INFORMED MACHINE LEARNING FOR CARDIOMEGALY DETECTION IN CHEST X-RAYS: A COMPARATIVE STUDY

Felix Hasse, Florian Leiser, Ali Sunyaev

Karlsruhe Institute of Technology
Department of Economics and Management

ABSTRACT

Recently, computer-aided disease detection from chest radiographs made considerable progress by using convolutional neural networks but issues like insufficient data quality or data availability remain. Informed machine learning (IML) combines domain knowledge and data-driven approaches and has been shown to improve results in many applications. However, there is limited research comparing and combining multiple IML approaches. This paper tackles this issue by implementing, combining, and evaluating three IML approaches for cardiomegaly detection. We find that curriculum learning and cropping images to regions of interest can improve prediction performance. With these results, we provide a reference for both implementing and evaluating multiple IML approaches as well as demonstrating methods to combine IML approaches.

***Index Terms*—** informed machine learning, cardiomegaly, CheXpert, curriculum learning

1. INTRODUCTION

Cardiomegaly, an abnormal enlargement of the heart, is a frequent indicator for other pathologies, like coronary artery disease, myocardial infarction, and ischemia [1]. Previously, clinicians investigated X-rays manually to identify these conditions. Machine learning (ML) and especially convolutional neural networks (CNNs) allowed for considerable progress in early and accurate disease detection from chest X-rays over the last years [2, 3].

However, challenges remain when using CNNs in practice. In particular, the models often perform badly if training data is limited [4]. Informed ML (IML) addresses this and other issues by incorporating external prior knowledge into ML models. Currently, however, practitioners do not know which IML approaches are best suited for their task [5]. Previous research implementing IML approaches only evaluated the approaches individually or compared them to a purely data-driven baseline [4]. Due to differences in knowledge, data, or model architectures [6], it is difficult to compare results across papers. This motivates the need for a detailed comparison between approaches on the same dataset using the same model architecture. Additionally, since different types of knowledge are incorporated into IML models, leveraging different approaches by combination could further improve performance. Therefore, we pose the following research question:

How do different IML approaches and their combination compare to each other?

To answer this question, we implement three different IML approaches and a purely data-driven baseline model and evaluate them on the CheXpert dataset [3]. The first approach is based on calculating the cardiothoracic ratio (CTR). CTR is the ratio of the heart diameter to the internal chest diameter on a posterior-anterior radiograph or computed tomography. Cardiomegaly is present when CTR is ≥ 0.5 [1, 7]. Here, the heart and lung diameters are segmented from X-ray images to predict an abnormal heart-to-lung diameter ratio. The second approach is curriculum learning where models are trained on easier examples first [8]. We use the predicted CTR as a proxy to score the difficulty of training samples. Our third approach investigates only the image region relevant to cardiomegaly diagnosis. To achieve that, we train a CNN on images cropped to bounding boxes extracted by generated lung segmentation masks. Finally, we implement a model combining curriculum learning and image cropping. Afterward, we evaluate all IML approaches' predictive performance for cardiomegaly detection on varying training dataset sizes.

With this study, we demonstrate new ways to combine IML approaches and provide meaningful comparisons between approaches allowing for easier future comparison. Furthermore, the used methods may provide a reference for future research implementing and evaluating IML methods.

2. BACKGROUND

IML is defined as "learning from a hybrid information source that consists of data and prior knowledge" [6]. This contrasts traditional ML approaches, where the model only learns from available data. The prior knowledge needs to stem from an independent source and is given by formal representation [6]. This prior knowledge is often referred to as domain knowledge (DK) [4].

An existing taxonomy of IML differentiates between three sources of knowledge, eight different representations, and four integration possibilities [6]. The sources of knowledge might stem from scientific, world, or expert knowledge, although previous research on IML in medicine found a predominant use of expert knowledge [5]. Different representations include algebraic equations like heart rhythms [9], or spatial invariances such as the proximity of predictions and ground truths [10]. This DK might be included as additional training data, with adapted ML models (either as hypothesis set or as learning algorithms), or as constraints in the final hypothesis.

In our research, we identified three promising approaches incorporating DK into ML models for cardiomegaly detection. The first is CTR calculation, where the diameter of lung segmentation masks is compared to the diameter of heart segmentation masks. Several papers implement automatic CTR calculation using CNNs. One approach used a modified U-Net architecture trained on a dataset of 5,000 images with heart and lung segmentation masks to predict novel masks and subsequently calculate CTR, achieving an accuracy of 95.3% for cardiomegaly detection [11]. A similar approach trained on the JSRT and Montgomery datasets augmented by 50 additional heart masks achieved an accuracy of 69.8% [7].

Various papers implement curriculum learning for cardiomegaly detection. In curriculum learning, ML models are presented with "easy" data points first and "difficult" data points later in the training process [8]. [12] used severity labels from radiology reports for cardiomegaly prediction to rank the difficulty of training samples. Another approach first trained a model on image patches close to lesions, which were considered easier. The model was then fine-tuned on whole images, which was considered a more difficult task [13].

The third approach focuses on regions where cardiomegaly usually occurs. One approach is implementing a two-stream collaborative network where, first, a segmentation model extracting lung regions is trained. The extracted lung regions and whole images are then fed into two distinct branches of the network, whose predictions are then fused for a final prediction [14].

3. METHODOLOGY

3.1. Dataset Description

We used three open-source datasets for our experiments, which all contain a sufficient amount of images. As our main dataset, we relied upon the CheXpert dataset containing 224,316 chest X-rays [3]. The data points contain either positive, negative, or uncertain labels for 14 conditions including cardiomegaly. In line with the results of [3], we found introducing a novel class as the best-performing policy to leverage uncertainty labels for cardiomegaly detection. Further, we used two datasets with segmentation masks. We relied upon the JSRT database of 247 X-ray images [15] in conjunction with heart and lung masks segmented by [16]. In addition, we used the Montgomery County Chest X-ray database containing 138 images with corresponding lung segmentation masks [17, 18].

3.2. Data Augmentation & Processing

Before training the classification models, we resized the images to 320x320 pixels. To ensure comparability between approaches, we only used frontal X-rays, since CTR can only be calculated using those [1]. Because of the limited number of training samples for segmentation masks, we adapted the data processing pipeline of [7] and augmented our datasets by randomly applying the following modifications during training:

- Random rotation between -8 and +8 degrees
- Gaussian blur with kernel size 5, randomly applied with probability $p = 0.3$
- Gaussian noise $N\sim(0, 1)$ randomly applied with $p = 0.3$
- Random horizontal flip of lung masks with $p = 0.5$
- Random image scaling with a factor between 0.7 and 1.3
- Random horizontal or vertical image shift (max. 20%)

Furthermore, we applied histogram equalization to all images to normalize their intensity [7] and resized them to 512x512 pixels. For data post-processing, we adopted the pipeline of [7] and applied binary erosion followed by binary dilation to the outputs to fill holes in the masks [7]. We then designated the largest two connected components from the lung segmentation model's output as the lung masks and discarded the rest [7]. For heart segmentation, we selected the largest connected component and designated it as the heart.

3.3. Implementation & Hyperparameter Optimization

We implemented our approaches using PyTorch [19] and trained all models on the high-performance computing BwUniCluster2.0. We tuned the learning rate and batch size for all models with the Optuna library [20] using a tree-structured Parzen estimator algorithm [21]. The search space for the learning rate spanned from 10^{-6} to 10^{-2}, with batch sizes ranging from 8 to 256 for the classification model and from 1 to 64 for the segmentation models. Each model ran 25 trials which resulted in the optimal parameter shown in Table 1.

Model	Learning rate	Batch size
DenseNet-121	7.3261e-05	42
Heart Segmentation	2.9436e-04	2
Lung Segmentation	3.5304e-05	5

Table 1. Best hyperparameters found

4. MACHINE LEARNING MODELS

4.1. Classification Model

We first implemented a purely data-driven baseline model as an evaluation reference. We adopted DenseNet-121 pretrained on ImageNet as baseline model architecture due to previous performance evaluations [3].

We extended the DenseNet-121 implementation provided by PyTorch [19] by replacing the classification layer with a linear layer with three output neurons, corresponding to the positive, negative, and uncertain labels, with a softmax activation function. We used cross-entropy loss and an Adam optimizer with the default parameters $\beta_1 = 0.9,$ and $\beta_2 = 0.999$. Building on previous approaches [3], we trained the model for three epochs, which we found to be sufficient in initial tests.

We applied an 85-14-1 split for the training, hyperparameter validation, and checkpoint validation set. As established by [3], we trained each model three times, saved model checkpoints every 4,800 images, and evaluated the checkpoints on a holdout validation set. For each run, we selected the ten best checkpoints and calculated the final predictions by averaging the predictions of the resulting 30 model checkpoints to rely on the best-performing checkpoints. For the baseline classification model, we achieved an AUC of 0.8504, which is in line with the results reported by [3].

4.2. Segmentation Model

Since DenseNet-121 does not provide image segmentation, we trained two separate U-Nets with VGG-16 encoder predicting heart or lung segmentation masks. We based our model on the implementation by [7] and adopted the code of [22]. Our loss function is a combination of binary cross-entropy with logits loss and soft dice loss [7]. Further, we used an Adam optimizer [23] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and applied a 70-10-10-10 split for the training set, checkpoint validation, hyperparameter validation, and test set. The models were trained three times for 300 epochs each. We saved model checkpoints after every epoch and evaluated the performance by using the intersection over union metric (IoU). We selected the best-performing model checkpoints and obtained the final prediction as described in the baseline training procedure. We achieved an IoU of 0.9623 for the lung and 0.9092 for the heart segmentation model.

5. INFORMED MACHINE LEARNING MODELS

With plenty of domain knowledge available for cardiomegaly detection, we implemented three IML approaches. First, we replicated the results of [7] using scientific knowledge by determining CTR directly based on heart and lung segmentation masks. Second, we used CTR to implement a curriculum learning approach, where we showed easier images to the model first, before presenting difficult images and, thereby, included feedback on the difficulty. Last, we cropped the provided images to regions of interest by using lung segmentation masks, thus making use of expert knowledge about regions of frequent cardiomegaly presence.

5.1. Automatic CTR Calculation

To determine CTR and thereby replicate the results of [7] we used the segmentation models. After predicting the segmentation masks, we calculated the internal chest diameter by measuring the horizontal distance in the segmentation masks between the leftmost point of the left lung and the rightmost point of the right lung. The heart diameter was calculated by measuring the horizontal distance between the leftmost and rightmost points of the heart segmentation mask. We then used both values to calculate CTR [7] and classified samples with a CTR greater than or equal to 0.5 as positive for cardiomegaly and others as negative. With this approach, we achieved an AUC of 0.7630. Using the 0.5 threshold for CTR, sensitivity was 0.7879 and specificity was 0.5376.

5.2. Curriculum Learning

We implemented a curriculum learning model relying on CTR as a proxy for training sample difficulty, as a larger CTR generally indicates a severe manifestation of cardiomegaly that is easier to detect. We used the predicted CTR calculations to split the dataset into easy, medium, and hard subsets by considering the accuracy of CTR calculation and the predicted disease severity. We classified samples with uncertain labels or incorrect predictions made based on CTR as hard. Additionally, we labeled all images with CTR > 0.8 or CTR < 0.2 as hard, since extreme CTR values may suggest inaccurate segmentation. We classified the remaining samples as medium if $0.4 <$ CTR < 0.6 and as easy if CTR > 0.6 for positive samples and CTR < 0.4 for negative samples.

To avoid bias, we balanced the subsets to achieve the same ratio of positive to negative samples in each subset. This was done by moving samples from easier to harder subsets until reaching consistent ratios. The resulting dataset sizes were 17,659 for the easy, 50,215 for the medium, and 93,661 data points for the hard subset. We used these subsets to train a classification model employing the same architecture and parameters as the baseline model. We trained our model on the easy subset first and then successively introduced the medium and hard subset. The best strategy was training for 3 epochs on each subset while keeping easier samples from previous sets. With these variations, we found that predictive performance was higher compared to the baseline model, with an AUC of 0.8592.

5.3. Cropping Images to Region of Interest

The third approach relied on extracting lung bounding boxes from the predicted lung masks. We cropped all images to

these regions of interest before resizing them to 320x320 pixels and feeding them into the model. Since cardiomegaly is an abnormal enlargement of the heart compared to the lungs, only the area inside the lung bounding boxes should be relevant for detecting the condition. We hypothesized that excluding irrelevant portions of the X-ray images could improve results, especially with limited training data. We used the pretrained DenseNet-121 with the previously described parameters. With the cropped images, we found that the AUC was 0.8523 and thus marginally higher than the baseline.

5.4. Curriculum Learning on Cropped Images

Based on these results, we also trained the curriculum learning approach on the cropped images to evaluate performance. We achieved an AUC of 0.8495, indicating that combining these IML approaches yields no benefit over using the approaches individually.

5.5. Comparison on Different Dataset Sizes

We evaluated the baseline model and the IML models on the CheXpert test set [3]. To allow a constant number of images shown to each network, we adjusted the number of epochs for varying dataset sizes. The results are shown in Figure 6.1. We found that in most instances, curriculum learning and cropping images outperformed the baseline, with stronger effects on limited training data. We again found no benefit in combining two approaches. Overall, curriculum learning achieved the best results. While performing worse than the other models on the full dataset, automatic CTR calculation achieved good performance relative to the limited dataset size.

6. DISCUSSION AND CONCLUSION

6.1. Principal Findings

In this work, we compare three IML approaches incorporating different forms of domain knowledge for cardiomegaly detection and modify these approaches to allow combination.

We evaluate DenseNet-121 as a purely data-driven baseline model on the CheXpert dataset and replicate previous results [3]. The first IML approach automatically calculates CTR using heart and lung segmentation masks and uses CTR to predict the presence of cardiomegaly. We find worse performance than the baseline model using all data, but comparable performance for smaller dataset sizes. Second, we combine CTR calculation with curriculum learning to split the dataset into three distinct difficulty levels where the easiest level is presented to the model first. This approach outperforms the baseline model in most circumstances. Third, we train a classification model on images cropped to lung bounding boxes which achieves slightly better performance than the

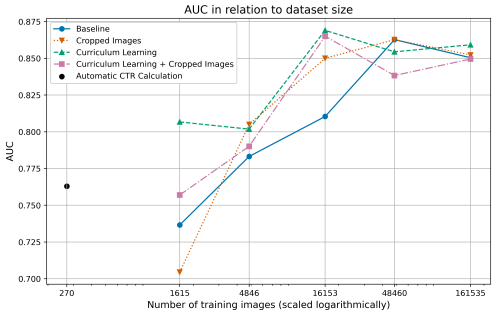

Fig. 1. AUC for the different models in relation to number of training images

baseline model. Finally, we evaluate a model combining curriculum learning and cropping images but don't find improvements compared to the baseline for this approach.

Overall, we see that IML approaches incorporating domain knowledge into ML models can improve classification performance for cardiomegaly detection, especially with limited training data. Using CTR as a proxy for the difficulty of images shows the most promising results and may provide a reference for future research on combining IML and curriculum learning. Improved performance with cropping images to lung bounding boxes suggests that extending the approach to detect further thoracic diseases may be beneficial [12]. Finally, our classification and evaluation of IML approaches may serve as a reference for future researchers investigating and comparing IML approaches.

6.2. Limitations & Future Research

There are several limitations in our work that could be addressed by future research. First, although our segmentation models achieve good performance on the test dataset, the performance of cardiomegaly detection using CTR is rather low suggesting insufficient generalization of the model. One likely cause is the limited size of the used segmentation datasets as other approaches included manually segmented masks [7, 11]. Therefore, future research could use additional data to improve results across all IML approaches. The comparatively low performance of the automatic CTR calculation may impact the difficulty scores for curriculum learning. Therefore, future research could explore alternatives to extract difficulties.

Additionally, we encourage future studies to extend the evaluation to either compare our approach with state-of-the-art methods or with qualitative evaluation (e.g., by investigating the perception of the approaches by practitioners).

7. COMPLIANCE WITH ETHICAL STANDARDS

This is a computational simulation study based on publicly available datasets for which no ethical approval was required.

8. ACKNOWLEDGEMENTS

No funding was received for conducting this study. The authors have no financial or non-financial interests to disclose.

9. REFERENCES

[1] H. Amin et al., "Cardiomegaly," in *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2022.

[2] C. Qin et al., "Computer-aided detection in chest radiography based on artificial intelligence: a survey," *BioMedical Engineering OnLine*, vol. 17, no. 1, pp. 113, Aug. 2018.

[3] J. Irvin et al., "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590–597, July 2019.

[4] X. Xie et al., "A Survey on Incorporating Domain Knowledge into Deep Learning for Medical Image Analysis," *Medical Image Analysis*, vol. 69, pp. 101985, Apr. 2021.

[5] F. Leiser et al., "Medical informed machine learning: A scoping review and future research directions," *Artificial Intelligence in Medicine*, vol. 145, pp. 102676, 2023.

[6] L. von Rueden et al., "Informed Machine Learning – A Taxonomy and Survey of Integrating Knowledge into Learning Systems," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[7] I. Chamveha et al., "Automated Cardiothoracic Ratio Calculation and Cardiomegaly Detection using Deep Learning Approach," Feb. 2020, arXiv:2002.07468 [cs, eess].

[8] Y. Bengio et al., "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, Montreal, Quebec, Canada, 2009, pp. 1–8, ACM Press.

[9] P. Lu et al., "KecNet: A light neural network for arrhythmia classification based on knowledge reinforcement," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–10, Apr. 2021.

[10] J. Lian et al., "A structure-aware relation network for thoracic diseases detection and segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 8, pp. 2042–2052, 2021.

[11] Z. Li et al., "Automatic Cardiothoracic Ratio Calculation With Deep Learning," *IEEE Access*, vol. 7, pp. 37749–37756, 2019.

[12] Y. Tang et al., "Attention-Guided Curriculum Learning for Weakly Supervised Classification and Localization of Thoracic Diseases on Chest Radiographs," in *Machine Learning in Medical Imaging*, Y. Shi et al., Eds., Cham, 2018, Lecture Notes in Computer Science, pp. 249–258, Springer International Publishing.

[13] B. Park et al., "A Curriculum Learning Strategy to Enhance the Accuracy of Classification of Various Lesions in Chest-PA X-ray Screening for Pulmonary Abnormalities," *Scientific Reports*, vol. 9, no. 1, pp. 15352, Oct. 2019.

[14] B. Chen et al., "Two-stream collaborative network for multi-label chest X-ray Image classification with lung segmentation," *Pattern Recognition Letters*, vol. 135, pp. 221–227, July 2020.

[15] J. Shiraishi et al., "Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule: Receiver Operating Characteristic Analysis of Radiologists' Detection of Pulmonary Nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, Jan. 2000.

[16] B. van Ginneken et al., "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Medical Image Analysis*, vol. 10, no. 1, pp. 19–40, Feb. 2006.

[17] S. Jaeger et al., "Automatic Tuberculosis Screening Using Chest Radiographs," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 233–245, Feb. 2014.

[18] S. Candemir et al., "Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 577–590, Feb. 2014.

[19] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Dec. 2019, arXiv:1912.01703 [cs.LG].

[20] T. Akiba et al., "Optuna: A Next-generation Hyperparameter Optimization Framework," July 2019, arXiv:1907.10902 [cs.LG].

[21] J. Bergstra et al., "Algorithms for Hyper-Parameter Optimization," in *Advances in Neural Information Processing Systems*. 2011, vol. 24, Curran Associates, Inc.

[22] Z. Zhou, "PyTorch-Unet," 2019, GitHub repository, https://github.com/zhoudaxia233/PyTorch-Unet.

[23] D. P. Kingma et al., "Adam: A Method for Stochastic Optimization," Jan. 2017, arXiv:1412.6980 [cs.LG].