation techniques that were used for our experiments including an example.

The augmentation is optional for the training and/or test data to evaluate various scenarios. In the first mode (M1), we evaluated the models, that are fine-tuned with unaugmented data and tested them on an unaugmented test set. In the second mode (M2), we augmented the test data to evaluate how the performance of the models may change due to the new, unseen data properties. In the third mode (M3), we just augmented the training set and tested the model with unaugmented data, to determine, whether augmented training reduces the performance compared to the normal training. In the fourth (M4), we augmented the training and test data. The four modes are depicted in Figure 3.

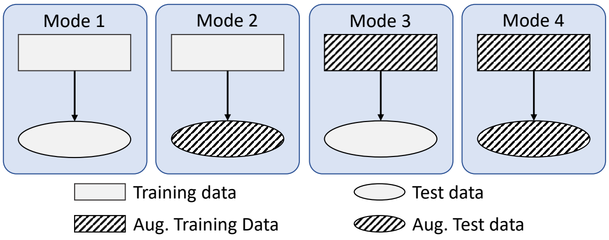

Fig. 3: Four modes to evaluate the robustness of DL models.

2.3 Model Training & Evaluation

In our experiments we used various (augmented) training data and (augmented) test data allocations to train a model in four different modes to detect ARDS in Chest X-rays, as depicted in Figure 1 and Figure 3. First, we pre-trained the models on pneumonia using the subset from the combined dataset to reduce the computational cost. Afterwards, we tested all augmentation techniques individually using the weighted dataset for ARDS. Subsequently, we selected 5 different augmentation techniques to analyze the overall impact of the methods on the training process. Therefore, we used the combined dataset for pre-training (see Table 1) and all augmentation methods concurrently. Overall, we had 55 training runs for the different modes and techniques. We evaluated the model using metrics such as accuracy, area under the receiver operating characteristics (AUROC) curve and F1-Score.

3 Results

The results of our experiments of evaluating the robustness of the DenseNet-121 for classification of ARDS in chest X-rays for the different training modes (see Figure 3) can be seen in Table 2. Overall, we can see that the performance of the model decreases in comparison to M1 when augmenting the test datasets (see M2), which indicates a low robustness of the trained models. For some augmentation methods however, the performance does not change (significantly), like p.e. Blur, Histogram Equalization or Sharpen. For others, like Color Invert, Dropout and Segmentation it decreases considerably. The metrics for M4 show, that except for Segmentation, all models adapt to the augmentation and perform better on the augmented test data. When training with augmented data (M3), no significant loss of performance was observed in the unaugmented test dataset (see Table 2). Based on the results in Table 2, we selected five techniques to analyze their combined influence on the training process. For this purpose, we focused on the methods that led to a performance loss in M2, which could be improved by augmenting the training data in M4. In particular, we focused on methods that change the intensity of the images or filtering methods. In the end, we used the following five methods: Color Invert, Color Jitter, Emboss, Gamma Contrast and Fog. These five methods are evaluated using the whole combined dataset for pneumonia training and the weighted dataset for ARDS (see Table 1), on which all augmentation techniques were applied. The results can be seen in Table 3. These show that the augmentation of the training data leads to better results for the unaugmented test data (see M3) and generally to better results (see M4). It can be seen that unaugmented training leads to less robust models (see M2).

Tab. 3: Resulting metrics for the DenseNet-121 using the resulting five augmentation methods on the data.

	Accuracy	**AUROC**	**F1-Score**
Mode 1	**0.960**	**0.940**	**0.926**
Mode 2	**0.891**	**0.862**	**0.783**
Mode 3	**0.990**	**0.982**	**0.981**
Mode 4	**0.982**	**0.969**	**0.966**

4 Discussion

Although our results show that some augmentation techniques led to a performance loss when only applied to test data (see Table 2), we found that data augmentation can improve the robustness of the model, when also used on the training data. This can be inferred from the comparison of M3 and M4 with M2. In our experiment, we included mainly intensity augmentations to simulate the differences in X-ray machines and showed that they were able to improve the model's robustness (see Table 3). Contrary to our initial impression that techniques such as Sharpen, Blur and Rotation are suitable for medically misrecorded radiographs, they did not provide significant improvement (M4) but also no deterioration compared to no augmentation. After observing the provided image data, it becomes apparent, that the training dataset already includes rotated, blurred and sharpened images. Therefore, unaugmented

Tab. 2: Results of the 17 Data Augmentation techniques for the different modes (M1 - M4) (see Figure 3). The orange-colored methods were selected for further evaluation of the model with concurrent application on the ARDS training data.

	No. Augmentation (M1)	Aug. Method	Blur	Brightness Contras	Color Invert	Crop	Dropout	Emboss	Histogram Equalisation	Fog	Gamma Contrast	Color Jitter	Noise	Rotation	Scale	Segmentation	Sharpen	Shear	Translate
Accuracy	0.990	M2	0.990	0.921	0.624	0.861	0.743	0.792	0.980	0.891	0.960	0.951	0.753	0.931	0.871	0.533	0.951	0.891	0.891
		M3	0.980	0.960	0.990	0.980	0.990	0.970	0.960	0.970	0.960	0.960	0.970	0.960	0.980	0.941	0.970	0.970	0.960
		M4	0.990	0.951	0.960	0.911	0.871	0.980	0.970	0.931	0.980	0.970	0.891	0.941	0.901	0.767	0.951	0.951	0.911
AUROC	0.982	M2	0.982	0.905	0.631	0.816	0.683	0.745	0.964	0.855	0.933	0.931	0.714	0.925	0.926	0.381	0.919	0.879	0.868
		M3	0.974	0.961	0.981	0.987	0.993	0.980	0.973	0.967	0.948	0.973	0.980	0.973	0.987	0.935	0.980	0.980	0.973
		M4	0.982	0.919	0.940	0.871	0.833	0.964	0.948	0.894	0.964	0.948	0.850	0.915	0.860	0.411	0.919	0.919	0.896
F1-Score	0.981	M2	0.981	0.840	0.513	0.741	0.567	0.667	0.963	0.793	0.929	0.906	0.627	0.857	0.667	0.000	0.912	0.766	0.776
		M3	0.962	0.926	0.980	0.963	0.981	0.946	0.929	0.943	0.923	0.929	0.946	0.929	0.963	0.889	0.946	0.946	0.929
		M4	0.981	0.912	0.926	0.848	0.800	0.963	0.946	0.881	0.963	0.946	0.820	0.889	0.833	0.000	0.912	0.912	0.816

training recognizes the pattern before augmenting.

Limitations: We only focused on basic data augmentation techniques, not considering generative augmentation methods or combining techniques in one image, which could also be beneficial in clinical environment. Thus, in future work, other techniques should be explored and evaluated to improve the robustness of deep learning models for the identification of ARDS in chest X-rays even further. In addition, it should be evaluated in cooperation with radiologists whether the augmented images still reflect clinical reality. Furthermore, due to computational costs required, we just included five different augmentation methods in our combined analysis. More concurrent methods and other combinations could be analyzed in future research. In addition, the dataset used for ARDS classification is relatively small. Although this was one of the motivations for our research, a larger dataset may further support and improve the results. In order to further test the generalizability of DL models, external data should be included, which unfortunately were not available at the time of this study.

5 Conclusions

In this research, we evaluated the robustness of DenseNet-121 when exposed to new and unseen data. Initially, we used 17 different augmentation techniques to test how they affect the performance of the models on augmented test data in different modes. Afterwards, we compared the performance when augmenting the training data to see how well we can improve the robustness by incorporating data augmentation in the training process. Overall, we could show, that the model had poor results on the augmented data, indicating, that they generalize poorly to unknown data. Furthermore, we were able to show that the augmentation of the training data does not have a major impact on the overall performance of a model, which implies that data augmentation has no direct drawbacks and can be included to improve robustness and enlarge the training data. We conclude that data augmentation is beneficial, when facing small amount of training data for CNN models, that may be used in a clinical setting.

Author Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest. Ethical approval: Not applicable. **Acknowledgements**: The authors thank Maximilian Schulze-Hagen, M.D., M.Sc., for annotating the images used for ARDS classification. Simulations were performed with computing resources granted by RWTH Aachen University under project thes1523.

References

[1] Ashbaugh, D.G., Bigelow, D.B., Petty, T.L., and Levine, B.E. (1967). Acute respiratory distress in adults. The Lancet. doi:10.1016/s0140-6736(67)90168-7

[2] Meyer, N. J., Gattinoni, L., and Calfee, C. S. (2021). Acute respiratory distress syndrome. Lancet (London, England), 398(10300), 622–637. doi:10.1016/S0140-6736(21)00439-6

[3] The ARDS Definition Task Force. Acute Respiratory Distress Syndrome: The Berlin Definition. JAMA. 2012;307(23):2526–2533. doi:10.1001/jama.2012.5669

[4] Bellani, G., Laffey, J. G., LUNG SAFE Investigators, ESICM Trials Group, et al. (2016). Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. JAMA, 315(8), 788–800. doi:10.1001/jama.2016.0291

[5] Bellani, G., Pham, T., and Laffey, J. G. (2020). Missed or delayed diagnosis of ARDS: a common and serious problem. Intensive care medicine, 46(6), 1180–1183. doi:10.1007/s00134-020-06035-0

[6] Rashid, M., Ramakrishnan, M., Chandran, V. P., Nandish, S., et al. (2022). Artificial intelligence in acute respiratory distress syndrome: A systematic review. Artificial intelligence in medicine, 131, 102361. doi:10.1016/j .artmed.2022.102361

[7] Sjoding, M. W., Taylor, D., Motyka, J., Lee, E., et al. (2021). Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation. The Lancet. Digital health, 3(6), e340–e348. doi:10.1016/S2589-7500(21)00056-X

[8] Fonck, S., Fritsch, S., Nottenkämper, G., Stollenwerk, André (2023). Implementation of ResNet-50 for the Detection of ARDS in Chest X-Rays using transfer-learning. Proc AUTOMED, 2(1), ID 742, www.journals.infinite-science.de/automed/article/view/742

[9] Nanni, L., Paci, M., Brahnam, S., Lumini, A. (2021). Comparison of Different Image Data Augmentation Approaches. Journal of Imaging, 7(12):254. doi:10.3390/jimaging7120254+

[10] Ding, N., Möller, K. (2023). The Image flip effect on a CNN modelclassification. Proc AUTOMED, 2(1), ID 755, www.journals.infinite-science.de/automed/article/view/755

[11] Chlap, P., Min, H., Vandenberg, N., Dowling, J., et al. (2021). A review of medical image data augmentation techniques for deep learning applications. Journal of medical imaging and radiation oncology, 65(5), 545–563. doi:10.1111/1754-9485.13261

[12] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, 33(1), 590-597. doi:10.1609/aaai.v33i01.3301590

[13] Kermany D, Goldbaum M, Cai W et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018; 172(5):1122-1131. doi:10.1016/j.cell.2018.02.010

[14] Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6, 317 (2019). doi:10.1038/s41597-019-0322-0

[15] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.