

A Comparative Study on Out of Scope Detection for Chest X-ray Images



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science

Department of Computer Engineering

FACULTY OF ENGINEERING

Chulalongkorn University

Academic Year 2022

Copyright of Chulalongkorn University

การศึกษาเปรียบเทียบการตรวจจับนอกสโคปสำหรับภาพเอกซเรย์ทรวงอก



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Thesis Title	A Comparative Study on Out of Scope Detection for Chest X-ray Images
By	Mr. Nuttapol Kamolkunasiri
Field of Study	Computer Science
Thesis Advisor	Ekapol Chuangsuwanich, Ph.D.
Thesis Co Advisor	Associate Professor Proadpran Punyabukkana, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

..... Dean of the FACULTY OF ENGINEERING
(Professor SUPOT TEACHAVORASINSKUN, Ph.D.)

THESIS COMMITTEE

..... Chairman
(Associate Professor ATIWONG SUCHATO, Ph.D.)

..... Thesis Advisor
(Ekapol Chuangsuwanich, Ph.D.)

..... Thesis Co-Advisor
(Associate Professor Proadpran Punyabukkana, Ph.D.)

..... External Examiner
(Sanparith Marukatat, Ph.D.)

CHULALONGKORN UNIVERSITY



จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ณัฐพล กมลคุณาสิริ : การศึกษาเปรียบเทียบการตรวจจับนอกสโคปสำหรับภาพเอกซเรย์ทรวงอก. (A Comparative Study on Out of Scope Detection for Chest X-ray Images) อ.ที่ปรึกษาหลัก : อ. ดร.เอกพล ช่วงสุวนิช, อ.ที่ปรึกษาร่วม : รศ. ดร.โปรดปราน บุญยพุกกณะ

แบบจำลองการจำแนกภาพในแอปพลิเคชันที่ใช้งานจริงนั้น อาจได้รับชุดข้อมูลที่อยู่นอกการกระจายของข้อมูลที่ต้องการ สำหรับการใช้งานที่สำคัญเช่นการตัดสินใจทางการแพทย์เป็นสิ่งจำเป็นอย่างยิ่งที่แบบจำลองสามารถรับรู้และรองรับข้อมูลที่อยู่นอกการกระจาย (out-of-distribution) ดังกล่าวได้ วัตถุประสงค์ของการศึกษานี้คือเพื่อตรวจสอบประสิทธิภาพของวิธีการต่างๆสำหรับการระบุข้อมูลที่อยู่นอกการกระจายในภาพทางการแพทย์ เราตรวจสอบวิธีการตรวจจับข้อมูลที่อยู่นอกการกระจายทั้งหมดสามประเภท (แบบจำลอง Classification , แบบจำลองConfidence-based และแบบจำลองGenerative) เกี่ยวกับข้อมูลของภาพเอ็กซเรย์เราพบว่าแบบจำลองClassificationและ HealthyGAN ทำงานได้ดีที่สุด อย่างไรก็ตาม HealthyGAN ไม่สามารถระบุข้อมูลที่ไม่เคยเรียนรู้มาก่อนได้ในขณะที่แบบจำลองClassificationยังคงรักษาความได้เปรียบด้านประสิทธิภาพไว้ได้ นอกจากนี้เรายังตรวจสอบประเภทของภาพที่อาจตรวจจับได้ยากกว่าว่าอยู่นอกการกระจายเราพบว่ากรอบตัดภาพ(crop-outs)นั้นมนุษย์สามารถระบุได้ง่ายแต่กับเป็นเรื่องที่ยากสำหรับแบบจำลองในการตรวจจับ

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา 2565

ลายมือชื่อนิสิต
ลายมือชื่อ อ.ที่ปรึกษาหลัก
ลายมือชื่อ อ.ที่ปรึกษาร่วม

6272040321 : MAJOR COMPUTER SCIENCE

KEYWORD: Image classification, Image preprocessing, Out-of-Distribution, Chest radiography

Nuttapol Kamolkunasiri : A Comparative Study on Out of Scope Detection for Chest X-ray Images. Advisor: Ekapol Chuangsuwanich, Ph.D. Co-advisor: Assoc. Prof. Proadpran Punyabukkana, Ph.D.

Image classification models in actual applications may receive input outside the intended data distribution. For crucial applications such as clinical decision-making, it is critical that a model can recognize and describe such out-of-distribution (OOD) inputs. The objective of this study is to investigate the efficacy of several approaches for OOD identification in medical images. We examine three classes of OOD detection methods (Classification models, Confidence-based models, and Generative models) on the data of X-ray images. We found that simple classification methods and HealthyGAN perform the best overall. However, HealthyGAN cannot generalize to unseen scenarios, while classification models still retain some performance advantage. We also investigate the type of images that might be harder to detect as out of scope. We found that image crop-outs, while being easily identifiable by humans, are more challenging for the models to detect.



Field of Study: Computer Science

Academic Year: 2022

Student's Signature

Advisor's Signature

Co-advisor's Signature

ACKNOWLEDGEMENTS

I sincerely thank Dr. Phalin Kamolwat, Director of the Tuberculosis Division of the Thai Ministry of Public Health, for their invaluable support throughout my thesis project. Their expertise in tuberculosis and their dedication to public health have played a significant role in shaping the direction of this research. I am genuinely grateful for their contributions and for providing the necessary resources and support to make this study possible.

Furthermore, I would like to acknowledge the generous support received from the Bureau of Tuberculosis, Thailand, the Health Systems Research Institute (HSRI 62-103), the Ratchadapiseksompotch Matching Fund, Faculty of Medicine, Chulalongkorn University (RA-MF-12/62), and the Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University.

I am also profoundly grateful to my thesis advisor, Ekapol Chuangsuwanich, Ph.D., and my co-advisor, Associate Professor Proadpran Punyabukkana, Ph.D. Their unwavering support, guidance, and technical expertise have been invaluable throughout every stage of this project. Their insightful feedback, constructive criticism, and mentorship have significantly contributed to the quality and rigor of this research endeavor.

Lastly, I would like to thank my friends, family, and loved ones for their unwavering support and understanding during the completion of this thesis. Your encouragement, patience, and belief in my abilities have motivated me throughout this journey.

Nuttapol Kamolkunasiri

TABLE OF CONTENTS

	Page
ABSTRACT (THAI)	iii
ABSTRACT (ENGLISH)	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vi
LIST OF TABLES	viii
LIST OF FIGURES.....	ix
1. Introduction.....	1
1.1 Motivation.....	1
1.2 Object	1
1.3 Scope 2	
2. Related work.....	3
3. Background.....	4
Out-of-distribution Detection	4
3.1 Out-of-distribution Detection Methods.....	4
3.1.1 Maximum Class Probability (MCP)[4]:	4
3.1.2 Mahalanobis Distance[22]:.....	5
3.1.3 Out-of-Distribution Detector for Neural Networks (ODIN)[23]	5
3.1.4 Supervised classification	5
3.1.4.1 Visual Geometry Group (VGG)[24]:.....	5
3.1.4.2 Residual Network (ResNet) [25]:.....	6
3.1.4.3 Densely Connected Convolutional Networks (DenseNet)[26]:.....	6

3.1.4.4 Pyramid Localization Network (PYLON)[27]:	6
3.1.5 Adversarially learned anomaly detection (ALAD)[28]:	6
3.1.6 f-AnoGAN[29]:	7
3.1.7 The Efficient GAN-Based Anomaly Detection (EGBAD) [31]:	7
3.1.8 GANomaly [33]:	8
3.1.9 HealthyGAN [34]:	8
4. Method	9
4.1 Dataset	9
4.2 Generating Corrupted Images	10
4.3 Models	11
4.4 Evaluation Metrics	11
5. Experimental results	12
5.1 Results	12
6. Further Analysis	13
7. Conclusion and future work	19
8. Appendix A. Details of datasets	20
REFERENCES	25
VITA	29

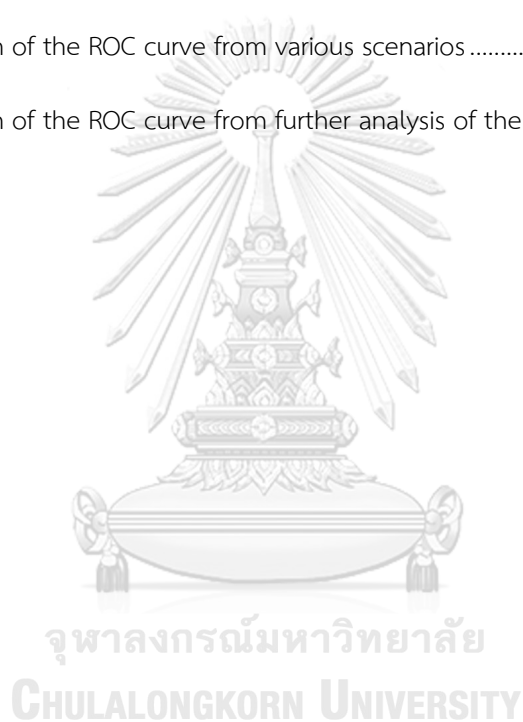
LIST OF TABLES

	Page
Table 1: Comparing the average AUROC and AUCPR scores with	12
Table 2: Out-of-distribution detection performance of models with unseen corruption classes..	14
Table 3: Performance of TB classification model (PYLON 50)	16
Table 4: Comparing the number of images in each class using various methods.....	17
Table 5: The percentages of corruption classes and example.....	20



LIST OF FIGURES

	Page
Figure 1: Sample corrupted images.....	10
Figure 2: Part of the training data	15
Figure 3: The model never encounters Coarse Dropout during training. The model completely fails to handle this.....	15
Figure 4: Comparison of the ROC curve from various scenarios	16
Figure 5: Comparison of the ROC curve from further analysis of the OOD data	18





จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

1. Introduction

1.1 Motivation

The number of medical imaging scans gathered over the previous decade now accounts for more than 90% of all medical data in hospitals worldwide [1]. As a result of the wide availability of imaging technology, automated methods for processing medical scans are increasingly more important to deal with the volume of pictures produced. Automation via machine learning, especially Neural Networks (NN), is critical to meeting this demand. On the other hand, deep neural networks have extremely limited extrapolation capabilities, and even fully trained models tend to act unpredictably when dealing with pictures that are outside the training data distribution. Given the possibility that sensory equipment, data transmission systems, and other software might fail unexpectedly, resulting in damaged images, these data should be flagged before sending to any automatic medical diagnosis model. This reduces the risk of producing erroneous, if not dangerously misleading, diagnoses.

1.2 Object

This thesis studies the out-of-distribution detection methods for validating data before sending it to the prediction models. The main hypothesis of this study is:

The corrupted data can reduce the robustness of the prediction models. Therefore, the out-of-scope detection techniques can be applied to screen the corrupted data and protect the robustness of the prediction models.

The objectives of this thesis are as follows:

1. To develop a machine learning technique for screening corrupted images prior to transmitting them to prediction models.
2. Comparing out-of-scope detection algorithms to evaluate picture data prior to prediction may increase the resilience of the prediction model.

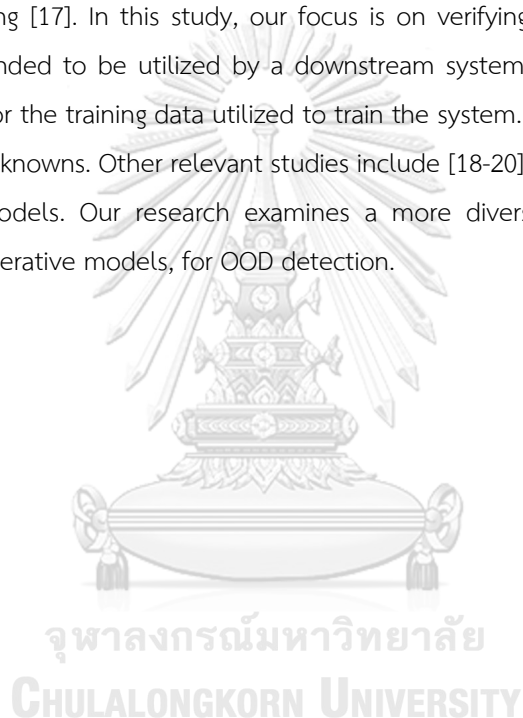
1.3 Scope

In this study, we assess the performance of various OOD detection methods on X-ray data. Identifying such distributional shifts are essential for securely deploying machine learning models in the medical industry. Several studies[2-4] have attempted to address this problem by training deep neural network classifiers to assign anomaly scores to inputs. Another set of techniques relies on reconstruction-based models for OOD detection[5]. A generative model, such as an auto-encoder, is trained using only in-domain (ID) data. The model should be able to adequately reconstruct any ID data. However, the reconstruction given by the generative model will have a large discrepancy when used on OOD samples, allowing their detection. This method is widely utilized in medical imaging for diagnostic purposes [6-8] because it gives an OOD score per pixel, allowing for unsupervised segmentation and localization of any abnormalities in the medical image. The majority of this line of research for OOD detection, however, has been conducted on toy datasets such as CIFAR-10 and CIFAR-100[9], TinyImageNet[10], LSUN[11], and MNIST[12]. Medical imaging data differs from these datasets in the variety of possible anomalies, higher fidelity in terms of pixel count and bit depth, and the tendency to have small localized anomalies. Moreover, some of these OOD differences might be caused by a difference in imaging technology or imaging apparatus, which can be so subtle that it is imperceptible to the human eye, but it can cause a significant drop in diagnosis performance by the machine learning

This work aims to survey and study the effect of different OOD detection methods on X-ray images to identify the type of anomalies that can be problematic for OOD detection and the best methods to tackle them. We start by creating multiple anomalies that can cause an X-ray diagnosis model fails, then tries to detect these anomalies via various methods in a systematic manner. To do this, the study re-implements and compares different approaches using a common testing framework for a fair and comprehensive comparison. Surprisingly, supervised methods outperform sophisticated OOD methods even on unseen anomalies by a large margin.

2. Related work

Despite its critical nature, the issue of image data validation has been largely ignored within the medical imaging field. On the other hand, many studies have been conducted on the control and assessment of picture quality [13, 14]. These techniques aim to determine the perceived image quality to identify issues like noise, compression artifacts, or lens distortion. However, there are no guarantees that these picture quality assessment models will not fail when dealing with OOD data. Previous research on out-of-distribution detection techniques mainly concentrated on finding aberrant or unhealthy structures through the use of statistical shape analysis [15], clustering [16], and one-class learning [17]. In this study, our focus is on verifying the authenticity of incoming images that are intended to be utilized by a downstream system, about which we possess no information except for the training data utilized to train the system. Thus, our problem setting has to deal with more unknowns. Other relevant studies include [18-20], which compare the results of confidence-based models. Our research examines a more diverse set of methods, including classification and generative models, for OOD detection.



3. Background

Out-of-distribution Detection

It has been shown that deep learning models can incorrectly categorize samples from distributions it has not been trained with. This gives rise to the topic of out-of-distribution (OOD) detection[21], which attempts to identify test samples with labels that do not overlap with the training data. It also requires models to reject label-shifted examples to ensure dependability and safety. In the OOD detection task, test samples are drawn from a distribution with a semantic shift away from in-distribution (ID), i.e., $P(Y) \neq P'(Y)$. The shift might be from a single class or a collection of classes. When several classes are present in training, OOD detection should not impair the capacity of ID categorization.

3.1 Out-of-distribution Detection Methods

OOD detection methods can be classified into three broad categories, namely confidence-based, generative, and classification. We list the methods considered below:

3.1.1 Maximum Class Probability (MCP)[4]:

MCP is a technique for identifying out-of-distribution (OOD) samples in a neural network. For the MCP method, the premise is that the classifier is more confident when accurately classifying in-distribution data and less confident when correctly classifying out-of-distribution samples. It is a straightforward method that uses the maximum class probability of the network prediction as a proxy for model confidence.

The approach compares the maximum class probability against a threshold to differentiate between in-distribution and out-of-distribution samples. The threshold can be determined using either the maximum class probabilities of the training samples or a validation set. In general, the maximum class probabilities of in-distribution samples are greater than those of OOD samples.

During the testing phase, it can also be altered dynamically based on the application and the required performance.

3.1.2 Mahalanobis Distance[22]:

Lee et al. proposed a technique for detecting anomalous samples, including adversarial examples and out-of-distribution data, by utilizing a pre-trained neural network classifier. The method is based on Gaussian discriminant analysis and uses the Mahalanobis distance to compute a confidence score for each sample. This method is more robust in situations where the training dataset has noisy labels or a small number of samples.

3.1.3 Out-of-Distribution Detector for Neural Networks (ODIN)[23]

ODIN is built on adding a small amount of noise to the input samples during the inference stage and assessing the improvement in confidence in the model's output. The idea is that out-of-distribution (OOD) samples will be less resilient to the added noise, resulting in a noticeable change in the model's confidence compared to samples from inside the distribution.

The approach applies a temperature-scaled noise to the model's input. The temperature scaling determines the noise level, and the output confidence of the model is calculated from the softmax output probability. By tracking the change in the output confidence, a threshold is applied to differentiate between data that belongs to the in-distribution and out-of-distribution categories.

3.1.4 Supervised classification

Another simple yet effective method is treating OOD detection as a binary classification between ID data and the rest. Multiple kinds of OOD data can be gathered and treated as positive data (OOD). This type of method has the possible caveat that it is impossible to envision all the possible anomalies beforehand, so the model might fail to detect a new kind of anomaly.

We consider four possible famous image classification models.

3.1.4.1 Visual Geometry Group (VGG)[24]:

VGG architecture is well-known for its extensive use of convolutional layers, often with small filters (3x3) and a relatively straightforward design. Each block of the VGG architecture consists of many convolutional layers, followed by max-pooling layers. The max-pooling layers limit the spatial resolution of the feature maps, making the network more resistant to input picture translations. The combination of convolutional filters and

pooling layers permits the network to learn a considerable number of distinct feature maps, which is advantageous for collecting fine-grained visual information.

3.1.4.2 Residual Network (ResNet) [25]:

ResNet's most significant innovation is the use of the so-called “residual connections,” which enable the model to learn a residual mapping between the input and output of a layer as opposed to the conventional strategy of learning the whole mapping. This alleviates the issue of disappearing gradients in deep neural networks, which can make training challenging. The ResNet design is composed of a succession of levels, with the possibility to use “shortcut connections” to bypass one or more layers. These shortcut connections permit the network to discover the residual mapping, which is the difference between a layer's input and output. By learning the residual mapping rather than the whole mapping, the network will learn more readily and converge more quickly.

3.1.4.3 Densely Connected Convolutional Networks (DenseNet)[26]:

The DenseNet architecture is recognized for its utilization of dense connections, which means that every layer is directly connected to all previous layers. In DenseNet, each layer receives the input feature maps from all preceding layers and combines them with its feature maps. This enables the network to learn a broader range of features and addresses the issue of disappearing gradients that can happen in deep networks. Additionally, the architecture includes transition layers that decrease the number of feature maps and control the expansion rate of the network. This makes the DenseNet architecture more efficient and less susceptible to overfitting.

3.1.4.4 Pyramid Localization Network (PYLON)[27]:

PYLON is a deep learning model that enhances the accuracy of identifying specific locations within images by using a technique called class activation maps (CAM) to generate high-resolution heat maps. It is particularly effective in identifying small objects and can be trained using only image-level labels, rather than requiring specialized annotations. PYLON has been shown to be effective in both general and medical image domains.

3.1.5 Adversarially learned anomaly detection (ALAD)[28]:

The ALAD method is similar to the AnoGAN[8] method, an early GAN-based anomaly detection method based on image reconstruction. ALAD improves on AnoGAN by creating bi-directional GANs that includes an encoder network for mapping data samples to latent variables. Unlike AnoGAN, which uses a standard GAN and requires computationally expensive inference procedures to recover latent variables, ALAD recovers these variables through a single feed-forward pass through the encoder network at test time, thus avoiding this computational cost. Furthermore, the anomaly scoring criteria used in ALAD is distinct from that in AnoGAN, and recent advances in GAN training are also incorporated to enhance model stability.

3.1.6 *f-AnoGAN*[29]:

Schlegl et al. proposed the *f-AnoGAN*, an anomaly detection system designed for real-time applications, that requires a two-step training process. The first step is the GAN training, where they utilized the Wasserstein GAN (WGAN) [30] architecture, which was state-of-the-art at that time. The second step involves training the encoder. During this step, the generator and discriminator parameters from the first step were frozen, while only the encoder was allowed to change. Three loss functions were used to enhance the model: the first loss function was based on the encoding of the real input images and their reconstructions, the second loss function resulted from the discriminator's classification of the real input images and their reconstructions, and the last loss function was based on a random latent space and the output from the encoder.

3.1.7 *The Efficient GAN-Based Anomaly Detection (EGBAD)* [31]:

The model presented by Zenati et al. in 2018 also utilized the bi-directional GAN (BiGAN) [32] concept, which includes an extra encoder compared to standard GANs. In BiGANs, the discriminator is responsible for classifying real latent codes and fake images or fake latent codes and real images.

The proposed model employs two additional discriminators, in contrast to ALAD.

The first discriminator attempts to differentiate between the pairs (x, x) and $(x, G(E(x)))$, while the second discriminator attempts to differentiate between the pairs (z, z) and $(z, E(G(z)))$. $x, z, G(), E()$ represent the in domain data, the embedding, the generator, and the encoder, respectively.

3.1.8 GANomaly [33]:

Akçay et al. proposed an enhanced GAN architecture by incorporating an encoder-decoder-encoder network and two additional losses (Adversarial Loss and Contextual Loss). The architecture comprises three sub-networks: an autoencoder network as the generator to learn the input data representation and reconstruct the input image, an encoder network for compressing the generator's output from the generator network, and a discriminator network for distinguishing between the input and output images as real or fake.

3.1.9 HealthyGAN [34]:

HealthyGAN is an image-to-image translation technique that works in a one-directional manner. The method comprises a generator network and a discriminator network, with the latter following the PatchGAN [35] architecture. The discriminator network determines whether an input image is real or fake. On the other hand, the generator network translates any images it receives, without requiring their labels, into healthy images. Once the model has learned how to perform the translation, it generates a difference map for each input image by computing the difference between the input image and its translated output. The difference map highlights regions with significant responses that correspond to potential anomalies.

4. Method

4.1 Dataset

The Montgomery dataset [36] was produced through a collaboration with the Montgomery County Department of Health and Human Services located in Montgomery, Maryland, USA. It consists of 138 frontal chest X-rays acquired as part of Montgomery County's Tuberculosis screening program. Among the 138 images, 80 were determined to be normal, and the remaining 58 exhibited symptoms of TB. The images are presented in Portable Network Graphics (PNG) format.

The Shenzhen dataset [36] was compiled in partnership with Shenzhen No.3 People's Hospital, Guangdong Medical College, and other institutions in Shenzhen, China. The dataset comprises 662 frontal chest X-rays, of which 326 are normal cases, and 336 are cases with TB symptoms, including pediatric X-rays (AP). The X-ray images are available in PNG format.

The RSNA Pediatric Bone Age Challenge data collection [37], which included 14,236 hand radiographs (12,611 training sets, 1,425 validation sets, and 200 test sets), was made accessible to challenge participants who enrolled.

The Pneumonia Detection Challenge dataset from the Radiological Society of North America (RSNA) [38] (hereafter, the RSNA dataset). has 30,000 frontal view chest radiographs, randomly assigned unique identification numbers, 16,248 posteroanterior views, and 13,752 anteroposterior views. It was extracted from the publicly available NIH CXR8 dataset containing only frontal views.

This paper employs several datasets, including the Montgomery and Shenzhen datasets[36] for chest X-rays and the RSNA Pediatric Bone Age[37] and Pneumonia Detection Challenge datasets[38]. The Montgomery dataset comprises 138 frontal chest X-rays, while the Shenzhen dataset contains 662 frontal chest X-rays. The RSNA Pediatric Bone Age Challenge data collection has 14,236 hand radiographs, while the RSNA Pneumonia Detection Challenge dataset contains 30,000 frontal view chest radiographs.

4.2 Generating Corrupted Images

We evaluate the models' performance using synthetic corruption, as we do not have access to corrupted image datasets. In our experiment, we established a criterion for selecting corruption classes by applying a corruption class to an image and submitting it to the trained model for Tuberculosis detection. The corruption class was considered a factor if it decreased the model's performance by more than five points in terms of AUROC. This results in 15 corruption/OOD classes. Test datasets were generated by applying the same corruption to images from the test splits. The applied corruptions can be classified into two categories: those that affect local image statistics while preserving the overall appearance and those that affect image-level statistics while preserving local image statistics. Examples of these corruptions are illustrated in Fig.1.

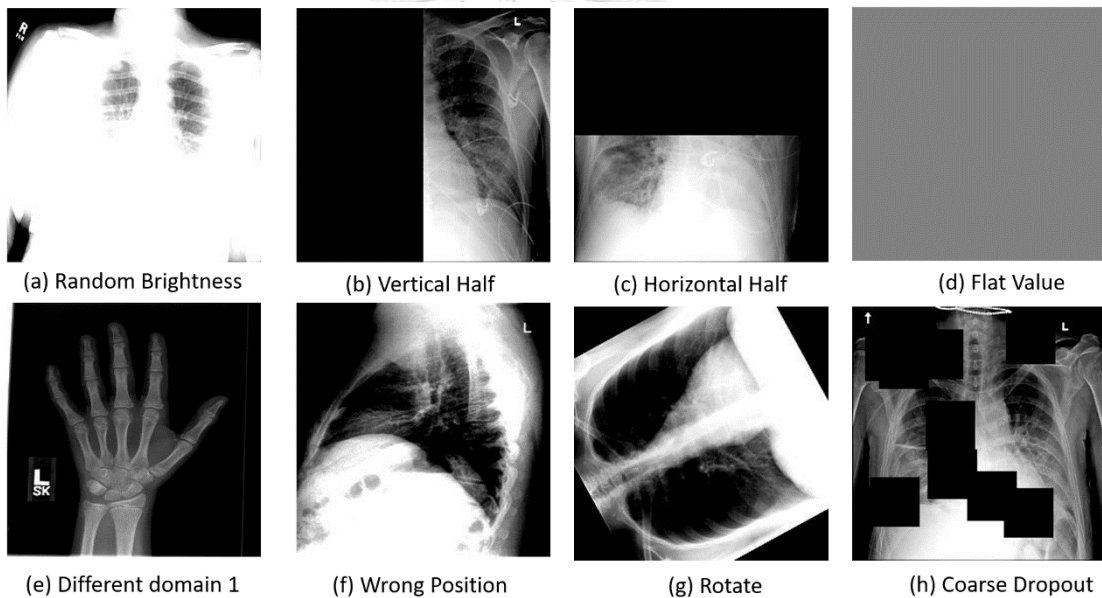


Figure 1: Sample corrupted images

In our study, we presented a summary of the different categories of corruption and their corresponding percentages of elements. These categories included In-Distribution (ID) at 40%, followed by various types of corruption such as Blur, Gaussian Blur (GB), Contrast, Brightness, Rotate, Gaussian Noise, Uniform Noise (UN), Flat Value (FV), Horizontal Half (H. Half), Vertical Half (V. Half), Coarse Dropout (CD), Wrong Position (WP), Different Source (DS), OOD1, and OOD2, each accounting for 4% of the total elements. In order to simulate a realistic out-of-distribution (OOD) detection task in a clinical setting, we separated the RSNA dataset into in-distribution (ID) and OOD images and applied various corruption classes. To represent the problem of “difference sources,” we

utilized the Montgomery and Shenzhen datasets, both of which contain images of the same disease. Additionally, the RSNA Pediatric Bone Age Challenge data was employed as “OOD1” to represent the problem of images coming from outside the domain. Lastly, a dataset of X-ray images from patients with HIV was used as “OOD2” to describe the problem of unseen diseases. This allowed us to simulate a realistic OOD detection task in a clinical setting and evaluate the performance of different OOD detection methods under various scenarios.

4.3 Models

In our study, we employed a selection of state-of-the-art models for OOD detection. Specifically, we used Resnet50, Densenet201, VGG-NETS19Bn, and PYLON as our classification models. For generative models, we used ALAD, EGBAD, f-AnoGAN, GANomaly, and HealthyGAN. And for confidence-based models, we used Mahalanobis Distance, MCP, and ODIN. In order to ensure consistency and fairness in our evaluation, we used the configuration values and hyperparameters as specified in the original papers for each of these models. This allowed us to accurately compare the performance of these different models under the same conditions and draw meaningful conclusions about their relative strengths and weaknesses.

4.4 Evaluation Metrics

We utilized the area under the receiver operating characteristic (AUROC) to measure the accuracy and performance of the model in classifying ID and OOD tasks. Moreover, in the real scenario, the imbalance problem is prevalent, and OOD is rarely encountered. Therefore, we employed the area under the precision-recall curve (AUCPR) to assess the model's performance from a precision and recall perspective.

5. Experimental results

5.1 Results

The performance comparison of out-of-distribution (OOD) detection between confidence-based, classification, and generative models is presented in Table 1. The results are an average of five runs. In all cases, the classification models consistently outperformed the baseline model in OOD detection, often by substantial margins. The generative models were more accurate than the confidence-based models. However, the generative models performed substantially worse than the classification models, except for HealthyGAN, which only lagged behind the classification model by a maximum of six percentage points.

Table 1: Comparing the average AUROC and AUCPR scores with different models

Model	AUROC score	AUCPR score
Classification		
— Densenet201	0.98	0.99
— PYLON	0.98	0.99
— Resnet50	0.98	0.99
— VGG-NETS19Bn	0.98	0.99
Confidence based		
— Confidence Branch	0.47	0.60
— Maximum Softmax Probability	0.38	0.52
— ODIN	0.44	0.58
Generative		
— EGBAD	0.79	0.87
— ALAD	0.75	0.84
— Ganomaly	0.73	0.82
— f-AnoGAN	0.60	0.74
— HealthyGAN	0.92	0.96

6. Further Analysis

We selected models with scores higher than 0.9 for further analysis. To provide a more realistic evaluation, we examined the performance of the out-of-distribution detection models under a condition where the models may not have seen certain corrupted classes before. We removed one class of corrupted data during the training phase and reintroduced it during the testing phase and repeated this process for all corrupted classes to determine which corrupted class has the most significant impact on the models.

In our study, we evaluated the impact of various corruption classes on the performance of out-of-distribution (OOD) detection models. The results of our analysis of the OOD data indicate that certain corruption classes significantly affect the performance of the models under investigation. Specifically, the “blur” class of corruption had a significant impact on the performance of the HealthyGAN model, while the “brightness” class had a significant impact on both the PYLON and HealthyGAN models. The “rotate” class had a significant impact on both the VGG-NETS19Bn and Resnet50 models, the “horizontal half” class significantly affected the Densenet201 model, and the “vertical half” class had a significant effect on the VGG-NETS19Bn model. The “difference source” and “OOD1” classes had a significant impact on the HealthyGAN model and the “OOD2” class had a significant impact on the PYLON model. Of all the corruption classes evaluated, Coarse Dropout was found to have the hardest type to detect. As shown in Fig. 2 and Fig. 3, the HealthyGAN model could not regenerate an image in the presence of Coarse Dropout if it had not been trained on this corruption class beforehand. Among all the models selected in our study, resnet50 had the best overall performance

Table 2: Out-of-distribution detection performance of models with unseen corruption classes.

Models	Corruption class														
	ID	Blur	GB	Brightness	Rotate	GN	UN	FV	H.Half	V.	CD	WP	DS	OOD1	OOD2
Densenet201	1.00	1.00	1.00	0.97	0.93	1.00	1.00	1.00	0.27	0.95	0.70	1.00	0.95	0.99	0.86
VGG-NETS19Bn	1.00	1.00	1.00	0.91	0.72	1.00	1.00	1.00	0.96	0.41	0.75	0.99	0.98	0.74	0.77
Resnet50	1.00	1.00	1.00	0.99	0.76	1.00	1.00	1.00	0.93	0.99	0.73	0.95	0.96	0.97	0.93
Pylon	1.00	1.00	0.99	0.76	0.97	1.00	0.94	0.97	0.97	0.90	0.67	0.97	0.88	0.90	0.60
HealthyGAN	0.95	0.47	0.99	0.78	0.87	1.00	1.00	1.00	0.90	0.99	0.59	0.89	0.69	0.92	0.85

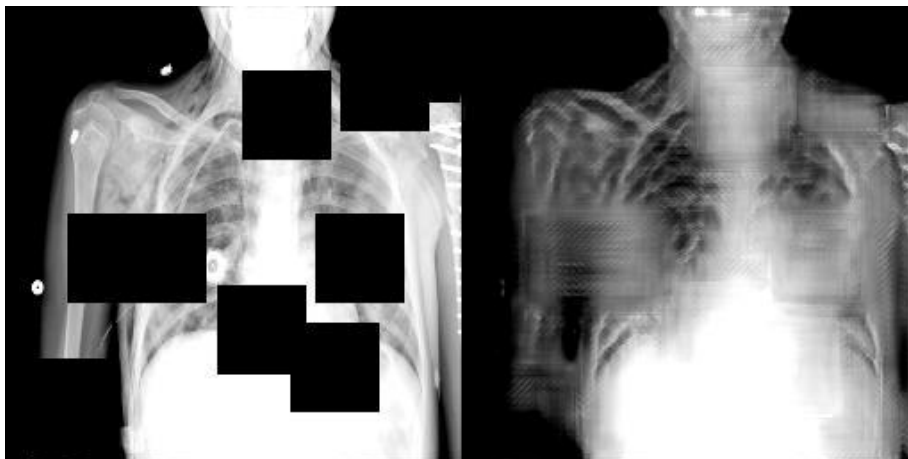


Figure 2: Part of the training data contains Coarse Dropout

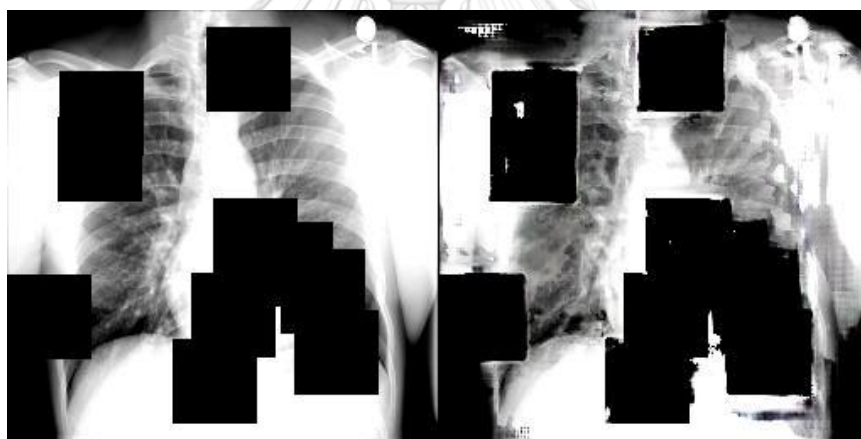


Figure 3: The model never encounters Coarse Dropout during training. The model completely fails to handle this kind of corruption if unseen.

To bridge the gap between testing and real-world scenarios, we conducted full-loop testing by simulating various scenarios and utilizing a new dataset, which is a private resource obtained from three hospitals in Thailand, namely Banglamung Hospital, Bureau of Tuberculosis, Thailand, and Maesot Hospital, Thailand. The dataset comprises two main classes: TB and non-TB. In the first scenario, we employed a TB classification model (PYLON 50) that had been trained using the standard process to classify a dataset and evaluate the model's performance. In the second

scenario, we randomly selected a dataset from the first scenario and converted 50% of the total data to out-of-distribution (OOD) data. We then allowed the model to classify this dataset again to examine the impact of OOD data on the model's performance. In the last scenario, we used the same dataset as the second scenario and employed an OOD detection model (Resnet50) to score the images and identify OOD images. We classified images with the top 20% OOD scores as OOD images, removed them from the dataset, and used this modified dataset to measure the model's performance and evaluate whether removing OOD data would improve the model's robustness. The results are displayed in Table 3.

Table 3: Performance of TB classification model (PYLON 50) with and without OOD data and OOD detection model (Resnet50)

Scenario	AUROC	FPR at 95% TPR
The dataset without OOD data	0.95	0.04
The dataset with OOD data	0.86	0.45
The dataset with filtering OOD data out 20%	0.88	0.27

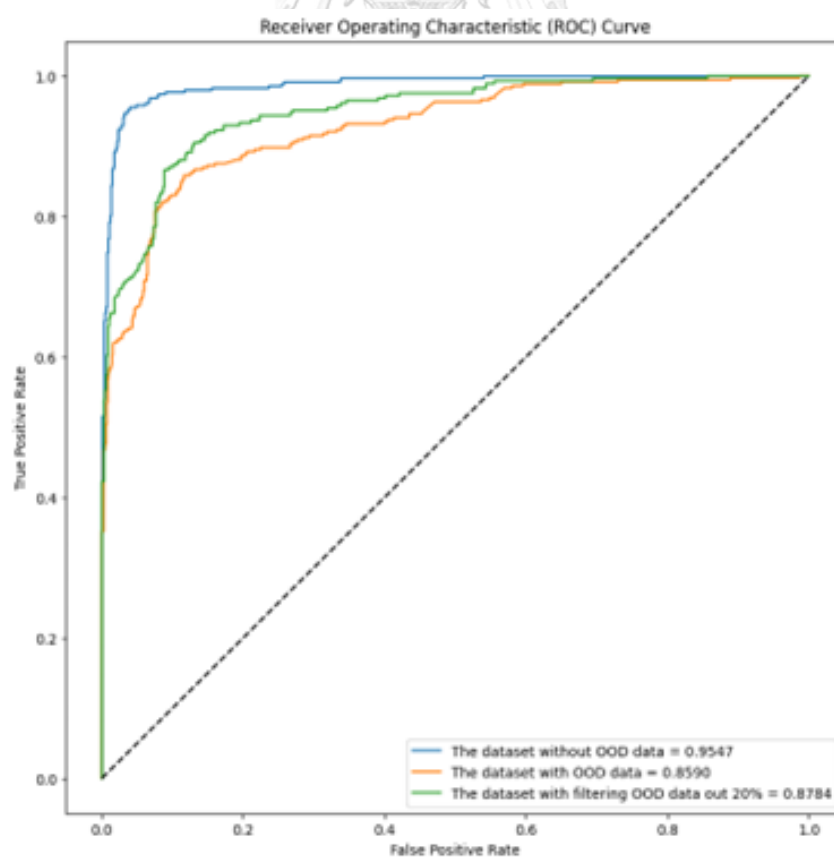


Figure 4: Comparison of the ROC curve from various scenarios

The first scenario showed that the model has the potential to classify TB patients with an accuracy of 95% and a false positive rate (FPR) of 0.04 at a true positive rate (TPR) of 95%. However, when the dataset contained OOD data, the accuracy dropped dramatically to 86%, and the FPR at 95% TPR increased significantly to 0.45, as shown in Table 3. Finally, the results of the third scenario demonstrated that detecting and removing OOD data before sending it to the model can improve its robustness. The model's accuracy with the OOD detector to screen the input was slightly higher than that of the model without the detector. Furthermore, the FPR at 95% TPR decreased dramatically by 18%.

We conducted further analysis of the OOD data to gain insights. When we exclusively fed the TB classification model with OOD data, the model's performance dropped significantly to 77.51% from its initial accuracy of 95%. This indicates that OOD detection can effectively identify data that adversely affects the model's performance.

However, we hypothesized that if we remove the output data for which the TB model lacks confidence (probabilities in the range between 40 and 60), would the model's performance remain the same? The results, as shown in Figure 5 and Table 4, demonstrate that the output from the OOD detection and the output obtained by removing the data within the 40 to 60 percentile range are identical. Moreover, it is important to note that the OOD detection approach removed 69 TB images, whereas the method of removing data at the specified percentile only removed 24 images. This implies that there is no need to feed the input data to the OOD detection method; instead, we can utilize the approach of removing the output data within the 40 to 60 percentile range to achieve the desired outcome and save on execution costs.

Table 4: Comparing the number of images in each class using various methods.

Method	Label		
	TB	Non-TB	Total
OOD data out 20%	69	169	238
Filtered data between 40 and 60 percentiles	24	215	239

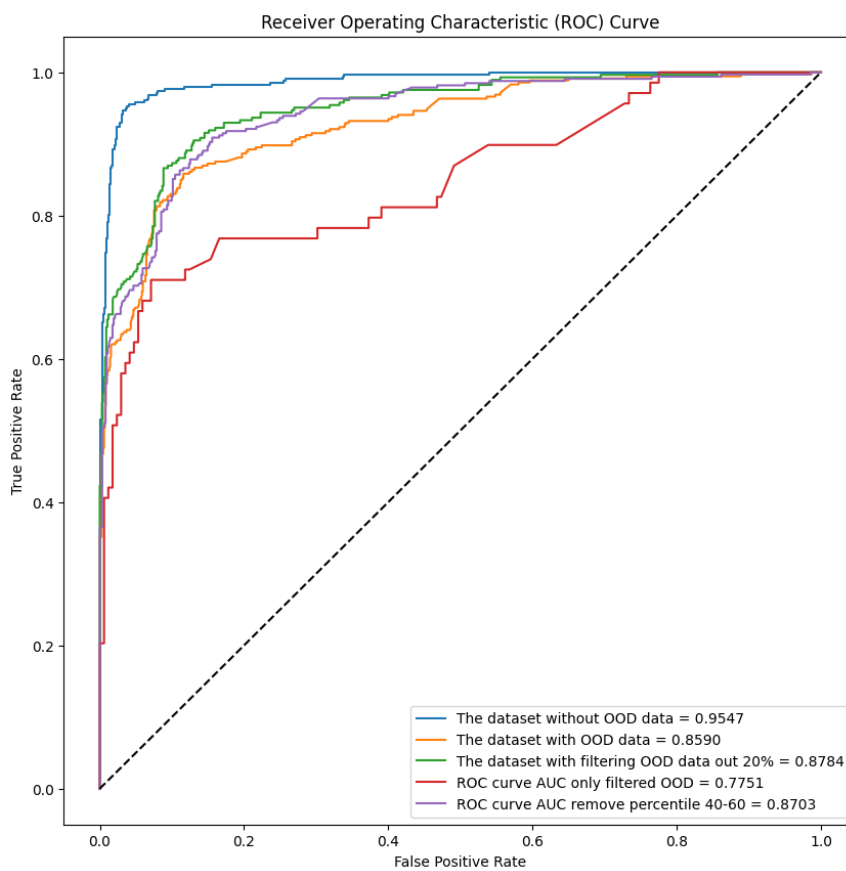


Figure 5: Comparison of the ROC curve from further analysis of the OOD data



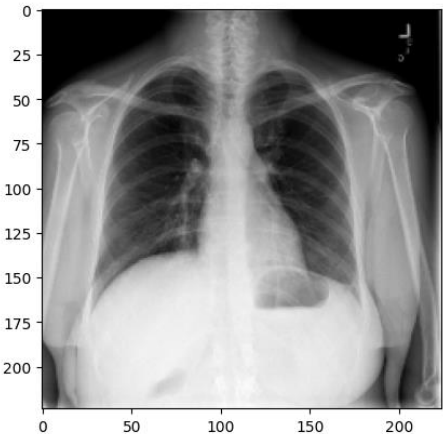
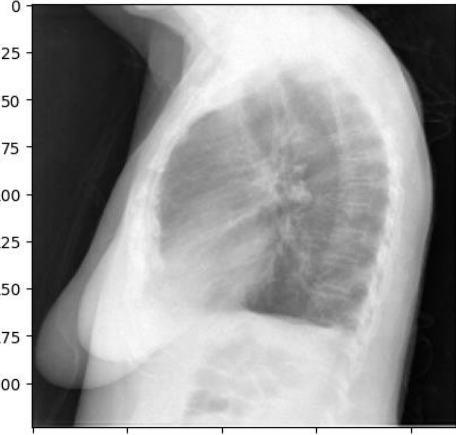
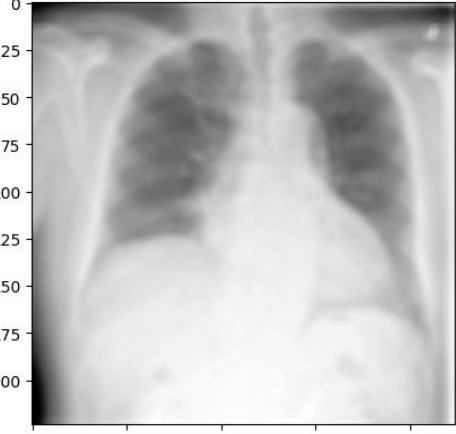
7. Conclusion and future work

This study aimed to evaluate various state-of-the-art methods for out-of-distribution (OOD) detection in medical imaging. Our findings indicated that methods based on simple classification models performed better on medical imaging tasks compared to those in the confidence-based and generative areas. Furthermore, we observed that the use of Coarse Dropout had a significant impact on the OOD detection performance for both classification models and HealthyGAN. Additionally, we conducted a real-world scenario to demonstrate the impact of OOD data and validated that screening the OOD data can enhance the model's robustness, including reducing the false positive rate. Further research should investigate OOD detection methods across other datasets and tasks, with a specific focus on the effects of various corrupted classes on the performance and reliability of these methods for real-world deployment.

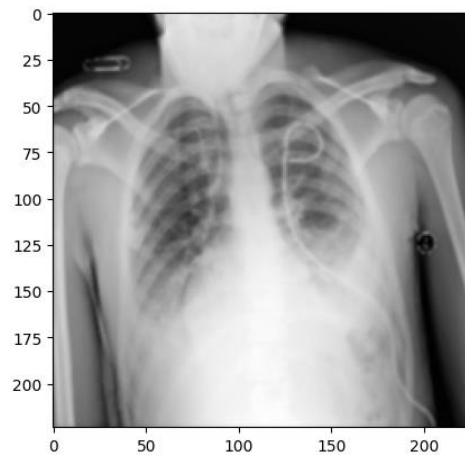


8. Appendix A. Details of datasets.

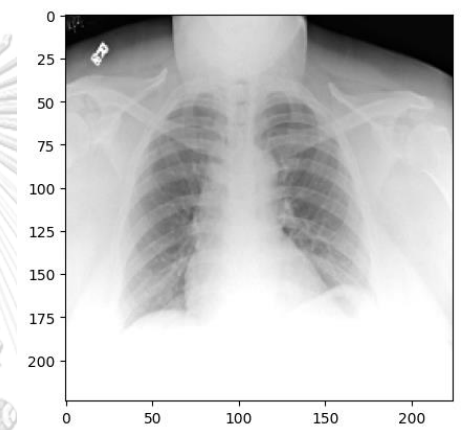
Table 5: The percentages of corruption classes and example.

Corruption class	Percentage	Example
No corruption	40%	
Wrong position	4%	
Blur	4%	

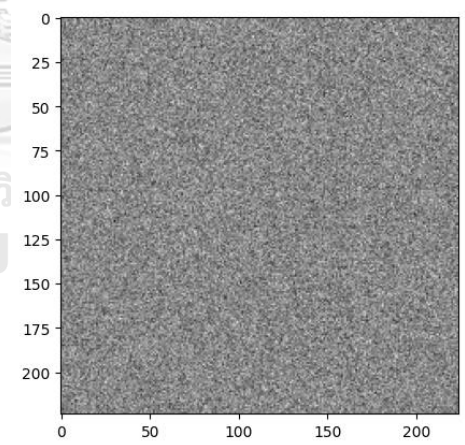
Gaussian Blur 4%



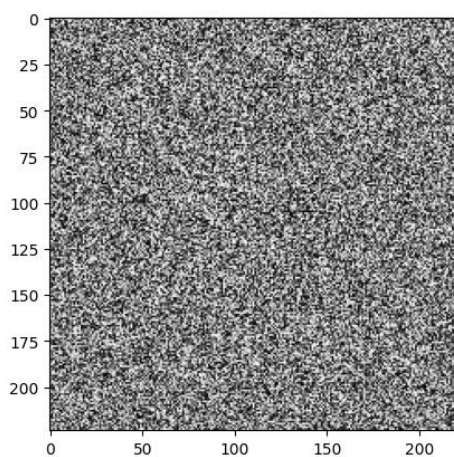
Contrast 4%



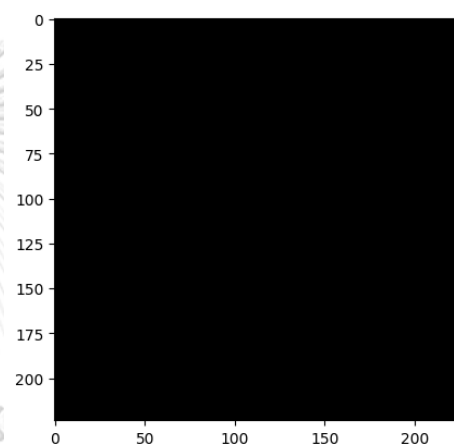
Gaussian Noise 4%



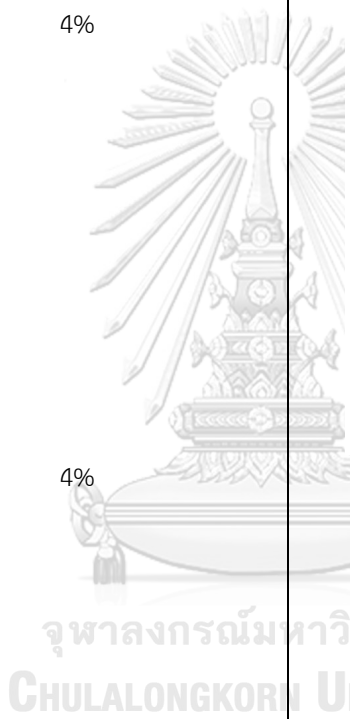
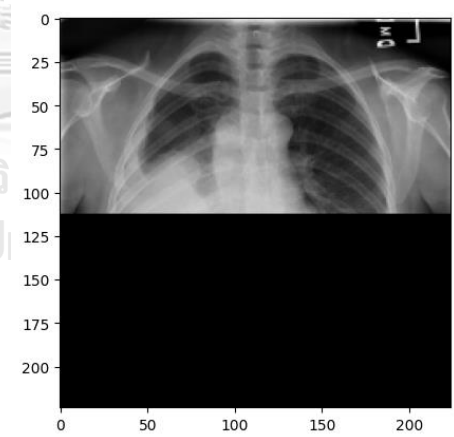
Uniform Noise 4%



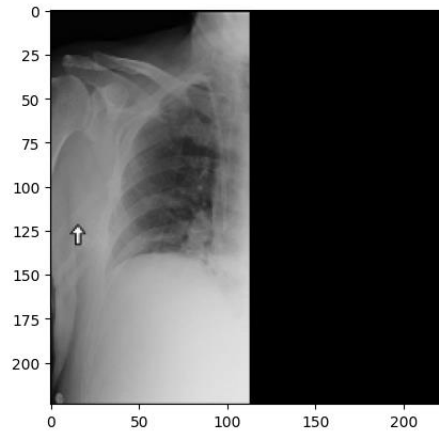
Flat Value 4%



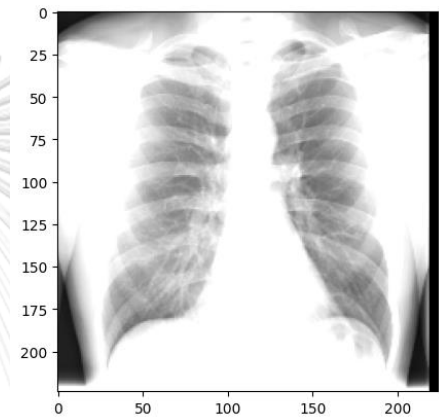
Horizontal Half 4%



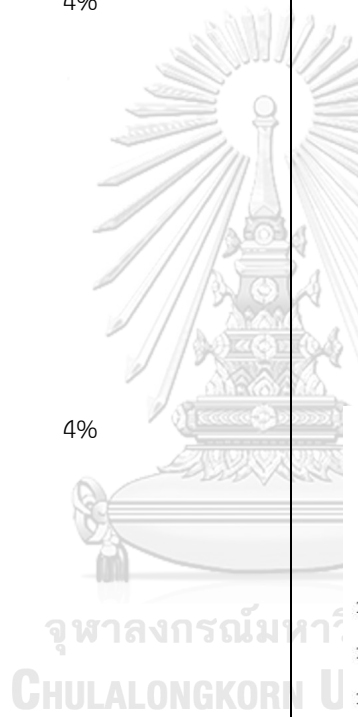
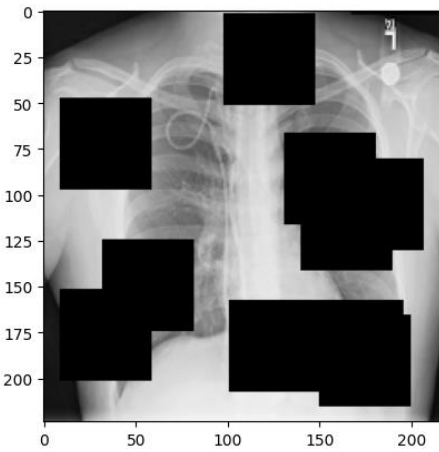
Vertical Half 4%



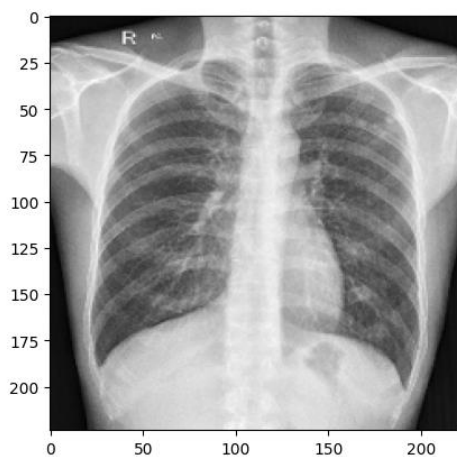
Brightness 4%



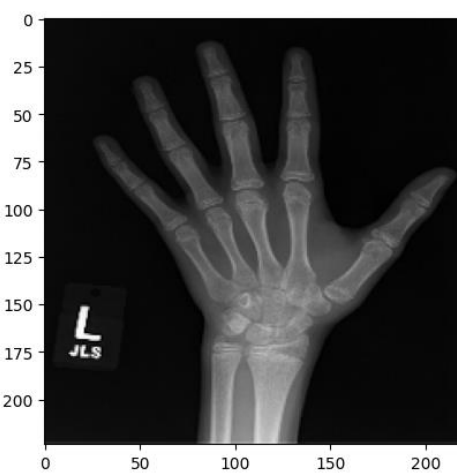
Coarse Dropout 4%



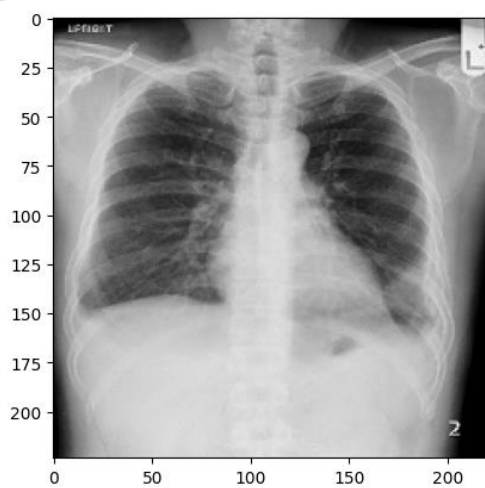
Different Source 4%



Out-of-Distribution 1
(Different Domain) 4%



Out-of-Distribution 2
(Different Symptom) 4%



REFERENCES

1. Márquez-Neila, P. and R. Sznitman. *Image data validation for medical systems*. in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV* 22. 2019. Springer.
2. Bendale, A. and T.E. Boulton. *Towards open set deep networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
3. DeVries, T. and G.W. Taylor, *Learning confidence for out-of-distribution detection in neural networks*. arXiv preprint arXiv:1802.04865, 2018.
4. Hendrycks, D. and K. Gimpel, *A baseline for detecting misclassified and out-of-distribution examples in neural networks*. arXiv preprint arXiv:1610.02136, 2016.
5. Japkowicz, N., C. Myers, and M. Gluck. *A novelty detection approach to classification*. in *IJCAI*. 1995. Citeseer.
6. Baur, C., et al., *Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study*. *Medical Image Analysis*, 2021. **69**: p. 101952.
7. Pawlowski, N., et al., *Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders*. 2018.
8. Schlegl, T., et al. *Unsupervised anomaly detection with generative adversarial networks to guide marker discovery*. in *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*. 2017. Springer.
9. Krizhevsky, A. and G. Hinton, *Learning multiple layers of features from tiny images*. 2009.
10. Wu, J., Q. Zhang, and G. Xu, *Tiny imagenet challenge*. Technical report, 2017.
11. Yu, F., et al., *Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop*. arXiv preprint arXiv:1506.03365, 2015.
12. LeCun, Y., et al., *Gradient-based learning applied to document recognition*. *Proceedings of the IEEE*, 1998. **86**(11): p. 2278-2324.
13. Woodard, J.P. and M.P. Carley-Spencer, *No-reference image quality metrics for structural MRI*. *Neuroinformatics*, 2006. **4**: p. 243-262.
14. Liu, Z., et al. *Quality control of diffusion weighted images*. in *Medical Imaging 2010: Advanced PACS-based Imaging Informatics and Therapeutic Applications*. 2010. SPIE.
15. Lekadir, K., R. Merrifield, and G.-Z. Yang, *Outlier detection and handling for robust 3-D active shape models search*. *IEEE Transactions on Medical Imaging*, 2007. **26**(2): p. 212-222.
16. Iakovidis, D.K., et al., *Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification*. *IEEE transactions on medical imaging*, 2018. **37**(10): p. 2196-2210.
17. Désir, C., et al. *A random forest based approach for one class classification in medical imaging*. in *Machine Learning in Medical Imaging: Third International Workshop, MLMI 2012, Held in Conjunction with MICCAI 2012, Nice, France, October 1, 2012, Revised Selected Papers* 3. 2012. Springer.

18. Berger, C., et al. *Confidence-based out-of-distribution detection: a comparative study and analysis*. in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. 2021. Springer.
19. Cao, T., et al., *A benchmark of medical out of distribution detection*. arXiv preprint arXiv:2007.04250, 2020.
20. Zhang, O., J.-B. Delbrouck, and D.L. Rubin. *Out of distribution detection for medical images*. in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*. 2021. Springer.
21. Yang, J., et al., *Generalized out-of-distribution detection: A survey*. arXiv preprint arXiv:2110.11334, 2021.
22. Lee, K., et al., *A simple unified framework for detecting out-of-distribution samples and adversarial attacks*. *Advances in neural information processing systems*, 2018. **31**.
23. Liang, S., Y. Li, and R. Srikant, *Enhancing the reliability of out-of-distribution image detection in neural networks*. arXiv preprint arXiv:1706.02690, 2017.
24. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556, 2014.
25. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
26. Huang, G., et al. *Densely connected convolutional networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
27. Preechakul, K., et al., *Improved image classification explainability with high-accuracy heatmaps*. *Iscience*, 2022. **25**(3): p. 103933.
28. Zenati, H., et al. *Adversarially learned anomaly detection*. in *2018 IEEE International conference on data mining (ICDM)*. 2018. IEEE.
29. Schlegl, T., et al., *f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks*. *Medical image analysis*, 2019. **54**: p. 30-44.
30. Arjovsky, M., S. Chintala, and L. Bottou. *Wasserstein generative adversarial networks*. in *International conference on machine learning*. 2017. PMLR.
31. Zenati, H., et al., *Efficient gan-based anomaly detection*. arXiv preprint arXiv:1802.06222, 2018.
32. Donahue, J., P. Krähenbühl, and T. Darrell, *Adversarial feature learning*. arXiv preprint arXiv:1605.09782, 2016.
33. Akcay, S., A. Atapour-Abarghouei, and T.P. Breckon. *Ganomaly: Semi-supervised anomaly detection via adversarial training*. in *Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. 2019. Springer.
34. Rahman Siddiquee, M.M., et al. *HealthyGAN: Learning from Unannotated Medical Images to Detect Anomalies Associated with Human Disease*. in *Simulation and Synthesis in Medical Imaging: 7th International Workshop*,

- SASHIMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings.* 2022. Springer.
35. Isola, P., et al. *Image-to-image translation with conditional adversarial networks.* in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2017.
 36. Jaeger, S., et al., *Two public chest X-ray datasets for computer-aided screening of pulmonary diseases.* *Quantitative imaging in medicine and surgery*, 2014. **4**(6): p. 475.
 37. Halabi, S.S., et al., *The RSNA pediatric bone age machine learning challenge.* *Radiology*, 2019. **290**(2): p. 498-503.
 38. Shih, G., et al., *Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia.* *Radiology: Artificial Intelligence*, 2019. **1**(1): p. e180041.





จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

VITA

NAME	Nuttapol Kamolkunasiri
DATE OF BIRTH	18 July 1983
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	Chulalongkorn University
HOME ADDRESS	5/207 Supalai Vill, Samrong Nue, Mueng, Samut Prakan, 10270

