Cascade Learning for Group-Specific Medicine: A Multi-Stage Approach

Ruggero Anello[1] and Florian Braunmiller[1]

Technical University of Munich, Germany
ruggero.anello@tum.de and florian.braunmiller@tum.de

Abstract. Computer-aided diagnosis systems are a powerful tool for physicians to support the identification of diseases in medical images. However, the 'one-model-fits-all' approach is not optimal for highly heterogeneous data, as it is common in medical imaging. This work proposes a cascading model for the group-specific classification of chest radiographs. The model consists of a classification stage, predicting the group, and multiple group-specific models, predicting the observations. The model is evaluated on the CheXpert dataset, split by view (FR/LAT) and the frontal view direction (AP/PA). The results show that the cascading models achieve comparable or slightly superior multi-label performance to 'one-model-fits-all' approaches while offering additional advantages, such as efficiency gains and selective retraining of specific modules. The single-label performance of the multi-stage models was slightly better than the baseline models in most conditions. The modular nature of these models allows for selective retraining of only specific modules after new data acquisition and faster training times. Future models could incorporate domain knowledge more explicitly into the model architecture and develop specialized architectures optimized for each view type.

Keywords: Computer-aided diagnosis (CAD) · Medical imaging · Cascading models · Chest radiographs · Multi-stage models.

1 Introduction and Background

Computer-aided diagnosis systems are considered a powerful tool for physicians to support the identification of diseases in medical images [6,1,8,9]. State-of-the-art models use one deep learning model to classify all images, regardless of their characteristics [1,8,9]. However, this approach is not optimal for highly heterogeneous data, as it is common in medical imaging due to variations in patients or imaging techniques.

Instead of a 'one-model-fits-all' approach trained on a highly heterogeneous dataset, we propose a cascading model consisting of a classification stage, predicting the group, and multiple group-specific models, predicting the observations. Patient-specific characteristics such as age, sex, and skin color, or sample-specific characteristics such as the direction or plane of an X-ray image or imaging modalities could define possible groups. This approach allows the model to learn distinct group-specific features within the data.

A cascaded deep learning classifier was previously used to detect COVID-19 and pneumonia diseases in chest X-ray images [6]. Instead of performing a multi-label classification, a series of binary classifiers for each tested case of the health status was used to mimic the clinical situation to diagnose potential diseases for a patient. The performance of the cascading model was superior to a multi-label classification model. Other concatenated classifiers consist of multiple stages, focussing on different aspects or regions of an image. By gradually eliminating non-relevant areas, the model can improve accuracy and reduce false positives [11].

This work aims to evaluate the performance of cascading models on the CheXpert dataset [5], split by view (FR/LAT) and the frontal view direction (AP/PA). We compare these cascading models to single models trained on the whole dataset.

2 Methods

2.1 The CheXpert Dataset

All experiments in the presented work were performed using the CheXpert Small dataset [5]. The dataset consists of 224,316 chest radiographs of 65,240 patients labeled for 14 common observations. It includes images acquired in the frontal (FR) and lateral (LAT) planes. The frontal plane is further divided into the anterior-posterior (AP) and posterior-anterior (PA) directions. Each observation is labeled as positive, negative, or uncertain. The original validation and test sets are very small, with only 234 and 668 images, and unbalanced concerning the views. Therefore, a new 90/5/5 split was created using all available images. The new test set was further processed to obtain multiple subsets to evaluate different models with balanced views. The subsets are the complete test set (AP/PA/LAT), the AP/PA test set, and the FR/LAT test set.

The CheXpert dataset was chosen because of its clear separation into different groups (AP/PA, FR/LAT) and its clinical relevance, as certain conditions are only diagnosable in specific data groups [7,10].

2.2 Models

Multiple models were trained and evaluated by classifying five observations from the CheXpert dataset: Atelectasis, Edema, Pleural Effusion, Cardiomegaly, and Consolidation. The models were trained to independently to predict each observation's presence, resulting in a multi-label classification task.

To keep the comparison fair, all models were trained with the same family of architectures, ResNet [4], and similar training parameters. ResNet was chosen because of its fast training times and availability of different sizes of pre-trained models.

Baseline Models The baseline models are used to compare the performance of the cascading models to a single 'one-size-fits-all' model. The baseline models are ResNet-50 architectures [4] pre-trained on ImageNet [3] and fine-tuned on the CheXpert dataset. Two models were trained, one on FR view images and one on the whole dataset.

Two-Stage Models Two two-stage models (2SM) were trained, one splitting the whole data into FR/LAT and the other splitting the FR images into AP/PA views. Both two-stage models consist of a classification model and a view-specific model. The classification model is a ResNet-18 [4] and is used to predict the view. Depending on the view, the view-specific model (ResNet-18 [4]) is used to predict the observation (Figure 1).

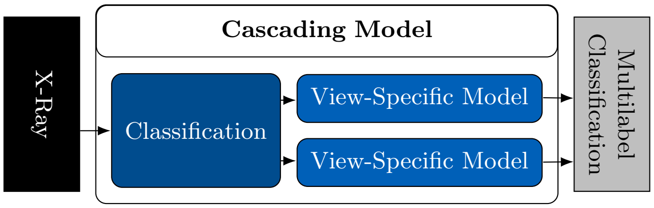

Fig. 1: Two-stage model architecture. The input image is first classified in the first stage; then the observation is predicted in view-specific models in the second stage.

Three-Stage Model The three-stage model (3SM) is an extension of the 2SM. It adds a third stage to the two-stage model, first predicting if the view is FR or LAT, then predicting if the view is AP or PA, and finally predicting the observation, again using a view-specific model (Figure 2).

Training All models preprocessed the images to a resolution of 256x256 pixels. Data augmentation techniques were applied during training, including random rotation, random affine and crop, Gaussian blur, color jitter, brightness, and saturation. The models were initialized using the ImageNet weights [3].

The multi-stage approach view-specific modules (AP/PA/FR/LAT) and classification modules (FR /LAT and AP/PA) were trained independently and only combined for inference. They are initialized with the weights of a pre-trained ResNet-18 (50 epochs, learning rate $10{}^{-}4$, Multi-Label Focal Loss) on the whole CheXpert dataset.

The view-specific and Baseline models were trained with the same parameters to ensure the comparability of the results. The learning rate was set to 5×10^{-5} and reduced by a factor of 0.1 every 10 epochs. The optimizer used was Adam. The baseline models were trained for 100 epochs, while the others were trained

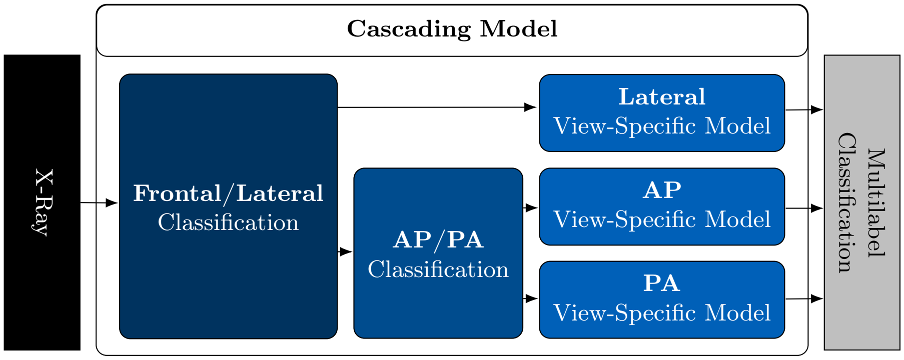

Fig. 2: Three-stage model architecture. The view of the input image is classified in the first stage in FR/LAT. Then, the second stage classifies all FR images in AP/PA. Finally, depending on the view, the images are classified using view-specific models in the third stage.

for 50 epochs. To handle the unbalanced dataset, a weighted sampler, uncertainty mapping and label smoothing, as described in [2], and Multi-Label Focal Loss with in-class (w_c) and inter-class (w_p) weighting were used (Equation 1).

$$FL(p_t) = (-\alpha_t(1-p_t)^\gamma \log(p_t)) \cdot w_p \cdot w_c, \text{ with } w_p = \frac{\#\text{ neg}}{\#\text{ pos}}, w_c = 1+|\log w_p| \quad (1)$$

For the observation classification, the best model was selected based on the Matthews Correlation Coefficient (MCC) score on the validation set; for the view classification, the best model was chosen using accuracy.

3 Results

In the first stage of the cascade approach, the ResNet-18 FR/LAT and AP/PA classification models predicted the plane and view perfectly with an accuracy of 1, Recall of 1, and Precision of 1.

Additionally, when it comes to the downstream classification, pre-training the models led to a significant increase in performance across all view-specific models. In particular, the MCC of the 2SM FR/LAT model improved from 0.615 to 0.661, in the 2SM AP/PA from 0.602 to 0.632, and in the 3SM model from 0.666 to 0.706. In the following, only the results of the pre-trained models are presented.

3.1 Multi-Label-Performance: Baseline vs. Multi-Stage Models

Both 2SM and 3SM show a strong overall performance, on par with the Baseline aggregate metrics (Figure 3).

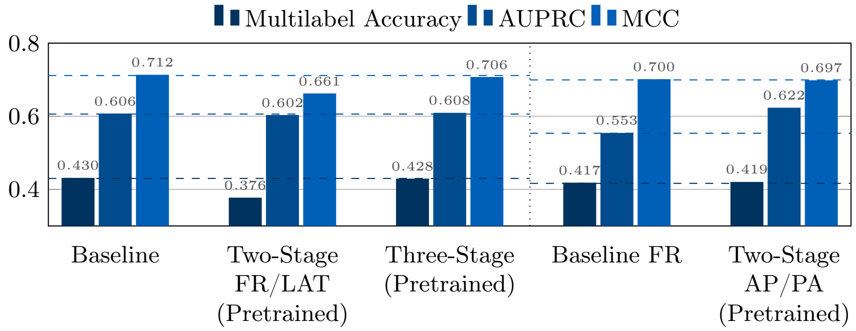

Fig. 3: Comparison Multi-label Metrics: Baseline vs. Multi-Stage Models. Baseline, Two-Stage FR/LAT (pre-trained), and Three-Stage (pre-trained) are tested on the complete test set. Baseline FR, Two-Stage AP/PA (pre-trained) are tested on the frontal test set.

Though improving aggregate metrics was not the primary objective of this work, the 2SM AP/PA showed an AUPRC improvement compared to the FR Baseline model from 0.553 to 0.622. Nevertheless, the performance of the multi-stage models is comparable to the baseline model.

3.2 Single-Label Performance: Baseline vs. Multi-Stage Models

Analyzing the single-label performance of the multi-stage models across the five target conditions, the specialized models slightly outperform the baselines in most conditions. Figure 4 illustrates the AUPRC values for each pathology, comparing the specialized models against the baselines. The multi-stage models achieved higher AUPRC scores in most conditions, with the Two-Stage AP/PA (pre-trained) model reaching an AUPRC of 0.807 for Cardiomegaly detection, outperforming the Frontal Baseline (0.776). For Pleural Effusion, the Three-Stage model demonstrated superior performance with an AUPRC of 0.837 versus the Baseline's 0.815. At the same time, Edema detection performed best with the Two-Stage FR/LAT model, achieving an AUPRC of 0.583.

4 Discussion

4.1 Advantages

The results demonstrate that cascade learning models achieve comparable or slightly superior multi-label performance to traditional "one-model-fits-all" approaches while offering several additional advantages.

The single-label performance of the multi-stage models was slightly better than the baseline models in most conditions. This aligns with the clinical reality,

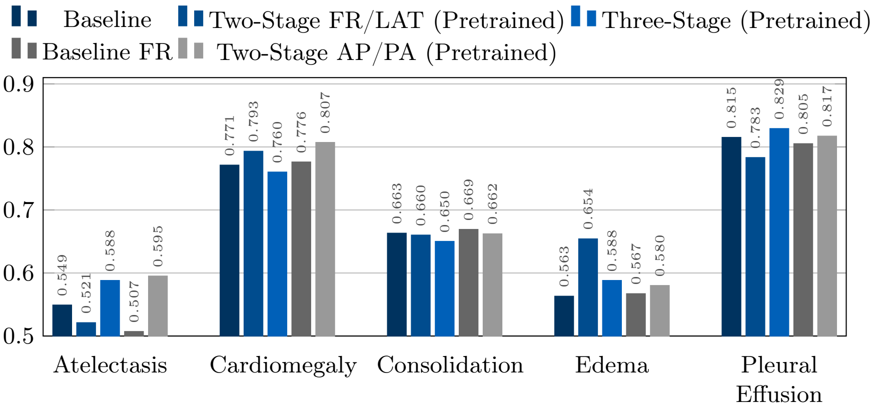

Fig. 4: Comparison Single-Label AUPRC: Baseline vs. Multi-Stage Models. Baseline, Two-Stage FR/LAT (pre-trained), and Three-Stage (pre-trained) are tested on the complete test set. Baseline FR, Two-Stage AP/PA (pre-trained) are tested on the frontal test set.

where certain conditions are better diagnosable in specific views. For example, the Two-Stage AP/PA model outperformed the Frontal Baseline in Cardiomegaly detection. Cardiomegaly assessment is optimally performed in PA view, where cardiac proportions can be more accurately evaluated [7].

Beyond performance metrics, cascade models offer significant efficiency gains. The computation time was much faster for the cascading models. The combined training time of all components in the 2SM $(4 : 58h)$ (classifier $(0 : 24h)$ and two view-specific-models $(1 : 35h + 0 : 24h)$) was almost twice as fast as the Baseline model $(8 : 59h)$. However, this speed improvement may be partially attributed to using ResNet-18 versus ResNet-50 architectures. Additionally, the training of the 2SM can be parallelized, as all models can be trained independently. The training was performed using a NVIDIA A100 GPU (80GB) with 48GB of memory for all models.

Moreover, the modular nature of these models allows for selective retraining of only specific modules after new data acquisition. This is particularly useful in clinical settings, where new data is continuously generated, and retraining the entire model is often infeasible. Also, different architectures can be used for the different view-dependent models, allowing for more specialized models.

4.2 Limitations

This work has several limitations. The unbalanced nature of the CheXpert dataset posed challenges. Using a Weighted Random Sampling and balanced Focal Loss improved the performance, but especially less common views with only a small percentage of positive labels were still challenging to predict.

To keep the results comparable, only models from the ResNet family [4]. Classification results could greatly benefit from more specialized architectures for the view-specific models.

The multi-stage models face the curse of dimensionality. Even if only binary splitting criteria are used, the number of splits increases the number of group-specific models required exponentially. Additionally, for each group-specific, large, representative, and high-quality training data must be available.

With the usage of ResNet-18 for the view and plane classification and the view-specific models, the total number of parameters in the cascade models is a lot higher (3SM: $11.4M \cdot 6 = 68.4M$ parameters, 2SM: $11.4M \cdot 3 = 34.2M$ parameters) than in the ResNet-50 baseline models ($23.9M$). The large imbalance in the total number of parameters between the Baseline and cascading models makes a fully fair comparison difficult.

A key concern with cascaded models is error propagation. An error in the first stage (view classification) will inevitably lead to errors in the subsequent stages. While perfect view classification accuracy was tested in this experiment, this is unlikely to be the case in a more realistic, noisy setting. This could be mitigated by passing the uncertainty of the view classification to the next stage and using it as a weighting factor for the final prediction.

4.3 Outlook

Future models using the cascading approach could benefit from incorporating domain knowledge more explicitly into the model architecture. For example, restricting certain conditions to be predicted only in views where they are clinically visible (e.g., predicting Cardiomegaly only in the PA view) could improve the classification performance. Additionally, developing specialized architectures optimized for each view type rather than using the same architecture across all components could further enhance the performance of the cascade models.

Changing the ResNet-18 for view and plane classification to a smaller model could further decrease the computational cost for training and inference of the cascading model. Finally, balancing the number of parameters of the baseline and cascading models could improve the comparability of the models.

Instead of training all modules of the multi-stage models independently, joint training of all modules could be considered. This could improve the performance of the models, as the view-specific models could learn from the classification model and vice versa.

5 Conclusion

This work demonstrated that cascading models for view-specific classification of chest X-ray images achieve comparable or moderately superior multi-label performance to traditional 'one-model-fits-all' approaches.

Besides, by structuring the classification task into multiple stages, these models can grasp view-specific features more effectively, leading to improved single-label performance, particularly in cases where anatomical differences between views significantly influence disease visibility. This is supported by comparing these models with a baseline model, which follows a 'one-size-fits-all' approach.

Most importantly, however, the results highlight the benefits of cascading models in terms of efficiency, as they offer faster training times, reduced computational costs, and the ability to selectively retrain specific stages without requiring a complete model retraining: advantages particularly relevant to dynamic clinical datasets.

While the approach introduces additional complexity and requires careful handling of potential error propagation, its modular nature provides flexibility for incorporating new pathologies, refining decision thresholds, and adapting to domain shifts with minimal effort.

Future developments could refine this strategy by integrating domain knowledge more extensively and explicitly, such as optimizing architectures for specific views and exploring joint training strategies to mitigate error accumulation.

The findings of this study suggest that cascading models represent a promising direction for medical image classification, balancing predictive performance with practical deployment advantages, particularly in dynamic clinical environments where continuous model improvement is essential.

References

1. Almaraz-Damian, J.A., Ponomaryov, V., Sadovnychiy, S., Castillejos-Fernandez, H.: Melanoma and Nevus Skin Lesion Classification Using Handcraft and Deep Learning Feature Fusion via Mutual Information Measures. Entropy **22**(4), 484 (Apr 2020). https://doi.org/10.3390/e22040484
2. Berrada, L., De, S., Shen, J.H., Hayes, J., Stanforth, R., Stutz, D., Kohli, P., Smith, S.L., Balle, B.: Unlocking Accuracy and Fairness in Differentially Private Image Classification. arXiv (2023). https://doi.org/10.48550/ARXIV.2308.10888, https://arxiv.org/abs/2308.10888
3. Deng, J., Dong, W., Socher, R., Li, L.J., Kai Li, Li Fei-Fei: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009). https://doi.org/10.1109/CVPR.2009.5206848, https://ieeexplore.ieee.org/document/5206848/
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE (2016). https://doi.org/10.1109/CVPR.2016.90, http://ieeexplore.ieee.org/document/7780459/
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 590–597 (2019). https://doi.org/10.1609/aaai.v33i01.3301590, https://ojs.aaai.org/index.php/AAAI/article/view/3834
6. Karar, M.E., Hemdan, E.E.D., Shouman, M.A.: Cascaded deep learning classifiers for computer-aided diagnosis of COVID-19 and pneumonia diseases in X-ray scans. Complex & Intelligent Systems **7**(1), 235–247 (Feb 2021). https://doi.org/10.1007/s40747-020-00199-4

7. Lee, K.H., Choi, J.W., Park, C.O., Han, D.H., Kang, M.S.: A Development and Validation of an AI Model for Cardiomegaly Detection in Chest X-rays. Applied Sciences **14**(17), 7465 (2024). https://doi.org/10.3390/app14177465, https://www.mdpi.com/2076-3417/14/17/7465
8. Mambou, S.J., Maresova, P., Krejcar, O., Selamat, A., Kuca, K.: Breast Cancer Detection Using Infrared Thermal Imaging and a Deep Learning Model. Sensors **18**(9), 2799 (Aug 2018). https://doi.org/10.3390/s18092799
9. Riquelme, D., Akhloufi, M.: Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans. AI **1**(1), 28–67 (Jan 2020). https://doi.org/10.3390/ai1010003
10. Simkus, P., Gutierrez Gimeno, M., Banisauskaite, A., Noreikaite, J., McCreavy, D., Penha, D., Arzanauskaite, M.: Limitations of cardiothoracic ratio derived from chest radiographs to predict real heart size: Comparison with magnetic resonance imaging. Insights into Imaging **12**(1), 158 (2021). https://doi.org/10.1186/s13244-021-01097-0, https://insightsimaging.springeropen.com/articles/10.1186/s13244-021-01097-0
11. Wang, J., Du, X., Farrahi, K., Niranjan, M.: Deep cascade learning for optimal medical image feature representation. In: Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M., Yeung, S. (eds.) Proceedings of the 7th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 182, pp. 54–78. PMLR (2022-08-05/2022-08-06)