Automated Detection of Cardiomegaly from Chest X-Ray Images

Khush Patel
School of Biomedical Informatics
UT Health Science Center at Houston
Houston, Texas, USA
Khush.a.patel@uth.tmc.edu

***Abstract*— Background:** Pattern recognition of different diseases from medical imaging studies using deep learning is evolving rapidly, with some algorithms performing better than expert radiologists in identifying these diseases. One area where deep learning algorithms could improve clinical workflows in a variety of medical settings is in automated cardiomegaly detection from chest X-ray images. Therefore, we developed and evaluated a series of deep learning algorithms for the classification of cardiomegaly from chest X-ray images. **Methods:** A subset of patients from the NIH Chest X-ray dataset consisting of positive (cardiomegaly) and negative (no cardiomegaly) samples were used to develop and validate deep learning classification algorithms. After image preprocessing, a variety of models with different architectures and parameters were constructed, evaluated, and compared. Model performance was predominantly measured via the area under the receiver operating characteristic curve (AUC), but sensitivity, specificity, F1 score, and Matthews correlation coefficient were also examined. Model interpretability was investigated using Grad-CAM. **Results:** Using independent training (N=21386) and test (N=600) sets, seven different deep learning models were developed and compared. Most of the models performed well in detecting cardiomegaly. The best model had an AUC of 0.905 in predicting cardiomegaly. Grad-CAM revealed models tended to focus on areas of the image containing cardiac tissue when classifying a case as having cardiomegaly. **Conclusion:** A pattern recognition algorithm for cardiomegaly classification was built using deep learning neural networks which showed good performance in identifying the presence of cardiomegaly from chest X-ray images. Model interpretability methods demonstrate findings consistent with human intuition. Integrating this pattern recognition algorithm in the workflow of X-ray image ordering and reporting may allow for early and accurate detection of cardiomegaly in different medical settings including primary and preventive healthcare.

Keywords—deep learning, image classification, cardiomegaly, X-ray

I. INTRODUCTION

Cardiomegaly is commonly suspected on chest X-ray images, in which the ratio of the widest transverse diameter of the heart to the widest diameter of the thoracic cage, i.e., cardiothoracic ratio, exceeds 50% [1], [2]. As chest X-rays are commonly ordered in medical practices, the development of an automated cardiomegaly detection tool could support healthcare providers in different settings to accurately recognize this warning sign in a timely fashion.

The application of machine learning and deep learning approaches to chest X-ray data provides an opportunity for accurate and improved detection of medical abnormalities, such as cardiomegaly [3]. Multiple studies have previously demonstrated the feasibility of automatic interpretation of chest X-rays using machine learning methods to predict the presence or absence of different abnormalities [4][5]. While many machine learning approaches could be used to design and train a pattern recognition algorithm to identify cardiomegaly in chest X-rays images, deep learning approaches show particular promise due to their excellent performance on a variety of classification tasks [6].

To determine the feasibility of using deep learning approaches for cardiomegaly classification, we developed multiple deep learning models and investigated their performance quantitatively and through examination of model predictions using interpretability approaches.

II. METHODS

A. Study Cohort

A subset of the "Chest X-ray" dataset was used for the development of the pattern recognition algorithm. The "Chest X-ray" dataset is an NIH database which consists of more than 112,120 frontal-view X-ray images for 30,805 unique patients [7]. The images contain annotations corresponding to labels of eight different diseases derived from natural language processing. For this study, only healthy patients or patients with cardiomegaly were extracted from the original Chest X-ray dataset.

B. Imaging Dataset and Preprocessing

In total, 21986 unique chest x-rays were provided in .png format along with corresponding clinical and demographic details in a .csv file. Data was previously partitioned into distinct training (N=21386) and test (N=600) sets. Variables in the clinical and demographic sheet of interest included age, cardiomegaly status, gender, and view orientation (posteroanterior [PA] vs. anteroposterior [AP]). As a pre-processing step, the age variable was converted into an integer by removing the trailing letter from the entry (i.e., "Y", "M", "D"). For age entries that contained "M" or "D", which corresponded to month or day, respectively, the value was rounded down to 0, thereby transforming all entries into years represented in integers. No preprocessing was applied to the other variables. To determine any differences in variables between the training and testing set, for continuous data we applied an omnibus test for normality [8] followed by a Mann-Whitney U test [9] if data were non-normally distributed. For categorical variables, we applied chi-squared tests [10]. Scipy v. 1.4.1 was used for all statistical tests [11].

The intensity of all images was scaled such that all images were in the range of [0,1]. Additionally, contrast limited adaptive histogram equalization (CLAHE) and histogram normalization were investigated in one sample patient, and the most appropriate method as determined by manual visualization was used for all further models. The CLAHE method was used with a (8,8) grid size. The histogram normalization method used a bin number of 256.

C. Data Augmentation

As part of the MONAI software package [12], we applied random rotations and random zooms to the training data to both amplify the number of training samples and increase model generalizability. These transformations were chosen based on the plausibility of the resulting images to be represented in the unseen test set, ie., realistic modifications of images that could be represented in chest x-ray datasets.

D. Models

A wide number of techniques were developed to optimize the training of the classifier. Different approaches were used to handle the high imbalance in the training set. We divide our approaches into three categories: I. direct classifier, II. proxy task-based classifier, and III. dual supervised and unsupervised loss-based classifier.

I. Direct classifier: We developed a classifier with the main task of identifying cardiomegaly (positive) or no cardiomegaly (negative). To address the dataset imbalance issue, we used four approaches:

1. Using a weighted cross entropy loss. We implemented class weights in a ratio of 44.75:1 (cardiomegaly: no cardiomegaly distribution in the training data) for the loss function.
2. Oversampling the cardiomegaly class with selective heavy augmentation for the cardiomegaly class.
3. Under sampling the non-cardiomegaly class.
4. Combining the oversampling and weighted loss approach.

We used Densenet 121 and Resnet 101 for the above approaches. Densenet121 [13] was selected due to its general ubiquity, high performance in similar medical image classification problems [14], and simple implementation. Resnet101 [15] was chosen since it is among the most cited deep learning architecture, with residual connections that improve performance for the deeper networks. The DenseNet model was developed using the MONAI software package [12], while the Resnet model was implemented from Pytorch [16].

II. Proxy task based classifier: Since an increase in heart size has been associated with aging [17], we hypothesized that developing a regressor to predict age based on chest X-Rays could enable the network to learn features to identify differences in the heart size. Importantly, age data was not imbalanced, so no corrections were applied to balance data. Once our regressor was built with the task of predicting age, we freeze the weights of all except the final convolution block consisting of three repeating layers of convolution followed by batch normalization and a Relu activation function with residual connections. The top regressor layer was replaced by a classifier. This allowed the training of the top convolutional block and classification layer. The network was fine tuned on the undersampled balanced dataset for the cardiomegaly classification task with a reduced learning rate and L2 regularization. An overview of the proxy task classifier architecture is shown in **Fig. 1**.

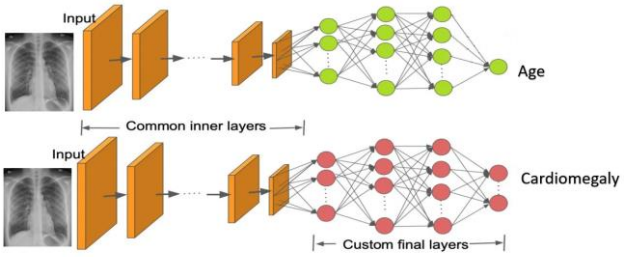

Fig. 1. Outline of proxy task classification method.

III. Dual supervised and unsupervised loss based classifier: We developed an autoencoder based architecture to train the model for two tasks: 1. Unsupervised image reconstruction, and 2. Supervised classification task of cardiomegaly. For the unsupervised image reconstruction, we used a mask excluding 0 values while training to not include background and the lungs in the mean square error loss function. For the supervised classification task, we used a weighted cross entropy loss. An overview of the architecture is shown in **Fig. 2**. Reconstruction results for a few select cases are shown in **Fig. 3**.

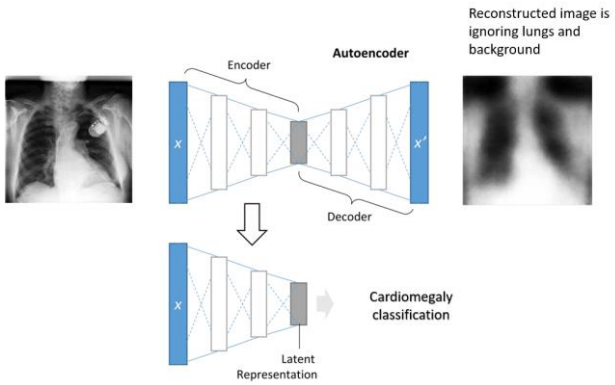

Fig. 2. Outline of the dual supervised and unsupervised loss based classifier.

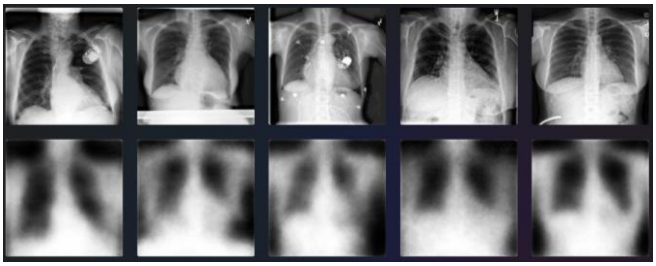

Fig. 3. Examples of reconstructed images for dual supervised and unsupervised loss-based classifier. The image reconstruction results ignore the lungs and the background.

E. Training Strategy

10% of the training data (N=2137) was randomly allocated to a validation set for model validation. We used Adam and stochastic gradient descent with and without

momentum and Nesterov momentum. We used L2 regularization when needed based on the training and validation loss curves. We tuned for hyperparameters like learning rate, weight decay, number of epochs, and loss weights. **Fig. 4** shows learning rate curves we obtained to tune the learning rate by changing learning rates successively over small intervals and observing the change in the loss function. We used learning rate schedulers to reduce the learning rate when the loss plateaued as shown in **Fig. 5**. We also used learning rate schedulers to vary the learning rate over one single epoch in cyclic fashion.

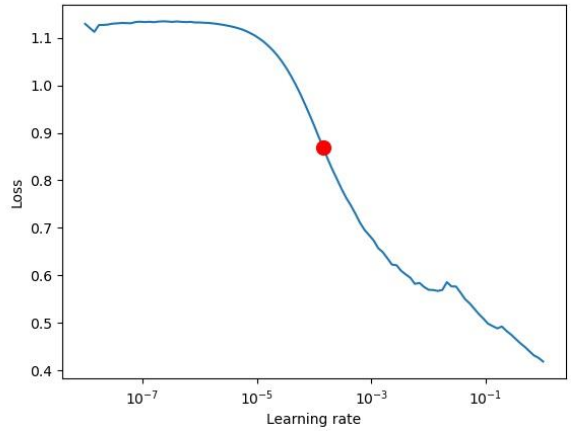

Fig. 4. Learning rate curves used to tune the learning rate.

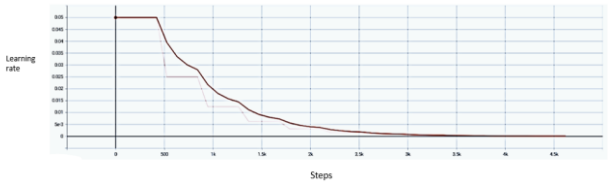

Fig. 5. Effect of learning rate scheduler on learning rate.

F. Method of Measuring Performance

Due to the highly imbalanced nature of our classification problem, we predominantly used the area under the receiver operator characteristic curve (AUC), i.e., the probability a randomly chosen positive instance (cardiomegaly) will be ranked higher than a randomly chosen negative instance (no cardiomegaly) [18], as the primary method of evaluating model classification performance. We also utilized other metrics to characterize additional facets of model performance, such as sensitivity, specificity, F1 score, and Mathew's correlation coefficient (MCC) since they could provide information about other facts of model performance.

G. Model Interpretability

Grad-CAM [19], a technique to highlight important regions in an image used for a prediction, was used to investigate representative predictions of the best models. Grad-CAM localization maps were superimposed on images of patients for the following 4 categories: correctly classified as cardiomegaly, correctly classified as no cardiomegaly, incorrectly classified as cardiomegaly, incorrectly classified as no cardiomegaly. A subset of 5 patients for each of the 4 categories were randomly selected from the test set. We utilized a Grad-CAM implementation previously developed for the MONAI software package [12].

H. Computational Resources

We used Google Colab computational resources for this project. GPUs utilized included a Tesla P100-PCIE-16GB.

IV. Results

A. Image Pre-Processing

An example of different normalization schemes are demonstrated for one patient in **Fig. 6**. The histogram equalization technique was selected for use in model development because it seemingly removed all noise except in the cardiac region, i.e., pulmonary opacities were not emphasized.

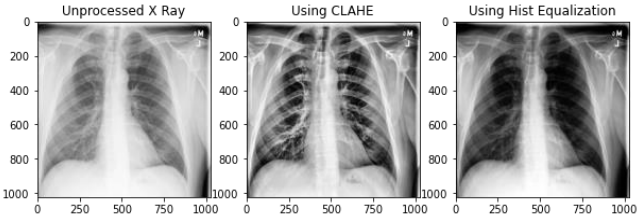

Fig. 6. Different methods of image normalization.

B. Patient Demographics

Using an omnibus test for normality revealed the age variable was not normally distributed, therefore we applied a nonparametric test (Mann Whitney U test) which revealed a significant difference in age between the training and test set ($p<0.05$). Chi-squared tests were applied to the categorical variables cardiomegaly, gender, and view orientation, which returned significant ($p<0.05$), non-significant ($p>0.05$), and non-significant ($p>0.05$) results, respectively. Full patient demographics and corresponding statistics are shown in **Table 1**.

TABLE 1. Patient demographics characteristics of training and test sets.

Characteristic	***Training Set (N=21386)***	***Testing Set (N=600)***	***p-value****
Age (median, IQR)	46.00 (25.00)	49.00 (26.00)	2.57x10-5
% Cardiomegaly	2.19	50.00	0.00**
% Male	53.50	49.67	0.07
% PA View***	89.39	88.33	0.45

* p-value for age corresponds to Mann Whitney U test. p-values for % Cardiomegaly, %Male, and %PA View correspond to chi square test. $p < 0.05$ indicates significantly different between training and test set.

** p-value of 0.00 returned from scipy.stats.chi2_contingency due to maximum precision threshold being crossed.

*** Patients were either in posteroanterior (PA) or anteroposterior view.

C. Model Performance

The performance of the different deep learning models is shown in **Table 2**. The best models in terms of AUC from best to worst were: Under sampled Resnet101 > DenseNet121 > Weighted Loss Resnet101 = Convolutional autoencoder classification and image reconstruction > Oversampling Resnet101 > Proxy task age >> Proxy task gender. Subgroup analysis (AP vs. PA view) is shown in the **Supplementary Materials.**

TABLE 2.. MODEL PERFORMANCE ON TEST SET. AUC = AREA UNDER THE RECEIVER OPERATOR CHARACTERISTIC CURVE; MCC = MATHEW'S CORRELATION COEFFICIENT.

Model	**AUC**	**Sensitivity**	**Specificity**	**F1 score**	**MCC**
Direct Classifiers					
DenseNet121	0.879	0.780	0.810	0.792	0.590
Under sampled, Resnet101	0.905	0.823	0.793	0.808	0.617
Weighted Loss, Resnet101	0.872	0.280	0.977	0.628	0.58
Oversampling, Resnet101	0.834	0.490	0.901	0.698	0.437
Dual supervised and unsupervised loss-based classifier					
Convolutional autoencoder, classification and image reconstruction.	0.872	0.690	0.850	0.770	0.547
Proxy task-based classifiers					
Proxy task age	0.807	0.717	0.733	0.725	0.450
Proxy task gender	0.471	0.473	0.483	0.478	-0.043

D. Model Interpretability

Correctly classified cases of cardiomegaly and no cardiomegaly are shown in **Fig. 7**. As can be seen, the model tended to focus on non-cardiac structures (often adjacent to the heart) for the cases without cardiomegaly while oppositely it focused on cardiac structures for cases with cardiomegaly. Analogously, incorrectly classified cases of cardiomegaly and no cardiomegaly are shown in **Fig. 8**. The results demonstrate for the misclassified cases the model tended to focus on non-cardiac areas that would not traditionally be associated with detection of cardiomegaly by human observers, e.g., some predictions focused on areas of background. Other methods for model interpretability are displayed in the **Supplementary Materials.**

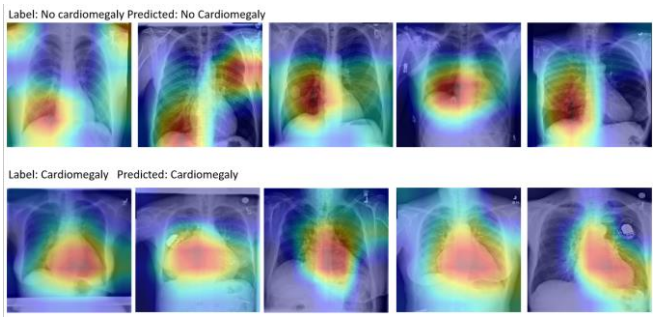

Fig. 7. Grad-CAM visualization of correctly classified cases. Redder colors indicate higher importance towards prediction while bluer colors indicate lower importance towards prediction.

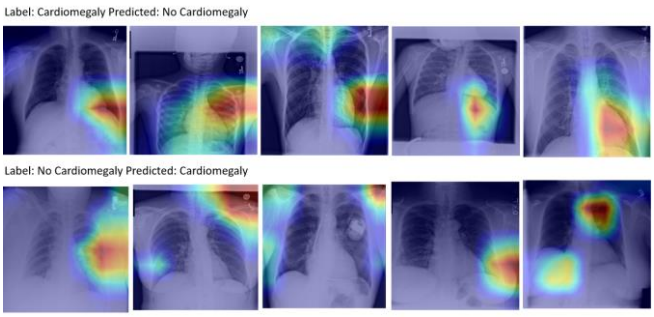

Fig. 8. Grad-CAM visualization of incorrectly classified cases. Redder colors indicate higher importance towards prediction while bluer colors indicate lower importance towards prediction.

V. DISCUSSION

Detecting cardiomegaly on chest X-ray images can be a challenging and time-intensive task as specific measurements need to be verified for proper diagnosis of this condition. Moreover, inter-observer variability is a well-known caveat of manual interpretation of medical images [20]. Automated approaches such as those afforded by deep learning enable the minimization of variability while maximizing classification performance. In this study we examined the performance of deep learning neural networks to correctly classify chest X-ray images for the presence or absence of cardiomegaly.

A subset of the NIH "Chest X-ray" dataset with annotations for cardiomegaly status was used to develop a pattern recognition algorithm for cardiomegaly classification. Our best model had an AUC of 0.905 in predicting cardiomegaly which corresponded to the Under sampled Resnet101 approach. While our model offers reasonable performance overall as shown through AUC, other metrics are relevant for model evaluation as well. Specifically, depending on the use-case, sensitivity or specificity may be deemed more important [21]. Under traditional circumstances for our domain application, a "true positive" would be considered more valuable than detecting a "true negative", as the presence of cardiomegaly can be potentially life threatening and must be intervened in a

speedy and efficient manner. Therefore, sensitivity would likely be an important metric for our detection model. Importantly, our Under sampled Resnet101 also offers reasonable sensitivity with a value of 0.823. The lowest performance model was the Proxy task gender model, which implies gender does not contain informative features for cardiomegaly prediction. Interestingly, most of the other models investigated demonstrated comparable performance, with AUCs ranging between 0.807 to 0.879.

Multiple previous studies examined the role of machine learning in identifying cardiomegaly from chest X-rays with several using the same database investigated in this study. In a cross-sectional study by Bougias et al., different transfer learning models were developed to identify the presence of cardiomegaly from chest X-rays by extracting 2048 deep features using different networks [22]. The best model based on the VGG19 network achieved an overall accuracy of 84.5% with high sensitivity and specificity. The other three methods (Inception V3, VGG16, and SqueezeNet) had lower overall accuracy ranging from 71% to 81.3%. A major strength of this study was the balanced number of cases and controls (1000 normal and 1000 cardiomegaly). A different study used convolutional neural networks based on U-Net after implementing adaptive histogram equalization and creating masks for cardiomegaly, a pipeline that yielded a diagnostic accuracy of 93% [23]. Importantly, our results are comparable to these existing investigations.

Interestingly, in a previous study the performance of deep learning models in detecting cardiomegaly was better in PA views than AP views [24]. As shown in our supplementary materials, a sub-group analysis based on AP vs PA view predictions showed no differences in model performance based on the X-ray view. We believe in our study, the low number of patients with cardiomegaly for each view contributed to the lack of consistency between our sub-group analysis and the previously established results.

Determining the severity of cardiomegaly using deep learning approaches was also examined in previous studies and showed that the accuracy of the classification increased with the severity of the disease [25]. For our prediction algorithm, the misclassification could be in patients with mild or borderline cardiomegaly. Severity labels were not available for our current analysis. However, future studies should investigate the incorporation of severity annotations generated by an expert radiologist to determine if our model performs better in severe cases when compared to mild or borderline cases.

As heart and lung segmentation annotations were difficult to obtain for this dataset, we only relied on an end-to-end deep learning approach to develop our pattern recognition algorithm. However, Sogancioglu et al. showed that an anatomical segmentation-based model is superior in terms of performance when compared to image-level classification (ResNet18, ResNet50, and DenseNet121) [26]. Interestingly, of the models used in the image-level classification, the ResNet model had the best performance, similar to what we observe in our data. A future research direction for our models could be the incorporation of segmentation information to determine if they improve predictive performance.

Deep learning models traditionally demonstrate superior performance when compared to classical machine learning frameworks for imaging classification tasks. However, a common caveat of deep learning approaches for medical image applications has been their “black box nature” [27]. In this study, we have investigated post-hoc model interpretability methods (i.e., Grad-CAM) to determine what models deem important for classification. Using these methods, we reveal our models prioritize areas of images consistent with human intuition, e.g., cardiac silhouette emphasized for cases with cardiomegaly. While our observations are not necessarily surprising, interpretability methods such as these are critical to foster increased transparency and trust in automated detection tools for their eventual clinical implementation. Moreover, our results for cases without cardiomegaly demonstrate a focus on regions often adjacent to but not containing the heart, which are interesting and may need to be investigated further. Finally, incorrectly classified cases often focused on areas completely non-informative for predictions, such as background, which likely explains their poor performance.

Certain limitations accompanied our study. First, the training and testing sets originated from the same database which may have limited the generalizability of our study. As the data were collected from the NIH Clinical Center and the personally identifiable information was removed, the use of our pattern recognition algorithm needs to be further validated in different settings/populations before it can be safely implemented and clinically used. Second, the training set was highly imbalanced towards healthy individuals with only a small number of images (300 out of the 21386 [1.40%]) being for patients with cardiomegaly. Though we have taken steps to address this imbalance during model training, the imbalance may have still negatively affected the learning step of the models in the development phase. Third, the presence or absence of cardiomegaly was labeled based on natural language processing of the imaging reports. Although natural language processing is an effective approach for such big data, we expect some of the images were mislabeled, which may have altered the accuracy of our approach. Finally, we have limited our analysis to the use of chest X-ray images since these are among the most ubiquitous image sources used for identifying cardiomegaly and were readily available from the NIH Chest X-ray database. However, alternative datastreams, such as ultrasound and magnetic resonance imaging, which are commonly used in cardiothoracic measurements [28], should be investigated as well.

VI. Conclusion

In summary, using large scale datasets and open-source tools, we have investigated the utility of deep learning approaches for classification of cardiomegaly using chest X-ray images. Our results show that most models accomplish this classification task with reasonable AUC performance. Our best model was a ResNet approach implementing under sampling of the training data, which led to an AUC value of 0.905. Importantly, we implement Grad-CAM as a model interpretability approach and demonstrate models that focus on cardiac tissue for classification. Our results are an important preliminary step towards the eventual clinical implementation of pattern recognition tools for cardiomegaly detection. Future studies should investigate

the utility of additional image modalities for cardiomegaly classification.

Supplementary Materials

Alternative interpretability techniques: We briefly investigated a variety of post-hoc interpretability techniques for prediction of cardiomegaly. These results are shown below.

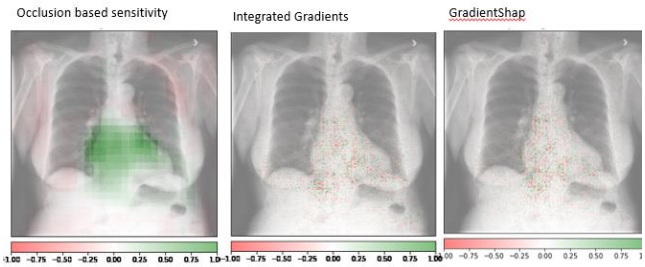

Fig 9. Original label and prediction Cardiomegaly

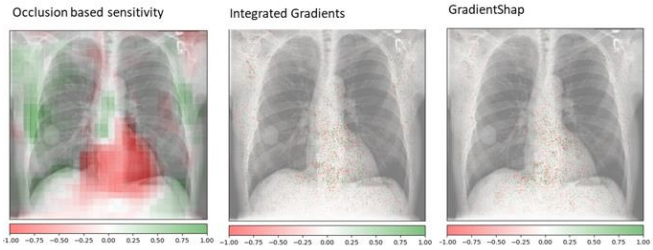

Fig 10. Original label and prediction: No Cardiomegaly

Sub-group analysis: As a test to evaluate if models had a preference towards a specific subset of images in the test set, we performed a subgroup analysis based on AP vs. PA view predictions. The number of correctly classified PA vs. AP view images in the test set were quantified and then compared through chi squared tests to determine if any preference was given to a certain view. The DenseNet model classified 419 PA correctly, 58 AP correctly, 111 PA incorrectly, and 12 AP incorrectly (p=0.56), therefore there was no difference in model performance based on the X-ray view.

References

[1] M. J. S. Zaman *et al.*, "Cardiothoracic ratio within the 'normal' range independently predicts mortality in patients undergoing coronary angiography," *Heart*, vol. 93, no. 4, pp. 491–494, Apr. 2007.

[2] Y. B. Mensah *et al.*, "Establishing the Cardiothoracic Ratio Using Chest Radiographs in an Indigenous Ghanaian Population: A Simple Tool for Cardiomegaly Screening," *Ghana Med. J.*, vol. 49, no. 3, pp. 159–164, Sep. 2015.

[3] M. Arsalan, M. Owais, T. Mahmood, J. Choi, and K. R. Park, "Artificial Intelligence-Based Diagnosis of Cardiac and Related Diseases," *J. Clin. Med. Res.*, vol. 9, no. 3, Mar. 2020, doi: 10.3390/jcm9030871.

[4] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification," *Sci. Rep.*, vol. 9, no. 1, p. 6381, Apr. 2019.

[5] P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Med.*, vol. 15, no. 11, p. e1002686, Nov. 2018.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[7] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[8] R. B. DIAgostino, "An omnibus test of normality for moderate and large sample sizes," *Biometrika*, vol. 58, no. 34, pp. 1–348, 1971.

[9] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *Ann. Math. Stat.*, vol. 18, no. 1, pp. 50–60, 1947.

[10] K. Pearson, "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling," *Springer Series in Statistics*. pp. 11–28, 1992. doi: 10.1007/978-1-4612-4380-9_2.

[11] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020.

[12] N. Ma *et al.*, *Project-MONAI/MONAI: 0.5.0*. 2021. doi: 10.5281/zenodo.4679866.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[14] M. Puttagunta and S. Ravi, "Medical image analysis based on deep learning approach," *Multimed. Tools Appl.*, pp. 1–34, Apr. 2021.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv [cs.CV]*, Dec. 10, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[16] A. Paszke *et al.*, "Automatic differentiation in PyTorch," Oct. 28, 2017. Accessed: Nov. 29, 2021. [Online]. Available: https://openreview.net/pdf?id=BJJsrmfCZ

[17] "Heart Health and Aging." https://www.nia.nih.gov/health/heart-health-and-aging (accessed Nov. 29, 2021).

[18] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1. pp. 29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Oct. 2016, doi: 10.1007/s11263-019-01228-7.

[20] S. E. Luijnenburg, D. Robbers-Visser, A. Moelker, H. W. Vliegen, B. J. M. Mulder, and W. A. Helbing, "Intra-observer and interobserver variability of biventricular function, volumes and mass in patients with congenital heart disease measured by CMR imaging," *Int. J. Cardiovasc. Imaging*, vol. 26, no. 1, pp. 57–64, Jan. 2010.

[21] R. Trevethan, "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice," *Front Public Health*, vol. 5, p. 307, Nov. 2017.

[22] H. Bougias, E. Georgiadou, C. Malamateniou, and N. Stogiannos, "Identifying cardiomegaly in chest X-rays: a cross-sectional study of evaluation and comparison between different transfer learning methods," *Acta radiol.*, vol. 62, no. 12, pp. 1601–1609, Dec. 2021.

[23] A. Bouslama, Y. Laaziz, and A. Tali, "Diagnosis and precise localization of cardiomegaly disease using U-NET," *Informatics in Medicine Unlocked*, vol. 19, p. 100306, Jan. 2020.

[24] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, and M. Xu-Wilson, "Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks," *arXiv [cs.CV]*, Apr. 20, 2018. [Online]. Available: http://arxiv.org/abs/1804.07839

[25] S. Candemir, S. Rajaraman, G. Thoma, and S. Antani, "Deep Learning for Grading Cardiomegaly Severity in Chest X-Rays: An Investigation," in *2018 IEEE Life Sciences Conference (LSC)*, Oct. 2018, pp. 109–113.

[26] E. Sogancioglu, K. Murphy, E. Calli, E. T. Scholten, S. Schalekamp, and B. Van Ginneken, "Cardiomegaly Detection on Chest Radiographs: Segmentation Versus Classification," *IEEE Access*, vol. 8, pp. 94631–94642, 2020.

[27] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy* , vol. 23, no. 1, Dec. 2020, doi: 10.3390/e23010018.

[28] K. Truszkiewicz, R. Poręba, and P. Gać, "Radiological Cardiothoracic Ratio in Evidence-Based Medicine," *J. Clin. Med. Res.*, vol. 10, no. 9, May 2021, doi: 10.3390/jcm10092016.