ANOMALY DETECTION ON CHEST X-RAY IMAGES

Yaron Aloni
Department of Electrical Engineering
Tel-Aviv University

Anat Cohen
Department of Electrical Engineering
Tel-Aviv University

ABSTRACT

Radiologists worldwide are experiencing difficulty with the rising workload. A feasible, possible solution to face this problem is using a computer aided triage system that would make physicians' work time more efficient and allow them to focus on more complex cases. One of the methods to aid the physicians in their job is using an approach called novelty detection, i.e. trying to identify whether a given sample is drawn from outside a certain distribution – in order to help distinguish between normal and abnormal imaging tests.

The Center for Artificial Intelligence in Medicine & Imaging provides us with the Chexpert dataset, which consists of more than 200,000 chest X-ray (CXR) images from more than 65,240 patients, expertly labeled with 12 different findings such as cardiomegaly, lung opacity, etc. Using this dataset, we developed an algorithm inspired by the work in 'CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances' which conducted a similar study on image benchmark datasets such as CIFAR-10, CIFAR-100, ImageNet-30, etc. As gathering abnormal imaging samples is somewhat harder than obtaining normal ones, we trained our network using only images that had no finding of illness or disorder. During the test phase, we used both the frontal and the lateral images of a given subject in order to decide whether the subject has any suspicious findings. Best results were achieved by detecting Pleural Effusion using both lateral and frontal views: AUROC of 0.82.

1 Introduction

In Radiology, "turn-around time is king" [Jackson, 2015]. Although sub-optimal, radiologists over the world are being evaluated based on their throughput rather than the quality of their reports. This unwanted attitude is a result of a growing workload on radiologists. This problem is prominent especially in rural areas where physicians rely mostly on teleradiology for the CXR interpretation, giving priority to turn-around time that might lead to lower quality reports. Thus may cause to confusion or misdiagnosis and potentially have life-changing and even life-threatening consequences [Eban, 2011].

Therefore – the need for computer-aided triage systems is clear. First, it would help radiologists prioritize their incoming work and focus their attention immediately on higher-risk cases. Second, it might give the radiologists a suggestion about what the diagnosis could be. Thirdly, it can give the primary care physician immediate information about the patient's condition – even before seen by the radiologist.

The input to our algorithm are jpg CXR images along with .csv files labeling them as normal/abnormal including the finding. We will then use a CNN to output a score of 'normality' for each sample, allowing us to choose a threshold which distinguishes between the labels. In accordance with [Tack et al., 2020] we will use CSI loss function, combining simple contrastive learning (SimCLR) loss [Chen et al., 2020] and Classifying shifted instances (cls-SI) loss [Tack et al., 2020] to enhance the performance.
We will use two similar models, and train them with different parts of the dataset. One model will be trained using frontal images whereas the other will be trained using lateral images. Lastly, when evaluating the scores from each network we will summarize them to get the conclusion.

2 Related Work

The concept of computer-aided diagnosis (CAD) once viewed as fiction is now part of the medical world. What began as rule-based survival prediction for lung x-rays has evolved into machine-learning approaches and deep learning [Ginneken, 2017] [Lodwick, 1966] [Doi, 2007]. [van Ginneken et al., 2011] make the argument that CAD in radiology is obligatory, as radiologists will soon be helpless with the quickly rising workload.
At first, CAD was sought for other types of diagnostic tasks such as breast cancer localization by GoogLeNet [Wang et al., 2016] and skin cancer classification [Esteva et al., 2017]. Both of these well-designed networks, along with others, have proven that CNNs can be utilized very successfully not just in natural image classification, but also for medical image classification and segmentation [Schmidhuber, 2014] [LeCun et al., 2015].

In the domain of CXR classification tasks, Rajkomar et al. used GoogLeNet along with image augmentation and pre-training on ImageNet to classify CXR images as either frontal or lateral with 100 percent accuracy [Rajkomar et al., 2016]. While this is not clinically relevant, it was an important proof of concept for the utilization of deep learning on CXR images. Anavi et al. aimed to create a network that could, given one query image, rank the remaining CXR images in the dataset by their resemblance to the query. They found that a 5 layer convolutional neural network was much more effective than similarity based on image descriptors [Anavi et al., 2015]. Such a network could be used by physicians to help them search for past cases and help inform their current or future diagnosis. In another study, Shin et al. used a neural network to detect specific diseases in CXR images and assign disease labels to them. They then used a RNN to describe the context of the annotated disease based on the features of the CNN and patient metadata [Shin et al., 2016]. They were able to reach a validation score of 0.698 on this ambitious task. This performance may be largely due to their relatively small data set size of 7470 images, the challenges of multi-class classification, and incorporating textual data from patient records. Wang et al. successfully designed a CNN to diagnose specific diseases by detecting and classifying lung nodules in CXR images with high accuracy [Elazab et al., 2016].

While all these approaches function as very good proof of concept, and may be useful for assisted diagnosis, they are not easily generalizable to different diseases as new training data and labels must be obtained to retrain models for

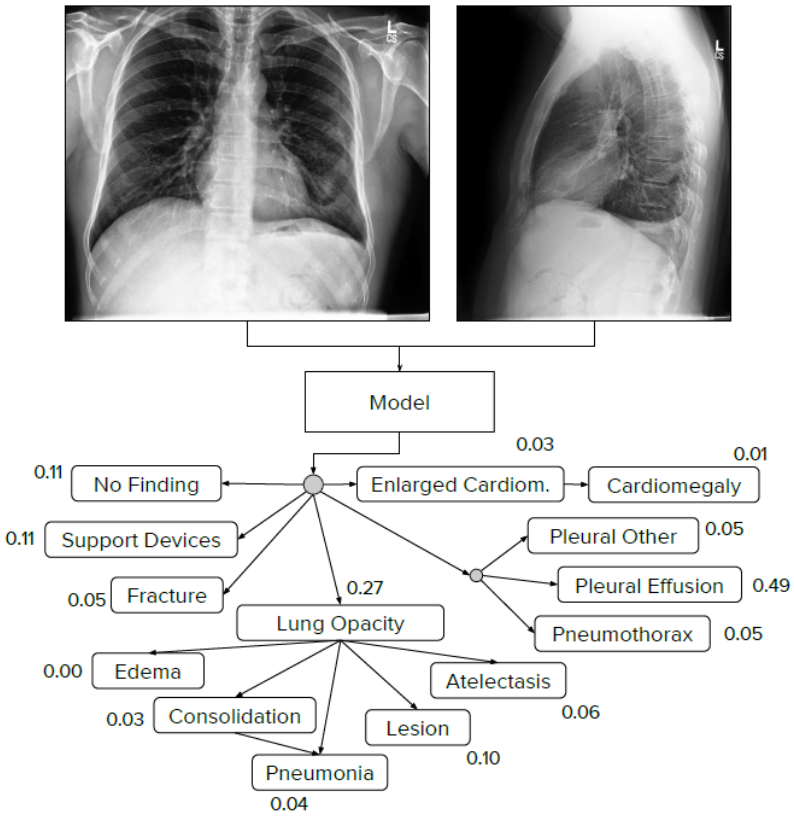


Figure 1: different observations from multi-view chest radiographs

each specific sub-question. Our work will classify CXRs as normal versus abnormal, in order to assist primary care physicians and radiologists to move more quickly and efficiently rather than render radiology obsolete. Given that we are not aiming to classify images to specific categories of disease, we will have a much more versatile framework that is applicable to a much larger patient population.

3 Dataset

Data was acquired from the Stanford Center for Artificial Intelligence in Medicine & Imaging (AIMI center), which maintains a large database of patient x-rays and includes 400,000 chest x-rays, Brain MRIs and several other imaging types. For this project we used AIMI's dataset CheXpert [Irvin et al., 2019] containing only labeled CXR images. The training dataset consists of only normal labeled images, 16,474 of them taken from frontal view, and 4907 of them from lateral view. The full training dataset was not taken in its entirety due to lack of required resources. Some patients may have more than one image associated with their file; in this case each image was treated as a separate patient for purposes of training and prediction. Each image has a matching line in the csv. files, containing several findings. We used 1000 samples for the algorithm evaluation, labeled as normal/abnormal with the type(s) of abnormality. Original images vary in size and range between 320 and 390 pixels in height and width.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Table 1: The CheXpert dataset consists of 14 labeled observations. We report the number of studies which contain these observations in the training set.

4 Methods

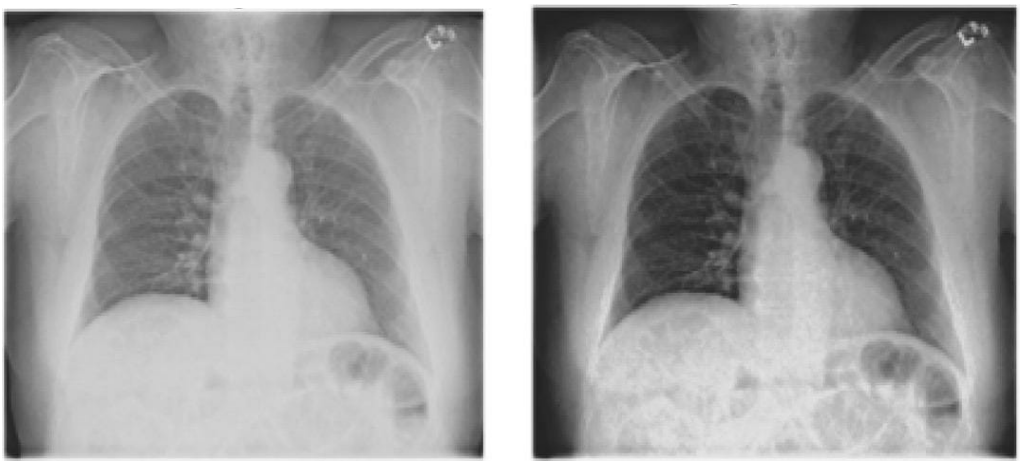
4.1 Preprocessing

There are many sources of variance in CXR data which may negatively affect the performance of classification tasks using feature-based methods or neural networks. Major sources of variance include contrast variance, positional variance and view angle variance (e.g Anterior-Posterior vs Posterior-Anterior vs Medial-Lateral). To face this challenge we process all images with histogram equalization, this increases contrast within each CXR image. Guided by previous works [Anavi et al., 2015] [Shin et al., 2016] [Elazab et al., 2016] regarding utilization of CNN on CXR images, we deemed it useful to enhance the difference between the bone and empty space of tissue depicted in x-rays so as to make relevant information more prominent (See figure 2). All image processing steps were carried out using the python scikit-image library [Buitinck et al., 2013].

4.2 Contrasting Shift Instances

4.2.1 Contrasting shifted instances loss

In their work facing the task of novelty detection, [Tack et al., 2020] innovated with an algorithm that achieves a proper feature map using training examples with only a single label. Their concept included applying several types of



(a) Original image (b) Image after histogram equalization

Figure 2: histogram equalization

image augmentations over the training dataset in order to obtain a feature map only from a single label. [Tack et al., 2020] mentioned a term called "OOD (out of distribution)-ness" which indicates the amount an image transformation changes the image in accordance to the distribution. An output image from a low OOD-ness transformation will have deeper-layer feature maps similar to the input image, while high OOD-ness transformation deeper-layer feature maps will have different values which should not resemble to the input image.
The augmentations in their work were split to two types: 'Soft' augmentations which have the lowest OOD-ness, and 'hard' augmentations that have higher OOD-ness. The soft augmentations they found to have the lowest OOD-ness are horizontal flips, inception crop, color jitter and gray scale layer. In our case, since our input is a single channel rather than RGB, we used a gray-scale jitter instead of color jitter. A gray-scale jitter includes changing the image brightness and contrast. The hard augmentations included rotations of $90°, 180°$ and $270°$. We took the same hard augmentations as in the original paper.
The idea expands on contrastive learning, where the aim is to pull towards each other samples from the same distribution while pushing different samples further away. Let x be a query image, and $\{x_+\}$ and $\{x_-\}$ be a set of positive and negative samples, hence samples from the same distribution and from outside this distribution respectively, $z(x)$ is the feature defined directly from the encoder, and finally $sim(z, z') := \frac{z \cdot z'}{\|z\|\|z'\|}$ Then the contrasting loss is defined as:

$$\mathcal{L}_{con}\left(x, \{x_+\}, \{x_-\}\right) := -\frac{1}{\left|\{x_+\}\right|} log \frac{\sum_{x' \in \{x_+\}} exp\left(sim(z(x), z'(x))/\tau\right)}{\sum_{x' \in \{x_+\} \cup \{x_-\}} exp\left(sim(z(x), z'(x))/\tau\right)}$$

[Tack et al., 2020] computed the final loss as the sum of all the losses. For each example x, x_+ will be generated using the soft transformations over x, while x_- is generated using hard transformations, or rotations, over x, and hard transformations followed by soft transformations over x. Using these augmentations, they used the SimCLR loss as proposed by [Chen et al., 2020]:

$$\mathcal{L}_{SimCLR}(B; T) := \frac{1}{2B} \sum_{i=1}^{B} L_{con}\left(x_i, T(x_i), B_{-i}\right) + L_{con}\left(T(x_i), x_i, B_{-i}\right)$$

Where T is a transformation taken from the domain of soft transformations, and $B_{-i} := \{x_j\}_{j \neq i} \cup \{T(x_i)\}_{j \neq i}$.

Finally, they introduced the contrasting shifted instances (con-SI) loss as follows:
For $\mathcal{S} := \{negative\,(rotation)\,transformtations\}$

$$\mathcal{L}_{con-SI} := L_{SimCLR}\left(\bigcup_{S \in \mathcal{S}} \mathcal{B}_S; T\right), \quad \text{where} \quad B_S := \{S(x_i)\}_{i=1}^{B}$$

4.2.2 Classifying shifted instances loss

In addition to contrasting shifted instances, [Tack et al., 2020] added an auxiliary task that predicts which shifting transformation is applied for a given input x, in order to facilitate $z(x)$ to discriminate each shifted instance. Specifically, they added a linear layer after $z(x)$ for modeling an auxiliary softmax classifier $p_{cls-SI}(y^S|x)$ as in [Golan and El-Yaniv, 2018]. Let $\tilde{B_S}$ be the batch augmented from B_S via SimCLR, then, the classifying shifted instances (cls-SI) loss is defined as follows:

$$\mathcal{L}_{cls-SI} := \frac{1}{2B}\frac{1}{K}\sum_{S\in\mathcal{S}}\sum_{\tilde{x_S}\in\tilde{B_S}} -log\, p_{cls-SI}\left(\tilde{x_S}\right), \quad \text{where} \quad K = |S|$$

The final loss proposed by CSI is defined by summing the two objectives:

$$\mathcal{L}_{CSI} = \mathcal{L}_{con-SI} + \lambda \cdot \mathcal{L}_{cls-SI}$$

4.3 Network architecture

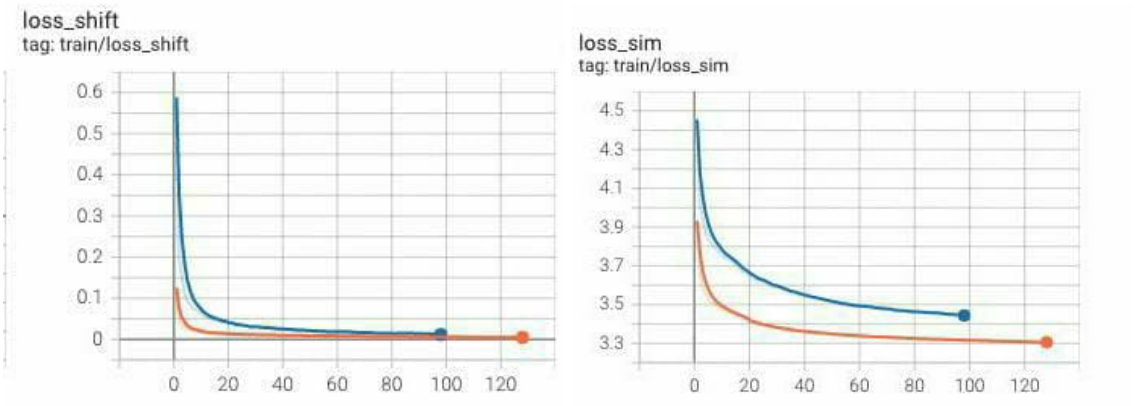
Following the best results from CSI paper, our work uses ResNet [He et al., 2015] as the main architecture. Residual Networks (ResNets) reformulate the concept of learning to learn residuals with respect to each layer rather than functions directly. This means each layer tries to learn RES where RES is the final loss minus the output of the previous layer. These networks are easier to optimize than traditional ones, and our implementation attains 152 layers without experiencing classical problems like vanishing gradients. For our work, due to computational limitations, we used ResNet18. The network trainings were run on a GCP server with NVIDIA Tesla K80.

We trained two different networks, one using the frontal CXR images, and the other using the lateral images. The concept behind this approach was that the two types of images are fundamentally different and training a single network on both might lead to sub-optimal results. Moreover, we believed that different types of diseases might be identified better using one type of image or the other. After training both networks and obtaining a stable error, we ran the evaluations phase on both networks, while the final score is the mean of the scores from both networks.

5 Experiments

5.1 Training details

We use ResNet-18 [He et al., 2015] as the base encoder network f_θ and 2-layer multi-layer perceptron with 128 embedding dimension as the projection head g_θ. All models are trained by minimizing the final loss $\mathcal{L}_{\mathcal{CSI}}$ with a temperature of $\tau = 0.5$. We follow the same optimization step of SimCLR [Chen et al., 2020]. For optimization, we train CSI Frontal View for 128 epoch and CSI Lateral View for 73 epochs, using stochastic gradient descent optimizer with momentum 0.9. The learning rate is set at 0.001. We use batch size of 10 where the batch is given by $\bigcup_{S\in\mathcal{S}} \mathcal{B}_S$.



(a) classifying shifted instances (cls-SI) loss

(b) contrasting shifted instances (con-SI) loss

Figure 3: Loss progression over epochs

5.2 Data augmentation details

We use SimCLR augmentations: Inception crop [Szegedy et al., 2014], horizontal flip, grayscale jitter, and gaussian blur for random augmentations $\mathcal{T}$, and rotation as shifting transformation $\mathcal{S}$. The detailed description of the augmentations are as follows:

- **Inception crop.** Randomly crops the area of the original image with uniform distribution 0.08 to 1.0. After the crop, cropped image are resized to the original image size.
- **Horizontal flip.** Flips the image horizontally with 50% of probability.
- **Grayscale jitter.** In accordance with the color jitter was used in [Tack et al., 2020] we aimed to create an augmentation with the same nature for gray-scale images. The augmentation, called gray-scale jitter, is a combination of changes in the brightness and contrast of the image.
- **Rotation.** We use rotation as $\mathcal{S}$, the shifting transformation, $\{0°; 90°; 180°; 270°\}$. For a given batch $\mathcal{B}$, we apply each rotation degree to obtain the new batch for CSI: $\bigcup_{S\in\mathcal{S}} \mathcal{B}_S$.

5.3 Evaluation metric

For evaluation, we measure the **Area under the receiver operating characteristic curve (AUROC)** that measures (a) the effectiveness of the proposed score in distinguishing in- and out-of-distribution images, (b) the confidence calibration of softmax classifier.
Let TP, TN, FP, and FN denote true positive, true negative, false positive and false negative, respectively. The ROC curve is a graph plotting true positive rate = TP / (TP+FN) against the false positive rate = FP / (FP+TN) by varying a threshold.

5.4 Main Results

We focus on the evaluation of 5 observations which are called the competition tasks [Irvin et al., 2019], selected based of clinical importance and prevalence in the validation set: (a) Atelectasis, (b) Cardiomegaly, (c) Consolidation, (d) Edema, and (e) Pleural Effusion. We also added a case with all observations together.
we conduct 5 one-class classification tasks, where each task sets one of the classes as out-of-distribution while the 'No Finding' case is always in-distribution. Table 2 summarizes the results, showing that the combined model for frontal and lateral view together outperforms all other models for every observation.

	Atelectasis	**Cardiomegaly**	**Consolidation**	**Edema**	**Pleural Effusion**	**All observations**
Frontal view	0.6876	0.6530	0.7090	0.6669	0.7418	0.6423
Lateral view	0.6973	0.6531	0.7368	0.7220	0.7929	0.6648
Combined (Frontal+Lateral)	0.7313	0.6814	0.7685	0.7465	**0.8242**	0.6830

Table 2: Test results

6 Conclusion

Anomaly detection on real-life medical images is a complicated task. Unlike supervised classification tasks, the positive observations are not used in training. This complicates the prediction task because the model doesn't learn the explicit features of each observation. Moreover, in medical diagnoses tasks acquiring large amounts of labeled data is often difficult.
In figure 5 we can see the top results for the Chexpert Competition on predicting different observations from multi-view chest X-rays. Our model did not exceed these results. While in most cases unsupervised learning achieves lower accuracy then supervised tasks [Niu et al., 2019], these models are still highly useful in cases with unlabeled data where a supervised approach is not feasible.

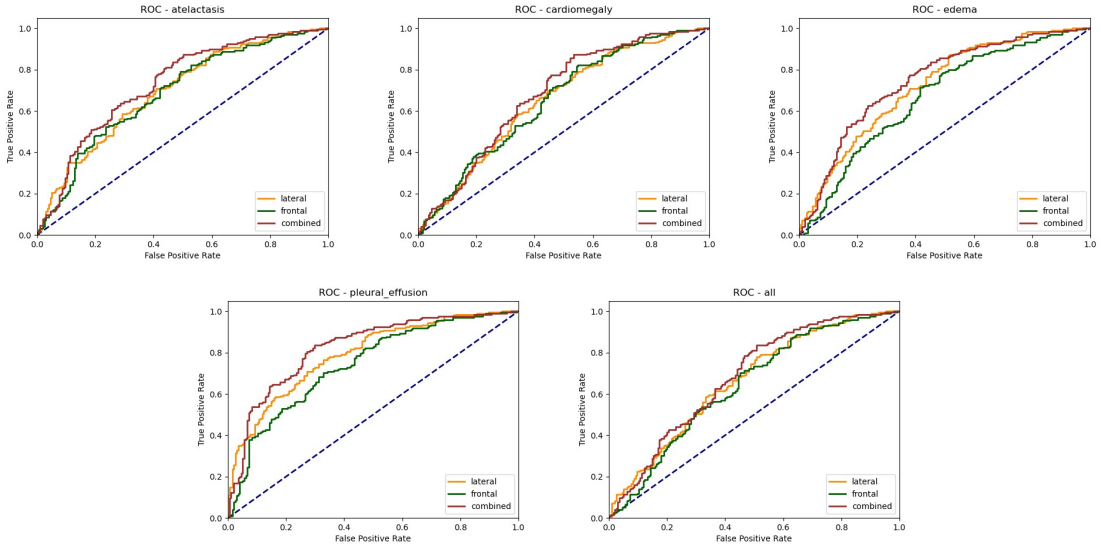


Figure 4: ROC curves of the three models for each observation

Leaderboard

Will your model perform as well as radiologists in detecting different pathologies in chest X-rays?

Rank	Date	Model	AUC	Num Rads Below Curve
1	Aug 31, 2020	DeepAUC-v1 *ensemble* https://arxiv.org/abs/2012.03173	0.930	2.8
2	Sep 01, 2019	Hierarchical-Learning-V1 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.930	2.6
3	Oct 15, 2019	Conditional-Training-LSR *ensemble*	0.929	2.6
4	Dec 04, 2019	Hierarchical-Learning-V4 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.929	2.6
5	Oct 10, 2019	YWW(ensemble) *JF&NNU* https://github.com/jfhealthcare/Chexpert	0.929	2.8

Figure 5: Chexpert Competition Leaderboard

7 Future Work

We believe future work should include a more complicated model with two input images: frontal and lateral view of the same patient. Each goes through the same pipeline as discussed here and concatenated into an additional layer that combines them together. Unfortunately due to lack of computing power this was not a feasible approach for us. Nonetheless we believe such approach will yield even better results than those shown here.

8 appendix

Code:

The code of this paper is available at https://github.com/anatcohen2/deep_learning_project.

References

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019. URL http://arxiv.org/abs/1901.07031.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. CSI: novelty detection via contrastive learning on distributionally shifted instances. *CoRR*, abs/2007.08176, 2020. URL https://arxiv.org/abs/2007.08176.

W. L. Jackson. In radiology, turnaround time is king, 2015.

K. Eban. Is a doctor reading your x-rays? maybe not, 2011.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. URL https://arxiv.org/abs/2002.05709.

Bram Ginneken. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiological Physics and Technology*, 10:1–10, 02 2017. doi: 10.1007/s12194-017-0394-5.

GS. Lodwick. Computer-aided diagnosis in radiology. a research plan. *Investigative radiology*, 1:72–80, 1966. doi: 10.1097/00004424-196601000-00032.

Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 31: 198–211, 06 2007. doi: 10.1016/j.compmedimag.2007.02.002.

B. van Ginneken, C. Schaefer-Prokop, and M. Prokop. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, 261 3:719–32, 2011.

Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H. Beck. Deep learning for identifying metastatic breast cancer, 2016.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115—118, February 2017. ISSN 0028-0836. doi: 10.1038/nature21056. URL https://europepmc.org/articles/PMC8382232.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *CoRR*, abs/1404.7828, 2014. URL http://arxiv.org/abs/1404.7828.

Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015. doi: 10.1038/nature14539.

Alvin Rajkomar, Sneha Lingam, Andrew G. Taylor, Michael Blum, and John Mongan. High-throughput classification of radiographs using deep convolutional neural networks. *Journal of Digital Imaging*, 30:95 – 101, 2016.

Yaron Anavi, Ilya Kogan, Elad Gelbart, Ofer Geva, and Hayit Greenspan. A comparative study for chest radiograph image retrieval using binary texture and deep learning classification. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2940–2943, 2015. doi: 10.1109/EMBC.2015.7319008.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. pages 2497–2506, 06 2016. doi: 10.1109/CVPR.2016.274.

Ahmed Elazab, Jianhuang Wu, and Qingmao Hu. Lung nodule classification using deep feature fusion in chest radiography. *Computerized Medical Imaging and Graphics*, 57, 11 2016. doi: 10.1016/j.compmedimag.2016.11.004.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. *CoRR*, abs/1309.0238, 2013. URL http://arxiv.org/abs/1309.0238.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9781–9791, Red Hook, NY, USA, 2018. Curran Associates Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. URL http://arxiv.org/abs/1409.4842.

Xuetong Niu, Li Wang, and Xulei Yang. A comparison study of credit card fraud detection: Supervised versus unsupervised. *CoRR*, abs/1904.10604, 2019. URL http://arxiv.org/abs/1904.10604.