Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2024 August 07.

Published in final edited form as:

IEEE J Biomed Health Inform. 2023 August ; 27(8): 3936–3947. doi:10.1109/JBHI.2023.3275104.

MedShift: Automated Identification of Shift Data for Medical Image Dataset Curation

Xiaoyuan Guo,
Department of Computer Science, Emory University, Decatur, GA 30322 USA

Judy Wawira Gichoya,
Emory University, Decatur, GA 30322 USA

Hari Trivedi,
Emory University, Decatur, GA 30322 USA

Saptarshi Purkayastha,
School of Informatics and Computing, Indiana University-Purdue University Indianapolis, IN 46202 USA

Imon Banerjee
Department of Radiology, Mayo clinic, Phoenix, and School of Computing and Augmented Intelligence, Arizona State University, AZ, 85259

Abstract

Automated curation of noisy external data in the medical domain has long been demanding as AI technologies should be validated on various sources with clean annotated data. To curate a high-quality dataset, identifying variance between the internal and external sources is a fundamental step as the data distributions from different sources can vary significantly and subsequently affect the performance of the AI models. Primary challenges for detecting *data shifts* are – (1) access to private data across healthcare institutions for manual detection, and (2) the lack of automated approaches to learn efficient shift-data representation without training samples. To overcome the problems, we propose an automated pipeline called *MedShift* to detect the top-level shift samples and evaluating the significance of shift data without sharing data between the internal and external organizations. *MedShift* employs unsupervised anomaly detectors to learn the internal distribution and identify samples showing significant shiftness for external datasets, and compared their performance. To quantify the effects of detected shift data, we train a multi-class classifier that learns internal domain knowledge and evaluating the classification performance for each class in external domains after dropping the shift data. We also propose a *data quality metric* to quantify the dissimilarity between the internal and external datasets. We verify the efficacy of *MedShift* with musculoskeletal radiographs (MURA) and chest X-rays datasets from more than one external source. Experiments show our proposed shift data detection pipeline can be beneficial for medical centers to curate high-quality datasets more efficiently. The code can be found at https://github.com/XiaoyuanGuo/MedShift. An interface introduction video to visualize our results is available at https://youtu.be/V3BF0P1sxQE.

xiaoyuan.guo@emory.edu .

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Index Terms—

Anomaly detection; dataset curation; medical shift data; X-ray; OOD detection

I. Introduction

While supervised deep learning has been promising in addressing various medical image tasks, the performance largely relies on the quality and quantity of annotations for training and evaluation, which primarily drives the necessity of generating high quality medical datasets from the academic healthcare and hospitals. Merely collecting a large-scale data from different sources is not sufficient [1], [40] because of the distribution shift and poor-quality data, which restricts the use of the data for both training and validation of the deep learning models, and are referred as *shift data* in this paper. The shift data introduces *out-of-distribution* (OOD) in the dataset, and should account for the performance dropping of well-trained models. Thus, identifying the shift data in advance is crucial for curating the datasets which could be extremely helpful in enhancing the model's generalization with future training. Unfortunately, sharing data from external sources can be relatively difficult due to the privacy concerns [26], complicated anonymization process and legal requirements. Thus, there still lacks an effective way to identify the reason for model generalization failure between various datasets from the same medical domain. Despite the efforts to pool and curate deidentified medical data for open-source research purposes [28], most medical data is still isolated and locally stored in hospitals and laboratories due to the challenges associated with sharing patient data [29]. Therefore, an efficient way of external dataset curation/cleaning without sharing data is needed to scale deep learning model validation and generalization.

To overcome the limitations, we propose *MedShift* pipeline to identify "shift" data among external datasets and allow model generalization. Instead of sharing external datasets with the internal site, we share models trained with internal distribution knowledge to external sources and detect outliers by computing the data quality differences from the internal. As observed by [7], [13], domain-discriminating approaches tend to be helpful for characterizing shifts qualitatively. Therefore, we utilize unsupervised anomaly detectors to learn the "normality" of indomain features. Given that the internal dataset has multiple classes, we suggest training an OOD detector for each class to learn the class-specific feature representation. This helps minimize the interference of variations from other classes and improve the sensitivity of intra-class variations from external classes. With the shared anomaly detectors, all the samples in the external dataset can acquire corresponding anomaly scores, which are indicators of how different a sample is from the in-distribution dataset. Still, it is desired to quantify the shiftness of external outliers. Inspired by the fact that the supervised deep learning suffers from the performance dropping when facing the distribution/dataset shifting, especially when training data and test data are from two sources, we propose to measure the *shiftness* of outlier data by checking the performance variance of a well-trained supervised model. Instead of inspecting the shift sample one by one, *MedShift* quantifies the *shiftness* for each class in small groups to save the efforts of domain experts for future analysis based on the observation that samples with close anomaly

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

scores often share heavy similarities (see Sec. IV-D). Based on the assigned anomaly scores, each class of the external datasets is clustered into multiple groups. Data samples with similar qualities will be grouped together. A multi-class classifier is then trained on the internal dataset and evaluated on the external datasets. Each group of each class in external datasets is gradually dropped in the decreasing order of anomaly scores. Meanwhile, the classification performance on the updated external data is recorded. The corresponding variation in performance, hence, reflects the significance of the distribution shift based on the fact that subtle changes in data distribution may affect the performance of well-trained classifiers. Additionally, we adopt a dataset quality metric (OTDD [37]) for helping facilitate the comparison of differences among a series of datasets coming from the same medical domain. Although sharing certain concept similarity with [8], our work solves more realistic medical applications and explores more possible data modalities.

We summarize our contributions as follows:

1. Propose an automatic pipeline of identifying shift data for medical data curation applications and evaluating the significance of shift data without sharing data between the internal and external organizations;
2. Employ different unsupervised anomaly detectors to learn the internal distribution and identify samples showing the significant *shiftness* for external datasets, and compared their performance;
3. Quantify the effects of the shift data by training a multi-class classifier that learns internal domain knowledge and evaluating the classification performance for each subgroup of each class in external domains after dropping the shift data;
4. Adapt a data quality metric to quantify the dissimilarity between the internal and external datasets;
5. Experiment on two pairs of representative medical datasets and show effective qualitative and quantitative results, which prove the usefulness of the suggested pipeline for future medical dataset curation.

II. Related Work

This paper focuses on the automated shift data identification of inaccessible external datasets for medical dataset curation, which involves two main parts: shift data identification and shiftness quantification. There are several research works of shift data identification for medical dataset curation [7], [8] and has been observed that OOD/anomaly detection methods are capable of capturing the shift in unseen OOD datasets. Generally, anomaly detection can be loosely categorized into unsupervised and supervised methods based on the object functions [43]. The supervised models [33] assume the external datasets are accessible for training using the preconceive anomaly classes, which is different from our experimental settings since the anomaly types in the external datasets is not known in advance. Unsupervised anomaly detectors can be summarized into reconstruction-based, generation-based and classifier-based [5]. Reconstruction-based methods [35], [36] utilize AutoEncoders (AEs), Variational AutoEncoders (VAEs) to reconstruct in-distribution data

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and treat the reconstruction error the anomaly score. Moreover, a recent work SSPCAB [34] can be integrate into different architectures with its designed reconstruction-based functionality to improve anomaly detection ability; generative models encourage the generated data to be similar with in-distribution data, the generation and discrimination errors together contribute to final anomaly score. f-AnoGAN [17] and GANomaly [42] belong to the category; classification-based models usually transform the original images into other formats and distinguish the transformed data from the original in-distribution data, representative methods.

Moreover, measurements of the data shiftness quantification are also important for comparison. A recent work [8] measures domain shift for histopathology data by computing the distance between the ID and OOD sources regarding to the intermediate deep features. By modeling the latent distributions, both Wasserstein distance and Kullback-Leibler (KL) divergence are considered to help quantify the representation shift. The other work Optimal Transport Dataset Distance [37] (OTDD) metric calculates distances between two classification datasets. It relies on optimal transport [38], which is a flexible geometric method for comparing probability distributions, and can be used to compare any two datasets, regardless of whether their label sets are directly comparable.

III. Methods

In Section III-A and III-B, we formulate the dataset shift identification problem and introduce the necessary notations. Then, we propose and illustrate the pipeline of shift identification in Section III-C. To complement, we introduce the details of our anomaly detection architecture used for *MedShift* pipeline in Section III-D; we further dive deep in the *shiftness* evaluation in Section III-E. Additionally we introduce the dataset quality measurement in Section III-F.

A. Problem Statement

In view of the fact that the digital healthcare research is hugely limited by the data sharing and privacy issues because of the regulation imposed by Health Insurance Portability and Accountability Act (HIPPA), *MedShift* aims to overcome the barrier by exploiting the advantage of sharing data quality evaluation models across the organizations and inspects the *shiftness* of external datasets based on the learnt internal domain.

B. Formulation and Notation

Given two datasets D_A and D_B of the same medical domain with the same classes (say $c_1, c_2, \ldots, c_n$, n is the total number of classes) from two intuitions A and B (e.g., a chest X-ray dataset from Emory University D_A and a chest X-ray dataset from Stanford University D_B), let D_A be the internal dataset and D_B be the external dataset. Dataset distribution shift is termed the situation where $P_{D_A}(Y|X) = P_{D_B}(Y|X)$ but $P_{D_A}(X) \neq P_{D_B}(X)$, where Y and X represent the class labels and input data respectively.

Suppose we are given an independent and identically distributed (i.i.d.) internal dataset $\{X_{c_i}^A\}_{i=1}^n$ with n classes, and input samples $\{x_j^{Ac_i}\}_{j=1}^{N_{c_i}^A} \subset X_{c_i}^A$ ($N_{c_i}^A$ is the sample number of dataset

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

A's class c_i) from the internal input distribution, and i.i.d. external data $\{X_{c_i}^B\}_{i=1}^n$ and input samples $\{x_j^{Bc_i}\}_{j=1}^{N_{c_i}^B} \subset X_{c_i}^B$ ($N_{c_i}^B$ is the sample number of dataset B's class c_i) from external distribution, the detection of class-wise distribution shift for dataset D_B based on D_A is to identify the anomalous samples $\overline{X}_{c_i}^B \subseteq X_{c_i}^B$. Take D_A class data as in-distribution (ID) data and train machine learning models (e.g. classification models), the models can learn the distribution of D_A's classes and make predictions $P(y_{c_i}^A | x_{c_i}^A)$ for some targets $y_{c_i}^A$ given data samples $x_{c_i}^A$ for class c_i. Theoretically, given the target model trained on the ID data $X_{c_i}^A$, the predictions over set $X_{c_i}^B \subseteq \overline{X}_{c_i}^B$ should produce more relevant results than on the whole set $X_{c_i}^B$.

C. Shift Data Identification

In this section, we introduce the methodology for identification of image data distribution shift to discriminate the poor-quality, noisy and under-represented samples from the external data in an automatic way. The whole pipeline is built on top of the anomaly detection architecture to leverage the anomaly score as illustrated in the framework in Fig. 1, which involves two separate phases - internal training (step (1)) and external test phase (step (2)).

During the training phase, only internal data samples and the anomaly detection models (see introductions in Sec. III-D) are involved. As shown in step (2) (the left blue part) of Fig. 1, a set of anomaly detectors $\mathcal{F}$s for each targeted category of D_A are trained on the internal dataset in an unsupervised fashion, considering the unavailability of external data sources. Each class will then obtain a unique OOD detector $\mathcal{F}$c. The anomaly detector learns to assign each data item with a specific anomaly score, a higher score means more possibility of being an anomalous data. Notably, the anomaly detectors are trained with accessible internal data, and then shared with the external validation sites.

In the test phase, no internal data will be shared but the trained anomaly detector model with shift identification capability will be exchanged. As represented with pink figures and dotted flows in Fig. 1 step (2), each trained anomaly detector is evaluated on each corresponding class of dataset D_B and assigns anomaly scores for the external dataset. To prepare for the *shiftness* quantification in Sec. III-E, an unsupervised clustering algorithm is subsequently applied to each class and clusters the data items into k groups based on the learnt anomaly scores shown in step (3) of Fig. 1. For each class, the optimal number of cluster k is determined by the Elbow Method. More details and results can be found in Sec. III-E and Sec. IV-D.2.

D. Anomaly Detection

Architecture.—As claimed in Sec. III-C, we propose to utilize anomaly detection models for not only identifying distribution shifts in the external dataset but also automated cleaning the external data without data exchange. First, we briefly describe our anomaly detection model - *Cascade Variational autoencoder-based Anomaly Detector (CVAD)* [30] used in *MedShift*. The reasons for selecting CVAD as our main anomaly detector are: (1) CVAD is an unsupervised anomaly detector, which only replies on in-distribution/internal data information; (2) it has previously been tested on both generic and medical image datasets

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and showed descent performance of detecting both intra- and inter-class OOD data; and (3) it detects anomalies by modeling internal dataset distributions benefiting from its cascade VAE architecture. As shown in Fig. 2, CVAD combines latent representation at multiple scales using the cascade architecture of variational autoencoders and thus, can reconstruct the in-distribution image x with high quality. Both the original image x and the reconstruction x' are then fed into a binary discriminator D to separate the synthetic data from the in-distribution ones.

Objective and Optimization.—A standard VAE's encoder $q_\phi(z|x)$ (parameterized by ϕ) maps the visible variables x to the latent variables z and the decoder $p_\theta(x|z)$ (parameterized by θ) samples the visible variables x given the latent variables z. Given a dataset drawn from some underlying data distribution $p^*(x)$, ϕ and θ are then learned by maximizing the variational lower bound (ELBO)$L(\phi, \theta)$, which is a lower bound to the marginal log-likelihood $\log p(x|\theta)$ [4].

To optimize the generator of CVAD, we minimize two objectives for the primary VAE part in (1) and the branch VAE part in (2), KL refers to Kullback-Leibler divergence.

$$L(x;\phi_1,\theta_1) = E_{z_1\sim q_{\phi_1}(z_1|x)}[log\ p_{\theta_1}(x|z_1)] + D_{KL}(q_{\phi_1}(z_1|x)\,||\,p_{\theta_1}(z_1)) \tag{1}$$

$$L(x;\phi_2,\theta_2) = -E_{z_2\sim q_{\phi_2}(z_2|f(x))}[log\ p_{\theta_2}(x|z_2)] + D_{KL}(q_{\phi_2}(z_2|f(x))\,||\,p_{\theta_2}(z_2)) \tag{2}$$

where $f(x)$ is the input of branch VAE, encoded by E_{11}. Therefore, the "generator" loss can be formulated as Eqn. 3. α_1 and α_2 to balance the weights of the two individual terms.

$$L_{\text{rec}} = \alpha_1 L(x;\phi_1,\theta_1) + \alpha_2 L(x;\phi_2,\theta_2) \tag{3}$$

The binary discriminator is trained to distinguish true/fake images using binary cross entropy loss (i.e.,L_{dis}).

Anomaly score.—The final anomaly score S is defined in Eq. 4 based on errors during inference and includes two parts: the reconstruction error L_{rec} output by the generator and the probability of being the anomaly class S_{dis} output by adding the two parts together.

$$S = L_{rec} + S_{dis} \tag{4}$$

This gives us the advantage that when dealing with heavy noisy data, the reconstruction error will be the dominant indicator for *shiftness*; when facing the hard distinguished cases the class probability plays the decision role.

Implementation.—We resize all the medical images to $256 \times 256 \times$ *channel* for simplicity considering the irregular image sizes. To train, we use the Adam optimizer with a batch size of 256 and 2,048 for MURA and chest X-ray dataset, respectively; we set the learning rate of 1×10^{-5} and 1×10^{-3} for the generator and the discriminator of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

proposed method(CVAD), respectively; we train the generator with 250–500 epochs and the discriminator with 10–20 epochs.

E. Shiftness Quantification

The above pipeline can be applied to detecting the shift data and assigning each data with an anomaly score to indicate its contribution to the dataset shift. Nonetheless, quantifying the *shiftness* of a single sample is not trivial due to the minor change of both the dataset quality and model performance variations. Instead, we suggest evaluating them in groups. As prepared in the first stage (step (3) in Fig. 1) of the whole pipeline, the clustering has split each class of dataset D_B into multiple groups according to the anomaly scores. For simplicity, we assume that each class has k groups. To evaluate the significance of detected outliers, we train a multi-class classifier $\mathcal{G}$ for D_A (step (4) in Fig. 1) and test on D_B (step (5)(6)…(4+k) of Fig. 1). We gradually drop one group that has the largest anomaly scores among current groups for each class until only one group remains. The corresponding class-wise classification performance is recorded. The performance variation thus is an indicator of the *shiftness* of the specific group.

Multi-class classifiers' details.—To quantify the *shiftness* of each clustered group for each class of external dataset D_B, we first train a multi-class classifier $\mathcal{G}$ for the internal dataset D_A. The classifier learns the class latent features of the internal domain and is able to predict class labels for test data. For MURA data, we train ResNet152 [22] on the Emory MURA dataset with the publicly available pretrained weights as initialization. We optimize the classifier using the Adam optimizer with a batch size of 512, a learning rate of $\times 10^{-3}$ for 50 epochs. For chest X-ray data, we utilize the model proposed by Ref. [16], which originally aims for multi-label classification of the CheXpert dataset, and modifies it for the Emory_CXR 14-class classification task. Following the same implementations in Ref. [16], we use DenseNet121 [32] as the feature extraction backbone and initialize it with the public pretrained model weights. We train the classifier with a batch size of 256 for 20 epochs. The corresponding classification performances, including the *Precision*, *Recall*, *F1-score* and *AUC* score are reported in Sec. IV-D.3.

F. Dataset Quality Measurement

To further quantify the efficacy of identifying the shift data among external datasets, we measure the quality of external datasets compared to the internal dataset and observe the difference after removing the shift data from the external sources in an iterative fashion. We apply OTDD [37] to measure distances between our internal and external datasets. Formally, the optimal transport dataset distance is defined as:

$$OTDD(\boldsymbol{D}_A, \boldsymbol{D}_B) = min_{\pi \in \prod(PA, PB))} \int_{Z \times Z} d(z, z') d\pi(z, z') \tag{5}$$

, of which

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$d(z, z') = (d(x, x')^2 + W_2(P_y, P_{y'})^2)^{\frac{1}{2}} \quad (6)$$

, where D_A, D_B are the two datasets, x, x' and y, y' are their samples and labels respectively, W_p denotes the p-Wassertein distance. Please refer Ref. [37] for more details.

IV. Experiments

A. Datasets

We evaluated our model on two diverse open source medical imaging datasets: (1) *Musculoskeletal radiographs* - Private Emory MURA dataset (internal) and Open-source Stanford MURA dataset [2] (external); (2) *Chest radiographs* - Private Emory Chest X-rays (internal, Emory_CXR in short), Open-source CheXpert dataset [41] (external_1) and MIMIC dataset [31] (external_2). We obtained the private datasets with the approval of Emory Institutional Review Board. More details about the datasets are presented in Table. I. Notably, each chest X-ray data may have multiple common diagnoses, different from the MURA dataset where class labels are mutually exclusive.

B. Anomaly Detectors in Use

OOD detection plays an important role in identifying shift data in external datasets. We adopt CVAD [30] as the main anomaly detector,which has been introduced in Sec. III-D. As this method poses no assumption on the input data and the applied situations, we utilize this anomaly detection architecture in our pipeline called *MedShift_w_CVAD* across all the experiments. The implementation code of CVAD is available at https://codeocean.com/capsule/3191573/tree/v1. For comparison, we select f-AnoGAN [17] as the baseline and apply the method in *MedShift* for comparison (*MedShift_w_fAnoGAN* in short). We also show the performance of SSP-CAB [34] by combining its self-supervised predictive convolutional attentive block with CVAD architecture, the corresponding shift data detection pipeline is named *MedShift_w_SSPCAB*.

C. Experimental Setup

For convenience, we have listed most of the symbols and the descriptions that have been used in Table II for quick query. We implement the pipeline using Pytorch 1.5.0, Python 3.7.3 and Cuda compilation tools V10.0.130 on a machine with 4 NVIDIA RTX A6000 GPUs with 48 GB memory. More details about the training of anomaly detectors and classifiers are introduced below.

D. Results

In this section, we evaluate the performance of our pipeline on three objectives - (i) shift data identification, (ii) shift data partition and (iii) shift data significance evaluation. To increase readability, one representative class is selected for explanation.

1) Shift Identification with Anomaly Detection: In the process of identifying the shift data from the external source, each class of the internal dataset will obtain its own anomaly detector. Figure 3 presents the anomaly score distributions of the *HAND*

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

class of MURA dataset obtained by *MedShift_w_CVAD*, *MedShift_w_fAnoGAN* and *MedShift_w_SSPCAB* architectures. The X-axis represents the anomaly score and Y-axis stands for the number of images that have the original anomaly scores in the corresponding range. In all cases, Emory data is considered as internal data. We normalize the anomaly scores of *MedShift_w_CVAD* and *MedShift_w_SSPCAB* results with a sigmoid function as min-max normalization fail to work for unseen data distributions. We show the original scores of *MedShift_w_fAnoGAN* as they give better distribution visualization.

For **MURA** dataset, the anomaly score distribution of *MedShift_w_CVAD* for *XR_HAND* is shown in the left of Fig. 3, with the blue curve for Emory *XR_HAND* and the orange distribution curve for Stanford *XR_HAND* data. As can be observed, the peaks of the two distributions are clearly separated, the Stanford data generally gets higher OOD scores than the internal Emory data. The difference between the internal and external anomaly score distributions can be easily observed. The closer and more similar the two distributions are, the less shift the external dataset has. Comparatively, the internal and external anomaly score distributions of *MedShift_w_fAnoGAN* heavily overlap with each other, indicating a limited discriminative ability of detecting shift data. *MedShift_w_SSPCAB* assigns higher anomaly scores to external sources, but still has a certain amount of overlapping with the internal source.

The similar phenomenon can also be seen in chest X-ray data when being tested on two external datasets. For **chest X-ray** dataset, the OOD detection for *Fracture* is shown in Fig. 4, with the blue histogram and curve for internal Emory_CXR dataset, the orange for CheXpert dataset and the green for MIMIC dataset. The differences in the distributions reflect how different the external chest X-ray data is from the internal domain. Both CheXpert and MIMIC *Fracture* distributions show significant shifts with the internal Emory_CXR distribution, which indicates that external *Fracture* shift data exists and can be identified by CVAD. For *MedShift_w_fAnoGAN*, the internal distribution curve is inside the two external curves, which means little variance has been identified by fAnoGAN. Differently, *MedShift_w_SSPCAB* shows similar performance with CVAD for MIMIC dataset but tends to treat all CheXpert data as outliers.

2) Shift Data Clustering Results: In this section, we showcase the clustering results based on anomaly scores for both MURA and chest X-ray datasets. Specifically, Stanford MURA dataset, CheXpert and MIMIC data are clustered into different groups according to their anomaly scores obtained in the previous step. The selection of group numbers is decided by the Elbow distortion curves. Take MURA *HAND* class as an example, Fig. 5 illustrates the curve plots of *MedShift_w_CVAD*, *MedShift_w_fAnoGAN* and *MedShift_w_SSPCAB*. For all the three situations, we pick 5 for group numbers. As observed during our experiments, the curve plots across all the classes showing similar patterns. Therefore, we keep k as 5 in the other datasets as well. The corresponding clustered examples can be seen in Fig. 6. There are 5 cluster groups in total, with each row representing one cluster. The groups are sorted in ascending order, namely, the top row is with the lowest anomaly scores and the bottom has the largest anomaly scores. For better understanding, their corresponding scores are labelled on top of each example item. As can be observed, the hand data of left figure gradually shows more and more variations in terms

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of image quality, positioning, and noise, as the anomaly score becomes large, especially when comparing the group 1 (first row with lowest anomaly score) to group 5 (last row with highest anomaly score). The variance exhibiting in the abnormal data indicates the existence of distribution shift in the external dataset. Nonetheless, the significance of the detected under-represented/shift data samples in affecting deep learning models' prediction/classification remains to be explored. In comparison, the results of fAnoGAN fail to demonstrate a clear variation pattern for each cluster group. The mixture of shift data across different groups hinders the detection of shift data identification.

Similarly, an example of **chest X-ray** *Fracture* is presented in the right of Fig. 7. Following the same arrange order, the difference for each group can be clearly captured by our model.

3) Classification Results for Shiftness Evaluation: As introduced in Sec. III-E, a multi-class classifier has to be trained on the internal dataset to quantify the effect of removing the *shiftness* of external datasets for the two targeted classification tasks. In this section, we report the classification training and testing performance on the internal dataset, and the performance on the external datasets after dropping the highest anomaly score group iteratively. The external group-wise *shiftness* is thus revealed by the performance variation. An evident decrease suggests a significant distribution shift in the dropped group. For comparison, we report the classification outcomes on external dataset based on the clustering results obtained with both anomaly scores computed with CVAD [30], fAnoGAN [17] and SSPCAB [34] architectures.

Figure 8 shows the classification results for the **MURA** data, including the test results of Emory MURA and evaluation on Stanford MURA groups. Both the class-wise and average *AUC scores* are reported. As the classification is evaluated in the order of TOP_k, TOP_k-2, …, TOP_1 order, which is TOP_5, TOP_4, TOP_3, TOP_2, TOP_1 for our experiments, meaning that we gradually drop the group that with the highest anomaly scores and evaluate the classification performance on the remaining data. There are five groups being clustered for each class. Therefore, the TOP 5 clusters constitute the whole external dataset and the corresponding classification results for CVAD version, fAnoGAN and SSPCAB version are the same. As can be observed, the classifier's predictions become more and more accurate as the groups are discarded gradually based on their anomaly score order. Look into the *AUC* scores of *XR_HAND* from TOP 5 to TOP 1, the values of CVAD, fAnoGAN and SSPCAB are growing, especially CVAD and SSPCAB, which means the removed group contains data with certain *shiftness* and will affect the in-domain model's ability. The extent of *shiftness* can be inferred via the change of classification measurements for a notable improvement indicates a severe shifting exists in the dropped group. The amount of data samples in the dropped group is the number difference between the adjacent groups. The sample numbers of different groups are also reported in Fig. 9. Take *XR_HAND* for example, group 5 of *MedShift_w_CVAD* has 753 samples by calculating the difference of total image number of TOP 5 clusters (3851) and TOP 4 clusters (3098), (i.e., 753 = 3851 − 3098) and group 5 of *MedShift_w_f-AnoGAN* has 13 samples (13 = 3851 − 3838). Notably, removing fewer samples with more improvements means more accurate shift data detection. Although the same trend is noted for all the three versions in general, the CVAD and SSPCAB versions can get more increase in performance after expelling the most anomalous group than the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

f-AnoGAN version, which demonstrates the effectiveness of our *MedShift* framework in determining shift data among external datasets. We report the classification performance on chest X-ray datasets in Table. III. Additionally, we also analyze the anomaly score differences between correctly and wrongly classified samples. Take MURA for example, we show the anomaly score distributions for *XR_HAND*, *XR_FOREARM* and *XR_HUMERUS* classes in Figure 10. Generally, wrongly classified samples have higher anomaly scores than correctly classified but experience the possibility of low anomaly scores for some hard cases.

4) Dataset Quality Measurement Results: We report the Stanford **MURA** dataset quality in the top left of Fig. 11 calculated via the OTDD metric (i.e., Eqn. 5 and Eqn. 6). We respectively evaluate the quality for TOP_5, TOP_4, TOP_3, TOP_2, TOP_1 cases as indicated by the X-axis values of the plots. To compare, we test our pipeline with CVAD, fAnoGAN and SSPCAB anomaly detection architectures and present the internal train and test dataset quality as the baseline. As can be seen, the distance between Stanford MURA and Emory MURA datasets is decreasing when the anomalous groups with shift data are removed gradually. Nevertheless, our CVAD version (in blue) and SSPCAB version (in green) shorten the distance more and faster than the fAnoGAN (in orange) version. The general external dataset quality achieves the best when it is composed by the group with the lowest anomaly scores and achieve nearly the same dataset quality as the internal baseline, which follows the same conclusion as the average classification performance in Fig. 8. An increase of classification accuracy indicates the dataset quality improvement.

For the reason that the OTDD method computes the distance values with label-data pairs, it was not designed for multi-label datasets. To adapt for the **chest X-ray** scenario, we report the class quality instead of the whole dataset. Due to the space limitation, we randomly select 5 representative classes (*Fracture, No Finding, Edema, Consolidation, Pleural Other*) and present the quality variations in Fig. 11. To compare, we show the two chest X-ray datasets (CheXpert and MIMIC) class-wise quality obtained by CVAD, fAnoGAN and SSPCAB versions. Generally, the distances between the internal and external are shortened in a limited way with *MedShift_w_CVAD* model, but the distance values are enlarged by the fAnoGAN version. SSPCAB exhibits better performance than fAnoGAN. Since the distance represents the dissimilarity between the evaluated dataset pair, an increase of distance indicates a failure of identifying shift data in the external domain. Here, the CVAD version shows better performance than the *MedShift_w_fAnoGAN* model and sometimes better than *MedShift_w_SSPCAB*.

Moreover, an increase of distance is also an indicator of stop sign for detecting shift data of a well-performed shift identification model. From the anomaly score distribution plots of Fig. 11, it is clear that external MURA *HAND* has more variance than the external chest X-ray *Fracture* data. Thus, shift data identification is relatively difficult for the chest X-ray dataset. Specifically, all the class-wise dataset qualities of chest X-ray datasets are improved in a limited way compared to the baseline. Subtle class variations and multi-label characteristics may account for the limited improvement as they lead to the difficulty of class-specific representation learning and degrade the distinguishing ability for external outliers with minor variation. Depending on the quality expectations, users can decide to remain the original *Fracture* class or remove one or two top groups from *Fracture*.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

V. Limitation

MedShift has been only validated on the medical image classification problem. Similar pipeline can also be evoked for segmentation and detection. For multi-class classification problem, the pipeline needs anomaly detectors trained for each class which ultimately increase the training time and computational complexity. The dataset quality metrics have only been computed on MURA and Chest Xray datasets. More evaluations need to be perform for generalizing these quality measures.

VI. Conclusion and Future Works

We have designed an automated pipeline - *MedShift*, for medical dataset curation based on anomaly score. Under-the-hood, *MedShift* identifies image data distribution shift based on anomaly detection and unsupervised clustering to discriminate the poor-quality, noisy and under-represented samples from the external data. The anomaly detection architecture involves two separate implementation phases - (1) internal training - time consuming and needs to be trained for each targeted class label, and (2) test phase - quick, only forward pass which needs minimal data pre-processing and cleaning from the external sites. Once trained, the anomaly detectors should be able to identify unknown anomalous patterns from an external dataset without ever seeing such anomalous data examples in training. This quality makes the proposed pipeline particularly suitable for medical image dataset curation since exchanging healthcare data among institutions and manually identifying noisy or anomalous data are both extremely challenging in the current healthcare situation.

Our pipeline is flexible towards the particular anomaly detector architectures. We evaluated two use-cases - diagnosis from chest X-ray and classifying anatomical joints from MURA and applied different anomaly detectors CVAD, fAnoGAN and SSPCAB.

Our experiments showed that being trained only on internal Emory datasets, deep learning models classification accuracy is gradually rising on the external dataset after removing the shift data items via *MedShift* and ultimately achieved performance close to the internal data. The improvement of classification accuracy represents the fact that the *MedShift* can identify relevant shift data that will degrade the performance of an in-domain model and be able to reproduce the internal performance on unseen external data without data sharing. Moreover, the brief cluster exploration on the external dataset showed that higher anomaly cluster groups contain more variations in terms of image quality, positioning, noise, and the pipeline correctly identified the shift data.

In its current state, the proposed pipeline *MedShift* can be evoke as a web-service and compute domain-specific quality checks and derive powerful and actionable insights from the datasets. The suggested workflow will be beneficial in future non-shareable healthcare collaboration where the *MedShift* pipeline will be set up as a browser-based service within the local firewall for automated dataset curation with multi-class labels. As an immediate future study, we plan to conduct a reader study with expert radiologists to interactively evaluate the proposed platform and quantify the performance based on user-feedback

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

matrices. In future, we are planing to incorporate novel proxy-based multi-classs similarity architecture for anomaly detection.

Acknowledgments

The work is supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) MIDRC grant of the National Institutes of Health under contracts 75N92020C00008 and 75N92020C00021 and the US National Science Foundation #1928481 from the Division of Electrical, Communication & Cyber Systems. Partially supported by NHLBI R01 HL155410-01.

REFERENCES

[1]. Ooijen P, “Quality and curation of medical images and data.” In Artificial intelligence in medical imaging, 2019, pp. 247–255. Springer, Cham.

[2]. Rajpurkar P et al. , “Mura: Large dataset for abnormality detection in musculoskeletal radiographs.” arXiv preprint arXiv:1712.06957

[3]. Gordon B et al. , “Development of a data utility framework to support effective health data curation.” BMJ health & care informatics, 2021, 28(1).

[4]. Daxberger E, and Hernández-Lobato JM. “Bayesian variational autoencoders for unsupervised out-of-distribution detection.” 2019, arXiv preprint arXiv:1912.05651

[5]. Ruff L et al., July. “Deep one-class classification.” In International conference on machine learning (pp. 4393–4402). 2018, PMLR.

[6]. Ni JC et al. , “Deep learning for automated classification of inferior vena cava filter types on radiographs.” Journal of Vascular and Interventional Radiology, 2020. 31(1), pp.66–73. [PubMed: 31542278]

[7]. Park C, Awadalla A, Kohno T, and Patel S. “Reliable and Trustworthy Machine Learning for Health Using Dataset Shift Detection.” Advances in Neural Information Processing Systems, 2021, 34.

[8]. Stacke K, Eilertsen G, Unger J, and Lundstrom C. “Measuring domain shift for deep learning in histopathology.” IEEE journal of biomedical and health informatics, 2020, 25(2), pp.325–336.

[9]. Rotemberg V et al. , “A patient-centric dataset of images and metadata for identifying melanomas using clinical context.” Scientific data, 2021, 8(1), pp.1–8. [PubMed: 33414438]

[10]. Wang X, Peng Y, Lu L, Lu Z, Bagheri M and Summers RM, 2017. 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 (pp. 2097–2106).

[11]. Guo X, Gichoya JW, Purkayastha S and Banerjee I, 2022. CVAD-An unsupervised image anomaly detector. Software Impacts, 11, p.100195.

[12]. Tang S, Ghorbani A, Yamashita R, Rehman S, Dunnmon JA, Zou J and Rubin DL, 2021. Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. Scientific reports, 11(1), pp.1–9. [PubMed: 33414495]

[13]. Rabanser S, Günnemann S and Lipton Z, 2019. Failing loudly: An empirical study of methods for detecting dataset shift. Advances in Neural Information Processing Systems, 32.

[14]. Sugiyama M, Nakajima S, Kashima H, Buenau P and Kawanabe M, 2007. Direct importance estimation with model selection and its application to covariate shift adaptation. Advances in neural information processing systems, 20.

[15]. Storkey A, 2009. When training and test sets are different: characterizing learning transfer. Dataset shift in machine learning, 30, pp.3–28.

[16]. Yuan Z, Yan Y, Sonka M and Yang T, 2020. Large-scale Robust Deep AUC Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification. arXiv preprint arXiv:2012.03173

[17]. Schlegl T, Seeböck P, Waldstein SM, Langs G and Schmidt-Erfurth U, 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis, 54, pp.30–44. [PubMed: 30831356]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[18]. Guo X, Gichoya JW, Purkayastha S and Banerjee I, 2022. Margin-aware intraclass novelty identification for medical images. Journal of Medical Imaging, 9(1), p.014004. [PubMed: 35127968]

[19]. Mann HB and Whitney DR, 1947. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, pp.50–60.

[20]. Akcay S, Atapour-Abarghouei A and Breckon TP, 2018, December. Ganomaly: Semi-supervised anomaly detection via adversarial training. In Asian conference on computer vision (pp. 622–637). Springer, Cham.

[21]. Li CL, Sohn K, Yoon J and Pfister T, 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9664–9674).

[22]. He K, Zhang X, Ren S and Sun J, 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

[23]. Quiñonero-Candela J, Sugiyama M, Schwaighofer A and Lawrence ND eds., 2008. Dataset shift in machine learning Mit Press.

[24]. Fang T, Lu N, Niu G and Sugiyama M, 2020. Rethinking importance weighting for deep learning under distribution shift. Advances in Neural Information Processing Systems, 33, pp.11996–12007.

[25]. Ng K, Kakkanatt C, Benigno M, Thompson C, Jackson M, Cahan A, Zhu X, Zhang P and Huang P, 2015. Curating and integrating data from multiple sources to support healthcare analytics. In MEDINFO 2015: eHealth-enabled Health (pp. 1056–1056). IOS Press.

[26]. DuMont Schütte A, Hetzel J, Gatidis S, Hepp T, Dietz B, Bauer S and Schwab P, 2021. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. NPJ digital medicine, 4(1), pp.1–14. [PubMed: 33398041]

[27]. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J and Cortes A, 2018. The UK Biobank resource with deep phenotyping and genomic data. Nature, 562(7726), pp.203–209. [PubMed: 30305743]

[28]. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M and Tarbox L, 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of digital imaging, 26(6), pp.1045–1057. [PubMed: 23884657]

[29]. Van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, Heymann D and Burke DS, 2014. A systematic review of barriers to data sharing in public health. BMC public health, 14(1), pp.1–9. [PubMed: 24383435]

[30]. Guo X, Gichoya JW, Purkayastha S and Banerjee I, 2021. CVAD: A generic medical anomaly detector based on Cascade VAE. arXiv preprint arXiv:2110.15811

[31]. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG and Horng S, 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1), pp.1–8. [PubMed: 30647409]

[32]. Huang G, Liu Z, Van Der Maaten L and Weinberger KQ, 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700–4708).

[33]. Hsu YC, Shen Y, Jin H and Kira Z, 2020. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10951–10960).

[34]. Ristea NC, Madan N, Ionescu RT, Nasrollahi K, Khan FS, Moeslund TB and Shah M, 2022. Self-supervised predictive convolutional attentive block for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13576–13586).

[35]. Guo X, Gichoya JW, Purkayastha S and Banerjee I, 2022. CVAD-An unsupervised image anomaly detector. Software Impacts, 11, p.100195.

[36]. Guo X, Gichoya JW, Purkayastha S and Banerjee I, 2022. Margin-aware intraclass novelty identification for medical images. Journal of Medical Imaging, 9(1), p.014004. [PubMed: 35127968]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[37]. Alvarez-Melis D and Fusi N, 2020. Geometric dataset distances via optimal transport. Advances in Neural Information Processing Systems, 33, pp.21428–21439.

[38]. Villani C, 2009. Optimal transport: old and new (Vol. 338, p. 23). Berlin: springer.

[39]. Wang R, Chaudhari P and Davatzikos C, 2022. Embracing the disharmony in medical imaging: A Simple and effective framework for domain adaptation. Medical Image Analysis, 76, p.102309. [PubMed: 34871931]

[40]. Yamoah GG, Cao L, Wu CW, Beekman FJ, Vandeghinste B, Mannheim JG, Rosenhain S, Leonardic K, Kiessling F and Gremse F, 2019. Data curation for preclinical and clinical multimodal imaging studies. Molecular Imaging and Biology, 21(6), pp.1034–1043. [PubMed: 30868426]

[41]. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K and Seekins J, 2019, July. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590–597).

[42]. Akcay S, Atapour-Abarghouei A and Breckon TP, 2018, December. Ganomaly: Semi-supervised anomaly detection via adversarial training. In Asian conference on computer vision (pp. 622–637). Springer, Cham.

[43]. Kirchheim K, Filax M and Ortmeier F, 2022. PyTorch-OOD: A Library for Out-of-Distribution Detection Based on PyTorch. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4351–4360).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

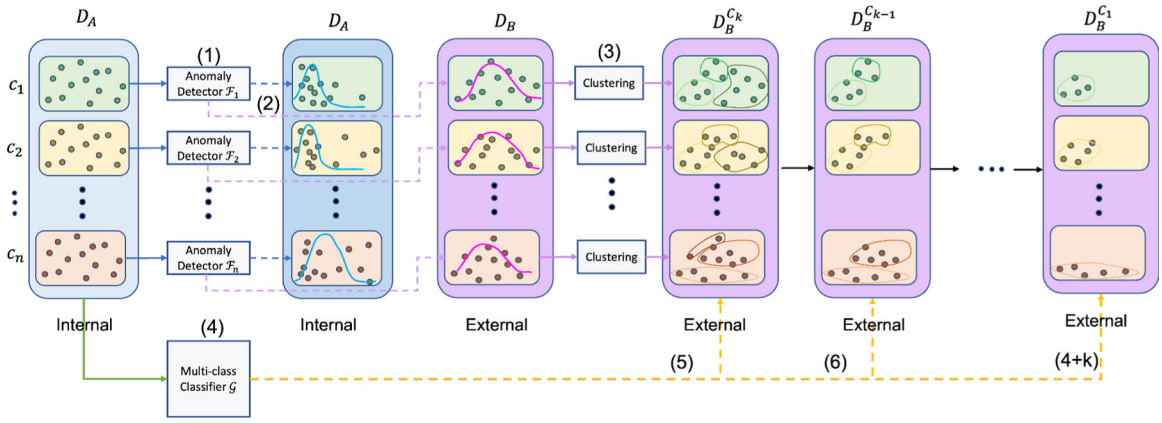

Fig. 1:

MedShift Pipeline for shift data identification (1)(2)(3) and quantification (4)(5)(6)…(4+k). (1) Train anomaly detectors for each internal class to learn in-distribution; (2) apply trained anomaly detectors to external classes for acquiring the anomaly scores for external data samples; (3) cluster each external class into multiple groups based on the anomaly scores in an unsupervised way; (4) train a multi-class classifier of internal dataset; (5) apply the trained classifier to the external classes to judge the generalizability on the external dataset without anomaly removal; (6) drop the group with the highest anomaly scores and apply the trained classifier to the updated classes; the previous process will proceed until (4+k) only one group left. Best view in color.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

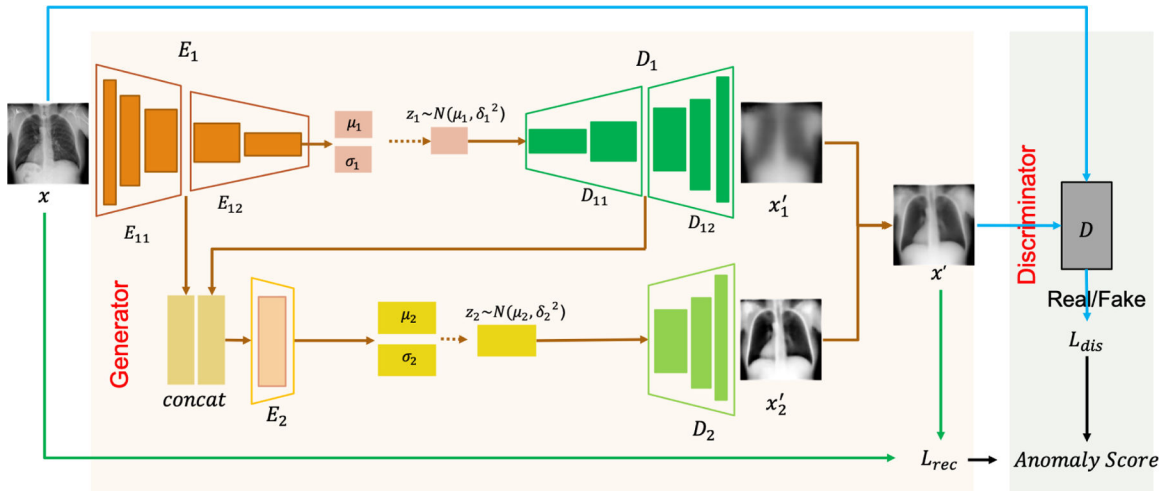

Fig. 2:

CVAD architecture - a cascade VAE as the generator (G) and a separate binary classifier (D) as the discriminator. The main VAE pipeline is composed by the encoder E_1 shown as the orange part and the decoder D_1 in the dark green part; the branch VAE has the pink part as the encoder E_2 and the light green for its decoder D_2. Given an input image x, the main VAE learns to reconstruct x'_1 via latent representations μ_1 and σ_1; the branch VAE takes the outputs of the results of the main VAE encoder intermediate part E_{11} and the intermediate decoder D_{11} as inputs and feeds the concatenated features to E_2 to formulate the branch latent variables μ_2 and σ_2, which gives a low-level reconstruction x'_2 via the corresponding decoder D_2. By adding the two reconstructions - x'_1 and x'_2 together with a sigmoid function, a final reconstruction x is generated and later treated as fake OOD data as compared to the original input x. The binary discriminator D will learn to distinguish them.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

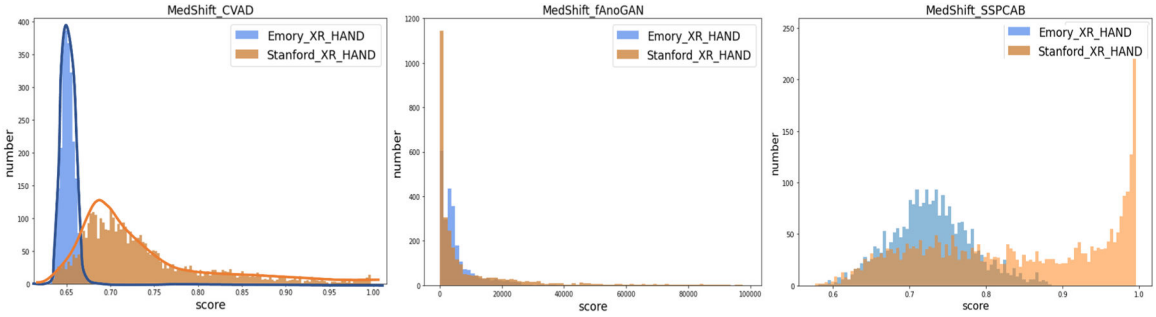

Fig. 3:

Shift identification with anomaly detection on Stanford_MURA *HAND* data - (left) anomaly score distributions of *MedShift_w_CVAD*; (middle) anomaly score distributions of *MedShift_w_fAnoGAN* and (right) anomaly score distributions of *MedShift_w_SSPCAB*. Distributions are truncated on samples with large anomaly scores for better visualization.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

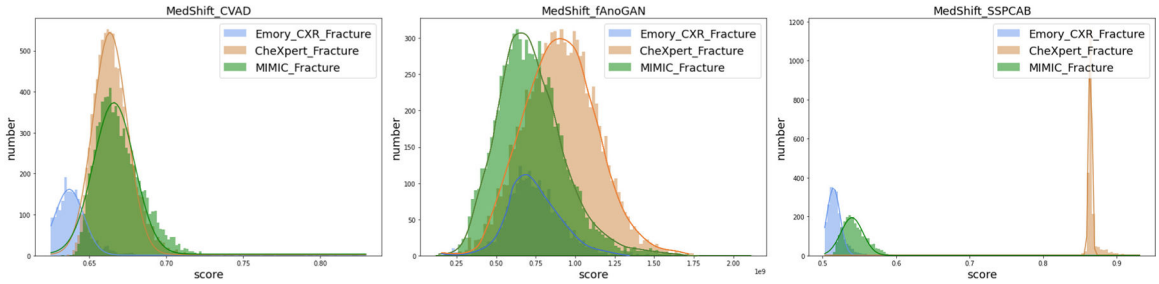

Fig. 4:

Shift identification with anomaly detection on CheXpert and MIMIC *Fracture* data - (left) anomaly score distributions of *MedShift_w_CVAD*; (middle) anomaly score distributions of *MedShift_w_fAnoGAN* and (right) anomaly score distributions of *MedShift_w_SSPCAB*. Distributions are truncated on samples with large anomaly scores for better visualization.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

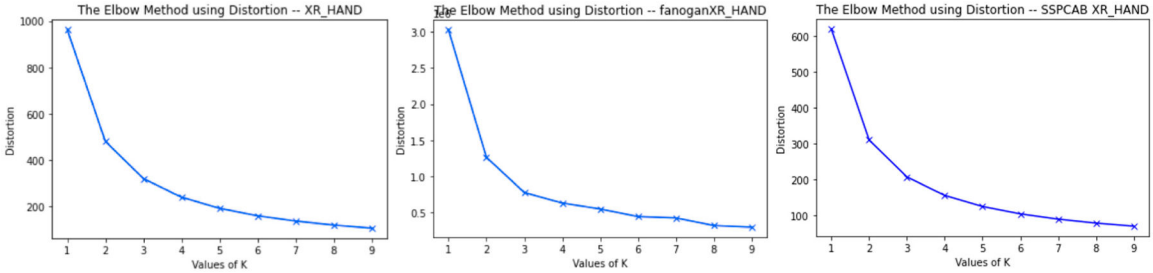

Fig. 5:

Elbow distortion curves for Stanford_MURA *HAND* data - (left) *MedShift_w_CVAD* results; (middle) *MedShift_w_fAnoGAN* results and (right) *MedShift_w_SSPCAB* results.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

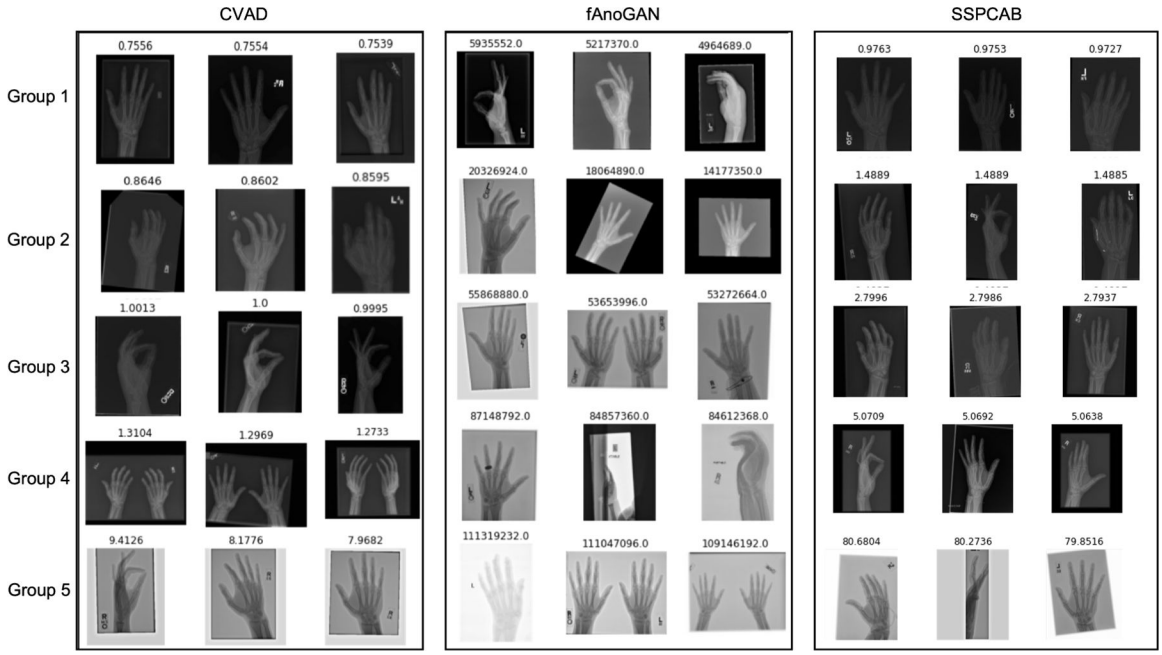

Fig. 6:

Clustering examples on Stanford_MURA *HAND* data - (left) *MedShift_w_CVAD* results; (middl) *MedShift_w_fAnoGAN* results and (right) *MedShift_w_fAnoGAN* results. Each row represents one group with five example images. The groups are illustrated in ascending order based on the anomaly scores from top to bottom. The corresponding original (not normalized) anomaly score is on top of each image.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

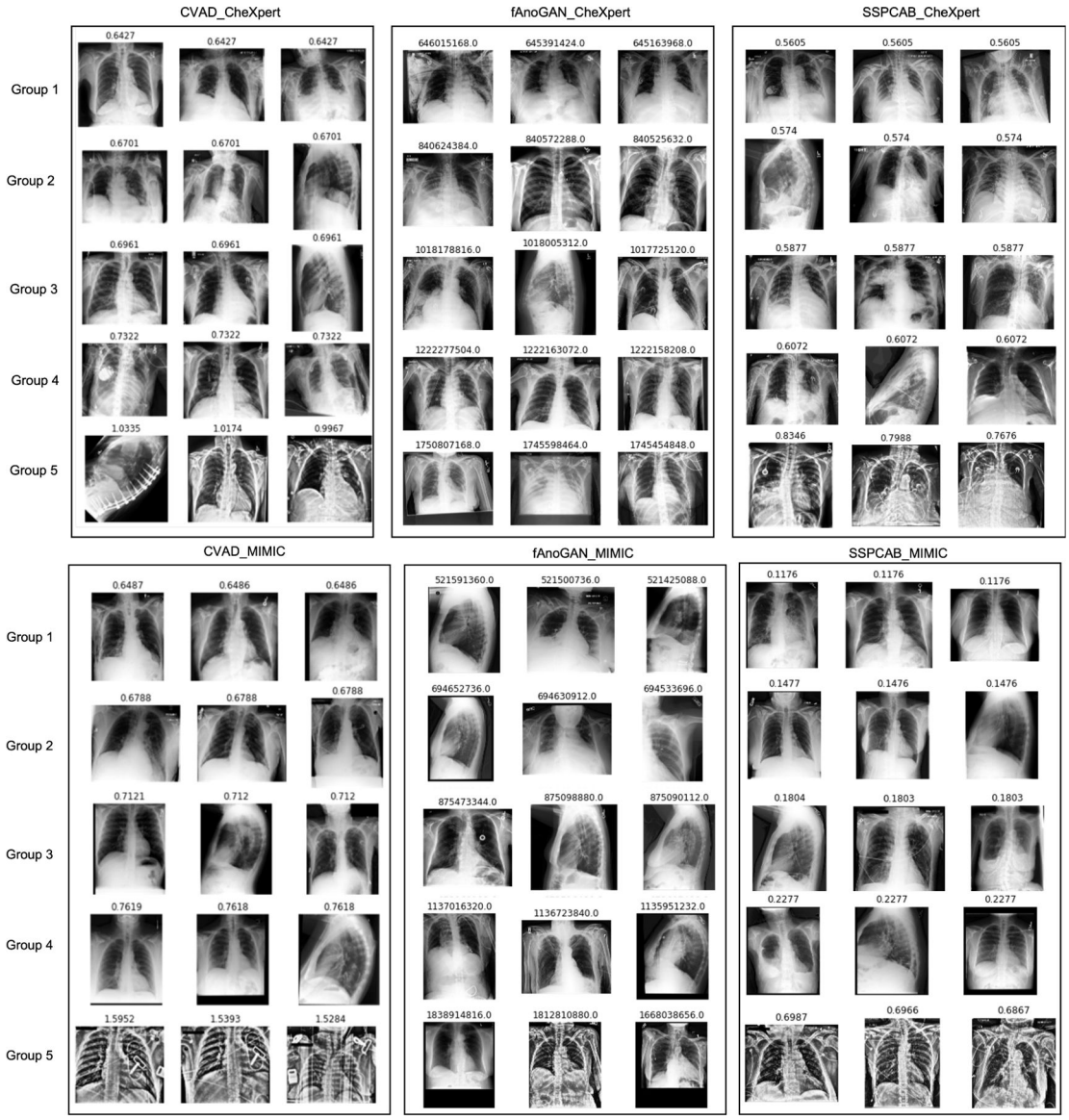

Fig. 7:

Examples of chest X-ray clustering results. The first row shows three clustering results of CVAD, fAnoGAN, SSPCAB on CheXpert *Fracture* data respectively; and the second row presents their clustering results on MIMIC *Fracture* data. Styles follow Fig. 6.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

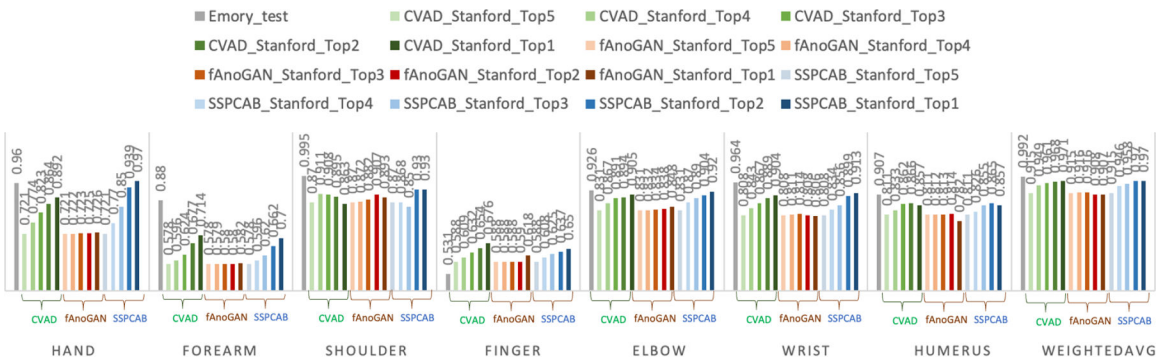

Fig. 8:
MURA classification performance.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

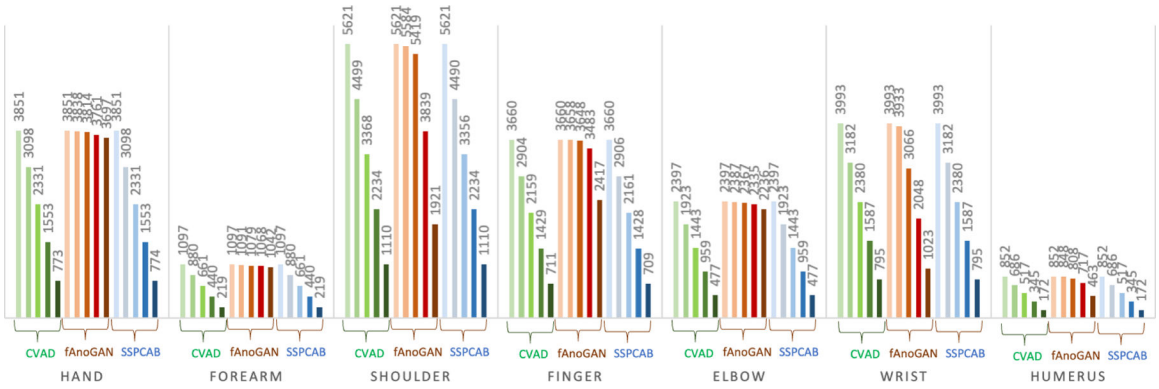

Fig. 9:

Number of images for MURA after removing groups in the order of Top_5, Top_4, …, Top_2. (Style follows Fig. 8.)

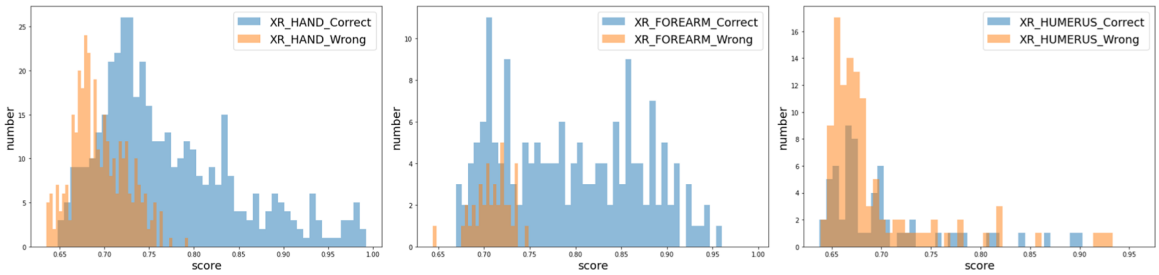

Fig. 10:

Anomaly score distributions of correctly(yellow) and wrongly(blue) classified external MURA data. From left to right, there are *XR_HAND*, *XR_FOREARM* and *XR_HUMERUS* classes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

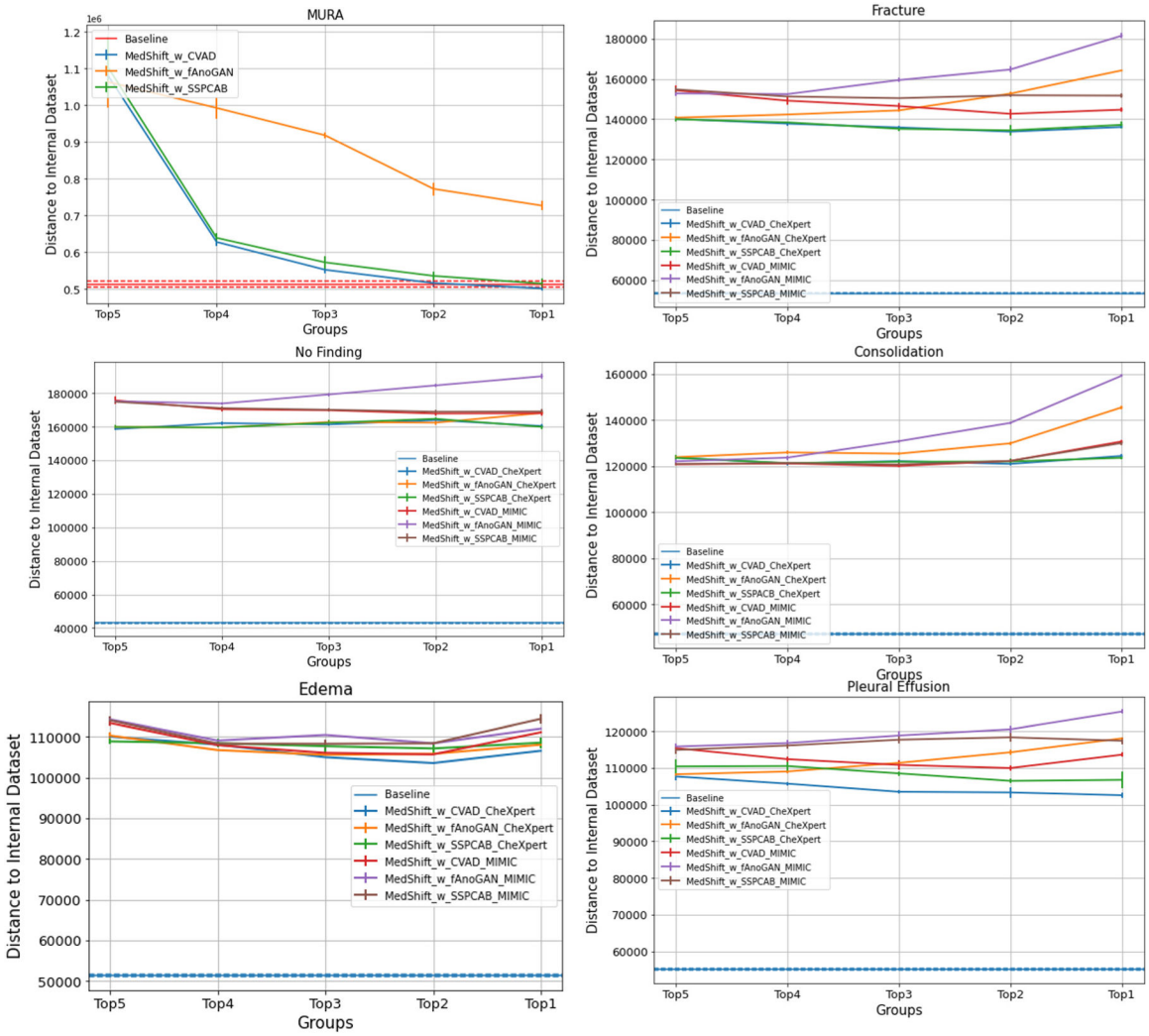

Fig. 11:

Dataset quality measurement results. From left to right, top to bottom, there are Stanford MURA whole dataset's quality, CheXpert and MIMIC *Fracture*, *No Finding*, *Consolidation*, *Edema* and *Pleural Effusion* class quality. Internal dataset quality baselines are also showed as *Baseline* with variations in dotted lines. X-axis values represent situations of the groups in use, and Y-axis values indicate the distance between the internal and external datasets (the lower the better). Distance mean and stdev values of ten rounds of evaluations are present in the plots.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE I:

Dataset details, with total image number and the percentage (in brackets) of each class presented. Upper part of the table present the MURA datasets and the lower is for Chest X-ray datasets.

	HAND	FOREAMR	FINGER	SHOULDER	ELBOW	WRIST	HUMERUS
Emory_MURA (internal)	**2,473 (21.33%)**	**368 (3.17%)**	**368 (3.17%)**	**3,451 (29.77%)**	**1,521 (13.12%)**	**2,858(24.65%)**	**553(4.77%)**
Stanford_MURA (external)	**3,851 (17.94%)**	**1,097 (5.11%)**	**3,660 (17.05%)**	**5,621 (26.18%)**	**2,397 (11.16%)**	**3,993(18.60%)**	**852(3.97%)**

	No Finding	Enlarged Cardiomediastinum	Cardiomegaly	Lung Lesion	Lung Opacity	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural Other	Fracture	Support Devices
Emory_CXR (train, internal)	57,973 (11.35%)	7,825 (1.53%)	27,019 (5.29%)	6,157 (1.21%)	64,439 (12.62%)	22,540 (4.41%)	6,906 (1.35%)	9,188 (1.80%)	66,150 (12.95%)	11,550 (2.26%)	51,828 (10.15%)	2,325 (0.46%)	2,114 (0.41%)	174,768 (34.22%)
Emory_CXR (test, internal)	7,962 (30.44%)	523 (2.00%)	1,256 (4.80%)	397 (1.52%)	2,141 (8.18%)	475 (1.82%)	151 (0.58%)	439 (1.68%)	1,684 (6.44%)	150 (0.57%)	711 (2.72%)	98 (0.37%)	177 (0.68%)	9,995 (38.21%)
CheXpert (external)	22,381 (4.34%)	10,798 (2.09%)	27,000 (5.24%)	9,186 (1.78%)	105,581 (20.48%)	52,246 (10.13%)	14,783 (2.87%)	6,039 (1.17%)	33,376 (6.47%)	19,448 (3.77%)	86,187 (16.72%)	3,523 (0.68%)	9,040 (1.75%)	116,001 (22.50%)
MIMIC (external)	143,352 (22.62%)	84,073 (13.26%)	76,957 (12.14%)	76,423 (12.06%)	65,047 (10.26%)	64,346 (10.15%)	36,564 (5.77%)	26,222 (4.14%)	14,675 (2.32%)	14,257 (2.25%)	10,801 (1.70%)	10,042 (1.58%)	7,605 (1.20%)	3,460 (0.55%)

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2024 August 07.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE II:

Notations and the descriptions used in the paper.

Notations	Descriptions
D_A,D_B	Internal and external dataset repectively
$P_{D_A}(Y \mid X)$, $P_{D_B}(Y \mid X)$	The distribution of internal dataset D_A, D_B
$X_{c_i}^A$	Input internal data samples of class c_i
$x_j^{A_{c_i}}$	An sample from class c_i of the internal dataset A
$N_{c_i}^A$	The number of the input data samples of class c_i
$\overline{X}_{c_i}^B$	OOD data in external dataset D_B
$\mathscr{F}_c$	Classwise anomaly detector
$q_\phi(z \mid x)$	VAE encoder, with x as input to learn the latent parameters z, ϕ as the encoder model parameters, q for the encoder.
$p_\theta(x \mid z)$	VAE decoder, aims to reconstruct original input x based on the latent parameters$z$, θ as the decoder model parameters and p for the decoder.
$L(\phi, \theta)$	The evidence lower bound with parameters ϕ and θ
L_{rec}	CVAD's generator loss (the reconstruction loss)
S_{dis}	The probability output by CVAD's discriminator

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE III:

se AUC score results of Chest X-ray datasets.

		Methods	No Finding	Enlarged Cardiomediastinum	Cardiomegaly	Lung Lesion	Lung Opacity	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural Other	Fracture	Support Devices
CXR	Test	–	0.724	0.561	0.788	0.652	0.742	0.783	0.697	0.697	0.743	0.681	0.824	0.732	0.607	0.540
	CheXpert		0.754	0.512	0.651	0.638	0.654	0.746	0.662	0.603	0.694	0.697	0.775	0.625	0.533	0.756
	MIMIC		0.757	0.515	0.656	0.639	0.650	0.764	0.661	0.602	0.693	0.693	0.786	0.631	0.531	0.759
	CheXpert	*CVAD*	0.760	0.518	0.709	0.655	0.611	0.691	0.622	0.623	0.589	0.607	0.740	0.623	0.546	0.661
		fAnoGAN	0.769	0.515	0.710	0.659	0.602	0.693	0.618	0.629	0.577	0.605	0.728	0.614	0.546	0.666
		SPCAB	0.763	0.510	0.707	0.655	0.610	0.690	0.620	0.622	0.570	0.600	0.723	0.612	0.540	0.660
	MIMIC	*CVAD*	0.755	0.513	0.653	0.639	0.651	0.754	0.661	0.602	0.694	0.693	0.779	0.627	0.532	0.757
		fAnoGAN	0.782	0.508	0.650	0.638	0.635	0.746	0.646	0.608	0.675	0.683	0.770	0.634	0.535	0.750
		SPCAB	0.742	0.510	0.651	0.637	0.652	0.756	0.666	0.608	0.695	0.693	0.777	0.637	0.530	0.751
	CheXpert	*CVAD*	0.752	0.520	0.709	0.656	0.604	0.691	0.617	0.625	0.585	0.610	0.735	0.629	0.538	0.653
		fAnoGAN	0.759	0.516	0.711	0.660	0.606	0.697	0.616	0.628	0.577	0.607	0.725	0.609	0.538	0.666
		SPCAB	0.750	0.519	0.701	0.650	0.606	0.690	0.617	0.623	0.583	0.617	0.720	0.619	0.538	0.661
	MIMIC	*CVAD*	0.757	0.515	0.656	0.639	0.650	0.764	0.661	0.602	0.693	0.694	0.786	0.631	0.531	0.759
		fAnoGAN	0.773	0.508	0.647	0.643	0.632	0.747	0.645	0.602	0.674	0.685	0.766	0.637	0.528	0.748
		SPCAB	0.760	0.518	0.657	0.633	0.650	0.767	0.667	0.605	0.683	0.695	0.786	0.635	0.538	0.758
	CheXpert	*CVAD*	0.744	0.521	0.710	0.655	0.600	0.690	0.612	0.626	0.581	0.619	0.729	0.630	0.532	0.648
		fAnoGAN	0.744	0.512	0.705	0.661	0.603	0.699	0.613	0.630	0.579	0.605	0.725	0.616	0.536	0.666
		SPCAB	0.743	0.515	0.706	0.651	0.603	0.694	0.613	0.628	0.589	0.615	0.728	0.626	0.531	0.646
	MIMIC	*CVAD*	0.761	0.517	0.660	0.640	0.650	0.776	0.664	0.603	0.691	0.699	0.793	0.633	0.526	0.761
		fAnoGAN	0.762	0.502	0.647	0.642	0.632	0.746	0.644	0.598	0.670	0.687	0.761	0.630	0.523	0.747
		SPCAB	0.760	0.512	0.667	0.640	0.642	0.778	0.663	0.601	0.690	0.697	0.794	0.634	0.530	0.757
	CheXpert	*CVAD*	0.736	0.527	0.709	0.655	0.596	0.681	0.606	0.632	0.572	0.632	0.725	0.640	0.521	0.644
		fAnoGAN	0.735	0.510	0.705	0.656	0.602	0.703	0.612	0.621	0.576	0.608	0.723	0.613	0.528	0.665
		SPCAB	0.733	0.516	0.703	0.656	0.592	0.683	0.612	0.631	0.576	0.621	0.720	0.643	0.520	0.645
	MIMIC	*CVAD*	0.765	0.519	0.669	0.636	0.651	0.783	0.659	0.606	0.685	0.693	0.797	0.630	0.519	0.765
		fAnoGAN	0.748	0.502	0.645	0.649	0.625	0.750	0.640	0.591	0.663	0.685	0.755	0.626	0.519	0.747
		SPCAB	0.768	0.512	0.665	0.635	0.652	0.780	0.650	0.601	0.683	0.698	0.795	0.636	0.519	0.767

IEEE J Biomed Health Inform. Author manuscript; available in PMC 2024 August 07.