Anomaly Detection for Safe Chest X-ray Interpretation

Nguyet Minh Phu & Jun Li

Introduction

Many deep learning based chest radiograph interpretation models have the **undesirable property of making predictions of pathology on all given input images**, including those that medical doctors (representing the gold standard) are unable to safely make a prediction on. A reliable model for clinical use must be able to **identify these bad inputs** and alert the user to retake the chest X-ray instead of proceeding to make a prediction. The purpose of this project is to ***build a model for detecting such images that should not be input into chest radiograph interpretation models***.

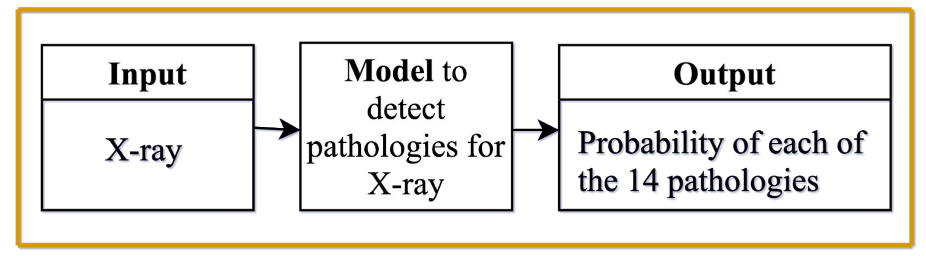


Fig 1. Pathology Detection Model

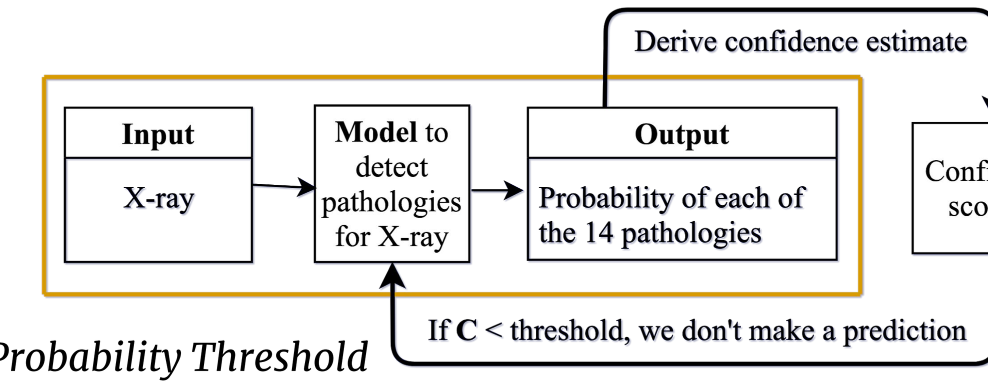


Fig 2. Probability Threshold

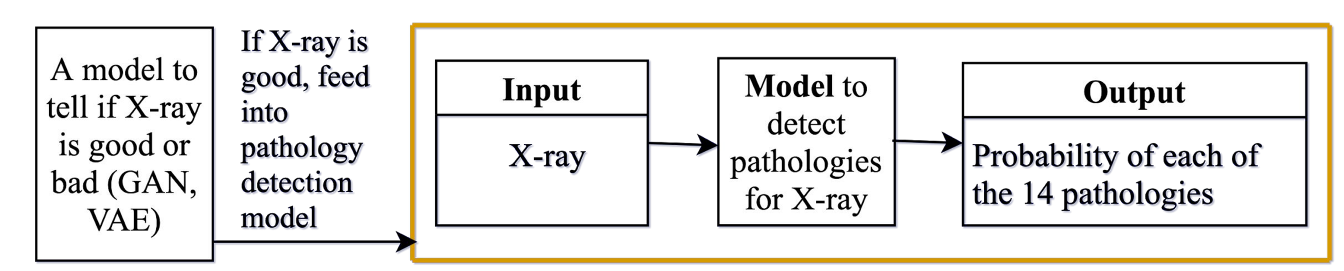


Fig 3. Cascading Approach

Dataset

- *In-distribution data:* **CheXpert**, a dataset consisting of 224,316 chest radiographs of 65,240 patients published by the Stanford Machine Learning group. We filter the dataset for only **frontal** chest X-rays.
- *Out-of-distribution data:* (1) original CheXpert X-rays with contrast and brightness changed to simulate **under/over-penetration**, (2) original CheXpert X-rays with Gaussian noise filters applied to to simulate **blurriness**, (3) CheXpert **lateral** chest X-rays (in contrast to frontal ones) for X-rays of another category.

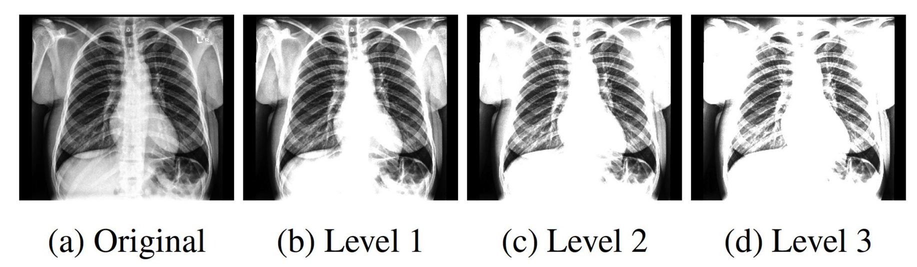


Fig 4. Examples of levels of under-penetrated X-rays

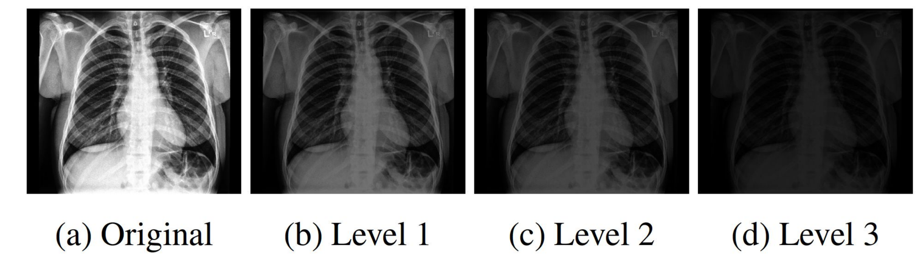


Fig 5. Examples of levels of over-penetrated X-rays

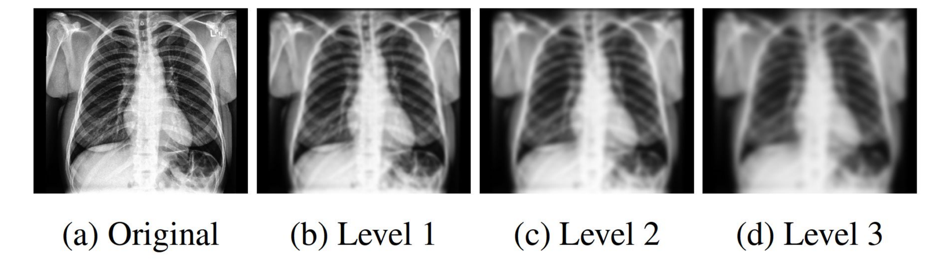


Fig 6. Examples of levels of blur X-rays

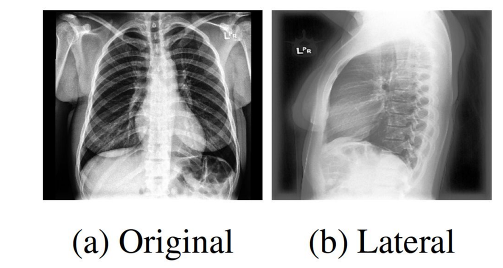


Fig 7. Example of lateral X-rays

Methods & Results

Pathology detection

We trained a **DenseNet** to **concurrently detect the presence of 14 pathologies** in frontal chest X-rays. Moreover, we showed in Table 1 that as **the quality of X-rays decreases, the model becomes less accurate**. Indeed, the model should not make a prediction when given a low-quality X-ray.

	Original	Level 1	Level 2	Level 3
Under-penetrated	0.7334	0.7302	0.7024	0.6664
Over-penetrated	0.7334	0.7119	0.6758	0.5717
Blurred	0.7334	0.7043	0.6935	0.6819
Lateral	0.7334	0.6109	N/A	N/A

Table 1. Mean AUROC for detecting 14 pathologies

Low-quality X-rays can affect the model predictions significantly. In Fig 8, the same frontal-view X-ray of a patient with No Finding is predicted as having Edema when the X-ray is under-penetrated, having Support Devices when the X-ray is over-penetrated, having Lung Opacity when the X-ray is blur or when a lateral-view X-ray of the same patient is used.

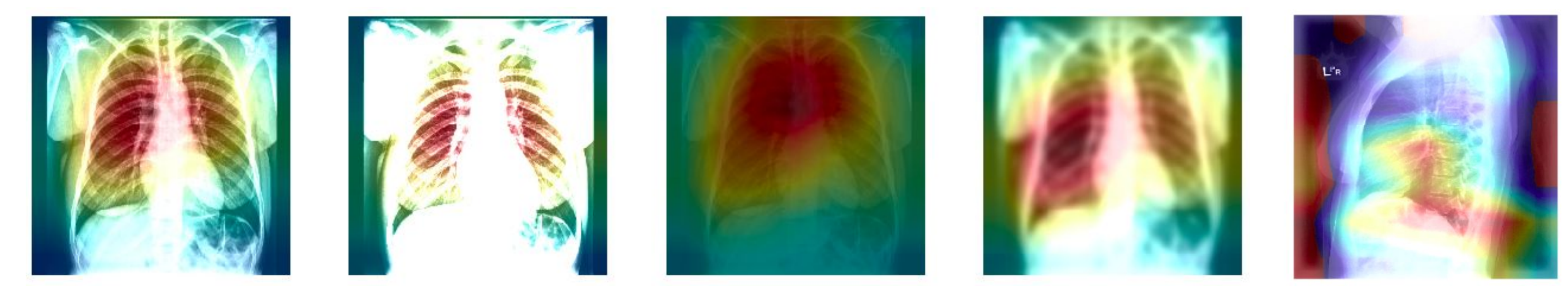


Fig 8. CAMs for in- & out-of-dist X-rays

Poor-quality X-ray detection

Probability Threshold

$$\mathcal{C} = 2 \times \left|\frac{1}{14}\left(\sum_i P_i\right) - 0.5\right|$$
$$\text{where } i \in \text{pathologies}$$

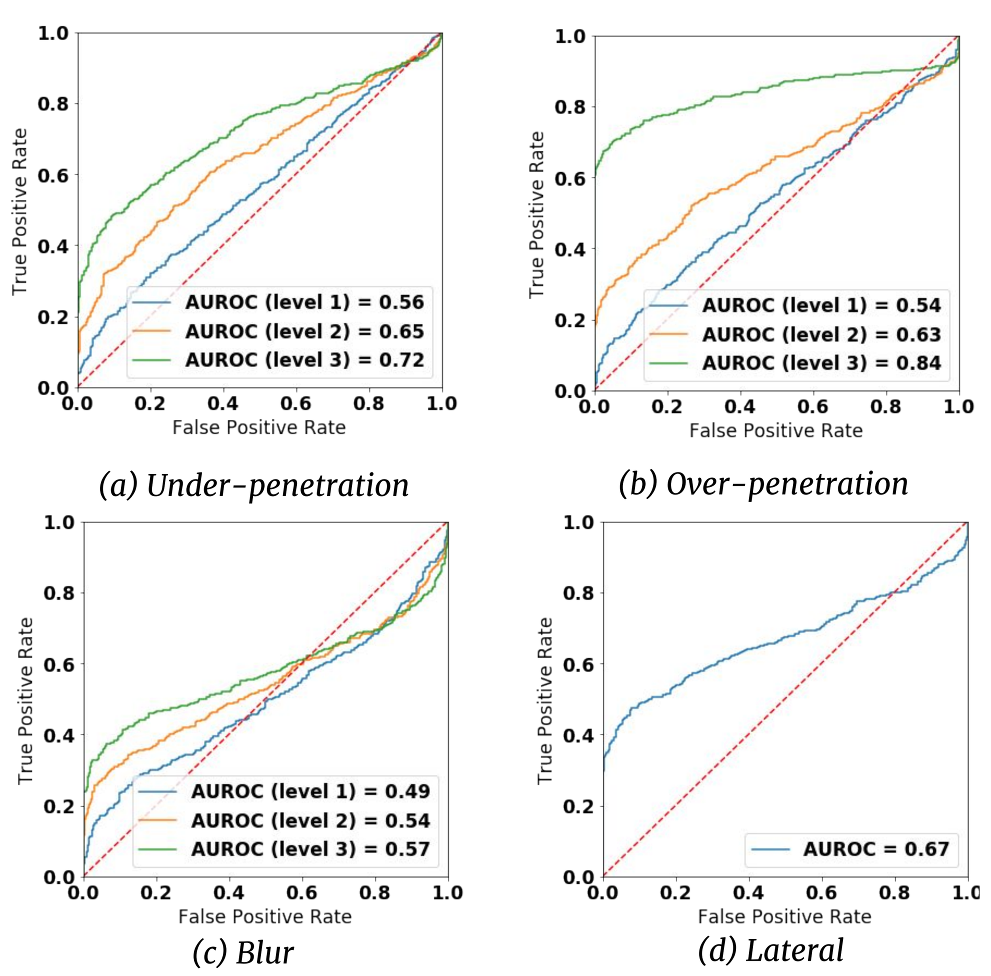
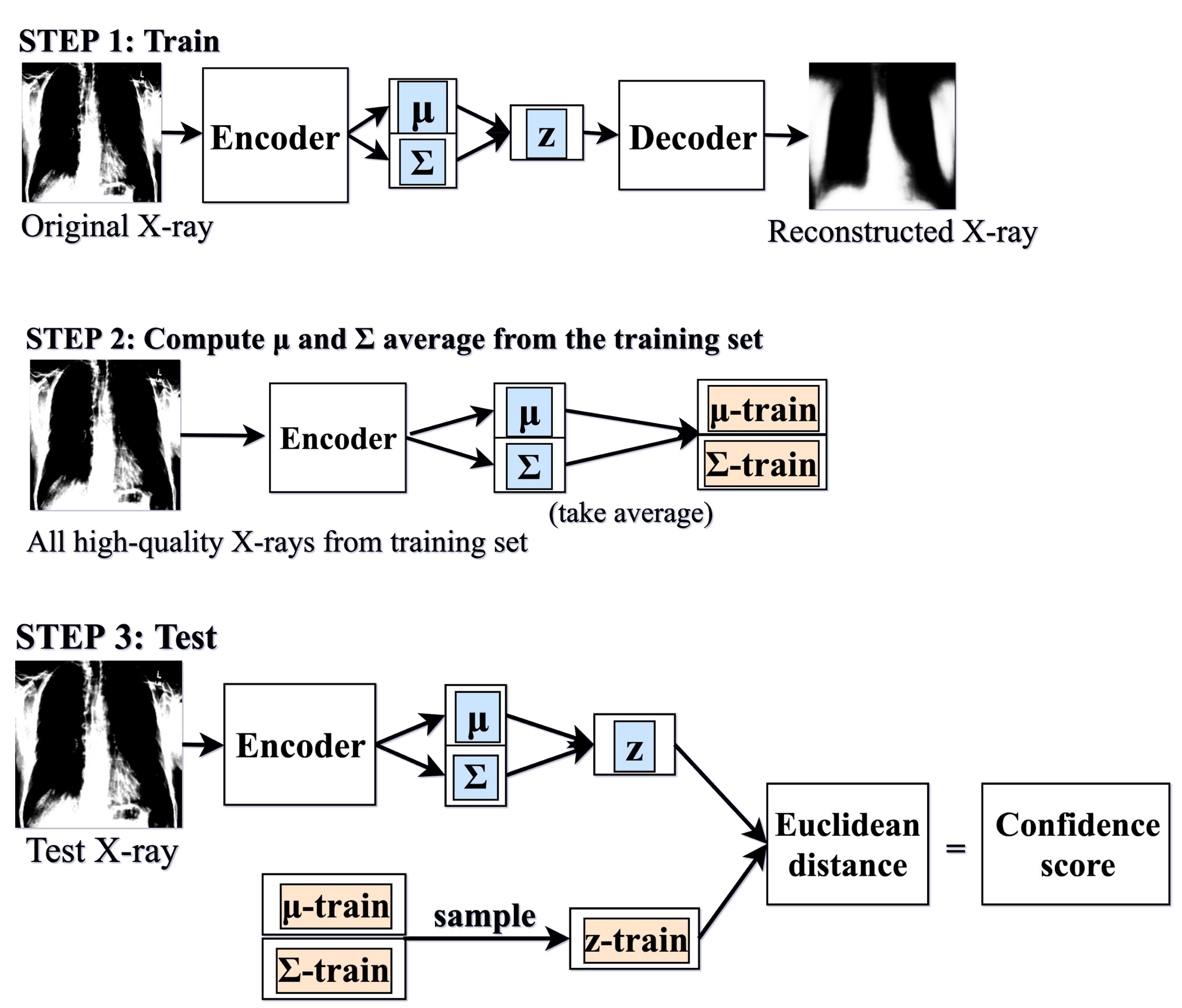


Fig 9. AUROC of Prob. Thres. poor-quality X-ray detection

Variational Autoencoder

$$\mathcal{L}(x^{(i)}, \theta, \phi) = E_z[\log p_\theta(x^{(i)}|z)] - D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z))$$

$$A(x) = d(\hat{\mathbf{z}}_{\mathbf{x}}, \hat{\mathbf{z}}_{\mathbf{train}}) = d(\hat{\mathbf{z}}_{\mathbf{train}}, \hat{\mathbf{z}}_{\mathbf{x}}) = \sqrt{\sum_i ((\hat{z}_x)_i - (\hat{z}_{\text{train}})_i)^2}$$



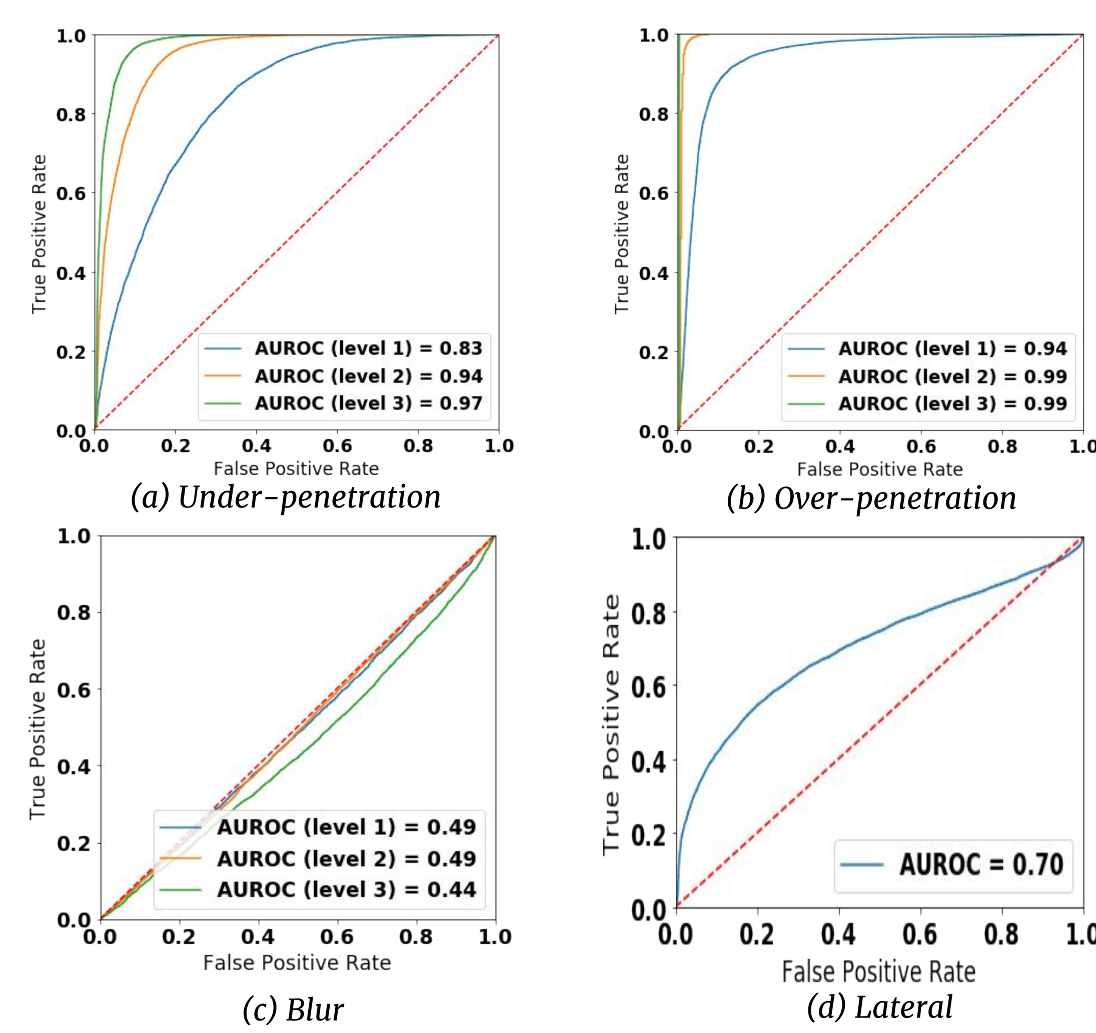
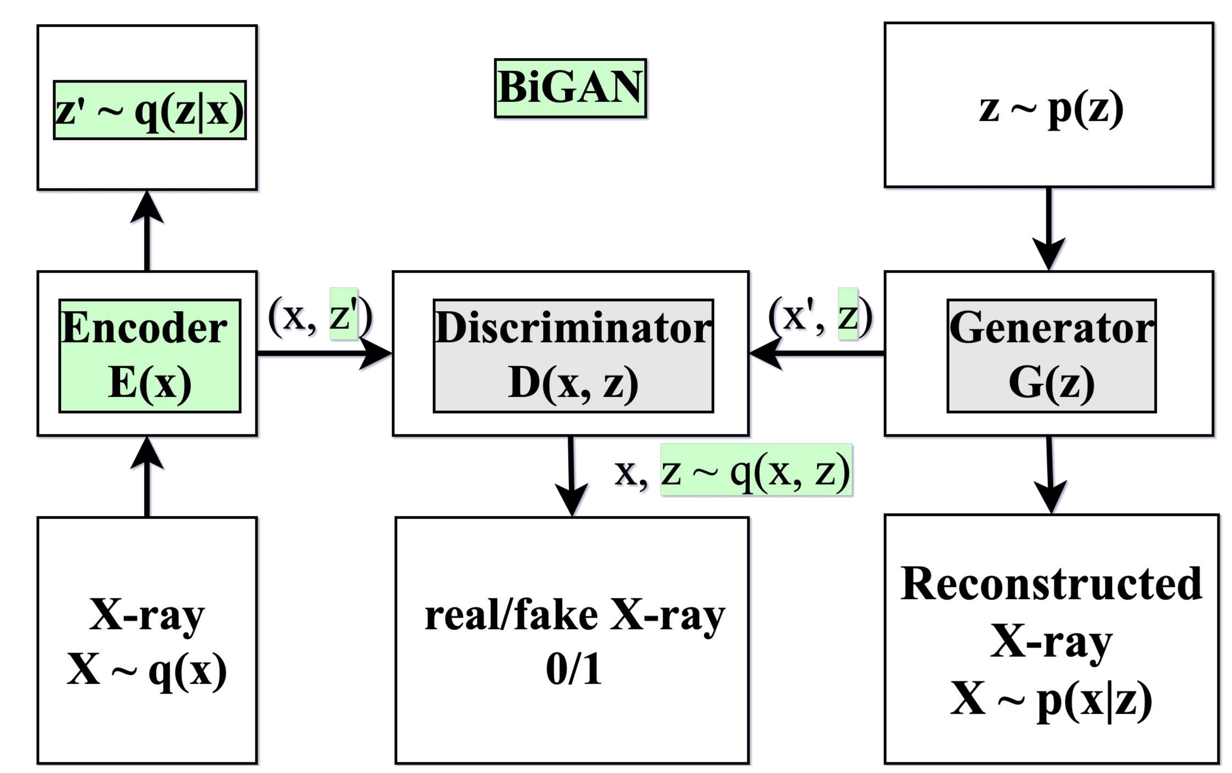


Fig 10. AUROC of VAE for poor-quality X-ray detection

Bidirectional GAN

$$\min_{G,E} \max_D V(D, E, G)$$
$$= \mathbb{E}_{x\sim pX}[\mathbb{E}_{z\sim pE(\cdot|x)}[\log(D(x,z))]]$$
$$+ \mathbb{E}_{z\sim pZ}[\mathbb{E}_{x\sim pG(\cdot|z)}[1 - \log D(x,z)]]$$



$$A(x) = \lambda_1 L_G(x) - \lambda_2 L_D(x)$$
$$L_G(x) = ||x - G(E(x))||_1 \quad L_D(x) = \sigma(D(x, E(x)), 1)$$

	Level 1	Level 2	Level 3
Under-penetrated	0.0031	0.0005	0.0002
Over-penetrated	1.0000	1.0000	1.0000
Blurred	0.5537	0.5001	0.6876
Lateral	0.7297	N/A	N/A

Table 2. AUROC of BiGAN for poor-quality X-ray detection

Conclusion

We built a model that can detect fourteen different pathologies from chest radiographs, then taught the model to refrain from making a prediction when given low-quality or bogus inputs. Our models could effectively detect low-quality X-rays. However, different types of poor quality are not equally easy to identify, and each model seems to do well on some types and fail on others. By combining these models, we could potentially create an even more powerful poor-quality X-ray detector for safe medical-image interpretation systems.