Distribution Agreement

Signature:

______________________________________ ______________

Xiaoyuan Guo Date

Medical Image Analysis with Deep Learning under Limited Supervision

By

Xiaoyuan Guo
Doctor of Philosophy

Computer Science and Informatics

Imon Banerjee, Ph.D
Advisor

Judy Wawira Gichoya, MD, MS
Committee Member

Jun Kong, Ph.D.
Committee Member

Xiaofeng Yang, Ph.D
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D, MPH
Dean of the James T. Laney School of Graduate Studies

Date

Medical Image Analysis with Deep Learning under Limited Supervision

By

Xiaoyuan Guo
B.S., TianJin University of Technology, Tianjin, CN, 2014
M.S., University of Chinese Academy of Sciences, Beijing, CN, 2017

Advisor: Imon Banerjee, Ph.D

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2022

Abstract

Medical Image Analysis with Deep Learning under Limited Supervision
By Xiaoyuan Guo

Medical imaging plays a significant role in different clinical applications such as detection, monitoring, diagnosis, and treatment evaluation of various clinical conditions. Deep learning approach for medical image analysis emerges as a fast-growing research field and has been widely used to facilitate challenging image analysis tasks, for example, detecting the presence or absence of a particular abnormality, diagnosis of a particular tumor subtype. However, one important requisite is the large amount of annotated data for supervised training, which is often lacking in medicine due to expensive and time-consuming expert-driven data curation process. Data insufficiency in medical images is also limited by healthcare data privacy requirements, which leads to barriers in the usage of deep learning methods across institutions. This thesis focuses on facilitating the applications of deep learning approaches to solve automatic medical image analysis tasks efficiently under limited supervision. Three situations are in consideration: **(1)** no annotated data; **(2)** limited annotated data; **(3)** curation of additional annotated data with minimal human supervision. The research covers multiple medical image modalities starting from fluorescence microscopy images (FMI), histopathological microscopy images (HMI) to mammogram images (MG), computed tomography (CT), chest radiographs (X-ray). A variety of medical image related tasks have been researched, including clumped nuclei segmentation in FMI, clustered liver steatosis segmentation in HMI, segmentation and quantification of breast arterial calcifications (BAC) on MG, out-of-distribution (OOD) detection for medical images, shift data identification from unseen external datasets, image retrieval in external datasets with OOD-awareness and accurate multi-label medical image retrieval. Due to the reality of limited supervision in medicine, unsupervised, weakly-supervised, and supervised deep learning techniques have been investigated in this thesis to solve the medical tasks under different situations. The diversity and concreteness of the thesis can be a guide to facilitate the efficient usage of deep learning approaches in future medical image analysis with minimal cost.

Medical Image Analysis with Deep Learning under Limited Supervision

By

Xiaoyuan Guo
B.S., TianJin University of Technology, Tianjin, CN, 2014
M.S., University of Chinese Academy of Sciences, Beijing, CN, 2017

Advisor: Imon Banerjee, Ph.D

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2022

Acknowledgments

I would like to thank my esteemed supervisor Dr. Imon Banerjee for her invaluable supervision, continuous support during my PhD study. My gratitude extends to Dr. Judy Wawira Gichoya, Dr. Hari Trivedi and Dr. Saptarshi Purkayastha for all the help, suggestions and advice in my academic research, especially the funding opportunity to support my work. I would also like to offer my special thanks to Dr. Jun Kong and Dr. Ashish Sharma for the awesome supervision and assistance during my early stage PhD study. It is their kind help that made my study and life in the United States a wonderful time. I would like to extend my sincere thanks to my committee memebers Dr. Judy Wawira Gichoya, Dr. Jun Kong and Dr. Xiaofeng Yang for their insightful suggestions during my thesis defense. My appreciation also goes out to my family and friends for their encouragement and support all through my studies.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

In the current clinical practice, medical imaging is one of the primary diagnostic tools for various diseases detection and characterization of abnormalities, applications such as pre-screening and triaging, cancer staging, treatment response assessment, recurrence monitoring, and prognosis or survival prediction [25]. Over the last decades, we have witnessed the quick development of medical imaging, e.g., Positron Emission Tomography (PET), Magnetic Resonance Imaging (MRI), Computed tomography (CT), Mammography. In the clinic, the medical image interpretation has mostly been performed by human experts such as radiologists and physicians. Researchers and doctors have recently begun to benefit from computer-assisted interventions. However, it is belated for the advances in computational medical image analysis due to the lack of enough supervision information, or in other word lack of labeled data [137]. Development of computational medical image analysis models usually requires large image datasets for training, validation and testing of algorithms. The need is underscored by the deep learning revolution and the dominance of machine learning in recent medical image analysis research. Nevertheless, due to ethical and legal constraints, commercial conflicts and the dependence on busy medical professionals, medical image analysis researchers have been described as “data starved”. Obtaining relevant medical images to create such large datasets is challenging and expensive, requiring the cooperation of medical professionals and institutes – thus in high demand and short supply [74].

Lack of annotations in medical images often cause data insufficiency and consequently a small sample learning dilemma when dealing with medical images [175]. However, the first and the major prerequisite to use deep learning methods is massive amount of training dataset as the quality and evaluation of deep learning models heavily rely on quality and amount of the data. As can be seen in Table 1.1, most of the deep learning models in natural images are trained on thousands of images except for U-Net, which is specifically designed for medical images. In comparison, the volume available of typical medical imaging datasets is clearly still several orders of magnitude behind.

Table 1.1: Dataset information used by different deep learning models.

Models	**Dataset**	**Train**	**Val**	**Test**	**Task**
Mask R-CNN [59]	COCO [87]	115,000	5,000	20,000	Instance Segmentation
DeepSnake [107]	KINS [114]	7,474	-	7,517	Instance Segmentation
	Cityscapes [34]	2,975	500	1,525	
SegNet [13]	SUN RGB-D [140]	5,285	-	5,050	Semantic Segmentation
U-Net [124]	ISBI cell tracking challenge 2015	30	-	-	Semantic Segmentation
Swin UNETR[57]	Decathlon[138]	2,633	-	-	Semantic Segmentation
Faster R-CNN [120]	PASCAL VOC 2007	5,000	-	5,000	Object Detection
SSD [93]	PASCAL VOC 2007&2012	21,503	-	10,991	Object Detection

1.2 Research contributions

The research work in this thesis primarily lies in the area of medical AI, utilising the key potential of state-of-the-art AI technologies to facilitate medical image analysis under limited supervision. Generally speaking, this thesis mainly contributes to three limited supervision situations: (1) no annotated data; (2) limited annotated data; (3) curation of additional annotated data with minimal human supervision. Several representative medical image analysis tasks (e.g., image segmentation, OOD identification, automated dataset curation and image retrieval) have been considered accordingly. By experimenting on diverse medical images, such as fluorescence microscopy images (FMI), histopathological microscopy images (HMI), mammogram images (MG), computed tomography (CT), radiographs (X-ray), different proposed strategies have demonstrated the high possibilities of solving difficult problems.

1.2.1 Medical image segmentation with limited supervision

Image segmentation has been a critical step for further medical image analysis. However, the task becomes challenging when supervision information is limited. We have explored (1) clumped nuclei segmentation; (2) liver steatosis segmentation; and (3) breast arterial calcification (BAC) segmentation under different supervision scenarios.

(1) Clumped nuclei segmentation: With the annotated data unavailable, we proposed an automatic algorithm for clumped nuclei segmentation in fluorescence microscopy image, which only relies on the object morphology and image intensity and can output accurate segmentation.

(2) Liver steatosis segmentation: Since the above nuclei segmentation method does not require annotations, we modified the algorithm and generated initial segmentation masks for liver steatosis patches obtained from the whole-slide images. By filtering the imperfect masks, we created a liver steatosis dataset with generated masks as weak supervision, and applied a powerful instance segmentation model Mask-RCNN to separate clustered steatosis. Experimental results show significant improvements from weak supervision.

(3) BAC segmentation: Breast arterial calcifications often occur along with vessels, which are extremely narrow and hard to annotate accurately. Therefore, the annotations are limited and inaccurate. To solve the problem, we proposed a lightweight segmentation model SCU-Net to distinguish BACs in patches cropped from the original mammograms and optimized the model with a dice loss function to maximize the overlapping part between the predictions and annotations. To evaluate and quantify the BAC detection more accurately, we have suggested five evaluation metrics which consider the prediction possibilities, area of prediction, intensities of prediction together. With the metrics, our predictions show high correlations with ground-truth.

1.2.2 Medical OOD identification with limited supervision

OOD identification helps exclude the unexpected data from input images, which greatly safeguards the correctness and reliability of deployed deep learning models. However, OOD samples can be numerous and difficult to enumerate all possible classes. Without OOD data available, we have researched on self-supervised learning and proposed two methods - TEND and CVAD, detecting intra-class OOD data and generic OOD data. TEND is built on top of autoencoder(AE), which can learn image features by reconstructing inputs; and CVAD takes the advantage of cascade variational autoencoder (VAE), extracting normal image features with latent Gaussian's distribution normalization. Both of the two algorithms have demonstrated effectiveness in various medical datasets.

1.2.3 Medical dataset curation with limited supervision

Medical dataset curation aims to help curate more annotations from various sources and contributes to future medical AI development. Without the knowledge of the source image qualities, we have suggested a dataset shift detection pipeline by utilizing the self-supervised anomaly detectors to identify shift data among noisy external datasets, which can automatically detect the outliers to clean external datasets.

1.2.4 Medical image retrieval with limited supervision

Medical image retrieval in noisy external dataset can be challenging as the outliers that show intra-class variations fail to be distinguished. Our OOD-sensitive image retrieval method learns intra-class variations from generated pseudo intra-class labels, which can help retrieve target samples, especially outliers, in the external dataset and facilitate the labeling process. Moreover, a multi-label medical image retrieval system is proposed to rank images with multi-class similarities. This can be beneficial for handling challenging datasets like chest X-ray.

1.3 Organization of the thesis

We organize the thesis according to the medical task types. In Chapter 2, we introduce the background knowledge of relevant research topics and the related works along each topic. In Chapter 3, we present our works for solving medical image segmentation tasks under limited supervision, including nuclei segmentation in Sec. 3.1, liver steatosis segmentation in Sec. 3.2 and BAC segmentation in Sec. 3.3. In Chapter 4, we show two medical OOD detection methods with self-supervised learning to overcome the limited supervision issue, with TEND (intra-class OOD detection) in Sec. 4.1 and CVAD (generic OOD detection) in Sec. 4.2. In Chapter 5, we propose a unified pipeline MedShift to help identify shift data in external datasets based on unsupervised anomaly detection. In Chapter 6, we design efficient medical image retrieval methods under limited supervision. The outlier-sensitive medical image retrieval work is in Sec. 6.1 and the multi-class radiology image retrieval is in Sec. 6.2. In Chapter 8, we conclude the thesis with future research directions and plans suggested.

Chapter 2

Background & Related Work

This chapter provides a comprehensive review of the current research works considering limited supervision situations in medical image related tasks, including **medical image segmentation**, **medical OOD identification**, **medical dataset curation** and **medical image retrieval** problems.

Training models from limited labeled data and readily available unlabeled data is crucial for the successful application of deep learning methods in clinical usage and health care. These challenges have inspired many research efforts on learning with limited supervision, where the training data only have a limited amount of annotated examples, accurate but sparse annotations, inaccurate annotations, coarse-level annotations, and combinations of them. Previous works have been introduced below for different medical image tasks under limited supervision respectively.

2.1 Medical image segmentation with limited supervision

Image segmentation draws boundaries of objects within an image at the pixel level. It has two sub-types: semantic segmentation and instance segmentation. Semantic segmentation denotes per-pixel classification without differentiating instances; whereas instance segmentation requires the correct detection of all the objects in an image while also precisely segmenting each instance [59]. The labeling costs for medical images are very high, especially in medical image segmentation, which typically requires intensive pixel/voxel-wise labeling. Therefore, the strong capability of learning and generalizing from limited supervision, including no annotations, a limited amount of annotations, sparse annotations, and inaccurate annotations, is crucial for the successful application of deep learning models in medical image segmentation [106].

Unsupervised segmentation: Traditional segmentation does not require annotations. Take the clumped nuclei segmentation for example, many efforts [160, 86, 47, 171, 173, 12, 15] have been done to analyze the shape, contour, intensity of the image for nuclei separation. These conventional approaches can segment objects automatically without the supervision information, yet they often output under/over-segmentation due to the variance of image density and object complexity.

Supervised segmentation: Supervised segmentation refers using annotated masks of objects to guide the model learning in separating the objects from background, which is a common strategy for mainstream deep learning-based models. Popular semantic segmentation models include SegNet [13], U-Net [124], FCN [97]; instance segmentation representatives are Mask-RCNN [59], PANet [92], DeepSnake [107]. These models have achieved significant improvement or performance on large public image datasets compared with classical image segmentation methods. Despite the good performance, it is often challenging to apply supervised segmentation methods to medical image tasks due to the requirements of abundant annotations (class masks for semantic segmentation and object masks for instance segmentation). Models designed for natural images possibly fail to perform well when facing medical images because of the image quality, modality, size, and variations among different classes.

2.2 Medical OOD identification with limited supervision

There have been a lot of research works that summarize state-of-the-art anomaly detection methods [23, 79, 104, 61, 22, 39]. Because OOD data samples are innumerate to enumerate and hard to define, they are commonly unavailable during training. Without any supervision information available, the current methods aiming for anomalous image data detection can be generally divided into the following three categories:

AutoEncoder-based methods AutoEncoder [35] (AE) models can help extract significant embedding features by reconstructing the original images unsupervised. Trained with ID data, the architectures learn the "normality" and should lead to large reconstruction error when working on OOD dataset. Thus, the reconstruction error acts as the anomaly score to separate ID and OOD data [130, 178, 17]. However, AutoEncoder risks learning the identity function by simply outputting the original inputs, which largely limits its discriminative ability of anomalies. Other improved versions of AE are also used for anomaly detection [146, 110, 98, 9], *e.g.,* Variational Autoencoders (VAE) [10] provides probabilistic way of describing the latent space to reconstruct input data. Nevertheless, the reconstruction is often blurry and not good enough for clear discrimination of outliers. Similar with VAE, UAV-AdNet [19] uses the Kullback-Leibler divergence to regularize losses for anomaly detection but focuses on autonomous surveillance systems with GPS label used.

Generative adversarial network (GAN) based methods Similar to the AE models, GAN [52] framework can also learn latent feature representations by training a fake image generator and a real-vs-fake image discriminator [169, 109]. With the adversarial feature learning, GAN-based anomaly detectors can acquire discriminative latent features that can be used for separating the ID data from the OOD data. To further improve the discriminative ability of latent representations, BiGAN [40] adopts a bidirectional mapping learning. GANomaly [4] minimizes the distance of the ID data and the generated ones in latent feature space to detect the OOD data with large distance. Even so, the performance of GAN-based anomaly detectors largely depends on the training of GAN models, which always require large amounts of training data for OOD and often fail to handle inputs with large image size. Instead of selecting AnoGAN [132], which detects pixel-wise anomalies rather than in image-level, we compare TEND with GANomaly [4] and f-AnoGAN [133], AnoGAN's extension, for experiments given the better performance.

Classifier-based methods As the novelty detection in medical images can be reduced to a one-class classification (OCC) [70] problem with the one-vs-rest setup, one-class classifiers are often used for identifying unseen classes, *e.g.,* OC-SVM [134], FCDD [94], DOC [108], DeepSVDD [128]. With only ID data as training inputs, one-class classifiers often optimize a kernel-based objective function and minimize a hyper-sphere to threshold out the anomaly data based on distance. In common, the one-class classifiers exploit in-distribution data with specific object functions to threshold out anomalies. Nonetheless, their detection abilities on intra-class OOD data are not effective as the intra-class OOD data shares a lot of similarity with the ID data. Except for the one-class classifiers, ODIN [84] works on multi-classes datasets by adding perturbations of the input and temperature scaling to the score function to distinguish in-distribution and OOD data. Despite the efficiency and sophisticated methodology, the prerequisites of multiple OOD classes of the dataset is not typical in the medical image area and thus classifier-based methods have limited applicability in healthcare.

2.3 Medical dataset curation with limited supervision

Supervised deep learning requires plenty of annotations to train models, which are often in a limited amount, even unavailable for lots of medical image tasks. To utilize deep learning methods, many institutions label their own data locally. Therefore, it will be beneficial to curate the annotated data from various sources and contribute to a large standard dataset for the research community usage.

Data curation aims to provide quality data as input for meaningful Deep Learning-based analysis, which can be a costly and challenging process while curating extensive volumes of disorganized data. Therefore, an automatic dataset curation pipeline is highly demanding. However, the medical datasets from different institutions can be heterogeneous and with distribution shifts. To address the problems, a whole dataset curation involves data denoising, outlier detection, imputation, balancing and semantic annotation of the noisy, incomplete, insufficient data [174]. Among the operations, outlier detection plays the most important role for data cleansing and dataset quality improvement. Current dataset outlier detection are classified as statistical methods [3], distance-based methods [2], density-based methods [18] and cluster-based methods [48]. These methods are limited when handling high-dimensionality and nonlinearity. Most recent deep learning-based methods have been reported to detect outliers(see Sec. 2.2).

However, there is still a huge challenge for medical applications since it lacks an effective way to identify the difference for a bunch of datasets from the same medical domain because of inaccessibility to external medical datasets. Privacy concerns around sharing personally identifiable information are a major barrier to data sharing in medical research [135]. To address these privacy concerns, there has been an impressive number of large-scale research collaborations to pool and curate de-identified medical data for open-source research purposes [32]. Nevertheless, most medical data is still isolated and locally stored in hospitals and laboratories due to the worries associated with sharing patient [150].

2.4 Medical image retrieval with limited supervision

Generally, there are two mainstream image retrieval categories: *real-value* based [153, 20, 157, 144] and *hash*-based approaches [122, 82, 155, 166]. Real-value based metric learning methods can be further classified into anchor-based [72] and pair-based methods such as pairwise (CircleLoss[144]), tripletwise (DeepRank [153]) and listwise (FastAP [20]) methods.

One limitation of existing approaches is that they focus on natural images and only consider single-label during sampling. The other limitation of existing retrieval methods is the intra-class similarity problem. These existing efforts all emphasize the inter-class similarity but neglect the intra-class similarity. It's not clear if they can be applied to the medical domain. To our best knowledge, most medical retrieval systems are also restricted to single-label similarity measure. For example, Qayyum et. al. [113] propose to retrieve images for a dataset that is composed of lungs, brains, livers, ect. The dataset studied exhibits clear inter-class variations and can be distinguished based on a single class label. In contrast, retrieval in a multimorbidity dataset is much more challenging as multiple diseases or conditions can present in one single patient. The exact matching for all the labels is difficult as the occurrence of multiple conditions should all be identified by a retrieval system. In another camp, deep supervised hashing methods learn compact similarity-preserving binary code [90] and can be applied to the multi-label image retrieval [176]. Popular works include DPSH [82], CSQ [166], DTSH [155], DBDH [177]. However, such approaches sacrifice the semantic information when encoding the learnt features into a fixed-length binary code to obtain fast image retrieval, which results in degradation of performance accuracy.

Chapter 3

Medical Image Segmentation with Limited Supervision

3.1 Clumped nuclei segmentation

Fluorescence microscopy images are commonly used in clinical and biomedical research to help visualize cellular components, such as membranes and nuclei. Analysis of fluorescence microscopy images often requires an accurate identification of individual cell nuclei. However, clumped nuclei due to a high cell density often make it challenging to achieve an accurate nuclei segmentation. Although a large number of clumped nuclei segmentation methods have been proposed for fluorescence microscopy image analysis, they are subject to salient defects [118, 159]. For example, shape-based methods are sensitive to clumped nuclei shape variance which often leads to under-segmentation [159, 152, 85, 46]. Similarly, under-segmentation occurs in concavity-based methods that heavily rely on the clumped nuclei contour concavity [46, 170, 172]. By contrast, existing watershed-based methods usually yield over-segmentation due to the variance of image intensity [11, 16]. A 3D method is developed to segment clumped nuclei for improved segmentation accuracy [11]. However, its computational cost is expensive. In this paper, we propose a segmentation algorithm that identifies point pair connection candidates and evaluates adjacent point connections with a formulated ellipse fitting quality indicator. After connection relationships are determined, we recover the resulting dividing paths by following points with specific eigenvalues from the image Hessian in a constrained searching space. We validate our algorithm with 560 image patches from two classes of tumor regions of seven brain tumor patients. Both qualitative and quantitative experimental results suggest that our algorithm is promising for dividing overlapped nuclei in fluorescence microscopy images widely used in various biomedical research.

3.1.1 Contribution

We summarize our contributions as below:

- We propose a new algorithm that fully exploits the boundary shape of nuclei clumps to identify connecting point pairs by local high curvature voting and point pair screening. This is followed by a novel method that connects point pairs with dividing curves derived from local shape-based intensity analysis.
- Our method shows significant improvements over an existing marker-controlled watershed method [16] on a set of 560 fluorescence microscopy image patches. We also make the code publicly available at `https://github.com/XiaoyuanGuo/EMBC2018_clumped_nuclei_segmentation`.

Publication:

- Guo, Xiaoyuan, Hanyi Yu, Blair Rossetti, George Teodoro, Daniel Brat, and Jun Kong. "Clumped nuclei segmentation with adjacent point match and local shape-based intensity analysis in fluorescence microscopy images." In 2018

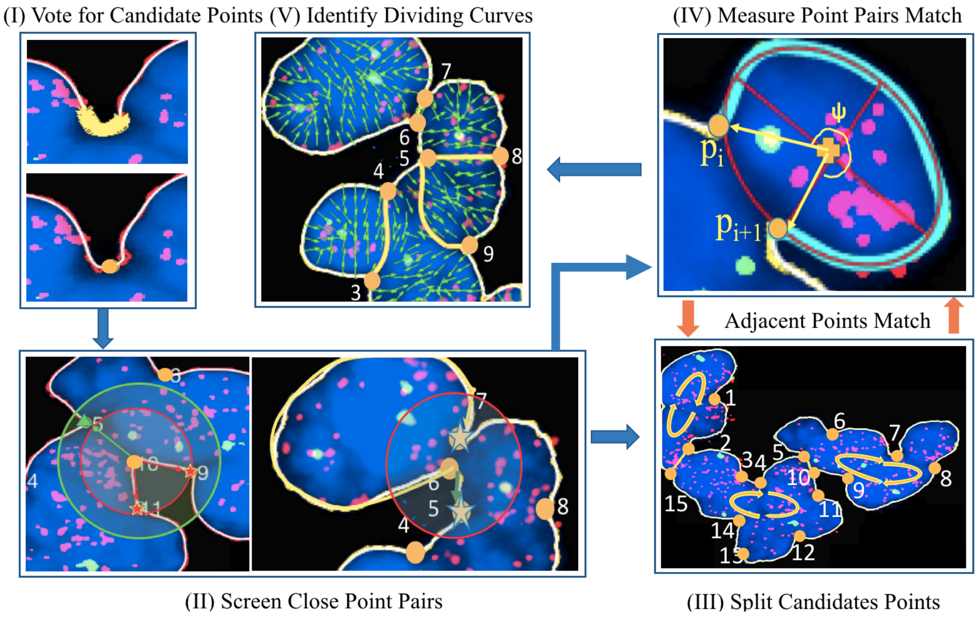

(I) Vote for Candidate Points
(V) Identify Dividing Curves
(IV) Measure Point Pairs Match
Adjacent Points Match
(II) Screen Close Point Pairs
(III) Split Candidates Points
p_i
p_{i+1}
ψ
3
4
5
6
7
8
9
10
11
12
13
14
15
1
2

Figure 3.1: The overall method flowchart for our proposed clumped nuclei segmentation is demonstrated. Nuclei are shown in blue with additional nuclear components shown as red and green interior regions.

40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3410-3413. IEEE, 2018.

3.1.2 Method

Our algorithm utilizes the contour shape of clustered nuclei to generate candidate point pairs and local shape based intensity analysis to separate overlapped nuclei with dividing curves. The work-flow consists of four main steps: (1) Voting for high curvature candidate points; (2) Screening close point pairs by a curvature-based distance metric for adjacent point pairs and a comprehensive metric taking into account curvature, distance, and matched normal vector angle for non-adjacent point pairs; (3) Assessing adjacent point pair connections via an ellipse fitting quality indicator; and (4) Identifying dividing curves by local shape based intensity analysis (see Figure 3.1).

High Curvature Candidate Point Voting

The blue-fluorescent DAPI nucleic acid stain preferentially binds to DNA within nuclei. Thus, we focus on the blue image channel for overall nuclear contour detection. We obtain the boundaries of such clustered nuclei using a global adaptive threshold. We propose a new method to find high curvature points on clump contours based on neighboring high curvature point votes. Curvature value of each point on the boundary is computed by $\kappa = \frac{x'y''-y'x''}{(x'^2+y'^2)^{3/2}}$, where x and y are coordinates of points on the denoised contour convolved with a smoothing Gaussian filter [159]. Different from high curvature point detection methods that select points with curvature values higher than a specified threshold [159], our method can avoid redundant candidate point detection with local high curvature point votes. As the following segmentation analysis depends on these initial candidate points, limiting the number of redundant candidate points can substantially reduce the computational cost and improve segmentation accuracy. In our data, it is quite common to have a long set of high curvature points along a concave contour segment defined by the curvature sign, as demonstrated in Figure 3.1(I). For those high curvature points in spatial proximity, it is unreasonable to mark all these points as candidates. Thus, we propose to vote for the optimal candidate point representation for a group of adjacent high curvature points by:

$$s^* = \frac{\int_{\partial C} \kappa(s) s ds}{\int_{\partial C} \kappa(s) ds} \tag{3.1}$$

where $s \in [0, 1]$ represents the normalized boundary arc length on a cell concave contour segment ∂C; $\kappa(s)$ is the curvature value at s; s^* is the normalized arc length of the optimal point representation for a segment of adjacent concavity points engaged in local voting. The bottom image in Figure 3.1(I) demonstrates the final high curvature point representation (in yellow circle) voted by all local high curvature points detected in the top image in Figure 3.1(I). With this procedure, we detect all high curvature point representations on each contour.

Close Point Pair Screening

Let us denote $P = \{p_i, | i = 1, 2, \ldots, N\}$ as a set of high curvature candidate points detected from Subsection 3.1.2. As demonstrated in Figure 3.1(II), we next identify all pairs of points (p_i, p_j) in spatial proximity by two circular searching regions of radius r_1 and r_2 $(r_1 < r_2)$ for under-segmentation reduction. Specifically, r_1 is used to detect the pairs of both adjacent and non-adjacent points, whereas the ring area defined by $r_2 - r_1$ is for detecting pairs of non-adjacent points in a neighboring area. In our analysis, they are set to 45 pixels and 70 pixels, respectively. In this way, we can alleviate under-segmentation suffered by numerous shape-based methods. The spatial proximity in our study is measured by the Euclidean distance. In addition, unlike the "bottleneck" points defined as close points presenting opposite gradient directions in previous research [152, 85], we relax this constraint and classify close points as either adjacent or non-adjacent point pairs. Plotted in the left panel in Figure 3.1(II), the search regions for close point pairs are represented by the red circle with radius r_1 and the green ring with inner radius r_1 and outer radius r_2, respectively. Here we use the searching region in the red circle of radius r_1 for detecting pairs of both adjacent and non-adjacent points in spatial proximity, whereas we only use the green ring for detecting pairs of non-adjacent points in a neighboring area. In the left plot of Figure 3.1(II), points 9 and 11 in the red circle region are close adjacent points of point 10, while point 5 in the green area is also close but not adjacent to point 10. Although these two types of close point pairs are processed in different ways in our proposed method, they both are important for an accurate segmentation.

(1) *Analysis for Pairs of Adjacent Points in Proximity*: Let us denote $C^+ = \{(p_i, q_i) | i = 1, 2, \ldots, N_1\}$ as the set of neighboring adjacent point pairs given the searching circular area of radius r_1. For each such pair demonstrated in the right panel

of Figure 3.1(II), we evaluate the associated Walking Energy to determine if we merge them (i.e. discard the undesired point and keep the point with the higher curvature value). Walking Energy is defined as $E_{p,q} = \int_{p(s)}^{q(s)} \kappa(s)ds$, where $\kappa(s)$ is the curvature of the corresponding curve segment $s \in [p(s), q(s)]$. Walking Energy represents the amount of "effort" required for a walk from a point $p(s)$ to its close adjacent point $q(s)$. In our formulation, we assume it consumes more energy to walk along a convex than non-convex contour. Demonstrated in the right panel of Figure 3.1(II), point 6 has two close point neighbors, i.e. point 5 and 7. The true walking route from point 6 to 7 is illustrated in a red curve, while that from point 6 and 5 is shown in green. Those pairs of neighboring adjacent points requiring low Walking Energy are merged. For each such pair of points under investigation, the point with higher curvature value is retained after points get merged. Adjacent points associated with a high Walking Energy would be retained, as shown in the right panel of Figure 3.1(II) where the Walking Energy for points 6 and 7 is sufficiently high to keep them separated. For any pair of such adjacent points presenting a high Walking Energy, it is highly likely that such points should be connected to separate overlapped nuclei. Therefore, we further evaluate the boundary segmentation quality in such cases by an ellipse fitting method discussed in Subsection 3.1.2.

(2) *Analysis for Pairs of Non-Adjacent Points in Proximity*: The resulting set of non-adjacent point pairs in close proximity can be represented as $C_1^- = \{(p_i, q_i)|i = 1, 2, \ldots, N_2\}$ when we use the circular searching area of radius r_1. To alleviate under-segmentation, we next expand the searching space to a ring searching region with radius between r_1 and r_2. To assess if such pairs of non-adjacent points (p_i, q_i) result from intersection of overlapped nuclei, we formulate and use the following measure to evaluate such point pairs:

$$V(p_i, q_i) = \frac{\alpha\theta(p_i, q_i)}{D(p_i, q_i) + \beta(\kappa(p_i) + \kappa(q_i))} \tag{3.2}$$

where θ is the angle between two normal vectors of the nuclear contour at p_i and q_i; D is the Euclidean distance between (p_i, q_i); α and β are weights set as 100 and 0.34, respectively. If $V(p_i, q_i)$ is larger than a specified threshold (e.g. 200 in our analysis), and (p_i, q_i) are non-adjacent points in a ring searching space between the radius of r_1 and r_2, such point pair is added to $C_2^- = \{(p_i, q_i)|i = 1, 2, \ldots, N_3\}$. After this evaluation analysis, the finalized set of non-adjacent point pairs in close proximity is $C^- = C_1^- \bigcup C_2^-$. For each pair in C^-, we next link the associated points to partition the whole contour into multiple subsets. Demonstrated in Figure 3.1(III), a whole nuclei cluster can be partitioned into multiple subsets by connecting points $(2, 15)$, and $(5, 10)$. As each candidate subset can produce new possible adjacent point pairs, the method for pairs of adjacent points in proximity discussed above is applied to each candidate subset.

Point Pair Segregation Assessment by Ellipse Fitting

We proceed with evaluating the segregation quality $Q_{i,i+1}$ of each adjacent point pair (p_i, p_{i+1}) by ellipse fitting. For each sub-contour segment under evaluation, an optimal ellipse, shown in red in Figure 3.1(IV), is fit to a candidate nucleus region with a closed sub-contour consisting of a contour arc and a dividing line between (p_i, p_{i+1}). An ellipse fitting quality measurement is defined as:

$$Q_{i,i+1} = \frac{\mu S^+_{i,i+1} + \nu \psi_{i,i+1}}{(\Delta x_{i,i+1} + \Delta y_{i,i+1}) + \gamma_1 \Delta L_{i,i+1} + \gamma_2 \eta_{i,i+1}} \tag{3.3}$$

where $S^+_{i,i+1}$ is the ratio of overlapped area between the candidate nucleus and fitting ellipse to their union; $\psi_{i,i+1}$ is the fitting angle formed by the line segment connecting the ellipse center and p_i and that connecting the ellipse center and p_{i+1}; $\Delta x_{i,i+1}$ and $\Delta y_{i,i+1}$ are the centroid coordinate difference between the fitting ellipse and the evaluated nucleus region; $\Delta L_{i,i+1}$ is the perimeter difference between the fitting ellipse and the evaluated nucleus contour; $\eta_{i,i+1}$ is the ratio of major to minor axis length derived from the fitting ellipse. μ, ν, γ_1 and γ_2 are weights set to 10.70, 10.70, 0.67 and 3.40 in our experiment.

With the formulated ellipse fitting quality measurement, we connect those neighboring adjacent point pairs when the associated Q is larger than a threshold (e.g. 0.7 in our analysis). This ellipse fitting evaluation process is carried out in an iterative manner. Each time, we connect the pair with the largest Q. In the post-processing step, we prune intersected connections and dividing lines that form a sharp angle to avoid over-segmentation. Finally, we obtain pairs of points to be connected in a set $C^* = \{(p_i^*, q_i^*)|i = 1, 2, \ldots, N_C\}$.

Identification of Dividing Curves

Given identified point pairs for connection, a vast majority of nuclei segmentation methods connect each such pair with a straight separating line. However, it is our observation that a better division is given by a curve following the image gradient information, as shown in Figure 3.1(V). Therefore, we next aim to recover a dividing curve for each such pair with local shape-based intensity analysis. Denoting $I(x, y)$ as the fluorescence microscopy image, we can represent it by Taylor series expansion as:

$$\begin{aligned} I(x, y) \approx I(x_0, y_0) + \bigtriangledown I(x_0, y_0) \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \\ + \frac{1}{2} \begin{bmatrix} x - x_0 & y - y_0 \end{bmatrix} H_I(x_0, y_0) \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \end{aligned} \tag{3.4}$$

where $H_I(x_0, y_0)$ is the Hessian matrix representing the second derivative values of I at the specific point (x_0, y_0). In order to have a better dividing path than a straight line connecting two contour points (p, q), we start off from one end point p and look for the next adjacent pixel with its local neighboring intensity change almost zero in one direction and drastically increased in its orthogonal direction. As a natural dividing curve passes along pixels with local intensity minimum, we can detect such adjacent pixels by computing two eigenvalues of Hessian matrix H_I satisfying such properties: $0 \approx \lambda_1 \ll \lambda_2$. The adjacent pixel with a near zero λ_1 and the maximum λ_2 is considered as the next pixel on the dividing curve.

To ensure that the final optimized connection path can merge to the other end point q smoothly, we search forward in a targeting region with a constrained deviation angle (i.e. $\pm 45°$ in our experiment) from the vector connecting the current dividing point to q at each step. Within such a sector-shaped searching space, we can force the resulting optimal connection path to remain in intensity "valley" with large gradient variance and to converge smoothly to the other end point q. This is demonstrated by the connection between point 5 and 9 in Figure 3.1(V) where our method is able to recover a dividing curve along a local minimum intensity route for overlapped nuclei segmentation.

3.1.3 Experiments

To validate the performance of the proposed algorithm, we tested our method with a dataset of 112 images (1024×1376 pixels) from two brain tumor regions of seven patients. As such images contain a large number of nuclei that are not overlapped, five image patches with overlapped nuclei clusters are randomly selected from each image. This results in 560 image patches for method evaluation.

We visually inspect the resulting nuclei segmentation results and compare our method with the marker-controlled watershed method [16]. Representative results of four overlapped nuclei clusters in our dataset are presented in Figure 3.2. We present the instances of clumped nuclei from original images in the first row, and ground-truth annotations on the second. Nuclei segmentation results from marker-controlled watershed segmentation and our proposed method are presented in the third and forth rows, respectively. It is noticed that our method is able to produce segmentation results similar to human annotations. In particular, those dividing curves recovered from local shape based intensity analysis can help improve segmentation results. By contrast, watershed method fails to identify and separate some overlapped nuclei clusters in the third row. To quantitatively evaluate the performance of our proposed method, five metrics including Jaccard index, Precision, Recall, F1 score, and Hausdroff distance are used. The resulting quantitative evaluation results are presented in Table 3.1. For each patient, both marker-controlled watershed method and our proposed method are tested on 40 images of *Bulk Tumor* (BT) and another 40 of *Tumor Margin* (TM). There are four rows of results associated with each patient in Table 3.1. The top two lines present quantitative results of watershed based method on BT and TM images, while the bottom two demonstrate results from our method on BT and TM images. Note that our method presents superior performance as measured by most of the metrics. Specifically, Jaccard index values from our proposed method are much higher than those of watered based method while Watershed method has better Precision as it tends to miss a large number of nuclei segregation events. One limitation of our method is that it could fail for cases where overlapped nuclei result in internal holes, as it uses contours of clustered nuclei. This will be improved in future work.

3.1.4 Conclusion

This paper presents a novel segmentation method for clumped nuclei in fluorescence microscopy images. Our analysis first generates precise candidate point representations based on a high curvature point voting method, followed by detection of connecting point pairs based on spatial proximity, shape convexity, and curvature information via close point pair screening. An ellipse model is proposed to fit to each

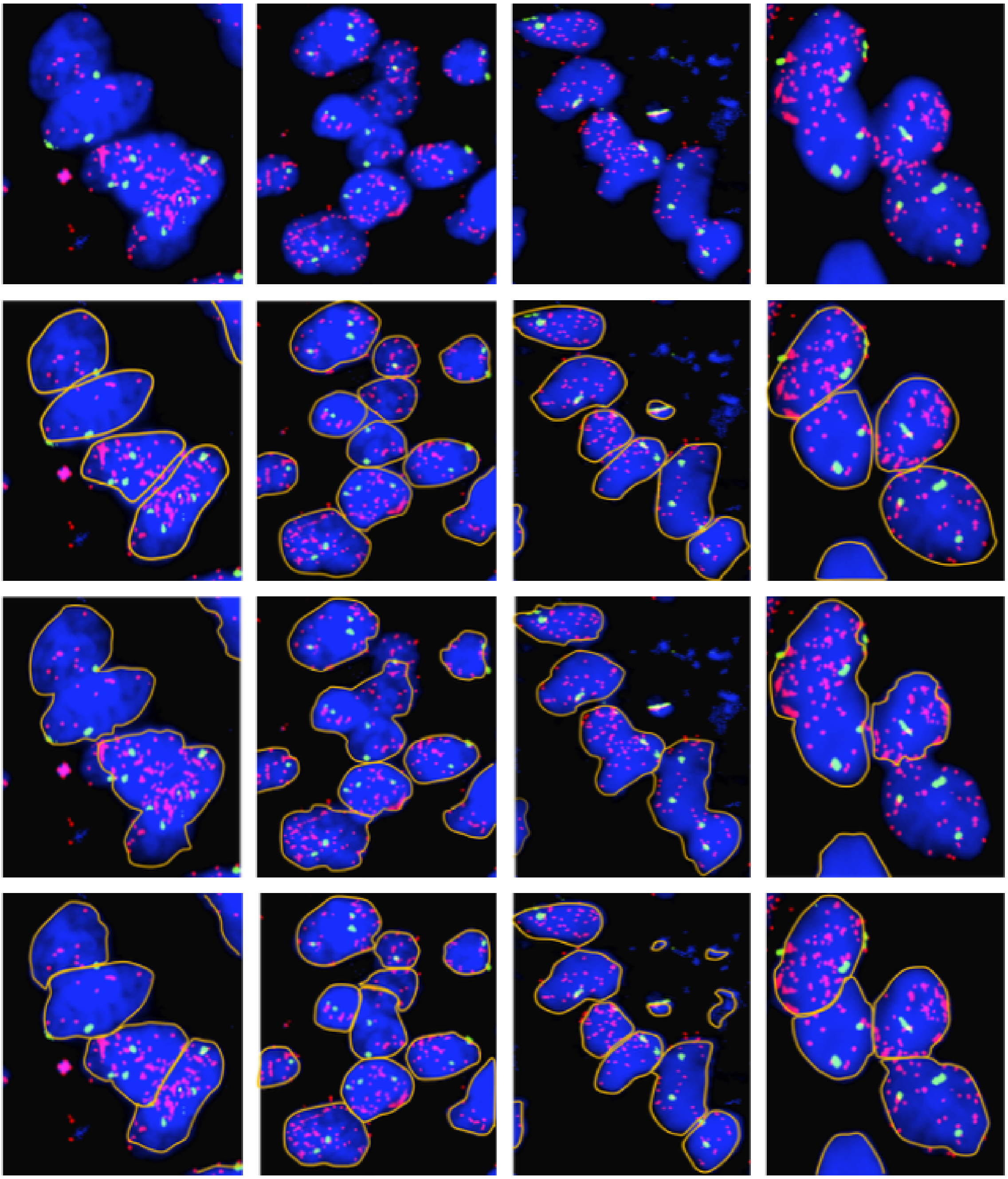


Figure 3.2: Comparisons of segmentation results over four representative overlapped nuclei regions are presented with original images, ground truth, results of marker-controlled watershed method, and outcomes of our proposed method from top to bottom rows, respectively.

Table 3.1: Quantitative method evaluation and comparison with 560 image patches from BT and TM tissue regions: (mean,std).

Patient	Jaccard	Precision	Recall	F1 score	HD (pix.)
1	0.44,0.19	**0.96,0.08**	0.46,0.21	0.58,0.22	34.3,19.1
	0.38,0.18	**0.97,0.07**	0.40,0.20	0.53,0.20	34.6,18.7
	0.75,0.20	0.86,0.18	**0.86,0.17**	**0.84,0.17**	**22.0,25.6**
	0.81,0.14	0.93,0.13	**0.87,0.11**	**0.89,0.11**	**15.0,19.6**
2	0.60,0.16	**0.96,0.11**	0.62,0.16	0.73,0.16	20.2,15.5
	0.64,0.16	**0.98,0.07**	0.66,0.16	0.77,0.13	24.3,17.9
	0.80,0.15	0.90,0.13	**0.89,0.12**	**0.88,0.11**	**13.8,18.4**
	0.83,0.15	0.91,0.12	**0.91,0.12**	**0.90,0.11**	**13.6,20.1**
3	0.49,0.18	**0.94,0.11**	0.52,0.21	0.63,0.19	32.2,23.5
	0.54,0.17	**0.97,0.10**	0.55,0.17	0.68,0.15	31.7,20.3
	0.70,0.23	0.89,0.14	**0.79,0.25**	**0.80,0.21**	**21.7,24.3**
	0.77,0.17	0.89,0.11	**0.85,0.16**	**0.86,0.13**	**13.8,18.2**
4	0.44,0.18	**0.94,0.13**	0.46,0.20	0.59,0.20	33.6,22.0
	0.47,0.15	**0.97,0.09**	0.48,0.16	0.62,0.16	34.5,17.9
	0.75,0.18	0.80,0.19	**0.93,0.11**	**0.84,0.14**	**27.1,29.3**
	0.80,0.19	0.87,0.15	**0.92,0.16**	**0.88,0.15**	**22.0,29.0**
5	0.57,0.16	**0.95,0.10**	0.59,0.17	0.71,0.15	28.9,19.3
	0.49,0.16	**0.97,0.10**	0.50,0.16	0.64,0.17	36.4,19.6
	0.80,0.17	0.85,0.16	**0.93,0.13**	**0.88,0.14**	**21.4,30.8**
	0.86,0.12	0.90,0.10	**0.95,0.09**	**0.92,0.08**	**15.5,20.4**
6	0.51,0.19	**0.93,0.13**	0.55,0.21	0.65,0.19	26.6,19.5
	0.67,0.20	**0.93,0.17**	0.71,0.20	0.78,0.19	22.2,24.7
	0.71,0.21	0.84,0.18	**0.84,0.20**	**0.81,0.17**	**23.8,25.4**
	0.82,0.14	0.88,0.13	**0.93,0.09**	**0.90,0.10**	**13.1,18.1**
7	0.58,0.20	**0.95,0.11**	0.61,0.21	0.71,0.19	33.3,25.0
	0.44,0.20	**0.96,0.08**	0.45,0.21	0.58,0.22	33.0,18.7
	0.78,0.18	0.86,0.14	**0.91,0.16**	**0.86,0.14**	**19.2,30.3**
	0.78,0.17	0.86,0.15	**0.90,0.14**	**0.86,0.14**	**18.5,26.1**

resulting area associated with each point pair candidate. We define a fitting quality as a function of fitting angle, fitting area, fitting ellipse center shift, fitting perimeter change, and fitting area elongation. Only point pairs presenting good fitting quality are connected. Instead of separating overlapped nuclei by a straight line, we recover dividing curves by local shape based intensity analysis in a sector-shaped searching space. We validate our algorithm with 560 image patches from two classes of tumor regions associated with seven brain tumor patients. Both qualitative and quantitative validation results suggest that our algorithm is promising for dividing overlapped nuclei in fluorescence microscopy images widely used in various biomedical research.

3.2 Liver steatosis segmentation

Due to abnormal retention of lipids in hepatocytes, liver steatosis can result from alcohol, obesity, and type II diabetes mellitus [80]. In addition, it serves as the hallmark of a large number of diseases, including non-alcoholic fatty liver disease (NAFLD), alcoholic fatty liver disease, and hepatotoxicity in diverse medical conditions [24]. Therefore, it is essential to achieve accurate quantification of steatosis droplet regions for an accurate disease diagnosis and liver transplantation evaluation [30]. The prevalent gold standard for steatosis assessment is via human visual inspections of liver tissue sections, a process known to be time-consuming and subject to observer variability [101]. Hailed as a new alternative solution, digital pathology is an emerging field that uses digital high-resolution images of tissue sections for machine-based image processing. Although multiple automated methods for liver steatosis measurement have demonstrated computational advantages over human reviewing process [101, 168, 88, 45, 126, 62, 75], they are not sufficiently accurate to support precise steatosis quantification, especially when overlapped steatosis regions with weak separating borders are in presence. As a result, it still remains challenging to develop a robust image analysis program to support precise liver steatosis analysis. As each tissue slide is projected to a two-dimensional microscopy image space, it is not unusual to identify a large number of tissue regions with overlapped steatosis droplets in clumps. Such spatial alignment nature, combined with substantial variations of size, staining color, and structure appearance, presents a technical barrier for individual steatosis droplet segmentation, leading to erroneous steatosis feature computation and size quantification. In this paper, a deep learning model Mask-RCNN is used to segment the steatosis droplets in clumps. Extended from Faster R-CNN, Mask-RCNN can predict object masks in addition to bounding box detection. With transfer learning, the resulting model is able to segment overlapped steatosis regions at 75.87% by Average Precision, 60.66% by Recall, 65.88% by F1-score, and 76.97% by Jaccard index, promising to support liver disease diagnosis and allograft rejection prediction in future clinical practice.

3.2.1 Contribution

We summarize our contributions of this work as follows:

- We adopt the Mask-RCNN based deep learning method [59] and successfully customize it to segment overlapped steatosis droplets in whole-slide histopathology images of live tissue sections.
- To establish a large training data efficiently, we propose to transfer our prior work on nuclei segmentation and have a domain expert to screen results for an accurate training data set generation [53].
- The proposed method can separate highly clumped steatosis droplets and recover their precise contours with promising accuracy.

Publication:

- Guo, Xiaoyuan, Fusheng Wang, George Teodoro, Alton B. Farris, and Jun Kong. "Liver steatosis segmentation with deep learning methods." In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 24-27. IEEE, 2019.

3.2.2 Method

Our work for steatosis analysis is enlightened by the Mask-RCNN segmentation method proposed for object instance segmentation [59]. Extended from Faster R-CNN [120], Mask-RCNN replaces the Region of Interest (ROI)-pooling operation with 'ROI-Align' for solving the misalignment problem. This change in architecture enables segmentation of individual objects from different categories and results in substantial improvement in the segmentation accuracy. Due to the promising performance of Mask-RCNN for instance segmentation, we propose to customize this architecture for steatosis segmentation. The method schema is presented in Figure 3.3 where three primary components are presented: training data preparation with our prior work on nuclei segmentation [53], model training with transfer learning, and overlapped steatosis segmentation in testing images.

Training Data Preparation

Due to the two-dimensional image space projection, a large amount of densely aligned steatosis droplets can touch together with blurred dividing boundaries. Therefore, it is not feasible for pathologists to annotate all steatosis masks for efficient training data production. This is a common problem for deep learning model training in a wide scope of research investigations.

To facilitate training data preparation, we modify our previous nuclei segmentation method [53] and generate initial segmentation masks for steatosis instances. As our previous method aims at separating clumped nuclei in fluorescence in-situ hybridization images, it can not be directly applied to bright field histopathology images for steatosis droplet segmentation. As a result, we modify such method as follows. First, we convert the input color image to its gray-scale representation that is further binarized by a normalized threshold. All non-tissue areas in the image background are excluded for further analysis. Next, overlapped steatosis candidates are identified by rejecting connected foreground regions where solidity is over 0.95. A high curvature point voting method is used to detect dividing candidate points. They are connected based on the fitting quality evaluation by an ellipse fitting model, spatial proximity, shape convexity, and curvature information. Finally, we recover dividing curves by local shape based intensity analysis in a sector-shaped searching space, and produce the corresponding isolated steatosis masks [53].

Although this process produces satisfied results for a large number of touching steatosis droplets, there are partitioned steatosis instances that are not matched with their histology structures as reviewed by the domain expert. Such results are removed from the training data set. As the number of such instances is limited, these unlabeled foreground regions would have little impact on the generalization ability of the trained model. In this way, a good training data set is established in an efficient manner. The resulting data set includes 451 images of $1024 \times 1024 \times 3$ with corresponding mask sets. Each image patch I has multiple masks $\{M_1, M_2, M_3, \cdots, M_n\}$, with each mask image containing one steatosis droplet, essential for solving the overlapped steatosis problem. With this training data set, Mask-RCNN is able to learn how to segregate overlapped steatosis droplets through image-mask pairs.

Deep Learning Model

There are three primary components in Mask-RCNN: the backbone, Region Proposal Network(RPN), and “ROI-Align”. The backbone is composed with Convolutional Neural Networks (CNNs) that can extract multi-level image features. We use modified resnet41, resnet50, and resnet65 [58] as our backbone CNNs. The second component RPN scans the input image with a sliding-window and detects steatosis droplet regions in our study. “ROI-Align” further analyzes ROIs from the RPN and interpolates the feature maps from the neural network backbone at multiple locations. In this way, it can handle the incorrect alignment from Faster R-CNN [120]. With these deep learning components, the resulting model can classify objects into different classes, provide object positions with bounding boxes, and produce a mask for each detected object. In the training process, these three components are orchestrated to minimize the following multi-mask loss function for each steatosis instance [59]:

$$Ł = Ł_{cls} + Ł_{bbx} + Ł_{mask} \tag{3.5}$$

where $Ł_{cls}$, $Ł_{bbx}$, and $Ł_{mask}$ are classification loss, bounding box loss, and mask prediction loss, respectively. More specifically, $Ł_{cls} = -\log(p_i)$ where p_i is the class probability of the instance i; $Ł_{bbx} = \sum_{c_i(j)\in\{x_i,y_i,w_i,h_i\}} SmoothL_1(c_i(j) - C_i(j))$, where c_i and C_i are the centroid coordinates, width, and height of predicted and ground-truth bounding box for the instance i. Additionally, we have the function $SmoothL_1(\cdot)$ defined as: $SmoothL_1(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}$. $Ł_{mask} = -\frac{1}{N^2} \sum\limits_{1\le i,j\le n} [P_{ij} \log p_{ij} + (1 - P_{ij}) \log(1 - p_{ij})]$, where p_{ij} is the predicted mask probability and P_{ij} is the ground-truth mask label at pixel (i, j) in a $N \times N$ region.

Typical steatosis segmentation results are demonstrated at the bottom part of Figure 3.3. The output image on the left presents the steatosis mask prediction with individual steatosis objects color-coded, whereas output images in the middle and right illustrate the predicted steatosis bounding boxes and the classification probabilities, respectively.

3.2.3 Experiments

With our prior segmentation method for nuclei [53], we generate segmentation masks at the highest image resolution. Our training data set is efficiently generated after a domain pathologist removes erroneous masks. The final data set contains 451 liver images of 1024 × 1024 × 3 with ground-truth masks of which 387, 45, and 19 images are used for training, validation, and testing, respectively.

Random neural network initializations can result in an overwhelmingly expensive time cost for model training. Demonstrating its strength for problem solving at a reduced computational cost, transfer learning [105] enables the pre-trained models to serve as the initial point for customized training, and has become popular in a large number of deep learning studies. In our experiment, network weights from the pre-trained COCO model are adopted to initialize our training process. Three backbones network structures, i.e. modified Resnet41, Resnet50 and Resnet65, are used, respectively. For these network backbones, we train the head layer for 30, 20, and 30 epochs, respectively. After head training, all layers are trained to achieve the best segmentation accuracy with 80, 50, 50 epochs, respectively. These epoch numbers are determined heuristically. To minimize the total loss in the training process, back-propagation and Stochastic Gradient Descent(SGD) are utilized. We run experiments on two GPUs (12GB RAM Tesla K80, NVIDIA Inc.) for 300 iterations, with six images per GPU for each mini-batch. The initial learning rate is 0.02, decreased by 10-fold for each 300 iterations. Additionally, online data augmentation techniques are used to further scale up data set size, improve the training performance, and increase generalizability and robustness of the trained model. These techniques include random affine transform, random flipping, and Gaussian blurring.

The process of applying the trained network to whole-slide microscopy images is presented in Figure 3.4 where the left image presents the overall view of a representative whole-slide image, with a green box illustrating the close-up view of a small tissue part. The image in the middle demonstrates clustered steatosis droplets in a small liver tissue region at the full image resolution, while segmentation results of steatosis droplets by our trained network are illustrated with distinct colors in the right image.

To make this method flexible for diverse liver disease diagnosis and transplantation evaluation settings, we characterize each recognized steatosis candidate by the eccentricity, size, and perimeter, and record the segmentation score resulting from the neural network prediction. We provide user-defined thresholds for these features, enabling customized steatosis droplets retention. In this way, domain pathologists can have a convenient way to select cutoff values for these features and get readily informed of the number and the morphological profiles of the retained steatosis objects. We demonstrate the neural network segmentation results before and after such post-processing in Figure 3.5 where we keep steatosis candidates with size, perimeter and eccentricity within $[0.001, 6]$, $[0.5, 4]$, and $[0.2, 1.5]$ times of the average steatosis, respectively.

Representative segmentation results are presented in Figure 3.6 where the original input images, ground-truth segmentation, results from our earlier method [53], outputs of presented deep learning approach with modified Resnet41, Resnet50 and modified Resnet65 backbones are presented in columns from left to right. It is noted that deep learning methods, especially the network with backbone Resnet50, present better results than our earlier work as they present less under-segmentation results. Additionally, deep learning methods present good performances on handling clumped steatosis clusters with complex topology.

To quantitatively evaluate the presented method, we compare results from deep learning algorithm with the ground-truth data. Table 3.2 presents evaluation results of averaged steatosis measures by four metrics, including Average Precision, Recall Ratio, F1-score, and Jaccard index. It is noticed that the trained network with Resnet50 achieves the best precision, recall ratio, F1-score, and Jaccard index.

3.2.4 Conclusion

We propose to use the deep learning method to solve the overlapped liver steatosis segmentation problem. Due to lack of labelled steatosis droplets from liver microscopy images for training, we modify our earlier nuclei segmentation method to generate liver steatosis training data after an efficient screening process by a domain expert. The trained neural network model is demonstrated to segment liver steatosis droplets, especially those in clumps with promising accuracy. Quantitative evaluations suggest that deep learning technology enables accurate and high-performance steatosis segmentation, a promising tool for enhancing liver disease diagnosis and transplantation assessment.

Table 3.2: Method performance evaluation and comparison.

Method	AP	Recall	F1-score	Jaccard
Clump_seg [53]	52.18%	45.50%	45.03%	67.42%
ResNet41	67.61%	58.37%	62.06%	73.18%
ResNet50	**75.87%**	**60.66%**	**65.88%**	**76.97**%
ResNet65	69.55%	55.60%	61.69%	74.38%

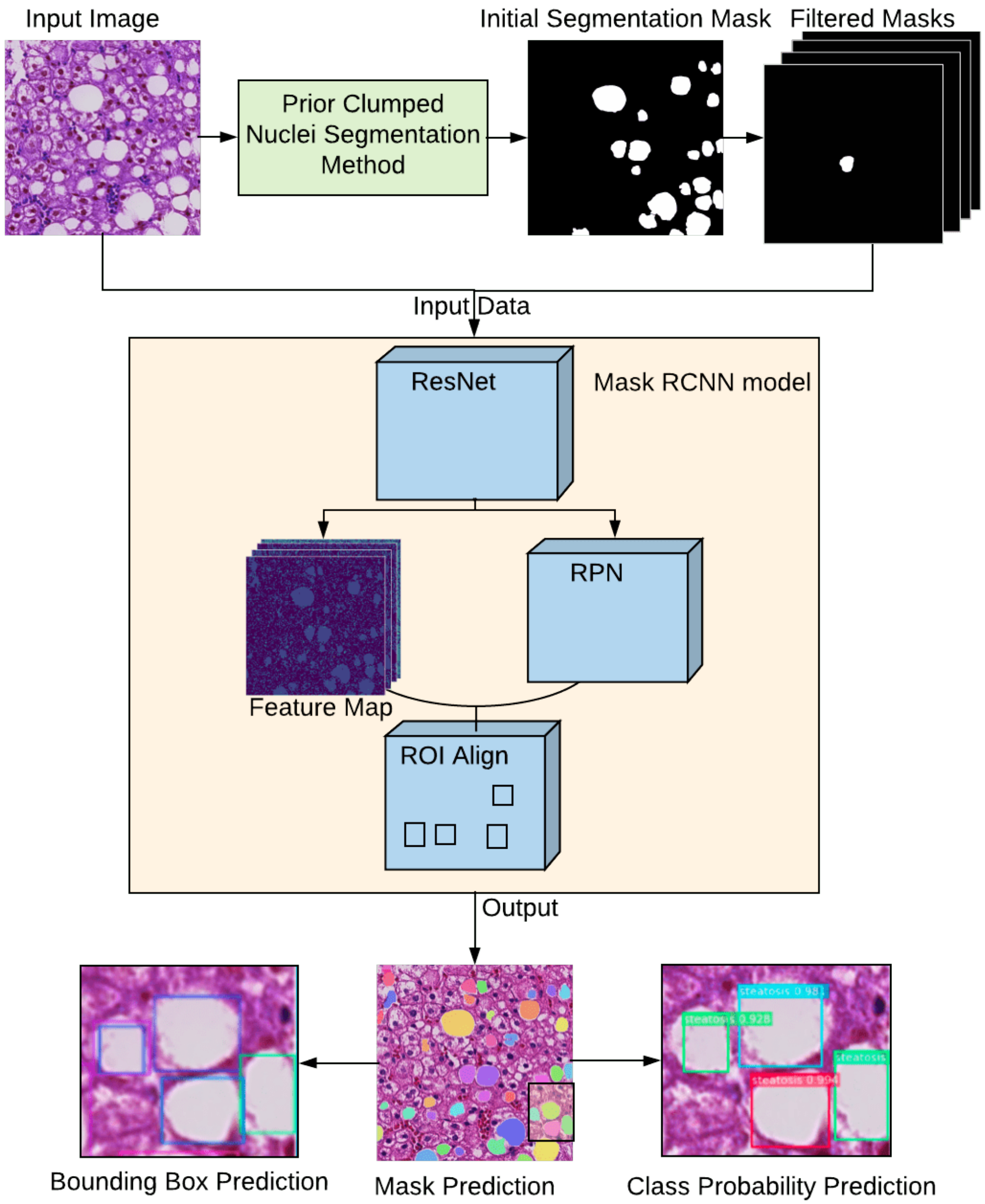


Figure 3.3: Schema of steatosis segmentation method.

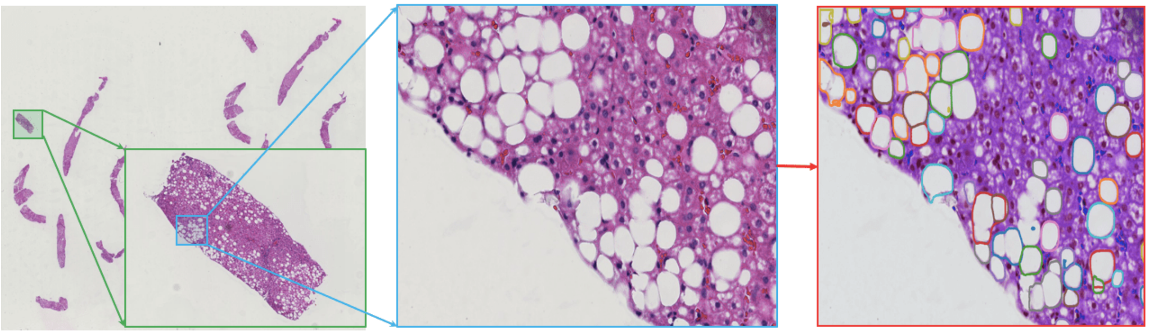


Figure 3.4: Segmentation process for a whole-slide microscopy image.

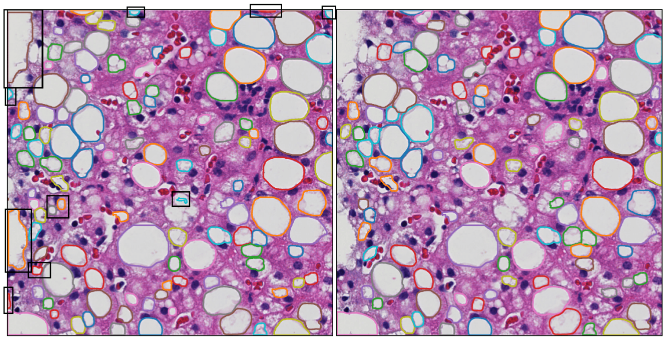


Figure 3.5: Segmentation result (Left) before and (Right) after post-processing, with black boxes in the left image indicating the discarded steatosis regions.

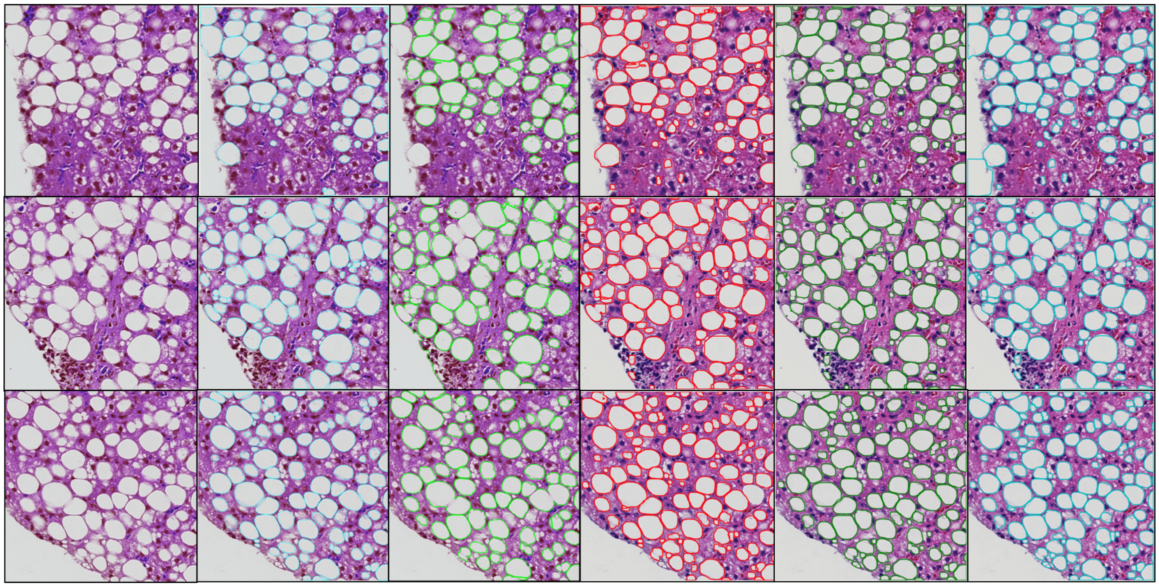


Figure 3.6: In columns from left to right, we demonstrate original images, ground-truth, segmentation results from method [53], Mask-RCNN with modified Resnet41, Resnet50, and modified Resnet65, respectively.

3.3 Breast arterial calcifications (BAC) segmentation

Cardiovascular disease is a source of high morbidity and mortality in women [49]. One of the barriers to improving diagnosis outcomes is the lack of a simple, inexpensive, and reliable method for screening and for assessing efficacy of therapies. Vascular disease commonly manifests as arterial calcifications, which are typically assessed by computed tomography (CT) or CT angiography of the coronary arteries and aorta [38]. However, these tests are expensive, usually performed only in symptomatic patients, and associated with additional radiation exposure. Calcification also occurs in breast arteries and can be readily observed on screening mammograms. The prevalence of breast arterial calcifications (BAC) correlates with calcifications in other arteries and is associated with an increased risk of cardiovascular disease events [60, 43, 1, 21]. We recently showed that quantification of BAC through manual measurements can more accurately stratify risk factors and provide a means to follow progression [7, 6, 100].

Each year, more than 40M women over age 40 undergo screening mammography for breast cancer screening [21]. Automatic detection and quantification of BAC in these women may be helpful in identifying patients at high-risk for cardiovascular events and following progression of vascular calcifications without additional cost or radiation exposure [66]. Stored digital mammograms over the past decade would also provide a vast dataset for robust retrospective research. Currently, there is no standardized method for accurate detection, segmentation and quantification of BAC on mammography, which limits the utility of this potential biomarker. There are many challenges in automated detection of BAC. First, BAC appear as slender, elongated regions of fragmented high pixel intensity on mammograms and typically represent fewer than 1% of a $4K \times 3K$ image. Moreover, the narrow appearance and potential variable lengths make precise segmentation of BAC much more challenging compared to general segmentation tasks. Second, there is no standard strategy for acquiring groundtruth BAC segmentations due to the variations in vessel width, severity of calcifications along the vessel, and tortuous vessel paths. Third, the large image size (over 12MP) adds significant difficulty in image processing.

Although there have been a number of existing works relevant to breast arterial calcifications, few have focused on accurate segmentation. Sulam et al. [143] examined only prevalence and Abriele et al. [148], Juan et al. [154] and Hossain et al. [63] all detected BAC with a patch-based method, but did not report detailed segmentation performance or quantification metrics. Since BAC segmentation can be considered as a type of semantic segmentation in the realm of general computer vision, current semantic segmentation models can be attempted for BAC segmentation. Generally, semantic segmentation models can be classified into two main categories: non-real-time and real-time segmentation models. Non real-time models such as U-Net [124], SegNet [13], DeepLabV3 [27] and LinkNet [26] usually have complex architectures and a high number of trainable parameters. Thus, they may achieve high accuracy but are slow to train and deploy. By contrast, real-time semantic segmentation models including ERFNet [123], ESNet [158], FastSCNN [112], ContextNet [111], DABNet [81], EDANet [96], FPENet [91], CGNet [161] have fewer trainable parameters but can still attain comparable performance with the non-real-time models. At our institution, up to 250 screening mammograms are performed daily constituting approximately 1,000 images. In live clinical deployment, it would be advantageous that BAC detection and quantification occur in near real-time so that the results are available to the interpreting radiologist in case patient referral is needed. Therefore, segmentation models with a high number of trainable parameters (*e.g.,* U-Net [124] has 13,395,329 parameters) would be prohibitive in their inference times, and lightweight models would enable more clinically viable.

3.3.1 Contribution

Our main contributions can be summarized as below:

- We propose Simple Context U-Net (SCU-Net), an automated lightweight segmentation model, to segment BAC in mammograms in a patch-based way. SCU-Net offers comparable performance of the most popular current segmentation architectures with an order of magnitude fewer training parameters. It achieves this by taking advantage of both dilated convolution operations and skip connections to learn and fuse global features with low-level information efficiently while maintaining far fewer trainable parameters.

- We demonstrate the efficacy of SCU-Net by visually and quantitatively presenting our BAC segmentation results as compared to a series of popular semantic segmentation models.

- Furthermore, we present five novel metrics to quantify the severity of BAC within the segmentation mask, compare our quantification metrics to breast CT, and demonstrate the ability to track a longitudinal increase in BAC in a cohort of patients with 10 years of retrospective mammograms. Thus, SCU-Net model may serve as a potential research and clinical tool for early detection and risk stratification of cardiovascular disease for women. The code is available at `https://github.com/XiaoyuanGuo/BAC_segmentation`.

Publication:

- Guo, Xiaoyuan, Judy Wawira Gichoya, Hari Trivedi, W. Charles O'Neill, Rhakur Priya, Weijia Sun, Manisha Singh, Kathiravelu Pradeeban, Thomas J. Kim, Tianen Christopher Yang and Imon Banerjee. "Deeper Thinner UNet (DT-UNet) for Fine Vessel Segmentation of Breast Arterial Calcification (BAC)." CMIMI2020. `https://cdn.ymaws.com/siim.org/resource/resmgr/mimi20/abstracts/deeper_thinner_unet_guo.pdf`

- Guo, Xiaoyuan, W. Charles O'Neill, Brianna Vey, Tianen Christopher Yang, Thomas J. Kim, Maryzeh Ghassemi, Ian Pan, Judy Wawira Gichoya, Hari Trivedi, and Imon Banerjee. "SCU-Net: A deep learning method for segmentation and quantification of breast arterial calcifications on mammograms." Medical physics 48, no. 10 (2021): 5851-5861.

3.3.2 Method

Preprocessing

Mammograms contain a wide variety of pixel intensities with varying breast shapes and proportions of breast tissue versus null background. Therefore, image preprocessing is critical to identify breast tissue and normalize the image to maximize the model performance. To this end, we first smooth the image using median filtering [50] with a disk kernel of size 5 for cleaning the noise but also avoiding causing serious blurring. This was chosen empirically among the evaluated range of [5-20] based on visual evaluation during preliminary experiments. To extract breast tissue only, we erode and then dilate the breast images with a disk kernel (size is 10 in our experiment) to erase the scanner labels of mammograms such as view type (*i.e.,* "RMLO" – right mediolateral oblique, "LMLO" – left mediolateral oblique, "RCC" – right craniocaudal, "LCC" – left craniocaudal). With the same setting, we dilate and then erode the binary mask to link together and smooth any nearby annotation segments, producing a continuous vessel mask. Finally, we enhance image contrast to maximize the difference between calcified vessel and background tissue. During training, we normalize input image patches with zero-means method to minimize the impact of variation contrast between vessels and background.

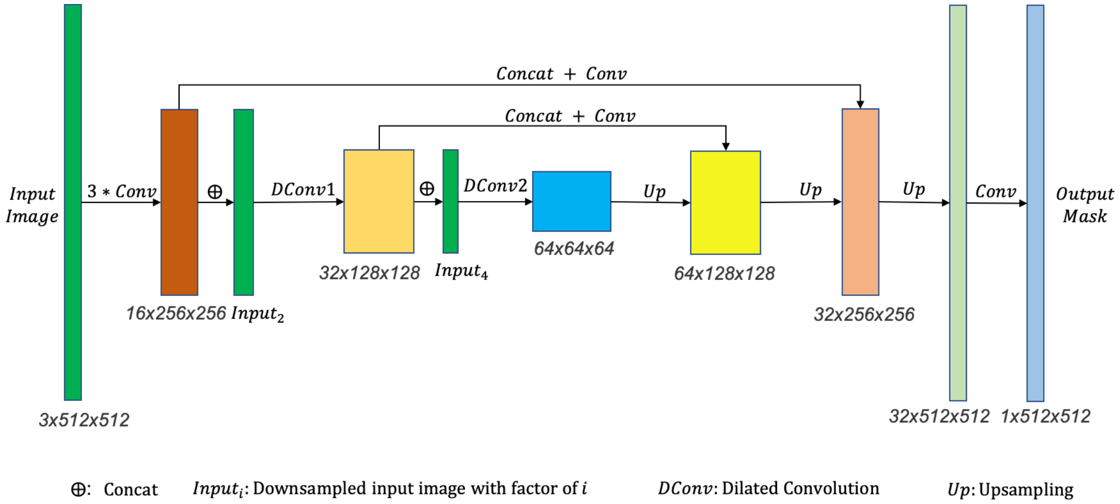


Figure 3.7: Network architecture of SCU-Net.

Network architecture

To overcome the issue of large image sizes and the inability to downsample images without data loss, we propose Simple Context U-Net (SCU-Net), whose inputs are patches cropped at the highest resolution of mammography images. The architecture of SCU-Net is shown in Figure 3.7. All the feature sizes in the figure are presented same as our experimental settings. SCU-Net is a symmetric, U-shaped model, similar to U-Net [124]. The model has input image patches with size of $3 \times 512 \times 512$. [1] The original input is fed into three 3×3 convolutional layers. To preserve the original image information, the input patch is downsampled with scale factor of 1 and 2. The obtained two downsampled input features are in size of $3 \times 256 \times 256$ and $3 \times 128 \times 128$ corresponding to the second and third green additional inputs of Figure 3.7. These two downsampled inputs will be concatenated with later high-level features. Each concatenation is followed by BatchNormalization and Parametric ReLU operations, enabling smooth fusion of high-level information with low-level

[1] Although the mammogram image is grayscale and has only one image channel, three duplicates of the mammogram patch are stacked together to form a three-channel image same as RGB image format. This setting ensures the model to work for both natural and grayscale images, and can be comparable with existing segmentation models.

features. Feature fusing is important, but the surrounding context is also very helpful for semantic segmentation [161]. Inspired by CGNet [161] and DilatedNet [164], SCU-Net adopts two different dilated convolutional layers (Dconv1 and Dconv2 in Figure 3.7) to aggregate multi-scale contextual information. In the decoder arm of the network, the learned image features are upsampled with bilinear interpolation and then concatenated with the corresponding encoder features of the same size. "Up" in Figure 3.7 means upsampling layer. Two 3 × 3 convoultional layers follow each concatenation. In total, there are three upsampling layers to get the network back to the original size. Finally, two 3 × 3 convolutional layers helps reduce the channel numbers to the class number, 1 in our case, and a Sigmoid layer is used to get the final mask prediction. All the convolutional layers including conv, Dconv1, Dconv2 and Up layers in Figure 3.7 are followed with BatchNormalization and Parametric ReLU operations. To avoid overfitting, we use online data augmentation techniques during training, including randomly vertical or horizontal flipping, randomly rotation by 90 or 270 degrees, and randomly changing the brightness, contrast and saturation of image.

Implementation details: In our experiments, binary cross entropy loss converges much more slowly than dice loss, therefore we adopt dice loss to optimize all the segmentation networks. For optimization, we use Adamw optimizer with a learning rate of 0.001 for model training. Each network is trained with 50 epochs. The pipelines are developed using Pytorch 1.5.0, Python 3.0. and Cuda compilation tools V10.0.130 on a machine with 4 NVIDIA Quadro RTX 6000 with 24GB memory.

Experimental setup

With the approval of Emory Institutional Review Board (IRB), three cohorts of subjects were identified from previous studies [7, 6, 100]. All mammograms extracted were 2D full-field digital mammograms (FFDM) obtained during routine screening exams on Hologic (Marlborough, PA) mammography scanners in accordance with Mammography Quality Standards Act (MQSA) requirements. Screening exams consisted of four standard views - LCC, LMLO, RCC, RMLO.

- Cohort A – 661 FFDM from 216 subjects were annotated and used for deep learning model training and validation. The mean age was 70 ± 11 and 37% were African-American. Because the previous studies focused on kidney disease, 35% had chronic kidney disease, end-stage renal disease (ESRD), or renal transplantation. Mean breast density was 2.23 ± 0.77 as reported according to Breast Imaging Reporting and Data System (BI-RADS) guidelines (A=1, B=2, C=3, D=4). The majority of patients were density B (scattered fibroglandular tissue - 43.6%) and C (heterogeneously dense - 41.7%) with a minority of density A (mostly fat - 7.2%) and D (extremely dense - 7.5%).

- Cohort B for comparison to breast CT calcification - A previously reported cohort of 10 subjects with contemporaneous measurement of BAC by breast CT. Mean age was 69 ± 11 and all but one were Caucasian. Mean breast density was slightly lower at 2.08 ± 0.76.

- Cohort C for longitudinal analysis - 26 additional subjects with BAC and at least 5 yearly mammograms were studied in order to assess the ability to detect progression of BAC. The mean age was 65 ± 12 and 54% were African-American. Of these, 9 had ESRD or had undergone kidney transplantation. Mean breast density was similar at 2.19 ± 0.70.

Groundtruth acquisition: Mammograms from Cohort A were annotated by four annotators - one physician (CO) with 15 years experience and three other annotators trained and monitored by CO. Groundtruth segmentations are performed manually on whole images using the online platform Md.ai[2] and standardized by annotating a

[2] `www.md.ai`

multi-segmented line down the center of any calcified vessel continuously until there is at least a 1cm length of non-calcified vessel, at which point a new segmentation is started where the calcification resumes. These annotations serve as groundtruth training and validation data.

Data preparation: To prepare high-quality datasets for training deep learning models, the whole mammogram dataset is randomly divided into training and validation parts with 527 mammography images for training and 134 for validation. The mammography images are either sized 4096 × 3328 pixels or 3328 × 2560 pixels, which require a large amount of memory to load and analyze. Therefore, we crop images into fixed-size patches of 512×512 with 64 pixels of overlap between adjacent patches. The overlapping ensures the ability to connect BAC segmentations from adjacent patches and improves the overall segmentation accuracy. We exclude black background image patches to eliminate unnecessary calculations. Moreover, only patches that contain calcifications are left for segmentation training given the fact that the calcification mask prediction is pixelwise classification. Ultimately, this yields 3,455 effective patches for training and 901 patches for validation.

Model comparison: Experiments are performed with SCU-Net and state-of-the-art deep learning models including SegNet [13], DeepLabV3 [27], U-Net [124], LinkNet [26], ERFNet [123], ESNet [158], FastSCNN [112], ContextNet [111], DABNet [81], EDANet [96], FPENet [91] and CGNet [161]. Their number of trainable parameters, including SCU-Net, are compared in Figure 3.8. The larger the circle area for a model is, the more parameters the model contains. As can be seen, SegNet [13] has the most parameters while FPENet [91] contains the least. Our model, SCU-Net, has the second fewest parameters (marked in blue). Models with fewer parameters have lower complexity, consume less memory, and achieve faster training. Since mammograms (along with most radiology images) are very large in size, the number of model parameters is an important factor for real-world implementation as it is directly related to speed.

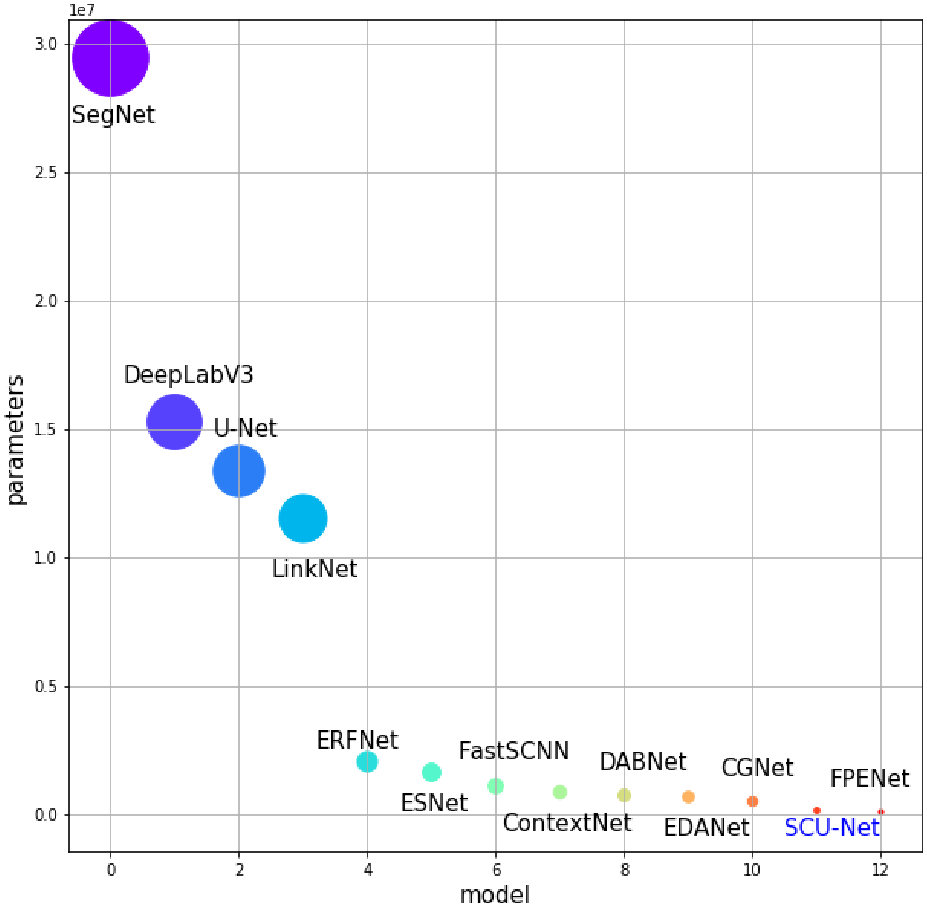


Figure 3.8: Trainable parameters comparison of segmentation models. The circle area is proportional to the total parameters of the model. Comparatively, SCU-Net is roughly two orders of magnitude smaller than other models.

Evaluation metrics for BAC segmentation: We evaluate both patch-wise segmentation results and final whole image segmentation results of all the models with five metrics: *Recall*, *Precision*, *Accuracy*, *F1-score/Dice score*, *Jaccard Index* value. The definitions are shown in Equations 3.6 and 3.7. In the equations, TP, FN, TN and FP calculations refer to pixelwise results.

$$Recall = \frac{TP}{TP+FN}, \quad Precision = \frac{TP}{TP+FP}, \quad Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.6}$$

$$F1 - score = \frac{2 \times Recall \times Precision}{Recall + Precision}, \qquad JaccardIndex = \frac{TP}{TP + FP + FN} \quad (3.7)$$

To further demonstrate the differences across all the models, we also perform pairwise t-test to compute the statistical significance of state-of-the-art models compared with SCU-Net. The p-value table is present in the supplementary material.

Evaluation metrics for BAC quantification: Beyond typical semantic segmentation evaluation metrics (*Recall, Precision, Accuracy, F1-Score/Dice Score and Jaccard Index*), we propose five BAC quantification metrics in Equations 3.8 and 3.9 to further measure the effectiveness of BAC detection in the predicted segmentation masks. Because of the segmentation challenges with BAC, we anticipated acceptable but imperfect segmentation results. However, unlike cancer detection where localization is extremely important, vessel segmentation can be considered an intermediate task to achieve BAC quantification. Slight differences in vessel segmentation region or width may have strong negative effects on standard evaluation metrics like Dice score and Jaccard index, but may still provide excellent results in terms of capturing clinically relevant calcifications. Therefore, we developed the following five metrics to capture the total segmented area, intensities of pixels within the segmented area, and thresholded pixel intensities and counts within the segmented area. Equations 3.8 and 3.9 show the definitions for Sum of Mask Probability Metric ($\mathcal{PM}$), Sum of Mask Area Metric ($\mathcal{AM}$), Sum of Mask Intensity Metric ($\mathcal{SIM}$), Sum of Mask Area with Threshold Intensity **X** Metric ($\mathcal{TAM}_X$) and Sum of Mask with Intensity Threshold **X** Metric ($\mathcal{TSIM}_X$). In the equations, m and n refer to the width and height of the mammogram, $p_{i,j}$ is the probability value at $< i, j >$ returned by the trained model, $\mathcal{I}_{i,j}$ means the intensity value of pixel at $< i, j >$ and **X** is the intensity threshold.

$$\mathcal{PM} = \sum_{i=0,j=0}^{m,n} p_{i,j}, \qquad \mathcal{AM} = \sum_{i=0,j=0}^{m,n} 1_{p_{i,j}>0.5}, \qquad \mathcal{SIM} = \sum_{0\leq i\leq m, 0\leq j\leq n | p_{i,j}>0.5} \mathcal{I}_{i,j} \tag{3.8}$$

$$\mathcal{TAM}_X = \sum_{0\leq i\leq m, 0\leq j\leq n | p_{i,j}>0.5} 1_{\mathcal{I}_{i,j}>\mathbf{X}}, \qquad \mathcal{TSIM}_X = \sum_{0\leq i\leq m, 0\leq j\leq n | p_{i,j}>0.5, \mathcal{I}_{i,j}>\mathbf{X}} \mathcal{I}_{i,j} \tag{3.9}$$

Specifically, $\mathcal{PM}$ summates all predicted probabilities for an image to evaluate the confidence of the model's prediction; $\mathcal{AM}$ is the total number of pixels that are classified as BAC in a mammogram; $\mathcal{SIM}$ is the sum of the intensities of the pixels classified as BAC; $\mathcal{TAM}_X$ is the total number of BAC-classified pixels greater than intensity threshold $\mathbf{X}$, as the BAC pixels usually have higher intensity values than background tissue area; $\mathcal{TSIM}_X$ is the sum of intensities for BAC-classified pixels with intensity value greater than the threshold $\mathbf{X}$. In our experiment, we set $\mathbf{X}$ to be 100 as the best threshold for $\mathcal{TAM}_X$ and $\mathcal{TSIM}_X$ metrics based on visual observations of threshold values of 50, 75, 100, 150, 200. Metrics $\mathcal{AM}$, $\mathcal{SIM}$, $\mathcal{TAM}_X$, and $\mathcal{TSIM}_X$ are all calculated with a model prediction cutoff of $p > 0.5$.

Comparison of BAC quantification metrics against breast CT measurements: To compare our quantification with a previously clinically validated measurement system [100], we evaluated our quantification metrics on mammograms of 10 patients in Cohort B who had contemporaneous breast CT exams. All BAC quantification metrics on mammograms were compared to calcified voxels and calcium mass as measured on breast CT.

Evaluation of BAC quantification metrics longitudinally: To evaluate the utility of BAC quantification metrics to track calcification longitudinally, we examined 26 new subjects (Cohort C) not included in the original dataset with serial mammograms. Each patient had 5$\sim$12 years imaging history with all four standard screening mammography views per exams, totalling 961 images across all subjects. SCU-Net was applied to each image to obtain the segmentation masks and $\mathcal{TAM}_{100}$ was calculated (based on top-performing correlation as shown in Figure 3.11). Plotting $\mathcal{TAM}_{100}$ over time *per view* initially yielded very noisy results in which calcification quantity appeared to oscillate over time, which typically would physiologically not occur. We then took the sum of the $\mathcal{TAM}_{100}$ for *all views* plotted against time, which somewhat decreased the fluctuation but did not eliminate it. Finally, we realized that each year the patient's breast position and magnification of the mammogram could vary, meaning that the raw number of pixels as counted in the $\mathcal{TAM}_{100}$ metric would be dependent on breast magnification. To normalize for this effect, we took $\mathcal{TAM}_{100}$ metric divided by the breast area for each image and then sum this result across all four views. This was the final method used for longitudinal analysis.

3.3.3 Experiments

Evaluation of BAC detection based on standard metrics - Figure 3.9 shows the patch-wise segmentation results of SCU-Net as compared to several semantic segmentation models including SegNet [13], ContextNet [111],U-Net [124], CGNet [161] and SCU-Net. The first row is of particular interest as it demonstrates ductal calcifications which are benign and unrelated to BAC, but can appear similar. SegNet [13], ContextNet [111], and U-Net [124] each erroneously detect these ductal calcifications to varying degrees, however SCU-Net correctly ignores these. Interestingly, SCU-Net demonstrates similar performance to CGNet [161] as they both utilize dilated convolution operations to learn context features. The second row of Figure 3.9 demonstrates a patch with overall lower image contrast and overlapping breast tissue which mimics linear calcifications. In this case, ContextNet [111] detects the most false positive pixels. The third and fourth cases contain less noise and a clear difference from the

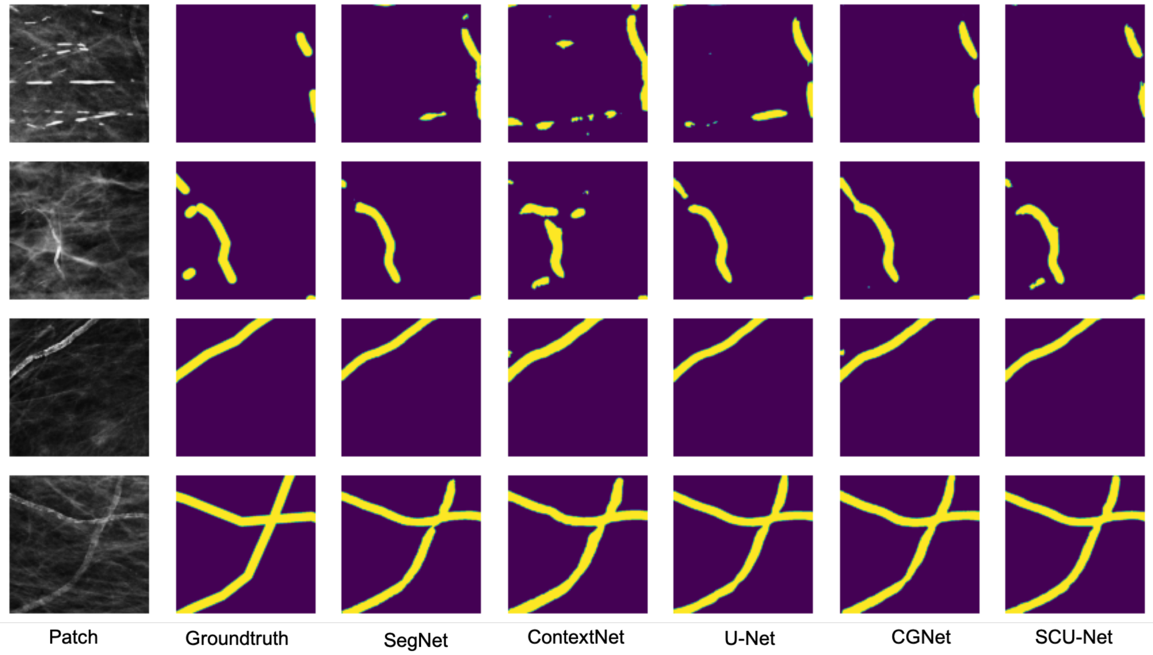


Figure 3.9: Examples of patch-wise segmentation results for BAC across multiple architectures as compared to the groundtruth. From left to right: original image patches, groundtruth mask, and prediction results of SegNet, ContextNet, U-Net, CGNet and SCU-Net.

background tissue, in which case all the models perform well at detecting BAC. In brief, image noise, low image contrast, and overlapping background tissue can all affect the quantitative accuracy of segmentation. The same types of errors are noticed on whole-image-size mask prediction as shown in Figure 3.10. For better visualization, only the breast region are kept by truncating the unnecessary background from the original mammograms.

In this figure, the dice scores for the predicted masks of each case are labelled in the top right corner. As can be seen, overall performance for BAC segmentation is quite good although each model suffers from varying degrees of false positives due to issues with image noise, tissue contrast, and lookalike findings. We also see that some images are intrinsically more difficult with lower dice scores across the board for rows 1 and 2 in as compared to rows 3 and 4 in Figure 3.10. In general, the

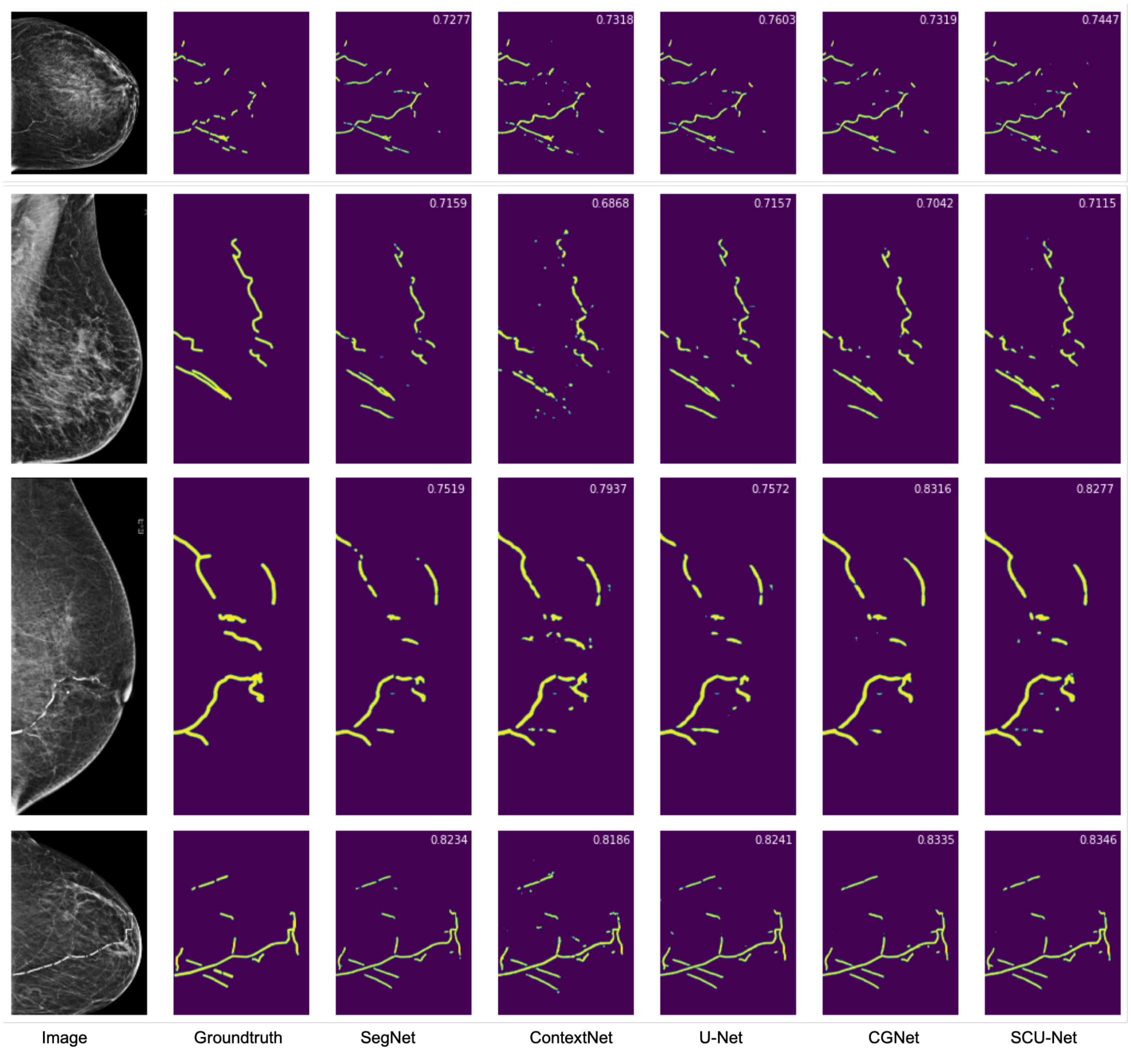


Figure 3.10: Examples of whole image segmentation results for BAC across multiple architectures as compared to groundtruth. From left to right: original mammography images (cropped to exclude background), groundtruth mask, prediction results of SegNet, ContextNet, U-Net, CGNet and SCU-Net. The F1-Score for each model is shown in the top right of the predicted mask. Higher F1-score means more overlap between groundtruth and the predicted mask.

segmentation masks of ContextNet [111] contain more false positive fragments than other results. Nevertheless, most of the BAC is captured by all the models. Notably, SCU-Net achieves comparable dice scores compared to SegNet [13], U-Net [124] and CGNet [161] despite significantly fewer parameters.

Furthermore, we evaluate the segmentation results for both patches and whole images to demonstrate the fine vessel calcification segmentation accuracy. Table 3.3 presents the quantitative performance metrics for all tested models including SCU-Net, for both invidual patches (columns with clear background) and whole mammography images (columns with gray background). For patch-wise quantitative results in Table 3.3, ERFNet [123] has the highest recall value, FPENet [91] achieves the best precision value, and SCU-Net has the best F1-score and ties with CGNet [161] for top Jaccard Index value. Accuracy values of all the models are relatively similar due to the high number of negative pixels in the image. Whole-image-size results are generated by concatenating the corresponding patches for each whole mammogram. Compared with patch-wise results, nearly all the evaluation metrics for the whole image are higher and are tightly grouped across all models. The reason lies in the overlapping 64 pixels with neighboring patches which helps enhance the segmentation accuracy by avoiding boundary effects[3]. On whole images, ERFNet [123], FPENet [91], DeepLabV3 [27] still maintain their advantages in recall, precision and accuracy respectively. U-Net [124] and DeepLabV3 [27] in Table 3.3 have the best F1-score/Dice-score (0.735) and Jaccard Index value (0.59) for full image segmentation. With many fewer parameters (79x less), SCU-Net also performs very well with 0.729 of F1-score and 0.581 of Jaccard Index value compared with SegNet [13] and FPENet [91].

Evaluation of BAC quantification based on defined metrics - Universal semantic segmentation evaluation metrics are helpful in evaluating segmentation results by performing pixel-to-pixel evaluation. However, the ultimate goal of this work is to quantify the amount of BAC within a mammogram for eventual correlation with cardiovascular outcomes. To evaluate the practical performance of SCU-Net's segmentations in capturing BAC, we computed the correlation for all metrics computed

[3] Cropped patches may only contain a very small piece of calcification along the cropped boarder, which is hard to segment accurately. However, the larger calcification can be more easily detected in the adjacent patches. Thus, concatenating the predictions of adjacent patches can eliminate the boundary effects.

Table 3.3: Quantitative evaluation results for **image patches** (columns without background) and **whole images** (columns with gray background) in the validation dataset, subscripts denote standard deviation.

Method	***Recall***		***Precision***		***Accuracy***		***F1-score***		***Jaccard***	
SegNet [13]	$0.707_{\pm0.100}$	$0.764_{\pm0.159}$	$0.704_{\pm0.095}$	$0.743_{\pm0.128}$	$\mathbf{0.981_{\pm0.005}}$	$\mathbf{0.998_{\pm0.002}}$	$0.676_{\pm0.084}$	$0.734_{\pm0.098}$	$0.554_{\pm0.079}$	$0.589_{\pm0.113}$
DeepLabV3 [27]	$0.742_{\pm0.099}$	$0.781_{\pm0.154}$	$0.709_{\pm0.088}$	$0.726_{\pm0.134}$	$\mathbf{0.981_{\pm0.005}}$	$\mathbf{0.998_{\pm0.002}}$	$0.692_{\pm0.084}$	$\mathbf{0.735_{\pm0.100}}$	$0.568_{\pm0.081}$	$\mathbf{0.590_{\pm0.118}}$
U-Net[124]	$0.738_{\pm0.092}$	$0.789_{\pm0.144}$	$0.704_{\pm0.088}$	$0.723_{\pm0.141}$	$\mathbf{0.981_{\pm0.005}}$	$\mathbf{0.998_{\pm0.002}}$	$0.689_{\pm0.074}$	$\mathbf{0.735_{\pm0.097}}$	$0.562_{\pm0.073}$	$\mathbf{0.590_{\pm0.112}}$
LinkNet [26]	$0.748_{\pm0.095}$	$0.801_{\pm0.151}$	$0.675_{\pm0.096}$	$0.690_{\pm0.137}$	$0.979_{\pm0.006}$	$0.997_{\pm0.002}$	$0.676_{\pm0.082}$	$0.720_{\pm0.101}$	$0.550_{\pm0.080}$	$0.572_{\pm0.114}$
ERFNet [123]	$\mathbf{0.788_{\pm0.088}}$	$\mathbf{0.826_{\pm0.133}}$	$0.669_{\pm0.086}$	$0.673_{\pm0.151}$	$0.979_{\pm0.006}$	$0.997_{\pm0.002}$	$0.694_{\pm0.075}$	$0.724_{\pm0.106}$	$0.568_{\pm0.077}$	$0.578_{\pm0.123}$
ESNet [158]	$0.757_{\pm0.096}$	$0.796_{\pm0.164}$	$0.684_{\pm0.091}$	$0.707_{\pm0.137}$	$0.980_{\pm0.005}$	$0.997_{\pm0.002}$	$0.687_{\pm0.083}$	$0.727_{\pm0.108}$	$0.563_{\pm0.081}$	$0.581_{\pm0.122}$
FastSCNN [112]	$0.687_{\pm0.105}$	$0.738_{\pm0.171}$	$0.662_{\pm0.100}$	$0.695_{\pm0.136}$	$0.979_{\pm0.006}$	$0.997_{\pm0.002}$	$0.647_{\pm0.096}$	$0.697_{\pm0.112}$	$0.522_{\pm0.092}$	$0.545_{\pm0.124}$
ContextNet [111]	$0.723_{\pm0.093}$	$0.765_{\pm0.165}$	$0.631_{\pm0.090}$	$0.628_{\pm0.150}$	$0.977_{\pm0.006}$	$0.997_{\pm0.003}$	$0.643_{\pm0.083}$	$0.671_{\pm0.123}$	$0.509_{\pm0.081}$	$0.517_{\pm0.130}$
DABNet [81]	$0.750_{\pm0.096}$	$0.804_{\pm0.143}$	$0.692_{\pm0.095}$	$0.706_{\pm0.142}$	$\mathbf{0.981_{\pm0.005}}$	$\mathbf{0.998_{\pm0.002}}$	$0.686_{\pm0.082}$	$0.734_{\pm0.102}$	$0.564_{\pm0.079}$	$0.589_{\pm0.118}$
EDANet [96]	$0.771_{\pm0.094}$	$0.810_{\pm0.150}$	$0.666_{\pm0.096}$	$0.682_{\pm0.137}$	$0.980_{\pm0.005}$	$0.997_{\pm0.002}$	$0.685_{\pm0.085}$	$0.723_{\pm0.102}$	$0.559_{\pm0.083}$	$0.575_{\pm0.117}$
CGNet [161]	$0.766_{\pm0.090}$	$0.798_{\pm0.149}$	$0.689_{\pm0.087}$	$0.703_{\pm0.138}$	$0.980_{\pm0.005}$	$0.997_{\pm0.002}$	$0.696_{\pm0.074}$	$0.730_{\pm0.102}$	$\mathbf{0.569_{\pm0.075}}$	$0.584_{\pm0.118}$
SCU-Net	$0.778_{\pm0.085}$	$0.789_{\pm0.137}$	$0.682_{\pm0.082}$	$0.708_{\pm0.140}$	$0.980_{\pm0.005}$	$0.997_{\pm0.002}$	$\mathbf{0.698_{\pm0.074}}$	$0.729_{\pm0.093}$	$\mathbf{0.569_{\pm0.074}}$	$0.581_{\pm0.110}$
FPENet [91]	$0.682_{\pm0.106}$	$0.730_{\pm0.173}$	$\mathbf{0.715_{\pm0.095}}$	$\mathbf{0.750_{\pm0.130}}$	$\mathbf{0.981_{\pm0.005}}$	$\mathbf{0.998_{\pm0.002}}$	$0.666_{\pm0.092}$	$0.721_{\pm0.114}$	$0.544_{\pm0.087}$	$0.575_{\pm0.129}$

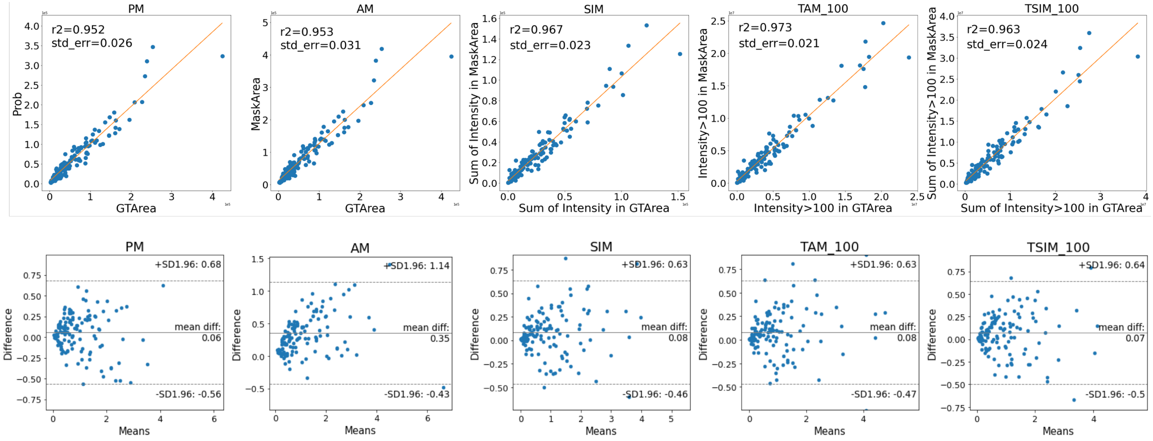


Figure 3.11: Statistical analysis on validation data for Cohort A. First row: R^2-correlation of whole-image SCU-Net calcification quantification results for predicted masks (Y-axis) as compared to the groundtruth (X-axis). All X-axis and Y-axis values are in scientific format. R^2-correlation values (r2) and standard errors (std_err) are also reported for each metric in each subfigure. Second row: Bland Altman test to compare each metric computed from SCU-Net against the groundtruth. There are 134 data elements in total for each subfigure, with each point representing one image in the validation dataset.

using SCU-Net segmentations against the same metrics computed on the ground truth segmentation. The upper row of Figure 3.11 shows the R^2-correlation of whole-image-size segmentation results of SCU-Net compared to the groundtruth based on the same metrics, demonstrating correlation >0.95 for all metrics. On the 134 validation scans, SCU-Net has the highest R^2-correlation value of 0.973 between the predicted mask and groundtruth when using the $\mathcal{TAM}_{100}$ metric, which measures the total number of pixels with intensity >100 in the segmented mask. The second row of Figure 3.11 indicates the Bland Altman test results [51] for the same validation data. The plots show the differences between quantitative metrics computed from the groundtruth and SCU-Net against the mean of the two measurements. Most metrics demonstrate very few outliers, and in particular $\mathcal{PM}$ does not have a single outlier.

Results of BAC quantification compared to breast CT: Evaluation of BAC quantification against breast CT in cohort B yielded good results. For calcification volume (voxels), R^2-correlation values were 0.834, 0.843, 0.832, 0.798, and 0.800 for the $\mathcal{PM}, \mathcal{AM}, \mathcal{SIM}, \mathcal{TAM}_{100}, \mathcal{TSIM}_{100}$ metrics, respectively. For calcium mass, R^2-correlation values were comparable at 0.866, 0.873, 0.840, 0.774, and 0.798 for the same metrics. Although breast CT is not performed clinically, this demonstrates that BAC quantification on mammography is comparable to a previously validated calcification quantification metric.

Results of BAC longitudinal analysis: Results of longitudinal analysis using the $\mathcal{TAM}_{100}$ metric showed the ability to automatically track BAC over time. Plots for five subjects shown in Figure 3.12 demonstrate a gradual increase in BAC over time. Figure 3.12 also shows five mammograms that demonstrate the progression of BAC in one subject over an 11 year period with predicted BAC masks highlighted in green.

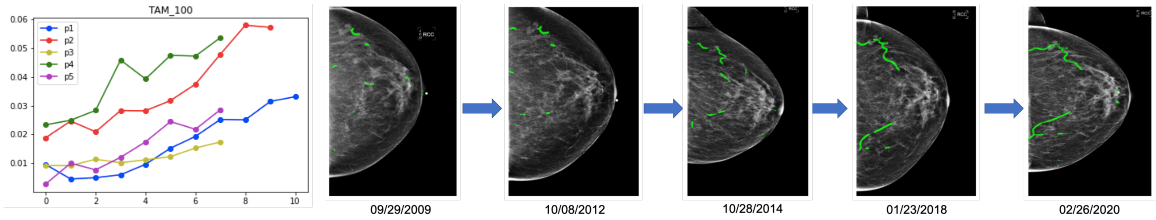


Figure 3.12: Longitudinal quantification of BAC in 5 patients. Left: The top-performing $\mathcal{TAM}_{100}$ metric applied to SCU-Net segmentations for five subjects plotted over time over time, wherein p1, p2, p3, p4, p5 represent different subjects. Right: Sampled mammograms from one subject over 11 years demonstrating an increase in detected BAC over time. BAC are highlighted in green. Each mammogram is cropped to exclude background with its exam date shown below.

3.3.4 Conclusion

We present a lightweight and accurate semantic segmentation model Simple Context U-Net (SCU-Net) designed for efficient vessel calcification segmentation on mammograms. It incorporates dilated convolution operations to learn context features and fuses multi-level features to enhance prediction accuracy. Due to the large size of mammograms, each image is processed in patches for both training and validation and the resultant masks are re-stitched to obtain whole-image predictions. Extensive experimental results for both patches and whole mammography images of 216 subjects showed comparable or better performance of SCU-Net as compared to current state-of-the-art models while maintaining far fewer training parameters. A further advantage of our model is that it does not require raw mammography data and can be applied retrospectively. This will enable analysis of the vast datasets of prior digital mammograms, allowing for large retrospective studies.

In addition to accurate segmentation of BAC, we applied quantification metrics to assess the extent of calcification and demonstrated excellent correlation between quantification values obtained on the predicted mask as compared to the groundtruth. Correlation was best using the $\mathcal{TAM}_{100}$ metric which counts all pixels above intensity 100 to differentiate between calcified and non-calcified portions of the vessel inside the mask. We also showed strong correlation of all metrics to calcium volume and mass obtained on breast CT for 10 subjects. Lastly, we were able to track and quantify the progression of BAC in 26 subjects longitudinally using this metric. Thus we believe this tool can accurately quantitatively measure and track BAC progression in patients and could be used to assess the efficacy of therapies and risk factors modification.

In summary, a robust, minimally complex, deep learning method for segmenting and quantifying breast arterial calcifications has been developed that can be applied retrospectively to routine screening mammograms. This will allow for analysis of large populations without additional imaging costs or radiation exposure. Future studies will determine the performance of this tool for predicting clinical outcomes and determining the efficacy of prevention approaches.

3.4 Discussions and future works

In this chapter, we have explored three different segmentation tasks - nuclei segmentation, liver steatosis segmentation and BAC segmentation - under limited supervision. To cope with the challenges in the process, we have utilized traditional segmentation, inaccurate annotation generation and transfer learning and supervised learning with post-processing methods. No universal standard approach could solve all the problems, thus, the strategy should be decided according to the task and label quantities and qualities. Although our segmentation methods have shown promising performance, there are some limitations and future works we can further work on.

Nuclei segmentation: Without supervised annotations, our clumped nuclei segmentation method has only experimented on a limited number of nuclei images with a certain number of patients involved. Extensive experimental results on more relevant images can be helpful for further demonstrating the method effectiveness. The other limitation of this approach is the requirement of clean background tissue and the round shape of objects. This is decided by nucleus's characteristics, and may fail to work on irregular object segmentation. In the future, it is worthwhile to utilize our segmentation model and generate initial clumped nuclei segmentation results. With the labels, we can try more advanced machine learning and deep learning techniques to handle clumped nuclei segmentation.

Liver steatosis segmentation: This work has exploited the potentials of deep learning segmentation methods. Even with inaccurate annotations, the model can eventually predict correct and accurate instance boundaries and exceed the traditional segmentation methods significantly. Nonetheless, the model can only take images with small sizes, 512×512×3 for instance and cannot deal with the original whole-slide images without pre-processing. This is caused by the original requirement of GPU memory and network architecture. In the future, it is suggested to develop a multi-scale segmentation method which can first identify the foreground tissue area from the background and then learn to process the large tissue areas.

BAC segmentation: The limitation is that the model is developed at a single institution using a single brand of scanners. It is possible that the model could underperform on external data, however we believe that the model can be successfully fine-tuned to re-optimized as needed, particularly due to its low number of parameters. The model is developed using only 661 images so fine-tuning can likely be achieved using an even smaller segmented dataset if needed. Another current limitation is that although our quantification metrics show strong correlation to breast CT data and track increases in BAC over time, they have not yet been validated against clinical outcomes in these patients. To address this in future work, we plan to evaluate our model and quantification metrics against outcomes data or existing validated risk assessment tools such as calcium scores on coronary CT.

Chapter 4

Medical OOD Identification with Limited Supervision

4.1 Medical novelty identification

With recent prominent developments of machine learning techniques in computer vision, integrating machine learning tools to solve medical image problems is becoming more and more popular due to the powerful computation and efficiency [89]. However, when deploying machine learning models in real-world applications, models trained on in-distribution (ID) data may fail to deal with out-of-distribution (OOD) inputs and assign incorrect probabilities [139]. This can severely contaminate the reliability of artificial intelligence models, especially in medical areas as the safety in clinical decisions is much more critical than other fields. For example, a classifier trained on existing bacterial classes wrongly classified a new type of bacteria as one of the classes from the training data with high confidence [119], which could be concerning for clinical usage but may be avoided by combining an OOD detection model. Thus, a successful open-world deployment with OOD detection should be sensitive to unseen classes and distribution-shifted samples and also be resilient to potential adversarial attacks [136].

However, medical OOD detection poses great challenges due to the heterogeneity and unknown data characteristics of medical data. 1) *Mutations can happen.* Different from natural objects with fixed attributes, known diseases may progress to other mutated versions and generate anomalous data; 2) *Heterogeneous data is a big concern.* Medical images collected from different race groups can introduce heterogeneity; 3) *Distribution shifting always exists.* Data scanned with different machines or institutes may have distribution shifting; 4) *Data with defects is common.* Medical images can be overexposed or scanned with incorrect positions/angles.

OOD data, also called *anomaly*, *outlier*, usually refers to data that shows dissimilarity from the training distribution. Given an image x, the goal of *OOD detection* is to identify whether x is from ID dataset D_{in} or OOD dataset D_{out}. There are two types of OOD data commonly targeted to identify - *(i) intra-class data:* OOD data belonging this type, which is also called *novelty data*, often shares severe similarity with the ID classes and is extremely challenging to distinguish, *e.g.*, the pneumonia chest X-ray presents close appearance with the normal images; *(ii) inter-class data:* this data is significantly different from ID samples, *e.g.,* a head CT image is much different in shape and color from the skin cancer image. Even though many anomaly detection methods have been proposed [131, 94], most of them focus on natural images and follow the one-vs-rest setup [145] for benchmark natural image datasets (*e.g.,* MNIST [78], Fashion-MNIST [162], CIFAR-10 [76], ImageNet [37], *etc.*). Thus, the performance reported on the benchmark datasets is actually for inter-class prediction due to the clear class variation and often trivial to detect. In contrast, the anomaly detection in medical images is more of an *intra-class* identification problem, which can be also called *novelty detection.*

To train a novelty detector with only ID data available, learning high-quality "normality" features is the fundamental step to identify the OOD samples during inference. AutoEncoder [35] architecture, as an unsupervised model to learn efficient data features through reconstruction, is the most straightforward way to extract features for ID data [128]. For anomaly detection, the reconstruction error is treated as the score of outliers based on the assumption that the AutoEncoder [35] is unable to reconstruct the anomalies well and causes large reconstruction errors. However, in the intra-class detection where the variations among the in-class and out-of-class medical images of the same category are very subtle, the AutoEncoder [35] often fails owing to the lack of discriminative ability for intra-class detection.

To enhance the discriminative ability of AutoEncoder [35], we propose **T**ransformation-based **E**mbedding learning of **N**ovelty **D**etection (**TEND**) to distinguish intra-class OOD inputs in an unsupervised fashion. Based on the vanilla AutoEncoder [35] model to learn the "normality" of ID data in the first stage and function as a feature extractor in the second stage, TEND utilizes distorted images generated by adding transformations on the ID data, and treats the data as non-ID data (marginal OOD, see Sec. 4.1.2). A binary classifier of TEND is trained with the ID data as normal class and the non-ID data as OOD class. Hence, the classifier is aware of the existence of outliers and gains certain identification ability of true outliers during inference without being trained on any true OOD data. To further separate OOD data from the ID ones, we learn a distance metric objective to encourage clustering of ID data during training and enforce a margin between OOD versus ID data in the embedding space.

4.1.1 Contribution

In summary, the main contributions of our paper are as follows:

- We propose a new novelty detection model TEND that utilizes the AutoEncoder's feature extraction and adds discrimination ability for outliers with transformations of in-distribution data and embedding distance as auxiliary. No out-of-distribution data is required for training the model.

- Although there have been a lot of anomaly detection research work done, the accurate detection performance results are lacking. We compare and report the novelty detection performance details of the unsupervised TEND model with state-of-the-art anomaly detection models and one supervised model on three public medical image datasets following two experimental settings - one-vs-rest and rest-vs-one.

- We validate our method on diverse image datasets and demonstrates our model's effectiveness. Extensive evaluations include the detection of intra-class out-of-distribution data from the original datasets and the corresponding generated with unused transformations on in-distribution data. Given the experimental observations, our model will be beneficial in discovering new anomaly cases in medical applications without any preconceived OOD training data. The corresponding code is available at `https://github.com/XiaoyuanGuo/TEND_MedicalNoveltyDetection`.

Publication:

- Guo, Xiaoyuan, Judy W. Gichoya, Saptarshi Purkayastha, and Imon Banerjee. "Margin-aware intraclass novelty identification for medical images." Journal of Medical Imaging 9, no. 1 (2022): 014004.

4.1.2 Method

TEND focuses on novelty identification for medical images. By following the one-vs-rest setup [94] and its revsered version - the rest-vs-one setup, one or more certain classes of the datasets in use are treated as normal classes. Unsupervised learning of feature embeddings for the normal classes is the fundamental step for outlier detection.

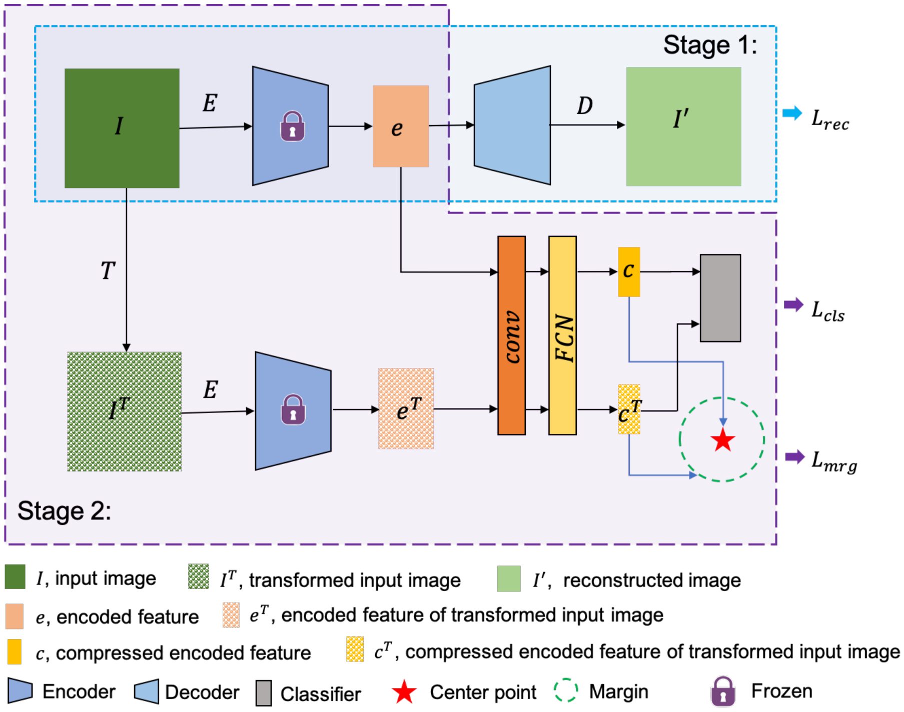


Figure 4.1: Network architecture of TEND. - Stage 1: Training AutoEncoder with in-distribution data; Stage 2: Joint training of the classifier and the margin learner.

GANs and AEs are all good options for this work. Nonetheless, GANs often require large amounts of data for training and are unstable for large images, we choose the vanilla AE [52] to encode the ID data. Moreover, as introduced in Sec. **??**, AEs are designed for compressing inputs and have no strong discriminative ability, which makes them inappropriate for medical novelty detection because of the minute intra-class variations of medical image datasets. Thus, to enhance the discriminative ability of TEND, we train a binary classifier and a margin-aware objective function (also called margin learner) jointly to separate the normal class data from the anomalies.

Architecture

Fig. 4.1 shows the network architecture of TEND, which is a two-stage novelty detector with an AutoEncoder [35] as the feature extractor backbone. In order to train the feature extractor with only ID data, the AutoEncoder [35] model (shown in the dotted blue box of Fig. 4.1) is optimized with a reconstruction loss function L_{rec}. The learnt bottleneck section will be frozen as indicated by the purple lock in Fig. 4.1 and used for encoding/extracting image features in the second stage. To train the following binary discriminator without OOD data available, we add transformations on the original images to construct distribution-shifted OOD samples based on the observation that some augmentations can be useful for OOD detection by considering them as fake OOD data [145]. The details of how to construct the transformations are explained in Sec. 4.1.2. The generated OOD data should be first fed to the trained encoder to obtain the corresponding deep features. Both of the encoded features of normal and transformed data are fed to the classifier simultaneously. With a convolutional (*conv*) layer and a fully connected layer (FCN), the classifier learns to identify the in-distribution data as normal class and the transformed images as outliers. A latent decision boundary between the two classes is optimized, the detection on true anomaly data is still not promising given the fact that the transformed images can not represent the true outliers' distribution. The decision boundary may not work for the anomalies in the feature space. To solve this problem, TEND adopts the margin-aware learning idea of DeepSVDD [128] to optimize a distance objective function simultaneously. Different from the objectives only for ID data [128], TEND works on both the ID data and the fake OOD data by enforcing the embeddings of ID data to cluster around a voted center O (see Sec. 4.1.2 for more details) whereas pushing out the embeddings of the generated abnormal class with a predefined margin R.

Transformations for generating fake OOD data

SimCLR [28] has performed an extensive study on which family of augmentations leads to a better self-supervised learning, i.e., which transformations should be considered as positives. The authors report that some of the examined augmentations (e.g., rotation), could lead to degraded performance. Based on the observation, such

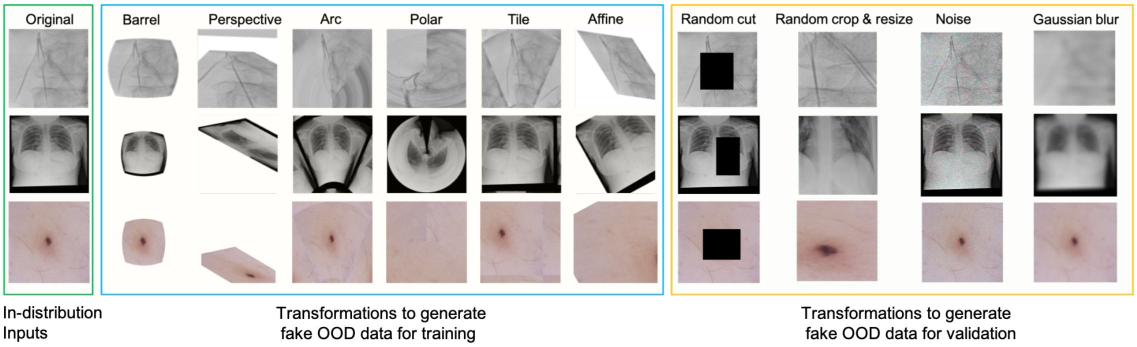


Figure 4.2: Examples of transformations used for generating fake OOD data. Three image examples from IVC-Filter (1st row), RSNA (2nd row) and ISIC2019 (3rd row) datasets are presented. The original data in the green box are inputs from in-distribution class, the transformed in-distribution images in the blue box are auxiliary data as anomalies feed to TEND's classifier during training, other possible transformations shown in the yellow box are for validation.

augmentations can be useful for OOD detection by considering them as fake OOD data. Therefore, we leverage a family of transformations and utilize more complex transformations and distortion functions that will change the visual features of the original inputs to generate fake abnormal data for training in OOD model. The generated auxiliary data are fed to the forehead of the TEND backbone and then to the classifier, which helps separate the embedding features of the ID data from those of the unknown OOD data. Different from the most common transformations, *e.g.,* rotation, used in classic data augmentation, we adopt a range of different distortions, *i.e., barrel, perspective, arc, polar, tile, affine* defined in the *Image.distort* method of *Wand* package[1]. The blue box in the middle part of Figure 4.2 shows the six different transformations on the three datasets. These transformations bring significant difference to the original inputs and generates intra-class OOD samples. We treated these extreme distortions of ID data as outliers for training. Expect for the six distortions used in this paper, there are more transformations worthwhile being explored. To further demonstrate the benefits of training the TEND model using

[1] https://docs.wand-py.org/en/0.6.5/guide/distortion.html

extreme transformations, we use moderate distortions, such as randomly cutting, randomly cropping and resizing, addition of noises, Gaussian blurring only for validation (shown in the right yellow box of Figure 4.2). The package usage and parameters selection for the six training distortions and the four validation transformation are present in our code repository.

Joint training

With an AutoEncoder [35] as the backbone, TEND incorporates a classifier and a margin-aware embedding mapping to gain discriminative ability for anomalies. In the first stage, the backbone is trained only on ID data. Suppose that the input image I and reconstructed image $I^{'}$ is with size of $M \times N$, a reconstruction objective f_{rec} defined in Eqn. 4.1 is used to optimize the learning embedding representations of the normal class. This first-stage training ensures the feature extractor to focus on learning the "normality" of in-class data.

$$f_{rec} = min \frac{1}{M}\frac{1}{N}\sum_{i=1,j=1}^{M,N} \left\| I_{ij} - I_{ij}^{'} \right\|^2 \tag{4.1}$$

$$L_{cls} = \frac{1}{S}\sum_{i=1}^{S} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{4.2}$$

With the distorted ID data as anomalies in the second stage, the binary discriminator is able to train with a final output indicating the data class. Notably, the inputs of this classifier are the encoded features extracted by the backbone. Here, the AutoEncoder model is fully frozen and only used for extracting image features. The encoded features e, e^T are processed by a following convolutional layer (conv) and a fully connected layer (FCN) of the classifier. Thus, the embeddings learnt by the encoder are mapped to a new compressed space as c, c^T with size of K (512 in our case). The classifier enables the separation of the compressed features of the ID data and the distorted data. A binary cross entropy loss function L_{cls} shown in Eqn.4.2 is utilized for optimizing, with the S to be the total number of the training data, y_i representing the ith data's binary label and $p(y_i)$ being the corresponding probability of the prediction. Nonetheless, the transformations T can only introduce limited class variations, hence the identification for real OOD data is still not ideal. Thus, a margin-aware objective is jointly trained to force the clustering of the compressed features of the ID data and the surrounding of the transformed ID data outside the margin as illustrated by Figure 4.2.

In experiments, we test three margin R values (150, 250 and 500). Similar to DeepSVDD [128], the compressed feature center O is calculated by the mean of all the ID data's compressed features. Before calculation, TEND's classifier block is trained with several warm-up epochs, (*e.g.,* 10 epochs), then the center O is defined with the same size of K as the compressed feature c. Since then, the margin learner of TEND is trained together with the discriminator. Importantly, the margin learner has different learning objectives for the normal class (g_{in}) shown in Eqn. 4.3 and the generated abnormal class (g_{out}) shown in Eqn. 4.4.

$$g_{in} = min\frac{1}{K}\sum_{i=1}^{K}\|c_i - O\|^2 \tag{4.3}$$

$$g_{out} = min\frac{1}{K}\sum_{i=1}^{K}max(R - \left\|c_i^T - O\right\|^2, 0) \tag{4.4}$$

In summary, TEND has two stage-wise losses. The first-stage loss is for the reconstruction of the AutoEncoder training, *i.e.,* $L_{1st} = L_{rec}$. The second-stage loss includes the binary classifier and the margin learner, *i.e.,* $L_{2nd} = L_{cls} + L_{mrg}$. In experiments, we use mean square error (MSE) loss for L_{rec} and binary cross entropy (BCE) loss for L_{cls}. Marginal loss L_{mrg} equals the summation of the mean of distance errors for ID data and the mean of the errors for distorted data.

Implementation details

An AutoEncoder architecture is trained as our baseline, the trained model later on is treated as the backbone of TEND. We report the encoder, decoder, Conv, FCN parts of TEND in Table 4.1. *FC* is fully connected layer, *Conv* stands for the convolutional layer , *TConv* means the transposed convolutional layer. *channel* indicates the image channel. All the *Conv* and *TConv* layer use kernel filter size 4, stride 2 and padding 1. The encoder encodes input images as e, while the *Conv* layer compresses e to c with smaller sizes. Each *Conv* and *TConv* is followed by a standard batch-normalization layer and a relu function.

Table 4.1: TEND architecture details.

Dataset	**Encoder**	**Decoder**	**Conv**	**FCN**

IVC-Filter/
RSNA/
ISIC
	Conv(channel,16)
	Conv(16,32)
	Conv(32,64)
	Conv(64,128)
	Conv(128,256)
		TConv(256,128)
		TConv(128,64)
		TConv(64,32)
		TConv(32,16)
		TConv(16,channel)
			Conv(256,512)	*FC(2048,512)*
				FC(512,1)

In our experiments, we use Adam optimizer with a learning rate of 0.001 for model training. Each network is trained with 50-150 epochs depending on the dataset size and the data complexity as datasets with more complex data or large amounts of samples often take more time to get the loss decreased to a satisfactory level. When training with the margin-aware metric, we run 10 warm-up epochs first and then calculate the embedding center O. The pipelines are developed using Pytorch 1.5.0, Python 3.0. and Cuda compilation tools V10.0.130 on a machine with 3 NVIDIA Quadro RTX 6000 with 24GB memory.

Anomaly score

As a standard evaluation procedure for anomaly detectors, the ID and outliers are mixed for computing the accuracy while different detectors have different anomaly score definitions. For the baseline AutoEncoder model, we set the reconstruction error as the OOD data score. TEND does not focus on the reconstruction, therefore, the final anomaly score of TEND is the classification probability adding the marginal distance. Giving the fact that the classification probability p is in range $[0-1]$ while the distance value d is in $[0, +\infty)$, we scale down the distance value d by dividing the predefined margin R, *i.e.,* $d^{'} = \frac{d}{R}$. Therefore, the final anomaly score for TEND is $S_i = \lambda p_i + (1-\lambda)d_i^{'}$. The value of λ is set as 0.5 in our experiments as default. To further demonstrate the effectiveness of each component of TEND, we have done ablation study of TEND and reported the results in Sec. 4.1.3. TEND without the binary classifier is called *MarginLearner* (the anomaly score is $d^{'}$).

Evaluation metrics

Having the anomaly prediction score, the detection accuracy largely depends on the threshold setting. To be fair, the detection evaluation should be threshold-invariant. Following the standard evaluation metrics used in other works [165, 83], we adopt AUROC (AUC in short) to showcase the performance difference among the models. AUROC is the Area Under the Receiver Operating Characteristic curve, which is a threshold independent metric. The AUROC can be interpreted as the probability that a positive example is assigned a higher detection score than a negative example. To find an optimal threshold for receiver operating characteristic (ROC) curve by tuning the decision thresholds, we use the Geometric Mean (G-Mean) as the metric to determine the best threshold values and report the resulted true positive rate ($TPR = \frac{TP}{TP+FN}$) and false positive rate ($FPR = \frac{FP}{FP+TN}$). The difference between the TPR and FPR given the optimal selection, $DIFF = TPR - FPR$, is also reported for model comparison. Large difference stands for better true and false positive predictions.

4.1.3 Experiments

In this section, we perform empirical evaluations of TEND on publicly available medical image datasets with varying complexity. For evaluating the accuracy in identifying novel class data, we compare our results with state-of-the-art unsupervised OOD models, starting from simple vanilla AutoEncoder (AE) [35] model and a variational AutoEncoder (VAE) [9], to DeepSVDD [128], GANomaly [4], f-AnoGAN[133] models. We also compare our unsupervised TEND model against a supervised binary classifier which was trained on both ID and OOD data for the detection task.

Datasets

In our experiments, we have three medical datasets in use, including inferior vena cava (IVC) filters on radiographs [103] and RSNA chest x-ray dataset [156], ISIC2019 [33]. IVC-filter dataset has 14 classes in total. The details are ALN (73 images), BardSimonNitinol (59 images), Optease (129 images), BardDenali (50 images), Celect (75 images), Option (196 images), BardEclipseG2X (84 images), CelectPlatinum (48 images), Trapease (100 images), BardG2 (45 images), Greenfield12Fr (122 images), Tulip (99 images), BardMeridian (55 images), GreenfieldTitanium (101 images). RSNA has 3 classes - normal, with opacity, not normal in total. ISIC2019 [33] consists of 8 classes, *i.e.,* Melanoma (MEL, 4148 images), Melanocytic nevus (NV, 11559 images), Basal cell carcinoma(BCC, 3323 images), Actinic keratosis (AK, 867 images), Benign keratosis (BKL, 2240 images), Dermatofibroma (DF, 239 images), Vascular lesion (VASC, 253 images), Squamous cell carcinoma (SCC, 628 images). The IVC-filter and ISIC2019 image are with varying sizes, with the width size ranging from 150 to 1500, height size ranging from 150 to 1500 roundly, *e.g.,* $469 \times 365 \times 3$. The RSNA dataset is in dicom format, each dicom file has the pixel array of size 1024×1024. To unify the training pipeline, we resize all the IVC-Filter, RSNA and ISIC data in $256 \times 256 \times channel$.

For the one-vs-rest setting, the in-class and rest classes data details are summarized in Table 4.2. Due to the data imbalance, we usually pick the class with the most data as our in-distribution data and all the left classes as intra-class OOD data. For IVC-filter, we select the *Option* type as the normal class; for RSNA dataset, we treat the *normal* class as ID data; for ISIC2019 dataset, we choose the NV class with the most samples as ID inputs. The total numbers of ID and OOD data for each dataset are reported in the column of **#images** in Table 4.2. Notably, the rest-vs-one setting experiments treat the classes conversely.

Table 4.2: Three publicly available dataset used in the study - total number of images in the dataset, In-distribution data (D_{in}) and out-of-distribution data (D_{out}) with one-vs-rest setting.

Dataset	**total classes**	D_{in}		D_{out}	
		class	**#images**	**class**	**#images**
IVC-Filter [103]	14	*Option*	196	*BardSimonNitinol, ALN...*	1,040
RSNA [156]	3	*normal*	8,851	*with opacity, not normal*	21,376
ISIC [33]	8	*NV*	11,559	*MEL, BCC...*	11,698

Training and evaluation settings

To train and evaluate OOD detectors' performance, we split the in-distribution data with 80% as training set D_{in}^{tr} and 20% as test set D_{in}^{te} and use all the left classes as D_{out}. For OOD detection evaluation, we mixed D_{in}^{te} and D_{out} by assigning the ID data with label 0 and OOD data with label 1. Since this paper focuses on intra-class OOD detection, we will report the OOD detection results within the same dataset instead of crossing different datasets.

Quantitative results

One-vs-rest results Following the one-vs-rest setting, Table 4.3 presents the AUC scores and the corresponding FPR, TPR values determined by the optimal thresholds for AutoEncoder [35], VAE [9], DeepSVDD [128], GANomaly[4], f-AnoGAN[133] and

TEND models with margin 150 (*i.e.*, TEND_150), 250 (*i.e.*, TEND_250) and 500 (*i.e.*, TEND_500). The difference between the TPR and FPR is also reported in the *DIFF* column in Table 4.3. ↓ means the lower the value the better the model is while ↑ stands for the higher the value the better the model performs. Thus, we expect the model to have high AUC score and prefer low FPR and high TPR values when deploying the models with the optimal threshold as decision boundary, which means the larger the difference between TPR and FPR the better. The best and second best *DIFF* and AUC results are highlighted by bold and underline respectively. Among the unsupervised anomaly detectors, our model TEND_150 attains the sub-optmial *DIFF* result 0.359 for IVC-Filter dataset and second best AUC score 0.616 for RSNA datasets; TEND_250 achieves the second highest AUC score 0.683 for IVC-Filter dataset and the highest *DIFF* 0.179 for RSNA dataset and second highest *DIFF* 0.344 for ISIC dataset2019. Meanwhile, TEND_250 reaches the second best AUC score 0.720 for ISIC2019 datasets compared to other methods with f-AnoGAN reaches the top *DIFF* 0.386 and AUC 0.740; TEND_500 reaches the highest AUC score 0.704 for IVC-Filter dataset and 0.627 for RSNA dataset and has the largest *DIFF* value 0.492 for IVC-Filter dataset and the second largest *DIFF* value 0.172 for RSNA dataset and 0.269 for ISIC2019. GANomaly performs better than DeepSVDD on IVC-Filter and RSNA datasets with higher *DIFF* and AUC values, while DeepSVDD exceeds GANomaly on ISIC2019 dataset. Observing the results on IVC-Filter, RSNA and ISIC2019 datasets, the performance of f-AnoGAN gradually improves as the training dataset becomes larger. Nevertheless, our model TENDs show advantages in acquiring better accuracy and exhibits competitive performances compared with other unsupervised models. Notably, we implement TEND with three different margins to show the difference with changing settings. By observing our results in Table 4.3, no unique margin in TEND provide the optimal result on all the datasets and thus it needs to be tuned for specific experiments. The effects of applying different radius are present in Sec. 4.1.3. The MarginLearner and the supervised model BinaryClassifier are also discussed in ablation study (see Sec. 4.1.3).

Table 4.3: FPR, TPR values, difference of TPR and FPR values, and AUC scores of various OOD detection methods trained on IVC-Filter [103], RSNA [156] and ISIC2019 [33] datasets **with the one-vs-rest setting**. Bold numbers are the best results and underlined numbers are the second best. Models with * are supervised and those without * are unsupervised.

Methods	IVC-filter				RSNA				ISIC2019			
	↓*FPR*	↑*TPR*	↑*DIFF*	↑*AUC*	↓*FPR*	↑*TPR*	↑*DIFF*	↑*AUC*	↓*FPR*	↑*TPR*	↑*DIFF*	↑*AUC*
AutoEncoder [35]	0.198 ± 0.104	0.350 ± 0.075	0.152 ± 0.067	0.436 ± 0.040	0.318 ± 0.014	0.461 ± 0.009	0.143 ± 0.010	0.566 ± 0.004	0.833 ± 0.060	0.186 ± 0.059	−0.648 ± 0.025	0.096 ± 0.003
AE_GMM	0.224 ± 0.138	0.153 ± 0.008	−0.071 ± 0.134	0.464 ± 0.067	0.496 ± 0.012	0.321 ± 0.003	−0.175 ± 0.013	0.412 ± 0.006	0.083 ± 0.006	0.211 ± 0.003	0.128 ± 0.006	0.564 ± 0.003
VAE [9]	0.489 ± 0.097	0.707 ± 0.076	0.218 ± 0.117	0.542 ± 0.080	0.473 ± 0.001	0.462 ± 0.001	−0.011 ± 0.012	0.487 ± 0.001	0.351 ± 0.011	0.395 ± 0.007	0.045 ± 0.007	0.471 ± 0.005
MarginLearner	0.426 ± 0.099	0.549 ± 0.033	0.123 ± 0.098	0.568 ± 0.055	0.475 ± 0.016	0.478 ± 0.013	0.003 ± 0.010	0.491 ± 0.005	0.517 ± 0.020	0.584 ± 0.024	0.067 ± 0.010	0.530 ± 0.005
DeepSVDD [128]	0.503 ± 0.106	0.672 ± 0.042	0.170 ± 0.130	0.500 ± 0.075	0.508 ± 0.021	0.413 ± 0.023	−0.095 ± 0.015	0.421 ± 0.009	0.348 ± 0.021	0.621 ± 0.021	0.273±0.006	0.677 ± 0.003
GANomaly[4]	0.446 ± 0.172	0.627 ± 0.227	0.181 ± 0.200	0.518 ± 0.103	0.524 ± 0.005	0.678 ± 0.015	0.154 ± 0.009	0.576 ± 0.005	0.396 ± 0.030	0.481 ± 0.027	0.086 ± 0.012	0.551 ± 0.009
f-AnoGAN[133]	0.419 ± 0.077	0.511 ± 0.070	0.092 ± 0.045	0.544 ± 0.022	0.365 ± 0.033	0.541 ± 0.029	0.176±0.008	0.614 ± 0.005	0.366 ± 0.007	0.600 ± 0.007	0.234 ± 0.005	0.647 ± 0.003
TEND_150 (ours)	0.219 ± 0.077	0.749 ± 0.086	**0.531±0.071**	**0.772±0.030**	0.425 ± 0.029	0.590 ± 0.026	0.165 ± 0.010	0.615±0.006	0.377 ± 0.016	0.596 ± 0.015	0.220 ± 0.009	0.650 ± 0.006
TEND_250 (ours)	0.160 ± 0.091	0.684 ± 0.035	0.524±0.082	0.752 ± 0.051	0.389 ± 0.045	0.561 ± 0.043	0.172 ± 0.009	0.615±0.006	0.326 ± 0.017	0.669 ± 0.020	**0.343±0.011**	**0.717±0.006**
TEND_500 (ours)	0.122 ± 0.099	0.639 ± 0.095	0.517±0.042	0.760±0.028	0.438 ± 0.040	0.616 ± 0.041	**0.178±0.008**	**0.627±0.005**	0.351 ± 0.012	0.618 ± 0.011	0.268 ± 0.009	0.678±0.006
BinaryClassifier*	0.280 ± 0.006	0.847 ± 0.003	**0.567±0.006**	**0.853±0.003**	0.417 ± 0.007	0.589 ± 0.006	0.172 ± 0.008	0.593 ± 0.003	0.497 ± 0.023	0.340 ± 0.015	−0.157 ± 0.010	0.363 ± 0.004

Rest-vs-one results To further compare the models' performances, the complementary experimental setting - rest-vs-one is implemented with the results reported in Table. 4.4. Same as the one-vs-rest experiments, we keep the tested models consistent, and change the in-distribution class as OOD classes and the previous OOD data as our in-distribution data. The training and testing processes are the same as reported in Sec. 4.1.3. Our model TEND_150 gets the best *DIFF* 0.298 and AUC score 0.658 for IVC-Filter dataset, and obtains the sub-optimal AUC score 0.584 for RSNA dataset. GANomaly performs the best for RSNA dataset. TEND_250 reaches the sub-optimal results for ISIC2019 dataset whereas f-AnoGAN can achieve the best. Generally the detection of anomalies under rest-vs-one setting is more challenging than the one-vs-rest setting and nearly no model can work well for all the situations. Still, TEND has satisfactory performances across the three datasets with the rest-vs-one setting.

Ablation studies

To further explore the effectiveness of each module in TEND, we perform the ablation studies with the settings of removing the binary classifier from TEND (Margin-

Table 4.4: FPR, TPR values, difference of TPR and FPR values, and AUC scores of various OOD detection methods trained on IVC-Filter [103], RSNA [156] and ISIC2019 [33] datasets **with the rest-vs-one setting**. Bold numbers are the best results and underlined numbers are the second best. Models with * are supervised and those without * are unsupervised.

Methods	IVC-filter				RSNA				ISIC2019			
	↓*FPR*	↑*TPR*	↑*DIFF*	↑*AUC*	↓*FPR*	↑*TPR*	↑*DIFF*	↑*AUC*	↓*FPR*	↑*TPR*	↑*DIFF*	↑*AUC*
AutoEncoder [35]	0.706 ± 0.163	0.312 ± 0.159	−0.394 ± 0.059	0.165 ± 0.027	0.760 ± 0.022	0.544 ± 0.046	−0.216 ± 0.024	0.321 ± 0.005	0.593 ± 0.023	0.383 ± 0.024	−0.210 ± 0.012	0.353 ± 0.007
AE_GMM	0.728 ± 0.053	0.748 ± 0.022	0.020 ± 0.058	0.510 ± 0.029	0.600 ± 0.008	0.584 ± 0.004	−0.016 ± 0.009	0.492 ± 0.005	0.159 ± 0.005	0.059 ± 0.002	−0.100 ± 0.006	0.450 ± 0.003
VAE [9]	0.359 ± 0.082	0.464 ± 0.088	0.105 ± 0.063	0.560 ± 0.035	0.518 ± 0.029	0.453 ± 0.027	−0.065 ± 0.007	0.461 ± 0.036	0.518 ± 0.032	0.658 ± 0.045	0.140 ± 0.016	0.575 ± 0.005
MarginLearner	0.617 ± 0.022	0.619 ± 0.045	0.003 ± 0.043	0.484 ± 0.025	0.510 ± 0.018	0.527 ± 0.016	0.017 ± 0.006	0.514 ± 0.004	0.510 ± 0.018	0.527 ± 0.016	0.017 ± 0.006	0.514 ± 0.004
DeepSVDD [128]	0.514 ± 0.043	0.475 ± 0.045	−0.039 ± 0.065	0.439 ± 0.043	0.514 ± 0.028	0.552 ± 0.032	0.038 ± 0.007	0.522 ± 0.004	0.530 ± 0.007	0.540 ± 0.013	0.010 ± 0.011	0.487 ± 0.007
GANomaly[4]	0.595 ± 0.060	0.622 ± 0.040	0.027 ± 0.051	0.449 ± 0.036	0.396 ± 0.014	0.638 ± 0.014	**0.242±0.004**	**0.656±0.003**	0.462 ± 0.0166	0.583 ± 0.019	0.121 ± 0.009	0.570 ± 0.005
f-AnoGAN[133]	0.419 ± 0.077	0.511 ± 0.070	0.092 ± 0.045	0.544 ± 0.022	0.295 ± 0.029	0.276 ± 0.012	−0.019 ± 0.019	0.406 ± 0.005	0.276 ± 0.004	0.677 ± 0.006	**0.401±0.007**	**0.718±0.004**
TEND_150 (ours)	0.359 ± 0.057	0.640 ± 0.031	**0.291±0.051**	**0.650±0.028**	0.452 ± 0.022	0.578 ± 0.024	0.126±0.007	0.584±0.003	0.336 ± 0.015	0.501 ± 0.006	0.164 ± 0.014	0.608 ± 0.007
TEND_250 (ours)	0.427 ± 0.061	0.582 ± 0.071	0.155 ± 0.058	0.573±0.039	0.492 ± 0.016	0.577 ± 0.015	0.084 ± 0.006	0.549 ± 0.004	0.386 ± 0.014	0.623 ± 0.011	0.237±0.011	0.637±0.008
TEND_500 (ours)	0.428 ± 0.069	0.584 ± 0.081	0.156±0.038	0.573±0.025	0.487 ± 0.015	0.550 ± 0.014	0.063 ± 0.008	0.541 ± 0.005	0.412 ± 0.018	0.533 ± 0.016	0.121 ± 0.013	0.582 ± 0.009
BinaryClassifier*	0.617 ± 0.022	0.619 ± 0.045	0.003 ± 0.043	0.484 ± 0.025	0.510 ± 0.018	0.527 ± 0.016	0.017 ± 0.006	0.514 ± 0.004	0.471 ± 0.014	0.599 ± 0.017	0.128 ± 0.005	0.584 ± 0.004

Learner) and training a supervised binary classifier (BinaryClassifier) respectively. For the one-vs-rest setting, the results are showed as *MarginLearner* with radius setting 150 in Table 4.3, with slight *DIFF* and AUC improvements compared to the baseline AutoEncoder on IVC-Filter and ISIC2019 datasets. Comparatively, TEND_150 enlarges the *DIFF* with 0.236, 0.037 and 0.147 improvements, and increases the AUC scores by 0.086, 0.051, 0.117 respectively on IVC-Filter, RSNA and ISIC2019 datasets. For the rest-vs-one setting, compared with the *MarginLearner*, TEND_150 achieves the *DIFF* with 0.25, 0.261, 0.074 improvements for IVC-Filter, RSNA and ISIC2019 dataset respectively; and enhances the AUC score with 0.156, 0.189, 0.065 for the three datasets. These observations indicate the effectiveness of TEND's architecture.

We also report the performance of an AE extension, AE_GMM, which clusters the embeddings from the AutoEncoder backbone and predicts the data classes - ID or OOD. From both Table 4.3 and Table 4.4, a GMM head can improve the discriminative ability of AutoEncoder to certain extent, however, when testing on transformed OOD data in Table4.5 and Table4.6, the advantages fail to remain. In comparison, TEND's heads on AE have more generalization ability and demonstrate consistent detection performance.

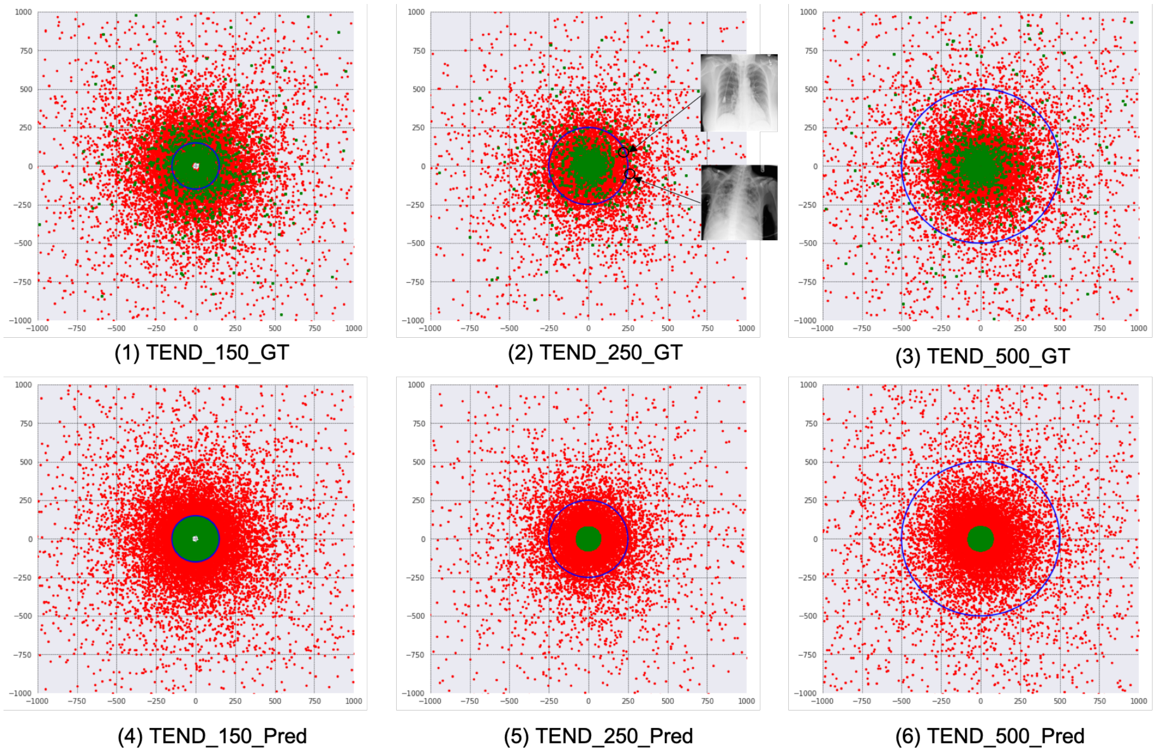
Instead of training the binary classifier of TEND model in an unsupervised fashion, we include partial true OOD data in training data. Since IVC-Filter and ISIC2019 datasets have multiple classes, we randomly select 2-3 OOD classes for training and the left classes for validation.

One-vs-rest setting: For RSNA datasets, we use the class *not normal* (see Table 4.2 for details) for known OOD data and test the model on the left *with opacity* data. The supervised *BinaryClassifer* is also evaluated with quantitative results appended in the end of Table 4.3. With prior knowledge about OOD data, the BinaryClassifer can achieve very high AUC scores, especially for IVC-Filter (+0.149 compared to the best of unsupervised results) and RSNA (0.178 compared to the best of unsupervised results) datasets. Similarly, BinaryClassifier has the largest *DIFF* values on IVC-Filter and RSNA datasets. Nonetheless, this advantage fails to remain on ISIC2019 dataset, which indicates the benefits from prior knowledge are limited.

Rest-vs-one setting: For RSNA datasets, we use the class *normal* as known OOD data and *not normal* as ID data, the left class is used for evaluation. Different from the observation above, the corresponding results in Table 4.4 for BinaryClassifier fail to exceed the unsupervised models, more results can be observed in Table 4.6. In conclusion, the supervised BinaryClassifier may lack generalization ability when dealing with unexpected data. Please refer Sec. 4.1.3 for more experimental results and discussions.

Qualitative results

As our model TEND has a margin learner module (see the L_{mrg} part of Figure 4.1) to enforce ID data inside of a predefined margin R (illustrated as the green dotted circle in Figure 4.1) as to the voted center O (represented as the red star in Figure 4.1) and OOD data outside of the region, we hereby visualize the data samples based on the obtained distance output by the *MarginLearner*. Take one-vs-rest setup results for illustration, the voted center O, whose calculation details were introduced in Sec. 4.1.2, is located at the origin of the 2D coordinate system. To visualize each data sample, we utilize their distance to the voted center O as their corresponding radius values to the origin. Each sample is represented by randomly picking one point along the circle that is defined with its corresponding radius. The x-axis and y-axis values help indicate how far the point is from the origin. Given an example with a distance value d_i, its corresponding coordinate (x_i, y_i) satisfies that $d_i^2 = x_i^2 + y_i^2$. The data samples with in-distribution labels are marked in green and the left data with OOD labels are in red. We draw the defined margin of the model with a blue circle for reference. Please refer to the Appendix code snippet for the visualization implementation details. Take RSNA dataset for example, in Figure 4.3, the voted center O is represented by the point with coordinates (0, 0) and the area defined by radius R is present with the plotted blue circles in each subfigure. For better visualization and comparison, each subfigure has both the x-axis and y-axis ranging from -1000 to 1000, those data points that have larger distance out of range will be ignored. The first row shows the distance distribution of data with ground-truth labels (*i.e., ID (in green) and OOD (in red)*) learnt by TEND with radius of 150 (1st column), 250 (2nd column) and 500 (3rd column), while the second row indicates the predictions after thresholding, with the green points for samples predicted as ID and red points for samples predicted as OOD. To help inspect the data points around the boundary, two cases based on the ground-truth information are illustrated for TEND_250_GT, with the upper one as an ID data and the lower case for OOD class. From the first row, the learnt distance distributions for ID and OOD data are similar for TEND with different radius values. But the ID data can be outside the circle with radius 150 (subfigure (1)) but will be inside the circle regions with radius 250 (subfigure (2)) and 500 (subfigure (3)) of Figure 4.3, which suggests that when using larger margin to divide ID and OOD data, ID samples will be easier to be included while more OOD data will be inside the region, leading to more false positive predictions. Therefore, it is not the larger the margin, the better the performance. After having the distance values predicted by the margin learner module, we apply the Gmeans method to find the optimal threshold considering both the distance predictions and the binary possibility. The second row illustrates the ID and OOD predictions of TEND after thresholding. We can see that the boundary of predicted ID data samples is very close to the margin circle of radius 150 (subfigure (4)), but much smaller compared to radius 250 (subfigure (5)) and 500 (subfigure (6)). As they are in the same scale, we can observe that the thresholding areas for ID are smaller when the margin values increase.



(1) TEND_150_GT

(2) TEND_250_GT

(3) TEND_500_GT

(4) TEND_150_Pred

(5) TEND_250_Pred

(6) TEND_500_Pred

Figure 4.3: 2D visualization of ID (green points) and OOD (red points) data distance distributions for RSNA dataset learnt by TEND's margin learner module with radius 150 (1st column), 250 (2nd column) and 500 (3rd column) **under the one-vs-rest setting**. The first row is for distance distribution with ground-truth labels; the second row shows the predicted results with the optimal threshold values. Blue circles are the plotted based on the radius in each subfigure for reference.

To further analyze the OOD detection ability of TEND, we take RSNA dataset for example and inspect part of the predications. As shown in Figure 4.4, four kinds of predictions, namely true positive, true negative, false positive and false negative

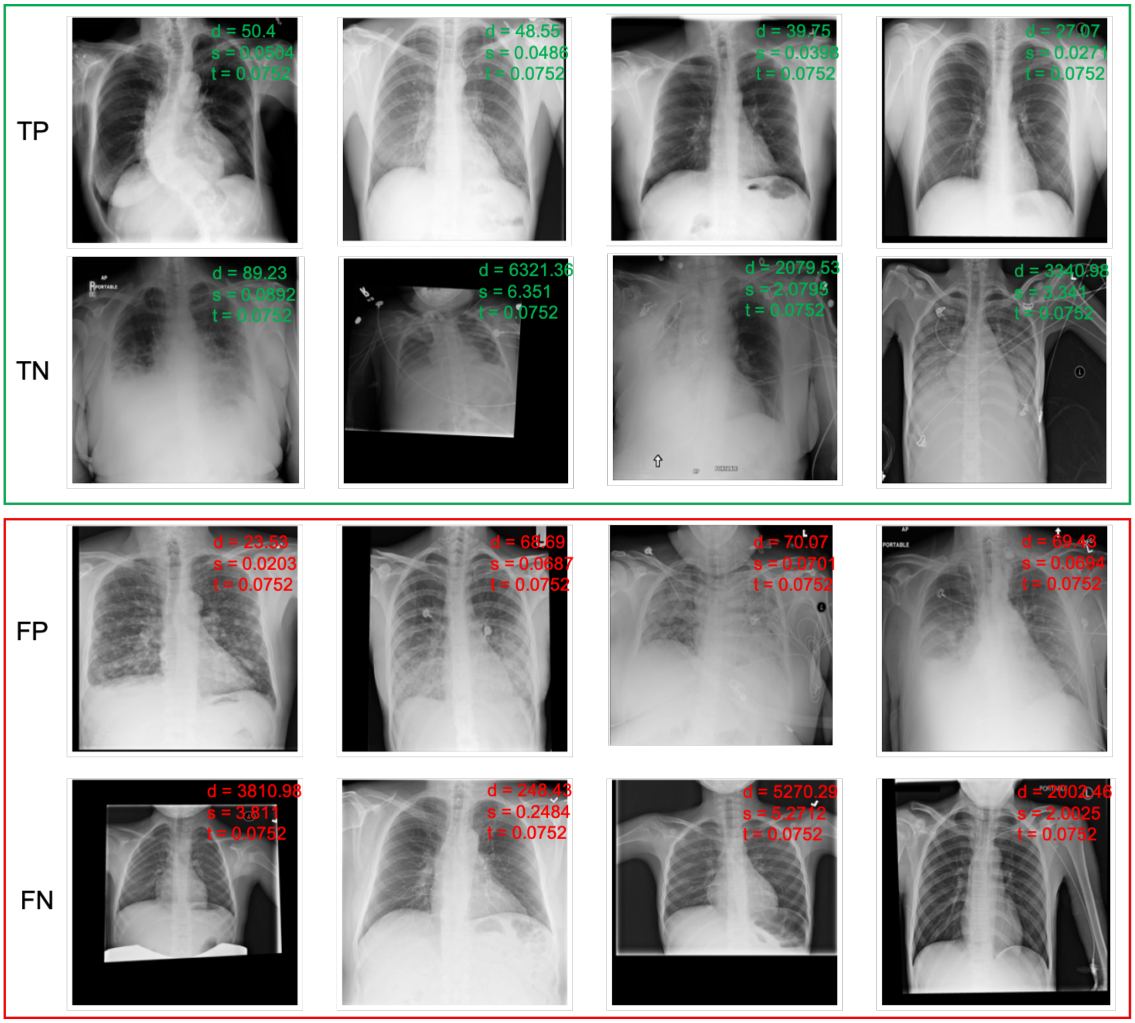


Figure 4.4: True Positive (TP, 1st row), True Negative (TN, 2nd row), False Positive (FP, 3rd row), and False Negative (FN 4th row) predictions of TEND_500 on RSNA datasets **following the one-vs-rest setting**. d: distance value from the margin learner module, p: probability outputted by the binary discriminator module, s: final score, t: optimal threshold (ID: $s < t$, OOD: $s >= t$).

predictions, predicted by TEND_500 are present, with four representative cases for each situation. TP means that true ID samples are correctly identified and TN is for correct identification of OOD samples. FP refers to the OOD data is mis-classified as ID data and FN stands for wrongly classified OOD data. From Figure 4.3, data points close to the center are more confident of being ID category, which means the smaller the distance, the lager the possibility the data being an ID sample. Observed the TP cases in Figure 4.4, most of them are with distance values less than 50, which is relatively small compared to the pre-defined margin 500; while the TN cases are often with larger distances. The first chest X-ray image of TN cases has final score 0.0892, close to the threshold 0.0752, which indicates this case is a challenging case. The third row of Figure 4.4 are the hard FP cases for TEND_500 to identify as they are all with both small distance values and probabilities. The FN cases shown in the fourth row of Figure 4.4 can be those ID data with irregular format or position shifting. With imperfections, TEND_500 will treat them as outliers and assign larger distance values by the margin learner module. Compared with others, the second FP case is much more challenging as the data is inside the pre-defined margin but classified wrongly due to the threshold setting. We also present the 2D distance visualization and detection results with examples for IVC-Filter and ISIC2019 datasets in the supplementary material.

Effects of transformations

To further compare the intra-class OOD detection ability, we generate validation data by applying four unseen transformations to all the ID data defined in Sec. 4.1.2 and showed in the right yellow box in Figure 4.2. As we have two experimental settings - the one-vs-rest and the rest-vs-one, we report them in Table 4.5 and Table 4.6 respectively. The best and the second best accuracy results are bolded and underlined respectively. As all the validation data are in OOD category, we calculate the OOD detection accuracy based on the optimal threshold t determined in Table 4.3 (corresponding to Table 4.5) and Table 4.4 (corresponding to Table 4.6) for each model and each dataset. Those data with score $s >= t$ are labeled as OOD (which are true negative samples, TN in short) and the data having score $s < t$ are classified as ID class (which are false positive samples, FP in short). Accordingly, the detection accuracy is formulated as $ACC_{val} = TN/(TN + FP)$.

Table 4.5: Accuracy of various OOD detection methods trained on IVC-Filter [103], RSNA [156] and ISIC2019 [33] **with the one-vs-rest setting**. Bold denotes the best results and underline shows the second best results. * indicates the model is supervised.

Methods	IVC-filter				RSNA				ISIC2019			
	Random Cut	*Random Crop&Resize*	*Noise*	*Gaussian Blur*	*Random Cut*	*Random Crop&Resize*	*Noise*	*Gaussian Blur*	*Random Cut*	*Random Crop&Resize*	*Noise*	*Gaussian Blur*
AutoEncoder [35]	**1.000±0.000**	0.371 ± 0.036	0.988 ± 0.007	0.064 ± 0.009	0.001 ± 0.000	0.029 ± 0.002	0.422 ± 0.004	0.000 ± 0.000	0.252 ± 0.004	0.581 ± 0.005	0.428 ± 0.004	0.187 ± 0.002
AE_GMM	0.110 ± 0.001	0.151 ± 0.000	0.142 ± 0.001	0.142 ± 0.001	0.660 ± 0.003	0.023 ± 0.001	0.577 ± 0.007	0.402 ± 0.007	0.055 ± 0.001	0.028 ± 0.001	0.086 ± 0.002	0.087 ± 0.002
VAE [9]	0.013 ± 0.006	0.137 ± 0.031	0.020 ± 0.013	0.008 ± 0.007	0.990 ± 0.001	0.288 ± 0.004	0.438 ± 0.005	0.424 ± 0.005	0.027 ± 0.001	0.241 ± 0.004	0.434 ± 0.003	0.364 ± 0.004
DeepSVDD [128]	1.000 ± 0.000	0.735 ± 0.039	0.607 ± 0.024	0.044 ± 0.018	0.604 ± 0.003	0.120 ± 0.005	0.642 ± 0.006	0.455 ± 0.005	0.985 ± 0.001	0.740 ± 0.003	0.567 ± 0.003	0.190 ± 0.004
GANomaly[4]	1.000 ± 0.000	0.792 ± 0.017	0.727 ± 0.030	0.690 ± 0.031	0.959 ± 0.003	0.910 ± 0.003	0.330 ± 0.005	0.313 ± 0.005	0.919 ± 0.003	0.608 ± 0.005	0.306 ± 0.002	0.348 ± 0.003
f-AnoGAN[133]	0.888 ± 0.024	0.699 ± 0.034	0.583 ± 0.035	0.501 ± 0.052	0.726 ± 0.005	0.729 ± 0.007	0.386 ± 0.003	0.413 ± 0.005	0.665 ± 0.007	0.431 ± 0.004	0.410 ± 0.005	0.391 ± 0.004
TEND_150 (**ours**)	0.951 ± 0.007	0.988 ± 0.006	0.921 ± 0.017	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**0.997±0.000**	**0.997±0.000**
TEND_250 (**ours**)	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.996 ± 0.001	0.942 ± 0.003	0.799 ± 0.005	0.741 ± 0.005
TEND_500 (**ours**)	0.752 ± 0.026	0.861 ± 0.026	0.797 ± 0.029	0.984 ± 0.008	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.950 ± 0.002	0.976 ± 0.001	0.905 ± 0.002	0.905 ± 0.003
BinaryClassifier*	0.963 ± 0.003	0.963 ± 0.005	0.509 ± 0.001	0.899 ± 0.001	0.499 ± 0.006	0.680 ± 0.003	0.281 ± 0.004	0.215 ± 0.004	0.271 ± 0.006	0.762 ± 0.004	0.498 ± 0.005	0.491 ± 0.006

One-vs-rest results of transformations Table 4.5 shows the accuracy of detecting the generated validation OOD data with different models with the one-vs-rest experimental setting. Among all the models present in Table 4.5, the AutoEncoder [35], VAE [9], DeepSVDD [128], GANomaly[4], f-AnoGAN[133] and our TENDs are all unsupervised methods, while the BinaryClassifier marked with an asterisk is a supervised model that is trained with both ID data and partial true OOD data. Compared with RSNA and ISIC2019 datasets, IVC-Filter objects are often with more background, thus the advantage of AutoEncoder, which is good at reconstruction, is eliminated for datasets like IVC-Filter but still remains for RSNA and ISIC2019 in detecting anomaly data with randomly cutting and noises. However, the randomly cropping and resizing and Gaussian blur transformations are much difficult for AutoEncoder to handle. VAE achieves the second best accuracy both for randomly cut RSNA data and blurred ISIC2019 data. DeepSVDD [128] and GANomaly[4] generally perform well in detecting randomly cut data for all the datasets, but TEND architectures with different margins nearly achieve all the best and the second best accuracy for IVC-Filter and RSNA datasets. TENDs still remain the similar performance for the randomly cutting and blurring of ISIC2019 datasets. In general, Gaussian blurring is the most difficult scenario of ISIC2019 dataset for all the models and adding noise to images is easy for AutoEncoder but potentially difficult for other models to deal with, including TENDs and the supervised model BinaryClassifier. In summary, although TEND is an unsupervised model, it can still obtain stronger intra-class OOD identification ability and even outperform other state-of-the-art models and the supervised model BinaryClassifer on both IVC-Filter and RSNA datasets. This advantage is due to the benefits of transformations during training.

Rest-vs-one results of transformations Table 6 presents the accuracy of detecting the generated validation OOD data with different models following the rest-vs-one experimental setting. AutoEncoder partially remains its sensitivity in random cut and noise transformations for both IVC-Filter and RSNA datasets. In general, VAE shows little advantages in transformed OOD detection except for the noise and gaussian blur OOD detection for ISIC2019 dataset. DeepSVDD, GANomaly, f-AnoGAN occasionally show advanced performance for different situations. Comparatively, TENDs show more stable results in accurate detection of the transformed OOD data, especially for both IVC-Filter and RSNA datasets. This stability for such intra-class OOD detection benefits from the learning process of training with transformation.

Table 4.6: Accuracy of various OOD detection methods trained on IVC-Filter [103], RSNA [156] and ISIC2019 [33] with the rest-vs-one setting. Bold denotes the best results and underline shows the second best results. * indicates the model is supervised.

Methods	IVC-filter				RSNA				ISIC2019			
	Random Cut	*Random Crop&Resize*	*Noise*	*Gaussian Blur*	*Random Cut*	*Random Crop&Resize*	*Noise*	*Gaussian Blur*	*Random Cut*	*Random Crop&Resize*	*Noise*	*Gaussian Blur*
AutoEncoder [35]	**1.000±0.000**	0.116 ± 0.009	0.627 ± 0.014	0.032 ± 0.005	0.999 ± 0.003	0.705 ± 0.004	0.901 ± 0.002	0.001 ± 0.000	0.782 ± 0.004	0.250 ± 0.005	0.388 ± 0.004	0.368 ± 0.004
AE_GMM	0.131 ± 0.010	0.206 ± 0.011	0.212 ± 0.010	0.220 ± 0.010	0.361 ± 0.003	0.319 ± 0.004	0.383 ± 0.005	0.396 ± 0.005	0.067 ± 0.002	0.054 ± 0.002	0.158 ± 0.003	0.157 ± 0.003
VAE [9]	0.036 ± 0.002	0.460 ± 0.008	0.476 ± 0.011	0.487 ± 0.010	0.188 ± 0.002	0.849 ± 0.003	0.603 ± 0.005	0.596 ± 0.004	0.174 ± 0.003	0.627 ± 0.006	0.555 ± 0.007	0.544 ± 0.007
DeepSVDD [128]	0.858 ± 0.011	0.529 ± 0.006	0.495 ± 0.008	0.496 ± 0.008	0.905 ± 0.001	0.415 ± 0.004	0.494 ± 0.004	0.425 ± 0.003	0.827 ± 0.003	0.294 ± 0.004	0.524 ± 0.005	0.541 ± 0.005
GANomaly[4]	0.785 ± 0.008	0.583 ± 0.009	0.577 ± 0.013	0.629 ± 0.009	0.999 ± 0.000	0.682 ± 0.003	0.792 ± 0.003	0.238 ± 0.005	0.979 ± 0.001	0.694 ± 0.003	0.464 ± 0.004	0.476 ± 0.004
f-AnoGAN[133]	0.934 ± 0.008	0.594 ± 0.013	0.361 ± 0.014	0.344 ± 0.012	0.380 ± 0.004	0.373 ± 0.004	0.716 ± 0.003	0.300 ± 0.004	0.989 ± 0.001	0.825 ± 0.002	0.460 ± 0.005	0.464 ± 0.006
TEND_150 (**ours**)	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
TEND_250 (**ours**)	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.995 ± 0.001	**1.000±0.000**	0.998 ± 0.000	0.997 ± 0.001
TEND_500 (**ours**)	**1.000±0.000**	0.997 ± 0.002	0.999 ± 0.001	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.984 ± 0.001	0.941 ± 0.002	0.902 ± 0.004	0.760 ± 0.002
BinaryClassifier*	0.025 ± 0.005	0.796 ± 0.010	0.659 ± 0.009	0.644 ± 0.012	0.927 ± 0.002	0.972 ± 0.001	0.984 ± 0.001	0.816 ± 0.003	0.100 ± 0.003	0.849 ± 0.003	0.470 ± 0.003	0.477 ± 0.003

4.1.4 Conclusion

In this paper, we introduced an unsupervised novelty detector - TEND, which can detect intra-class OOD data for medical applications in an open-world environment. TEND is a two-stage anomaly detector with a vanilla AutoEncoder trained on in-distribution data in the first stage to serve as feature extractors in the second stage and two modules - a margin learner module and a binary discriminator module - jointly trained in the second stage for separating in-distribution inputs from the non-linearly transformed counterparts. With no OOD data used in training, TEND is able to learn nuances from intra-class variations in medical image analysis problem and provide a stepping stone for developing rare disease diagnosis model with no sample images. Extensive results with the one-vs-rest and rest-vs-one experimental settings on multiple public medical image datasets demonstrate the effectiveness of our model. More general evaluations on data with unseen transformations further evince our model's generalization ability and robustness. In summary, an efficient novelty detection method for medical images has been developed that can be applied to discover unknown classes with only predefined normal data. We plan to extend this work by integrating TEND into real time imaging pipelines for inference of medical imaging models.

4.2 Generic medical anomaly detection

Despite recent advances in deep learning that have contributed to solving various complex real-world problems [36], the safety and reliability of AI technologies remain a big concern in medical applications. Deep learning models for medical tasks are often trained with data from known distributions, and fail to identify out-of-distribution (OOD) inputs and possibly assign high probabilities to the anomalies during inference

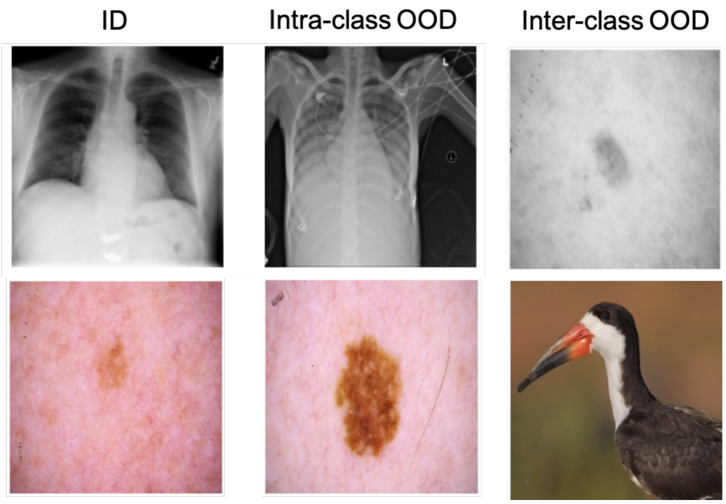


Figure 4.5: ID, Intra- and Inter-class OOD examples for medical images. Compared to natural images, medical OOD samples exhibit more subtle intra-class variations (e.g., normal vs pneumonia in the 1st row and benign vs malignant in the 2nd row).

because of the insensitivity to distribution shifting. Medical anomalies, *a.k.a., OOD data, outliers*, can arise due to various reasons such as noise during data acquisition, changes in disease prevalence and incidence (*e.g.*, the evolution of rare cancer types), or inappropriate inputs (*e.g.*, different modalities unseen during training) [44]. To ensure the reliability of deep models' predictions, it is necessary to identify unknown types of data that are different from the training data distribution. A good anomaly detector should be able to capture the variations between the in-distribution (ID) data used in training and the OOD data from open word and thus identify the outliers. However, the core challenges for medical anomaly detection are – (1) the OOD data is usually unavailable at the time of model training; (2) in theory, there are infinite numbers of variations of OOD data; and (3) different types of OOD data can be identified with varying difficulties. In general, the OOD classifications [22] can be refined based on the variation difference by summarizing them as **inter-class** OOD data and **intra-class** OOD data. Inter-class OOD data has larger variations from the ID data, whereas the intra-class OOD data is close to ID data, as observed in Figure 4.5. Thus, identifying intra-class OOD data is more difficult than the inter-class OOD data given subtle differences with ID data. To cope with the OOD unavailability and uncertainty challenges, we adopt an unsupervised way to design our anomaly detector. To acquire high identification of hard OOD cases, we expect our model can learn both coarser and finer features to screen the various dissimilar inputs. Inspired by [77, 14], we propose a generative anomaly detector – **C**ascade **V**ariational autoencoder based **A**nomaly **D**etector (CVAD), which is built on top of a branch-cascaded VAE – pchVAE [179]. With the cascade VAE architecture to model the in-distribution representations, CAVD gains superior reconstructions and learns good-quality features to threshold out the OOD data. The ability of CVAD to detect anomalies is further enhanced through training a binary discriminator with the reconstructed data with random perturbations on aforementioned cascade VAE's latent parameters as OOD category.

4.2.1 Contribution

In this paper, our contributions are three-fold:

- We propose a novel OOD detector – CVAD. By utilizing a cascade VAE to learn latent variables of in-distribution data, CVAD owns good reconstruction ability of in-distribution inputs and obtains discriminative ability for OOD data based on the reconstruction error.

- We adopt a binary discriminator to further separate the in-distribution data from the OOD data by taking the reconstructed image as fake OOD samples. We add minor random disturbance in VAE latent parameters during fake data generation to enrich data variations. Thus, our model has better discriminative capability for the inter-class as well as intra-class OOD cases.

- We conduct extensive experiments on multiple public medical image datasets to demonstrate the generalization ability of our proposed model. We evaluate

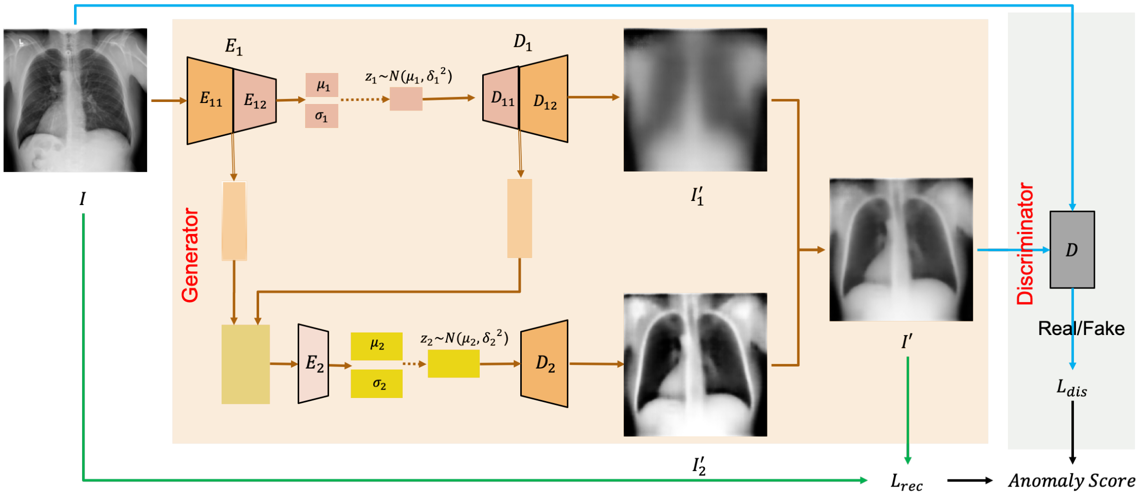


Figure 4.6: Proposed CVAD architecture - CVAE as the generator and a separate binary classifier (C) as the discriminator.

comprehensively against state-of-the-art anomaly detectors in detecting both intra-class and inter-class OOD data, showing improved performance. The code is available at `https://github.com/XiaoyuanGuo/CVAD`.

Publication:

- Guo, Xiaoyuan, Judy Wawira Gichoya, Saptarshi Purkayastha, and Imon Banerjee. "CVAD-An unsupervised image anomaly detector." Software Impacts 11 (2022): 100195.

- Guo, Xiaoyuan, Judy Wawira Gichoya, Saptarshi Purkayastha, and Imon Banerjee. "CVAD: A generic medical anomaly detector based on Cascade VAE." MICCAI workshop (2022).

4.2.2 Method

Anomaly detection includes both intra- and inter-class OOD identification, of which medical intra-class OOD data is much more challenging because of the minute dissimilarity compared to ID data. With no prior knowledge available and no sophisticated pre-processing, we utilize a variation autoencoder to learn the "normality" of in-distribution inputs via image reconstruction and enhance the discriminative ability for both two OOD classes via a binary discriminator. Both the reconstruction and discrimination contribute to accurate intra- and inter-class OOD detection.

CVAD architecture

Figure 4.6 shows the design of CVAD. Inspired by the GAN's architecture, we adopt the VAE architecture as the "generator" for modeling ID representations and a separate classifier as the "discriminator" to strengthen OOD discrimination.

A standard VAE module consists of two neural networks: an encoder and a decoder [73], with the encoder $q_\phi(z|x)$ (parameterized by ϕ) mapping the visible variables x to the latent variables z and the decoder $p_\theta(x|z)$ (parameterized by θ) sampling the visible variables x given the latent variables z [65]. Given a dataset $D = \{x_i\}_{i=1}^N$ with N input vectors 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 $\log p(x|\theta)$ [36]. However, a vanilla VAE exhibits limited potential in distinguishing unseen distributions due to the blurry reconstructions for large-size images. Thus, we adopt a modified VAE architecture – pchVAE [179] for high-quality reconstruction and better latent representations, which improves the reconstruction by adding a branch VAE on the standard VAE pipeline and then cascade the two representations for final outputs. For convenience, we use pchVAE and CVAE interchangeably.

Generator: Different from the standard VAE, CVAE has two encoders E_1, E_2 and two decoders D_1, D_2. To learn the high-level features, a deep and standard VAE architecture constructed by E_1 and D_1 formulates the deep latent variables z_1 by sampling parameters μ_1 and σ_1 of size K. Meanwhile, the low-level features are learnt by the branch VAE. Instead of using the original input, branch VAE utilizes the concatenation of two intermediate features from E_{11} and D_{11}. Given original input variables x, the input of branch VAE can be represented as $f(x)$. The encoder of branch VAE E_2 is simpler than E_1 whereas the decoder D_2 owns the same architecture as D_{12}. This branch VAE formulates latent Gaussian distributions with parameters μ_2, σ_2 of size $4K$. After sampling, two sets of latent variables, i.e., z_1, z_2 are acquired and decoded to image contexts $I_1^{'}$ and finer details $I_2^{'}$ respectively. I is the combination of $I_1^{'}$ and $I_2^{'}$.

Discriminator: Since the CVAE itself has no awareness of distinguishing outliers, we add a binary discriminator C to distinguish the reconstructed image $I^{'}$ and its counterpart with minor disturbance $I^{''}$ from the original input image I. As $I^{'}$, $I^{''}$ share very similar features with I after the first-stage training of the image generator, the discriminator is much more sensitive to minor differences from the in-distribution data, enhancing the accuracy of identifying both intra-class OOD data and inter-class OOD data.

Network training

Instead of training CVAD in an adversarial way, we train the generator and the discriminator in two stages. The reason is that training with adversarial losses often leads to much sharper reconstructions but ignores the low-level information of ID data, incurring high reconstruction errors and potential dangerous decisions for medical applications. Therefore, CAVD is designed to first train the image generator and then the binary discriminator to detect OOD data. This non-adversarial training enables CVAD to inherit the merit of VAEs [73] and avoid the instability of GANs [52].

To optimize CVAE, we minimize two objectives for the primary VAE part in Eqn. 4.5 and the branch VAE part in Eqn. 4.6, KL refers to Kullback-Leibler divergence.

$$L(x;\phi_1,\theta_1) = -E_{z_1 \sim q_{\phi_1}(z_1|x)}[\log \mathrm{p}_{\theta_1}(\mathrm{x}|\mathrm{z}_1)] + \mathrm{D}_{\mathrm{KL}}(\mathrm{q}_{\phi_1}(\mathrm{z}_1|\mathrm{x})||\mathrm{p}_{\theta_1}(\mathrm{z}_1)) \tag{4.5}$$

$$L(x;\phi_2,\theta_2) = -E_{z_2 \sim q_{\phi_2}(z_2|f(x))}[\log \mathrm{p}_{\theta_2}(\mathrm{x}|\mathrm{z}_2)] + \mathrm{D}_{\mathrm{KL}}(\mathrm{q}_{\phi_2}(\mathrm{z}_2|\mathrm{f(x)})||\mathrm{p}_{\theta_2}(\mathrm{z}_2)) \tag{4.6}$$

Therefore, the CVAE loss can be formulated as Eqn. 4.7. α_1 and α_2 to balance the weights of the two individual terms.

$$L_{\mathrm{cvae}} = \alpha_1 L(x;\phi_1,\theta_1) + \alpha_2 L(x;\phi_2,\theta_2) \tag{4.7}$$

The binary discriminator is trained to distinguish true/fake images using binary cross entropy.

Anomaly score: An anomaly score is defined in Eqn. 4.8 based on errors during inference and includes two parts: the reconstruction error S_cvae and the probability of being the anomaly class S_dis. Instead of simply adding the two parts together, we first scale the CVAE reconstruction errors into [0,1] and get the average score value to avoid assigning imbalanced weights between the two parts:

$$S = 0.5 * (\frac{S\_cvae}{S\_cvae_{max} - S\_cvae_{min}} + S\_dis) \tag{4.8}$$

Network Details

As illustrated in Figure 4.6, CVAE has a standard VAE part which consists of E_{11}, E_{12}, D_{11} and D_{12} and a branch VAE composed by a shallow encoder E_2 and a decoder D_2. The primary VAE is a symmetric network with five 4×4 convolutions with stride 2 and padding 1 followed by five transposed convolutions. Respectively, E_{11} stands for the first three convolution layers; E_{12} refers the last two convolution layers; D_{11} is for the first three transposed convolution layers and D_{12} means the last two transposed convolution layers. The input of the branch VAE is the intermediate features of E_{11} and the middle decoded features of D_{11}. E_2 here is a convolution layer which has a same 4×4 kernel with stride 2 and padding 1. D_2 shares the same decoder architecture as the standard VAE, namely, $D_2 = D_{11} + D_{12}$. All convolutions and transposed-convolutions are followed by batch normalization and leaky ReLU (with slope 0.2) operations. We used a base channel size of 16 and increased number of channels by a factor of 2 with every encoder layer and decreased the number of channels to half for each decoder layer. The latent dimension K of z_1 is set as 512 and z_2 is with $4K$, i.e., 2048 dimensions.

The binary discriminator is composed of five convolution layers with the same settings as above and a final fully connected layer to make a binary prediction. After a sigmoid function, the final ID/OOD class probability is obtained.

4.2.3 Experiments

Datasets

We conducted extensive experiments, verifying the generalizability and effectiveness of our approach on multiple open-access medical image datasets for intra- and inter-class OOD detection. In total, we used four independent datasets, including three medical image datasets – RSNA Pneumonia dataset [156], inferior vena cava filters (IVC-Filter in short) on radiographs [102] and SIIM-ISIC Melanoma dataset [125] (identify melanoma in lesion images) and one natural image datasets – Bird Species[2]. Among the medical datasets, RSNA and SIIM datasets have binary classes – normal and abnormal, whereas IVC-Filter dataset has 14 distinct types (classes). Table 4.8 lists the class information and number of images for each dataset and the corresponding usage in the **Details** column. Bird dataset, which contains 270 bird species with 38,518 training images, was only used as inter-class OOD for detection validation.

[2] https://www.kaggle.com/gpiosenka/100-bird-species

To unify the OOD detection pipeline and facilitate evaluation, we resized both the medical images and the validation inter-class OOD images to a unified $256 \times 256 \times$ *channel* size, where IVC-Filter and RSNA datasets are in gray scale with *channel* as 1 and the SIIM images are in RGB format and have *channel* 3.

Implementation

We implemented our model using Pytorch 1.5.0, Python 3.6. α_1, α_2 were equal to 1. We ran the models on 4 NVIDIA Quadro RTX 6000 GPUs with 24 GB memory each. In our model training, we used Adam optimizer with a learning rate of 0.001, and each network was trained for 100-350 epochs.

Evaluation Metrics

We evaluated our anomaly detection model performance in terms of standard statistical metrics - (i) area under the receiver operating characteristic (AUROC, AUC in short): a performance metric for "discrimination" between ID and OOD data (close to 1 gives optimal discrimination); (ii) True Positive rate (TPR): number of samples correctly classified as OOD (higher yield indicates better performance); (iii) False positive rate (FPR): number of samples wrongly classified as OOD (lower is better). To classify ID and OOD classes, a threshold should be defined for the anomaly scores. Notably, the AUC value is threshold-invariant, while the TPR and FPR are determined by the selection of the anomaly threshold. We adopted the Geometric Mean (G-Mean) method to determine an optimal threshold for the ROC curve by tuning the decision thresholds and reported the resulted FPR and TPR values. We also reported the corresponding DIFF, which is the difference of TPR and FPR under optimal selection, i.e., DIFF=TPR-FPR (larger is better). To be fair and thorough, we ran all the experiments on both intra-class OOD and inter-class OOD to further analyze the performance of anomaly detectors on the specific type of OOD detection.

Quantitative results

To demonstrate the model's effectiveness, we set the vanilla AE and VAE architectures as baselines and compared our CVAD model with three state-of-the-art models with varying architectures – pchVAE [179], a classifier-based approach DeepSVDD [127], and a GAN-based method GANomaly [4]. Table 4.7 shows the models' performance for the intra-class OOD detection and Table 4.8 primarily presents the inter-class OOD performance. The selection of in-class data, intra-class OOD and inter-class OOD data are summarized in the **Details** column of Table 4.8.

Results for Intra-class OOD Detection Intra-class OOD images are the most challenging outliers to identify since they often share similarity to the ID data but belong to a different class with unique characteristics. This similarity leads to the difficulty in identifying this type of OOD data, especially for medical images. As illustrated in Figure 4.5, e.g., the variations of benign and malignant skin cancer images are not as obvious as the natural objects. Still, CVAD exhibits its superiority in detecting intra-class OOD for medical images. On the RSNA dataset, CVAD achieves the best DIFF value 0.322 and AUC score 0.699 (+0.129 from DeepSVDD's AUC score 0.570, +0.123 from GANomaly's AUC score 0.576); for IVC-Filter, though GANomaly obtains the highest DIFF and AUC values, CVAD shows competitive performance and improves its AUC score 0.582; and for the RSNA dataset, DeepSVDD has the largest DIFF value 0.407 but CVAD reaches the second best DIFF 0.393. Moreover, CVAD acquires the optimal AUC score 0.750. Overall, CVAD performs stably and effectively for intra-class OOD detection except the sub-optimal results for IVC-Filter dataset. The reason behind this is the training data size. With 196 training images of IVC-Filter, CVAD may not be able to learn enough ID feature representations. Nevertheless, CVAD still outperforms GANomaly on IVC-Filter dataset.

Table 4.7: Intra-class OOD detection results (FPR, TPR, DIFF and AUC values) of various anomaly detectors trained on RSNA, IVC-Filter and SIIM datasets. Best results are highlighted.

Methods	RSNA				IVC-Filter				SIIM			
	↓*FPR*	↑*TPR*	↑*DIFF*	↑*AUC*	↓*FPR*	↑*TPR*	↑*DIFF*	↑*AUC*	↓*FPR*	↑*TPR*	↑*DIFF*	↑*AUC*
AE [129]	0.318	0.461	0.143	0.566	0.443	0.526	0.083	0.520	0.403	0.685	0.282	0.673
VAE [9]	0.381	0.611	0.230	0.614	0.426	0.525	0.099	0.524	0.442	0.740	0.298	0.676
pchVAE [179]	0.498	0.737	0.239	0.604	0.475	0.567	0.092	0.529	0.399	0.568	0.169	0.616
DeepSVDD [127]	0.399	0.509	0.110	0.570	0.545	0.713	0.168	0.522	0.276	0.683	**0.407**	0.740
GANomaly [4]	0.524	0.678	0.154	0.576	0.409	0.603	**0.194**	**0.584**	0.553	0.495	-0.058	0.418
CVAD (ours)	*0.321*	*0.643*	***0.322***	***0.699***	*0.541*	*0.706*	*0.165*	*0.582*	*0.381*	*0.774*	*0.393*	***0.750***

Table 4.8: AUC scores predicted by OOD detectors for inter-class identification on RSNA, IVC-Filter and SIIM datasets. The total number of samples of each dataset is reported in the bracket of **Details** column. Bold indicates the best performance.

Dataset	**Details**	**Methods**	**AUROC score**		
			InterClass1	*InterClass2*	*InterClass3*
RSNA	In-class: normal (8,851) Intra-class: pneumonia (9,555), abnormal (11,821) InterClass1: BIRD (38,518) InterClass2: SIIM (33,125) InterClass3: IVC-Filter (1,258)	AE [129]	0.680	0.608	0.616
		VAE [9]	0.752	0.604	0.613
		pchVAE [179]	0.795	0.776	0.619
		DeepSVDD [127]	0.838	**0.834**	0.604
		GANomaly [4]	0.775	0.819	0.594
		CVAD (ours)	***0.865***	*0.806*	***0.706***
IVC-Filter	In-class: type 11 (196) Intra-class: type 0-10, 12,13 (1,062) InterClass1: BIRD (38,518) InterClass2: SIIM (33,125) InterClass3: RSNA (30,227)	AE [129]	0.372	0.353	0.237
		VAE [9]	0.666	0.400	0.706
		pchVAE [179]	0.775	0.321	0.846
		DeepSVDD [127]	0.864	**0.979**	**0.889**
		GANomaly [4]	0.829	0.525	0.740
		CVAD (ours)	***0.916***	*0.705*	*0.844*
SIIM	In-class: benign (32,541) Intra-class: malignant (584) InterClass1: BIRD (38,518) InterClass2: IVC-Filter (1,258) InterClass3: RSNA (30,227)	AE	0.572	0.013	0.752
		VAE [9]	0.712	–	0.759
		pchVAE [179]	0.943	0.992	0.684
		DeepSVDD [127]	0.986	0.992	0.804
		GANomaly [4]	0.686	0.989	0.442
		CVAD (ours)	***0.993***	***0.993***	***0.831***

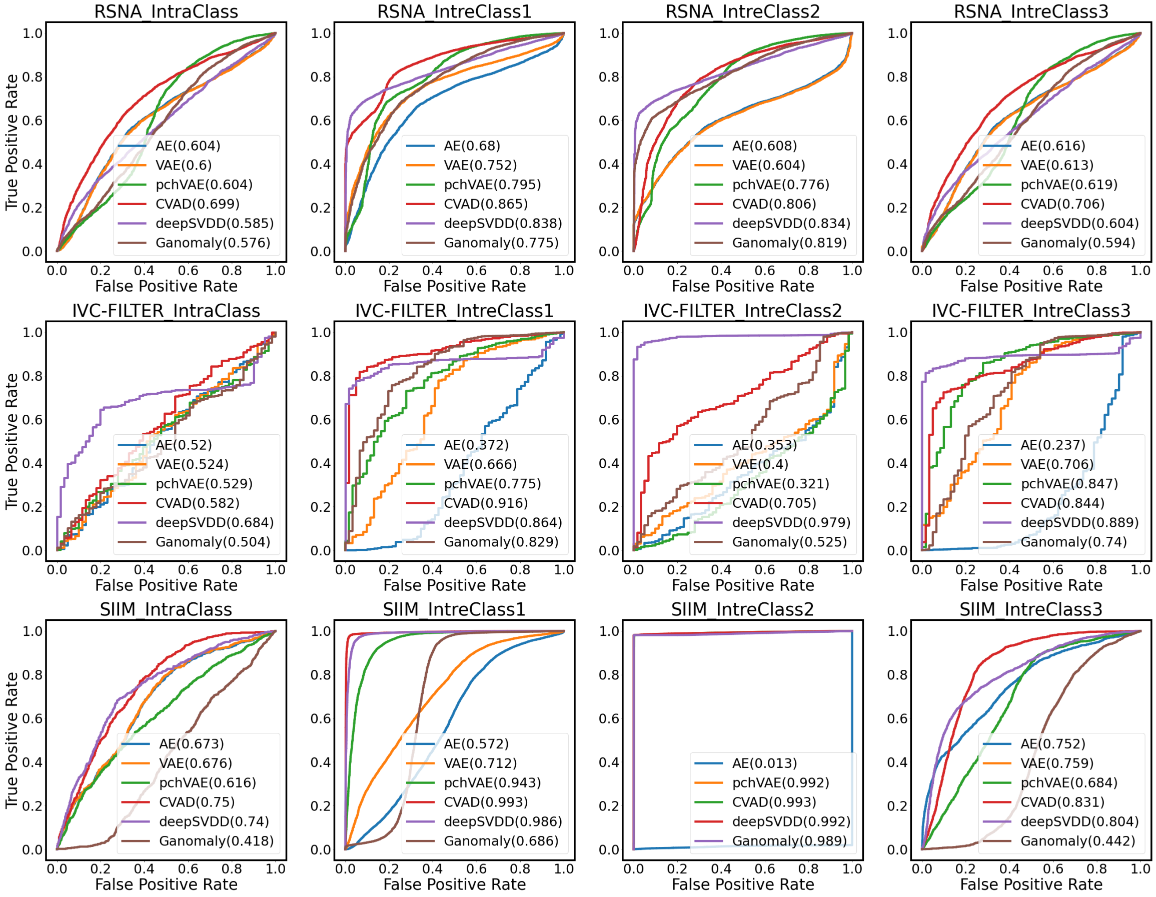


Figure 4.7: ROC curves of different models for intra- and inter-class OOD identification on RSNA, IVC-Filter and SIIM dataset. Performance of different models are highlighted with different colors with the corresponding AUC scores labeled in the brackets.

Results for Inter-class OOD detection To fairly evaluate all the models, we tested them on multiple inter-class OOD data types and presented the corresponding AUC scores in Table. 4.8. As the OOD image datasets may have different image channels and image sizes from the ID training images, we adjusted the image channels and resized the images to ensure consistent input data format for evaluation[3]. CVAD obtains the highest AUC values on RSNA (except for inter-class2) and SIIM datasets across three inter-class OOD detection evaluations. The inter-class OOD detection of CVAD on IVC-Filter is also satisfying with stable performance.

To further show the models' performance difference, we plotted the Receiver operating characteristic (ROC) curves of all the datasets for all the models evaluated on four OOD situations – intra-class, inter-class1, inter-class2, inter-class3 OOD data. Figure 4.7 shows the plots for RSNA, IVC-Filter and SIIM datasets with the corresponding AUC scores included. Notably, the difficulties in detecting intra-class and inter-class OOD data are reflected on the AUC scores, with most scores on inter-class OOD data are much higher than detection on intra-class OOD samples, especially in the RSNA results of Figure 4.7.

Ablation Study Generally, CVAD can exceed the baseline's performance with certain improvements and show competitive performances for both intra- and inter-class OOD detection. Here we analyze the functionality of the "generator" and "discriminator" of CVAD. As CVAD utilizes pchVAE to learn latent ID representation, we also report the performances of pchVAE itself on detecting intra- and inter-class OODs in Table. 4.7 and Table. 4.8 respectively.

For intra-class OOD detection, CVAD improves DIFF value from pchVAE's 0.239 to 0.322 (+0.083) and AUC score from pchVAE's 0.604 to 0.699 (+0.095) on the RSNA dataset; similarly on IVC-Filter dataset, CVAD enhances the DIFF from pch-

[3] For example, to evaluate trained models on RSNA, we converted the BIRD and SIIM images to grayscale mode and resized them to the same in-distribution image size.

VAE's 0.092 to 0.165(+0.073), AUC from pchVAE's 0.529 to 0.582 (+0.053); for SIIM dataset, CVAD also increases DIFF from pchVAE's 0.169 to 0.393 (+0.224), AUC from pchVAE's 0.616 to 0.750(+0.134). The same observation also exists in the inter-class OOD detection results in Table. 4.8. This performance improvement can be attributed to the discriminator's learning with the exposure of generated OOD data samples, which enables CVAD to gain better discriminative ability than pchVAE itself.

Additionally, the standard AE and VAE are evaluated as baselines for various OOD detection. Although pchVAE reconstructs image with higher quality than VAE, it fails to exceed VAE in OOD detection. Autoencoder, which can also output good reconstruction, exhibits the weakest OOD detection accuracy according to the results reported in Table. 4.7 and Table. 4.8. In conclusion, good image reconstruction does not ensure strong OOD identification ability and adding a discriminator can be functional and contribute to discriminative learning.

Qualitative Results

Here we provide visualizations for anomaly detection of CVAD on different datasets and the reconstruction effects of CVAE.

Anomaly Detection Figure 4.8 shows experimental results for RSNA dataset. Each column represents a specific type of input data. From left to right, they are in-distribution data, intra-class OOD data, inter-class OOD1 data, inter-class OOD2 data and inter-class OOD3 data, respectively. There are two examples for each type of data. The corresponding anomaly score predicted by CVAD is on top of each example. A high anomaly score means high possibility the data is with to be in OOD category. As can be seen in Figure 4.8, the two intra-class OOD samples are alike as the in-distribution data but the inter-class OOD examples show very

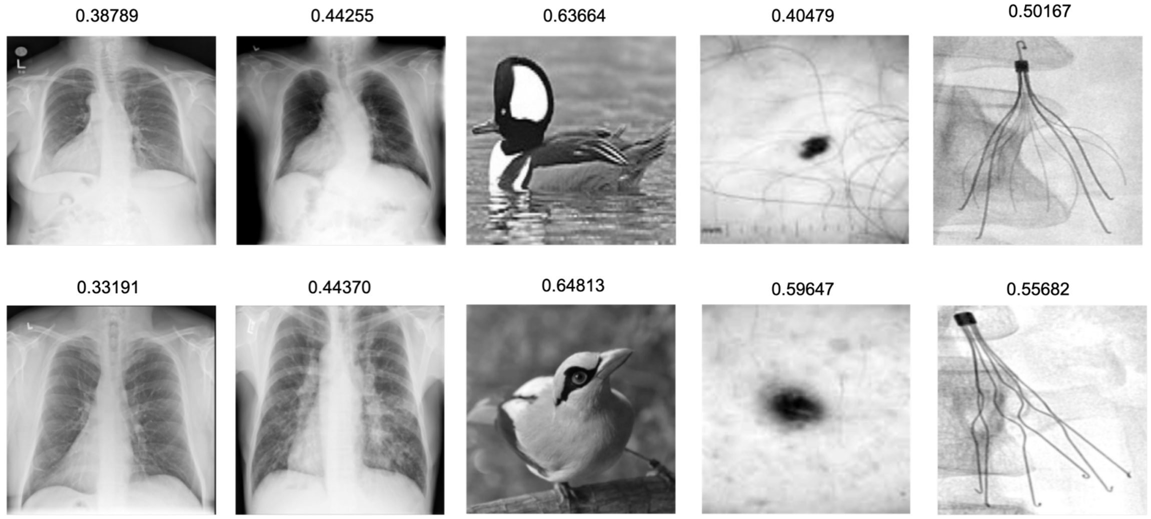


Figure 4.8: Anomaly scores output by CVAD for different types of input data (experiments for RNSA dataset). Columns from left to right, ID, intra-class OOD, inter-class OOD1, inter-class OOD2, inter-class OOD3.

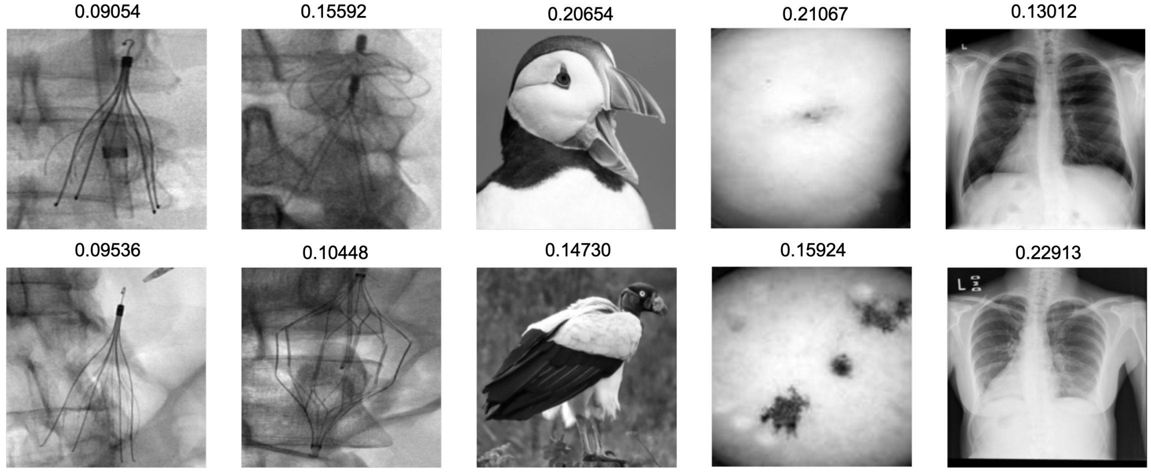


Figure 4.9: Anomaly scores for IVC-Filter dataset, from the left to right: in-distribution data, intra-class OOD, inter-class OOD1, inter-class OOD2, inter-class OOD3

different appearance from in-distribution data. Correspondingly, the anomaly scores of intra-class OOD are close to the scores of ID samples and difficult to separate whereas the intra-class OOD cases with clear variations are assigned higher anomaly scores and easy to identify. This phenomenon further demonstrates the challenges of identifying intra-class OOD data. The predicted anomaly scores for IVC-Filter and SIIM experiments are present in Figure 4.9 and Figure 4.10, respectively.

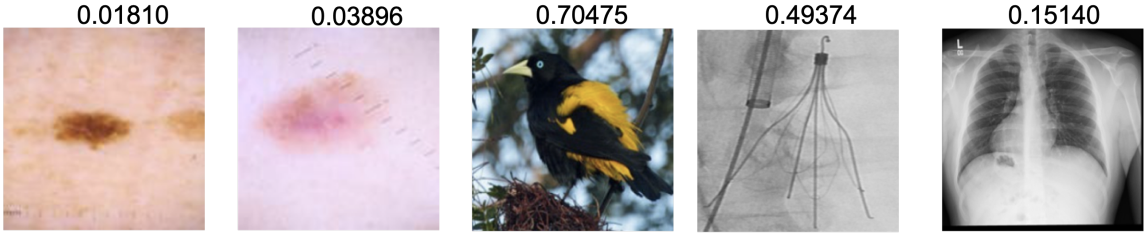


Figure 4.10: CVAD prediction examples of SIIM dataset. From left to right, ID, intra-class OOD, inter-class OOD1, inter-class OOD2, inter-class OOD3 respectively. Anomaly scores are labeled on top of each case.

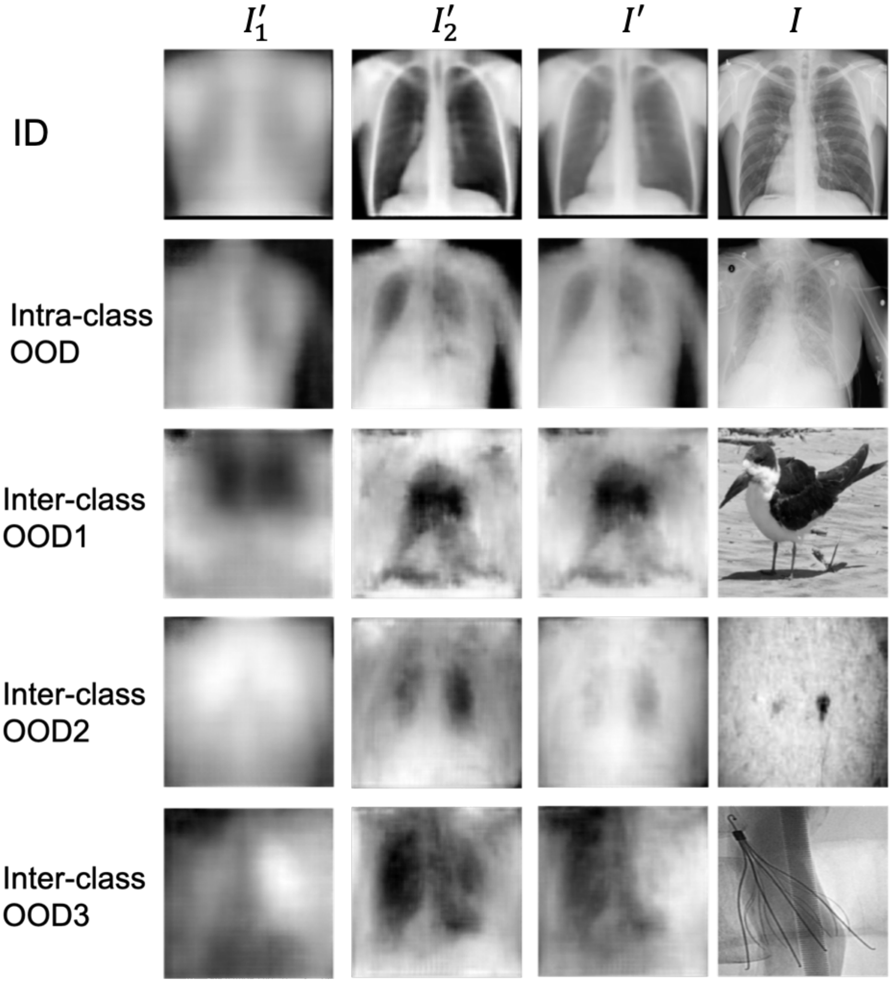


Figure 4.11: Reconstruction details visualization of CVAE trained on RSNA dataset for different data types.

Visualization of reconstruction effects CVAD gains good latent in-distribution features via its "generator" – CVAE, which learns both low-level and high-level representations. To demonstrate the effectiveness, we took RSNA dataset as a representative and showcased the reconstruction details in Figure 4.11, with the first column for branch VAE reconstruction $I_2^{'}$, the second column for standard VAE part reconstruction $I_1^{'}$, the third column for ultimate reconstruction $I^{'}$ and the last column for the original input image I (following the same notations indicated in Figure 4.6). To further reveal the effects of CVAE on different OOD samples, we also presented example images for ID (i.e., normal class, 1st row), intra-class OOD (i.e., pneumonia or with opacity, 2nd row), inter-class OOD1 (i.e., gray-scale bird images, 3rd row), inter-class OOD2 (i.e., skin cancer images from SIIM dataset,4th row) and inter-class OOD3 (i.e., images from IVC-Filter dataset, 5th row) in Figure 4.11. Compared with the intra-class medical OOD data, reconstructions on inter-class OOD inputs are more messy and dissimilar to the original OOD data, which leads to larger reconstruction errors and thus easier to distinguish. This observation reveals the varying difficulties of detecting different types of OOD data – intra-class OOD is much more challenging than inter-class OOD.

4.2.4 Conclusion

We propose an effective medical anomaly detector CVAD that can reconstruct coarse and fine image components by learning multi-scale latent representations. The high quality of generated images enhances the discriminative ability of the binary discriminator in identifying unknown OOD data. We demonstrate the OOD detection efficacy for both intra-class and inter-class OOD data on various medical and natural image datasets. Our model has no prior assumptions on the input images and application scenarios for OOD, thus can be applied to detect OOD samples in a generic way for multiple scenarios.

4.3 Discussions and future works

In this chapter, we have researched on medical image OOD detection under limited supervision. Specifically, we have proposed TEND and CVAD to handle intra-class OOD detection and generic OOD detection, respectively.

We implement TEND with three different margins and show our results across various medical datasets under different settings. Although our models show competitive performance and surpass other methods under certain situations, the margin parameter has to be tuned for specific usages. Depending on the data complexity and variance across classes of a dataset, 250 is a good starting point. The ability of separation OOD from ID does not always improve as the margin increases due to the data complexity. For datasets with clear class variations, the margin can be set larger accordingly and vice versa. Besides, TEND utilizes transformation to generate fake OOD samples for discriminative learning. Due to the large amount of possibilities, this work only exploits a limited number of possible transformations. In the future, more variations of fake OOD generations can be explored to check the effectiveness of each different transformation.

We design CVAD based on a cascade VAE model to learn the normality of in-distribution data. Because of the characteristics of VAE, the generator of CVAD requires an extremely small learning rate to avoid the gradient explosions and a long time to get the model converged. Our work only considers 2D X-ray images and natural images, more data modalities and application scenario should be researched, including the common used MRI image data, 3D images, etc.

Chapter 5

Medical Dataset Curation with Limited Supervision

Supervised deep learning has been promising in solving various medical image-related tasks, and often requires well-annotated datasets for training and validation which must be extracted and curated to a high quality standard before being usable for model development [149]. Medical datasets from different institutions can be heterogeneous due to equipment, acquistion techniques, and patients, resulting in data distribution shifts between sets. Models trained on an internal dataset A from a specific institute may show degraded performance on an external dataset B from other sources due to the possible noisy data, distribution shift and poor-quality data, which are called *shift data* in this paper. Dataset/Distribution shift is a common problem in predictive modelling and present in most practical applications, for reasons ranging from the bias in introduced by experimental design the irreproducibility of the testing conditions at training time [115], of which imbalanced data, domain shift, source component shift, may be the most common forms [142]. 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 is crucial for cleaning the datasets and helpful in enhancing the model's generalization with future training. Unfortunately, it still lacks an effective way to identify the difference for a bunch of datasets from the same medical domain. The main challenge lies in the inaccessibility to external medical datasets. Privacy concerns around sharing personally identifiable information are a major barrier to data sharing in medical research [135]. To address these privacy concerns, there has been an impressive number of large-scale research collaborations to pool and curate de-identified medical data for open-source research purposes [32]. Nevertheless, most medical data is still isolated and locally stored in hospitals and laboratories due to the worries associated with sharing patient data [150]. Therefore, an efficient way of external dataset curation/cleaning without sharing data is desired.

To overcome the obstacle, we propose *MedShift*, a pipeline for identifying shift data, which takes advantage of the accessible models trained on the internal dataset to gain the in-distribution knowledge. As observed by Ref. [116], domain-discriminating approaches tend to be helpful for characterizing shifts qualitatively and determining if the are harmful. Therefore, we utilize unsupervised anomaly detectors to learn the "normality" of in-domain features. Suppose the internal dataset has multiple classes, the feature representation of each class is learnt by an OOD detector. Without sharing the internal dataset with others, the shift data is theoretically under-represented and should be detected by the accessible anomaly detectors as outliers from the external datasets. Since 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, two intuitions for example, the *shiftness* of the identified data can be reflected via the performance variance of a well-trained model. Instead of checking the shift sample one by one, *MedShift* quantifies the *shiftness* for each class in small groups. 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 adapt a dataset quality metric (OTDD [8]) for helping facilitate the comparison of differences among a series of datasets coming form the same medical domain.

5.1 Contribution

We summarize our contributions as follows:

1. We 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. We employ two unsupervised anomaly detectors to learn the internal distribution and identify samples showing the significant *shiftness* for external datasets, and compared their performance;

3. We 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 sub-group of each class in external domains after dropping the shift data;

4. We adapt a data quality metric to quantify the dissimilarity between the internal and external datasets;

5. We 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. The code is available at `https://github.com/XiaoyuanGuo/MedShift`. An interface introduction video to visualize our results is available at `https://youtu.be/V3BF0P1sxQE`.

Publication:

- Guo, Xiaoyuan, Judy Wawira Gichoya, Hari Trivedi, Saptarshi Purkayastha, and Imon Banerjee. "Shift data identification for external medical datasets." SIIM 2022.

- Guo, Xiaoyuan, Judy Wawira Gichoya, Hari Trivedi, Saptarshi Purkayastha, and Imon Banerjee. "MedShift: identifying shift data for medical dataset curation." Joural of Biomedical and Health Informatics 2022 (under 2nd round review).

5.2 Method

In Section 5.2.1 and 5.2.2, we formulate the dataset shift identification problem and introduce the necessary notations. Then, we propose and illustrate the pipeline of shift identification in Section 5.2.3; we further dive deep in the *shiftness* evaluation in Section 5.2.5. To complement, we introduce the details of our anomaly detection architecture used for *MedShift* pipeline in Section 5.2.4. Additionally we introduce the dataset quality measurement in Section 5.2.6.

5.2.1 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.

5.2.2 Formulation and notation

Given two datasets D_A and D_B of the same medical domain with the same classes (say $c_1, c_2, ..., c_n$, 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 i.i.d. internal data $\{X^A_{c_i}\}^n_{i=1}$ with n classes, and input samples $\{x^{Ac_i}_j\}^{N^A_{c_i}}_{j=1} \subset X^A_{c_i}$ ($N^A_{c_i}$ is the sample number of dataset A's class c_i) from the internal input distribution, and i.i.d. external data $\{X^B_{c_i}\}^n_{i=1}$ and input samples $\{x^{Bc_i}_j\}^{N^B_{c_i}}_{j=1} \subset X^B_{c_i}$ ($N^B_{c_i}$ 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 $\bar{X}^B_{c_i} \subseteq X^B_{c_i}$. 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 $P(y^A_{c_i}|x^A_{c_i})$ for some targets $y^A_{c_i}$ given data samples $x^A_{c_i}$ for class c_i. Theoretically, given the target model trained on the ID data $X^A_{c_i}$, the predictions over set $X^B_{c_i} - \bar{X}^B_{c_i}$ should produce more relevant results than on the whole set $X^B_{c_i}$.

5.2.3 Shift 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 pipeline is built on top of the anomaly detection architecture to leverage the anomaly score as illustrated the framework in Figure 5.1, which involves two separate phases - internal training and test phase. An interesting challenge of shift identification is that the anomaly detectors should be able to identify unknown anomalous patterns of an external dataset without including any anomalous data samples in training since in the real situation, exchanging healthcare data among institutions and manually identifying noisy or anomalous data are not trivial tasks.

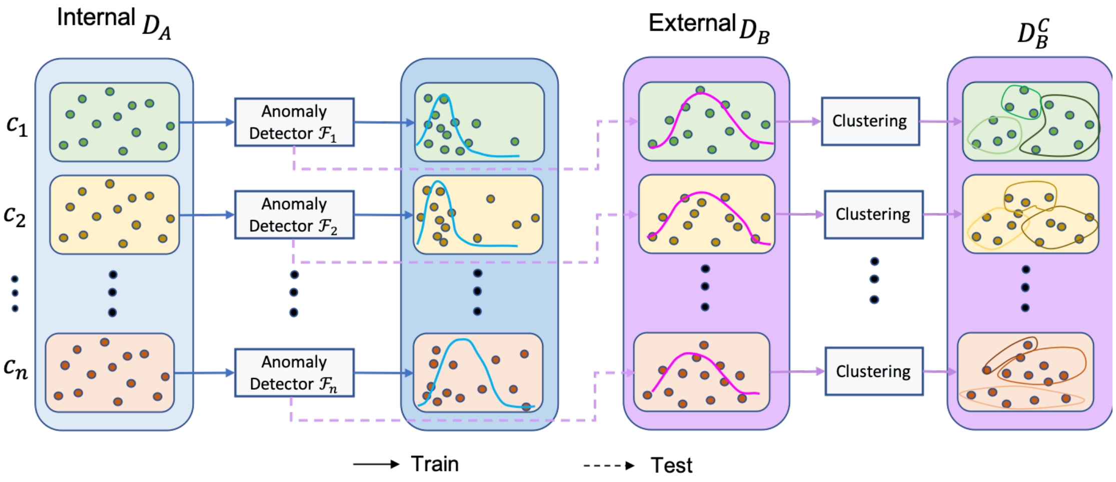

Internal D_A
External D_B
D_B^C
c_1
c_2
c_n
Anomaly Detector $\mathcal{F}_1$
Anomaly Detector $\mathcal{F}_2$
Anomaly Detector $\mathcal{F}_n$
Clustering
Clustering
Clustering
Train
Test

Figure 5.1: Shift data identification pipeline

During the training phase, only internal data samples and the anomaly detection models (see introductions in Sec. 5.2.4) are involved. As shown in the left blue part of Figure 5.1, a set of anomaly detectors $\mathcal{F}$s for each targeted categories 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 Figure 5.1, 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. 5.2.5, an unsupervised clustering algorithm is sub-sequentially applied to each class and clusters the data items into k groups based on the learnt anomaly scores. For each class, the optimal number of cluster k is determined by the Elbow Method. As observed during our experiments, data collected from the same source usually presents similar distributions. Therefore, we keep k as same across all the classes.

5.2.4 Anomaly detection

Architecture. As claimed in Sec. 5.2.3, we propose to utilize anomaly detection models to not only identify distribution shifts in the external dataset but also automated cleaning of the external data without any data sharing. First, we briefly describe our anomaly detection model - *Cascade Variational autoencoder-based Anomaly Detector (CVAD)* [54] used in *MedShift*, which was previously been tested on both generic and medical image datasets. As shown in Figure 5.2, CVAD is a self-supervised variational autoencoder-based anomaly detection model which 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.

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 $D = \{x_i\}_{i=1}^N$ with N input vectors 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 $\log p(x|\theta)$ [36].

To optimize the generator, we minimize two objectives for the primary VAE part in (5.1) and the branch VAE part in (5.2), KL refers to Kullback-Leibler divergence.

$$L(x; \phi_1, \theta_1) = -E_{z_1 \sim q_{\phi_1}(z_1|x)}[log\, p_{\theta_1}(x|z_1)] + D_{KL}(q_{\phi_1}(z_1|x)||p_{\theta_1}(z_1)) \tag{5.1}$$

$$L(x; \phi_2, \theta_2) = -E_{z_2 \sim q_{\phi_2}(z_2|f(x))}[log\, p_{\theta_2}(x|z_2)] + D_{KL}(q_{\phi_2}(z_2|f(x))||p_{\theta_2}(z_2)) \tag{5.2}$$

where $f(x)$ is the input of branch VAE, encoded by E_{11}. Therefore, the "generator" loss can be formulated as Eqn. 5.3. α_1 and α_2 to balance the weights of the two individual terms.

$$L_{\text{rec}} = \alpha_1 L(x; \phi_1, \theta_1) + \alpha_2 L(x; \phi_2, \theta_2) \tag{5.3}$$

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 includes two parts: the reconstruction error S_{rec} in the first stage and the probability of being the anomaly class S_{dis} in the second stage. To adapt the application of detecting abnormal data for multiple unknown external sources, we modified that anomaly score computation by simply adding the two parts together $S = S_{rec} + S_{dis}$. 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 256×256×*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 proposed method(CVAD), respectively; we train the generator with 250-500 epochs and the discriminator with 10-20 epochs.

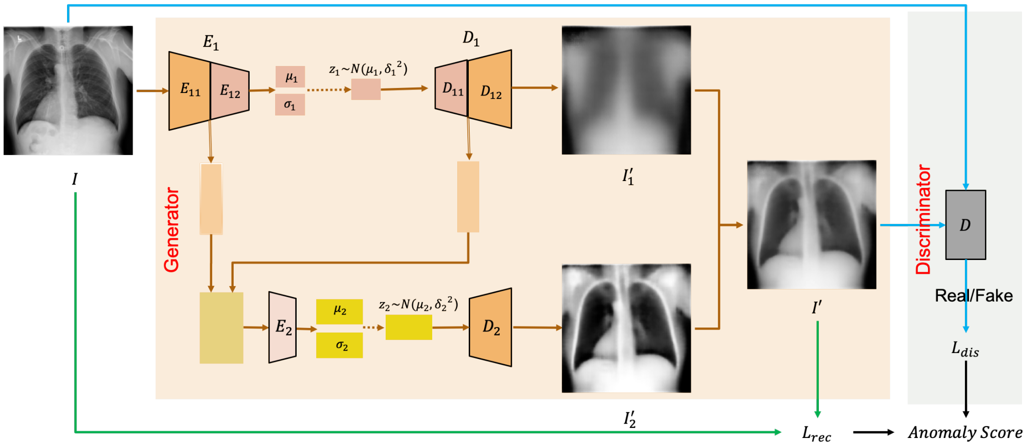


Figure 5.2: CVAD architecture - a cascade VAE as the generator (G) and a separate binary classifier (D) as the discriminator. The main VAE pipeline is composed by the encoder E_1 shown as the orange part and the decoder D_1 in the dark green part; the branch VAE has the pink part as the encoder E_2 and 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.

5.2.5 Shiftness quantification

The above pipeline can be applied to detect the shift data and assign each data with an anomaly score for indicating its contribution to the dataset shift. Nonetheless, the *shiftness* of the identified data is not simple and straightforward to evaluate in relation with the targeted task. We suggest to evaluate them in group. As prepared in the

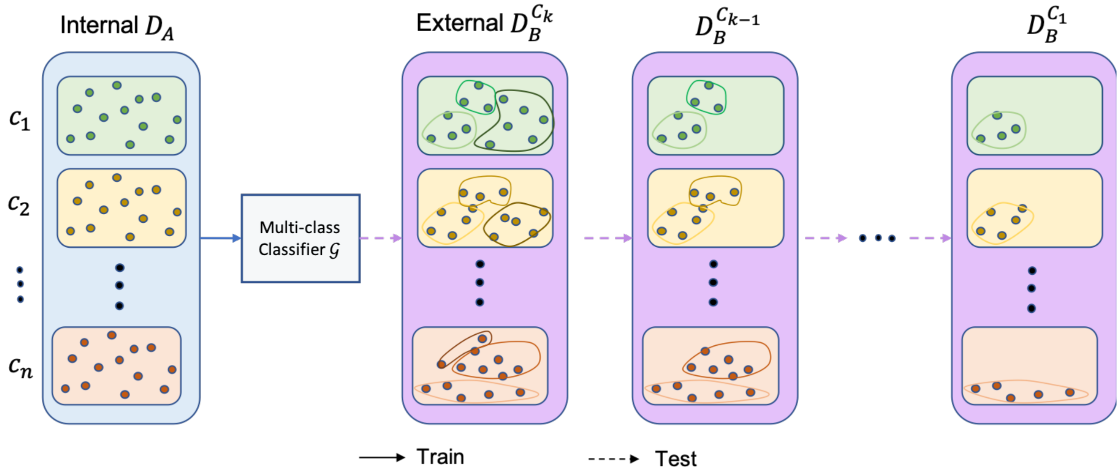


Figure 5.3: *Shiftness* quantification pipeline

first stage 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 and test on D_B. As presented in Figure 5.3, 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 [58] on the Emory MURA dataset with the publicly available pre-trained weights as initialization. We optimize the classifier using the Adam optimizer with a batch size of 512, a learning rate of 1×10^{-3} for 50 epochs. For chest X-ray data, we utilize the model proposed by Ref. [167], 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. [167], we use DenseNet121 [64] 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. 5.3.4.

5.2.6 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 the Optimal Transport Dataset Distance [8] (OTDD) measure to calculating similarities, or distances, between classification datasets. It relies on optimal transport[151], 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. Formally, the optimal transport dataset distance is defined as:

$$OTDD(\boldsymbol{D}_A, \boldsymbol{D}_B) = min_{\pi \in \prod(P_A, P_B))} \int_{\boldsymbol{Z} \times \boldsymbol{Z}} d(z, z^{'}) d\pi(z, z^{'}) \tag{5.4}$$

, of which

$$d(z, z^{'}) = (d(x, x^{'})^2 + W_2(P_y, P_{y'})^2)^{\frac{1}{2}} \tag{5.5}$$

, where D_A, D_B are the two datasets, W_p denotes the p-Wassertein distance. Please refer Ref. [8] for more details.

	HAND	**FOREAMR**	**FINGER**	**SHOULDER**	**ELBOW**	**WRIST**	**HUMERUS**
Emory_MURA	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	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 Cardiome-diastinum**	**Cardio-megaly**	**Lung Lesion**	**Lung Opacity**	**Edema**	**Consoli-dation**	**Pneu-monia**	**Atele-ctasis**	**Pneumo-thorax**	**Pleural Effusion**	**Pleural Other**	**Fracture**	**Support Devices**
Emory_CXR	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%)
CheXpert	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	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%)

Table 5.1: 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.

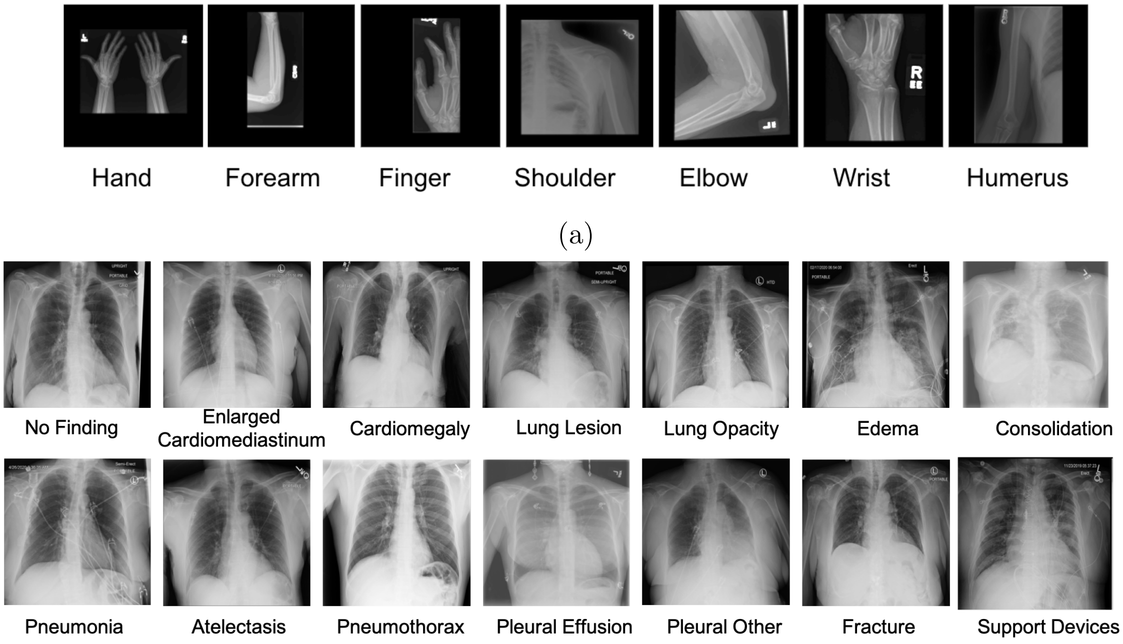
5.3 Experiments

5.3.1 Datasets

There are two categories of medical datasets used in this paper: (1) *Musculoskeletal radiographs* - Emory MURA dataset (internal) and Stanford MURA dataset [117] (external); (2) *Chest radiographs* - Emory Chest X-rays (internal, Emory-CXR in short), CheXpert dataset [68] (external_1) and MIMIC dataset [69] (external_2).

MURA (musculoskeletal radiographs) is a large dataset of bone X-rays. Each MURA dataset has seven classes, *XR_HAND*, *XR_FORARM*, *XR_FIGER*, *XR_SHOULDER*, *XR_ELBOW*, *XR_WRIST*, *XR_HUMERUS*. Image examples are illustrated in Figure 5.4a for each class. To demonstrate the effectiveness of detecting shift data, we have Emory MURA and Stanford MURA datasets as a pair and treat Emory MURA as the internal dataset with Stanford MURA as the external one. More class-wise details of the datasets are presented in the upper of Table. 5.1.

For chest X-ray, we used three dataset - Emory-CXR (XX images retrieved from Emory Healthcare system), CheXpert and MIMIC datasets. The bottom part of Table. 5.1 shows the details of the three datasets. All the chest X-ray dataset has 14 classes (or diagnosis) in total. The classes are *No Finding*, *Enlarged Cardiomediastinum*, *Cardiomegaly*, *Lung Lesion*, *Lung Opacity*, *Edema*, *Consolidation*, *Pneumonia*, *Atelectasis*, *Pneumothorax*, *Pleural Effusion*, *Pleural Other*, *Fracture*, *Support Devices*. Image examples are displayed in Figure 5.4b. Different from MURA dataset



(a)

(b)

Figure 5.4: Sample images from the datasets: (a) Musculoskeletal radiographs examples for each anatomical joint class. (Intensity contrasts are changed for better visualization); (b) ChestXray examples for each class. (Image are resized for better visualization);

where class labels are mutually exclusive, each chest X-ray data may have multiple common diagnosis.

5.3.2 Anomaly detectors in use

Main anomaly detector. OOD detection plays an important role in identifying shift data in external datasets. We have proposed a self-supervised anomaly detector - CVAD [54], whose introduction will be present in Sec. 5.2.4. 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`.

Baseline anomaly detector. Apart from our anomaly detection model *CVAD*, we select f-AnoGAN [133] as the baseline and apply the method in *MedShift* for comparison (*MedShift_w_f-AnoGAN* in short). f-AnoGAN is a generative adversarial network to identify anomalous images unsupervised. Two training steps are necessary for its anomaly detection - (1) GAN training, and (2) encoder training based on the trained GAN model. Only normal images are fed into the GAN model to learn in-distribution representations. And the encoder is trained to learn difference in feature space level. During reference, a combination of image reconstruction residual and the discriminator residual in feature level yields the anomaly score to detect anomalies.

To train f-AnoGAN, we use the default Adam optimizer with a learning rate of 2×10^{-4} and the same batch sizes as CVAD for the corresponding datasets; we run the generative adversarial training for 1000-1500 epochs and the encoder training for 300-500 epochs.

The reason for comparison with f-AnoGAN [133] is its superiority over experiments on three public medical image datasets (RSNA Pneumonia dataset [156], inferior vena cava filters (IVC-Filter in short) on radiographs [102] and SIIM-ISIC Melanoma

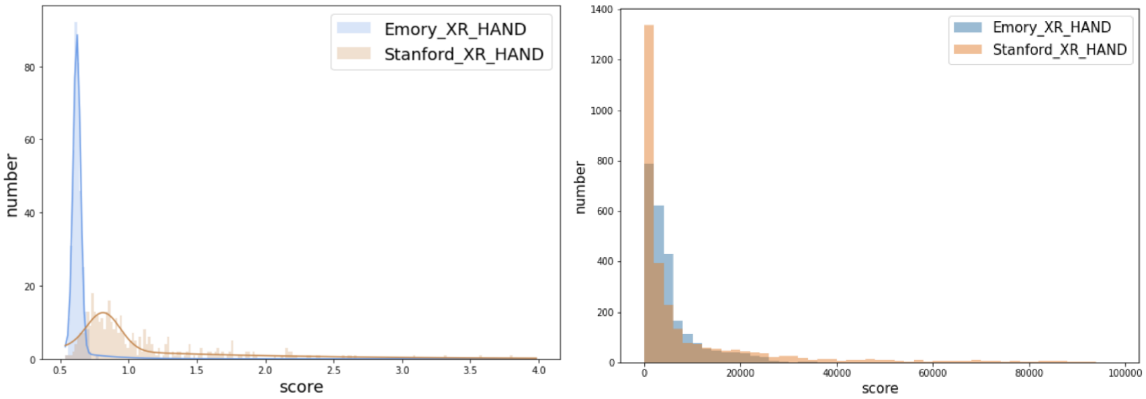


Figure 5.5: Shift identification with anomaly detection on Stanford_MURA *HAND* data - (left) anomaly score distributions of *MedShift_w_CVAD*; (right) anomaly score distributions of *MedShift_w_fAnoGAN*. Distributions are truncated on samples with large anomaly scores for better visualization.

dataset [125] (identify melanoma in lesion images)) with other representative anomaly detectors (DeepSVDD [127], GANomaly [4], etc.) evaluated in the meantime. More experimental details can be found in [54, 56].

5.3.3 Experimental setup

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.

5.3.4 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.

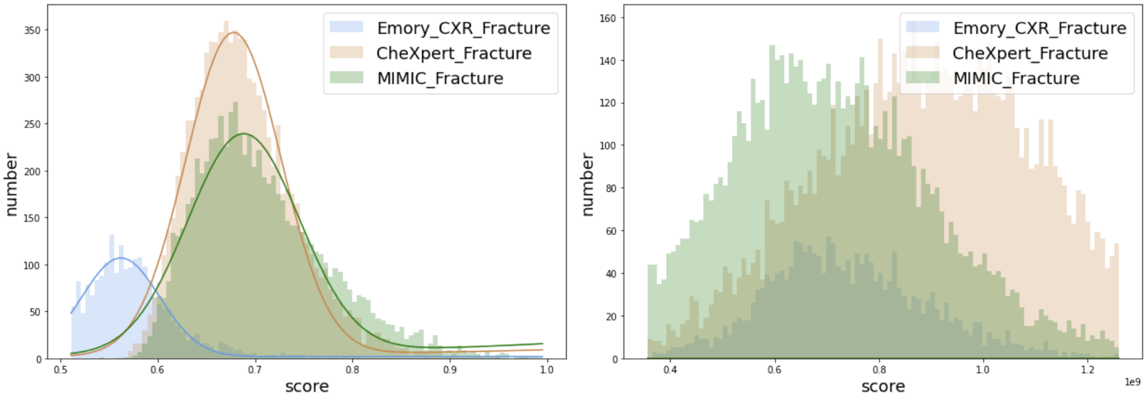


Figure 5.6: Shift identification with anomaly detection on CheXpert and MIMIC *Fracture* data - (left) anomaly score distributions of *MedShift_w_CVAD*; (right) anomaly score distributions of *MedShift_w_fAnoGAN*. Distributions are truncated on samples with large anomaly scores for better visualization.

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 5.5 presents the anomaly score distributions of the representative class from both MURA and Chest X-ray. The X-axis represents the anomaly score and Y-axis stands for the number of images that have anomaly scores in the corresponding range. In both cases, Emory data is considered as internal data.

For **MURA** dataset, the anomaly score distribution of *MedShift_w_CVAD* for *XR_HAND* is shown in the left of Figure 5.5, 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.

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 Figure 5.6, 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.

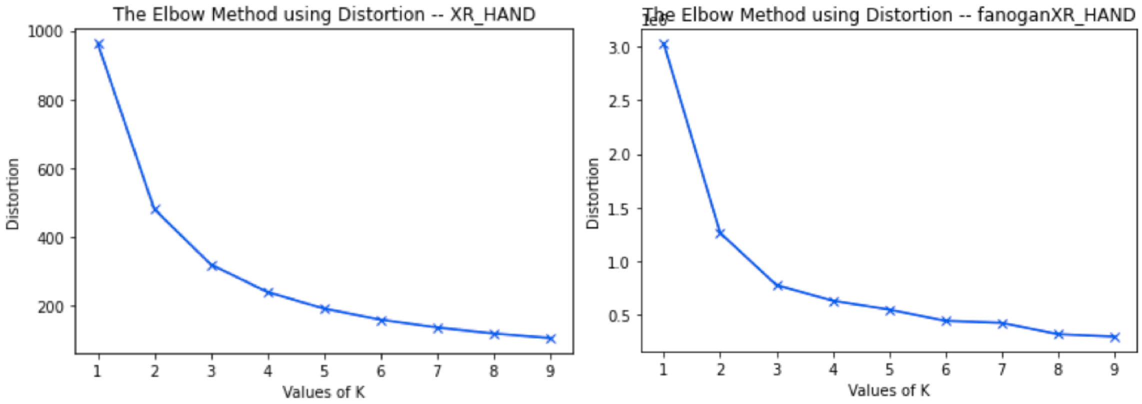


Figure 5.7: Elbow distortion curves for Stanford_MURA *HAND* data - (left) *MedShift_w_CVAD* results; and (right) *MedShift_w_fAnoGAN* results.

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, Figure 5.7 illustrates the curve plots of *MedShift_w_CVAD* and *MedShift_w_fAnoGAN*. For both situations, we pick 5 for group numbers.

The corresponding clustered examples can be seen in Figure 5.8. There are 5

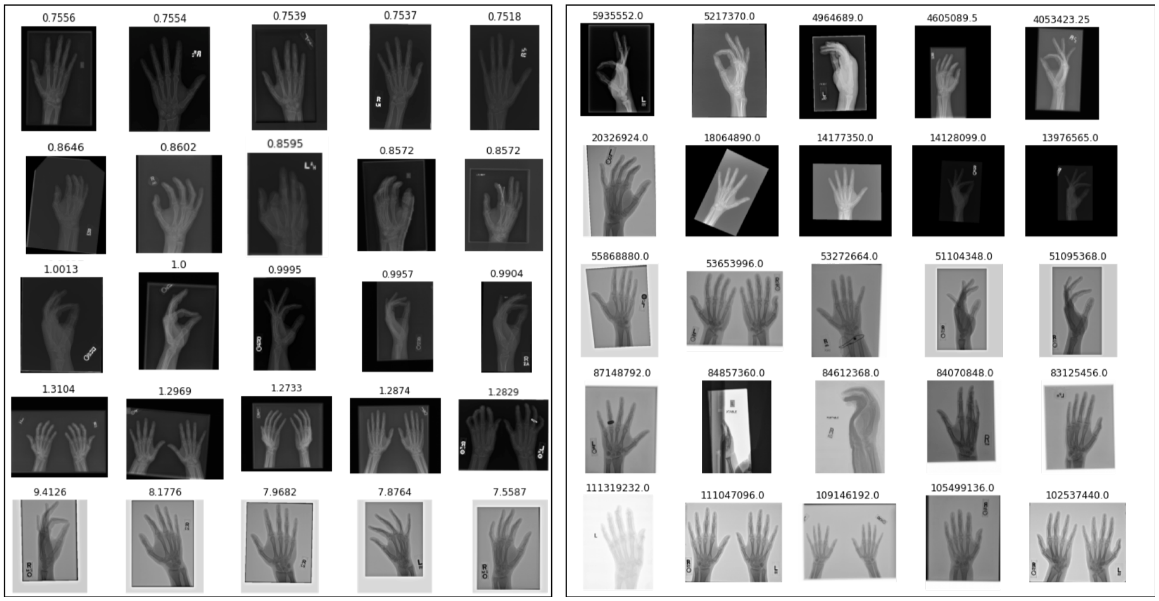


Figure 5.8: Clustering examples on Stanford_MURA *HAND* data - (left) *MedShift_w_CVAD* results; and (right) *MedShift_w_fAnoGAN* results. Each row represents one group with five example images. The groups are illustrated in ascending order based on the anomaly scores from top to bottom. The corresponding anomaly score is on top of each image.

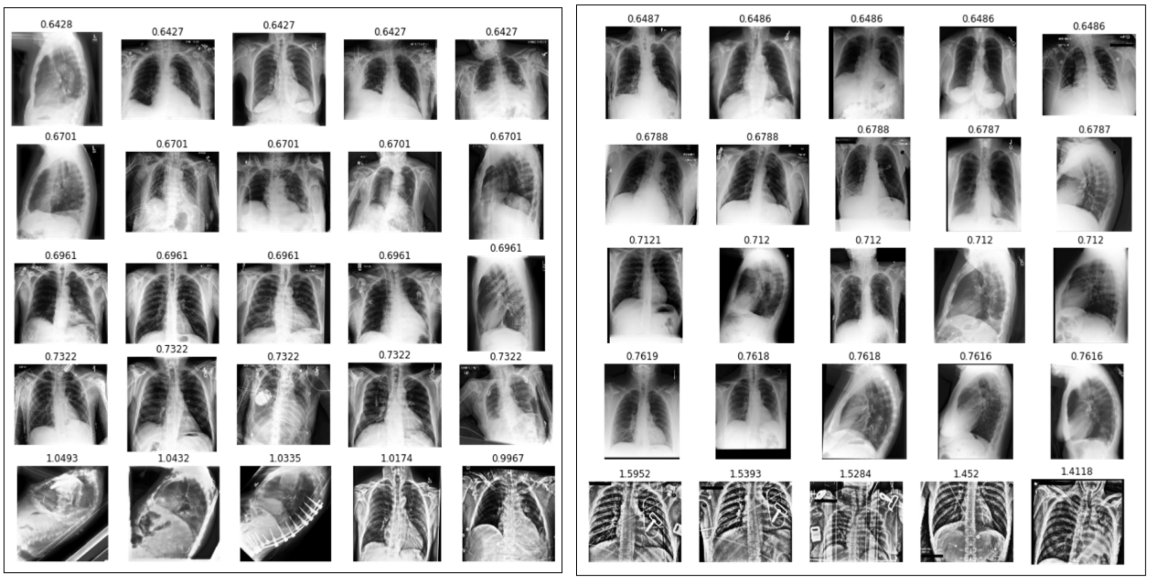


Figure 5.9: *MedShift_w_CVAD* examples of clustering results - (left) clustering results on CheXpert *Fracture* data; and (right) clustering results on MIMIC *Fracture* data. Styles follow Figure 5.8.

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 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 f-AnoGAN 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 Figure 5.9. Following the same arrange order, the difference for each group can be clearly captured by our model.

Classification results for shiftness evaluation

As introduced in Sec. 5.2.5, 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 gradually. 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 score computed with CVAD [54] and

Table 5.2: MURA classification class-wise results with CVAD (left) and f-AnoGAN (right). The best classification values are in bold for each method.

Dataset	Metric	HAND	FOREARM	SHOULDER	FINGER	ELBOW	WRIST	HUMERUS	Average	
									Macro	Weighted
Emory_test	*#images*	495	74	691	74	305	572	111	2,322	
	Precision	0.842	0.704	0.979	0.312	0.929	0.957	0.875	0.800	0.903
	Recall	0.970	0.770	0.999	0.068	0.862	0.942	0.820	0.776	0.915
	F1-score	0.901	0.735	0.989	0.111	0.895	0.950	0.847	0.775	0.905
	AUC	0.960	0.880	0.995	0.531	0.926	0.964	0.907	0.984	0.992
Stanford_TOP 5	*#images*	3,851	1,097	5,621	3,660	2,397	3,993	852	21,471	
	Precision	**0.921**	0.758	0.977	0.765	0.695	0.380	0.395	0.699	0.754
	Recall	0.450	0.160	0.746	0.188	0.701	0.983	0.664	0.556	0.604
	F1-score	0.605	0.264	0.846	0.301	0.698	0.548	0.496	0.537	0.594
	AUC	0.721	0.578	0.870	0.588	0.831	0.808	0.811	0.902	0.915
Stanford_TOP 4	*#images*	3,098 / 3,838	880 / 1,091	4,499 / 5,584	2,904 / 3,658	1,923 / 2,387	3,182 / 3,933	686 / 848	17,172 / 21,339	
	Precision	0.921 / 0.921	0.758 / 0.758	0.986 / 0.978	0.768 / **0.765**	**0.695** / 0.695	0.426 / **0.379**	0.545 / 0.404	0.728 / 0.700	0.772 / 0.755
	Recall	0.558 / 0.452	0.195 / 0.160	**0.827** / 0.750	0.233 / 0.188	0.777 / 0.704	**0.990** / 0.987	0.691 / 0.665	0.610 / 0.558	0.665 / **0.605**
	F1-score	0.695 / 0.606	0.311 / 0.265	**0.899** / 0.849	0.358 / 0.302	0.734 / 0.700	0.596 / **0.548**	0.609 / 0.503	0.600 / 0.539	0.654 / **0.596**
	AUC	0.774 / 0.722	0.596 / 0.579	**0.911** / 0.872	0.609 / 0.588	0.867 / 0.832	0.843 / 0.811	0.833 / 0.812	0.938 / 0.903	0.949 / **0.916**
Stanford_TOP 3	*#images*	2,331 / 3,814	661 / 1,079	3,368 / 5,419	2,159 / 3,648	1,443 / 2,367	2,380 / 3,066	517 / 808	12,859 / 20,201	
	Precision	0.913 / 0.920	0.759 / 0.758	0.986 / **0.981**	0.789 / 0.765	0.690 / 0.698	0.471 / 0.329	0.589 / 0.414	0.743 / 0.695	0.784 / 0.764
	Recall	0.661 / 0.455	0.253 / 0.162	0.821 / 0.769	0.279 / 0.188	0.831 / 0.710	0.988 / 0.991	0.747 / 0.666	0.654 / 0.563	0.701 / 0.595
	F1-score	0.767 / 0.609	0.379 / 0.267	0.896 / 0.862	0.413 / 0.302	**0.754** / 0.704	0.638 / 0.494	0.659 / 0.511	0.644 / 0.535	0.692 / 0.593
	AUC	0.823 / 0.723	0.624 / 0.580	0.908 / 0.882	0.632 / 0.588	0.891 / 0.834	0.867 / **0.814**	0.862 / 0.813	0.953 / 0.905	0.961 / 0.916
Stanford_TOP 2	*#images*	1,553 / 3,761	440 / 1,068	2,234 / 3,839	1,429 / 3,483	959 / 2,335	1,587 / 2,048	345 / 717	8,547 / 17,251	
	Precision	0.894 / 0.921	0.763 / 0.771	0.984 / 0.979	0.801 / 0.765	0.666 / 0.724	0.520 / 0.262	0.592 / 0.483	**0.746** / **0.701**	**0.788** / 0.770
	Recall	0.748 / 0.461	0.359 / 0.164	0.795 / **0.818**	0.324 / 0.195	0.842 / 0.719	0.986 / 0.991	**0.754** / 0.658	0.687 / 0.572	0.724 / 0.575
	F1-score	0.815 / 0.614	0.488 / 0.270	0.879 / **0.891**	0.461 / 0.311	0.744 / 0.722	0.681 / 0.414	**0.663** / 0.557	0.676 / **0.540**	0.717 / 0.582
	AUC	0.864 / 0.725	0.677 / 0.580	0.895 / **0.907**	0.654 / 0.590	0.894 / 0.838	0.889 / 0.808	**0.866** / 0.814	0.963 / 0.904	0.968 / 0.908
Stanford_TOP 1	*#images*	773 / 3,697	219 / 1,042	1,110 / 1,921	711 / 2,417	477 / 2,236	795 / 1,023	172 / 463	4,257 / 12,799	
	Precision	0.855 / **0.925**	**0.779** / **0.799**	**0.989** / 0.966	**0.816** / 0.751	0.612 / **0.767**	**0.575** / 0.184	0.559 / 0.467	0.741 / 0.694	0.788 / **0.784**
	Recall	**0.814** / **0.468**	0.434 / 0.168	0.730 / 0.791	**0.368** / **0.257**	**0.881** / **0.743**	0.974 / **0.997**	0.738 / 0.590	**0.705** / **0.573**	**0.732** / 0.547
	F1-score	**0.834** / **0.622**	**0.557** / **0.278**	0.840 / 0.869	**0.508** / **0.382**	0.722 / **0.755**	**0.723** / 0.310	0.637 / 0.521	**0.689** / 0.534	**0.726** / 0.580
	AUC	**0.892** / **0.726**	**0.714** / **0.582**	0.863 / 0.893	**0.676** / **0.618**	**0.905** / **0.848**	**0.904** / 0.806	0.857 / 0.782	**0.969** / **0.909**	**0.971** / 0.907

f-AnoGAN [133] architectures. Their results are present with the style of CVAD's/f-AnoGAN's.

Table. 5.2 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 performances are reported, including *Precision*, *Recall*, *F1-score* and *AUC* scores. 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 and f-AnoGAN version are the same. For simplicity, only one version is present (see Table. 5.2 *Row* **Stanford_ MURA_TOP 5**). The total number of images being evaluated on is listed in the row *#images* for each class. The amount of data samples in the dropped group is the number difference between the adjacent groups. 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 ($13 = 3851 - 3838$). As can be observed in the table, 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 both CVAD and f-AnoGAN are growing, 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. Although the same trend is noted for both CVAD and f-AnoGAN versions in general, the CVAD version can get more increase in performance after expelling the most anomalous group than the 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. 5.3.

Dataset quality measurement results

We report the Stanford **MURA** dataset quality in the top left of Figure 5.10 calculated via the OTDD metric (i.e., Eqn. 5.4 and Eqn. 5.5). 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 both CVAD and f-AnoGAN anomaly detection architectures. 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) shortens the distance more and faster than the f-AnoGAN (in orange) version. The general external dataset quality achieves the best when it is composed by the group with the lowest anomaly scores, which follows the same conclusion as the average classification performance in Table 5.2.

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 Figure 5.10. To compare, we show the two chest X-ray datasets (CheXpert and MIMIC) class-wise quality obtained by both the CVAD and f-AnoGAN 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 f-AnoGAN version. 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_f-AnoGAN* model.

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 Figure 5.10, 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, and the quality improvement is limited when little *shiftness* exists in the external dataset. Depending on the quality expectations, users can decide to remain the original *Fracture* class or remove one or two top groups from *Fracture*.

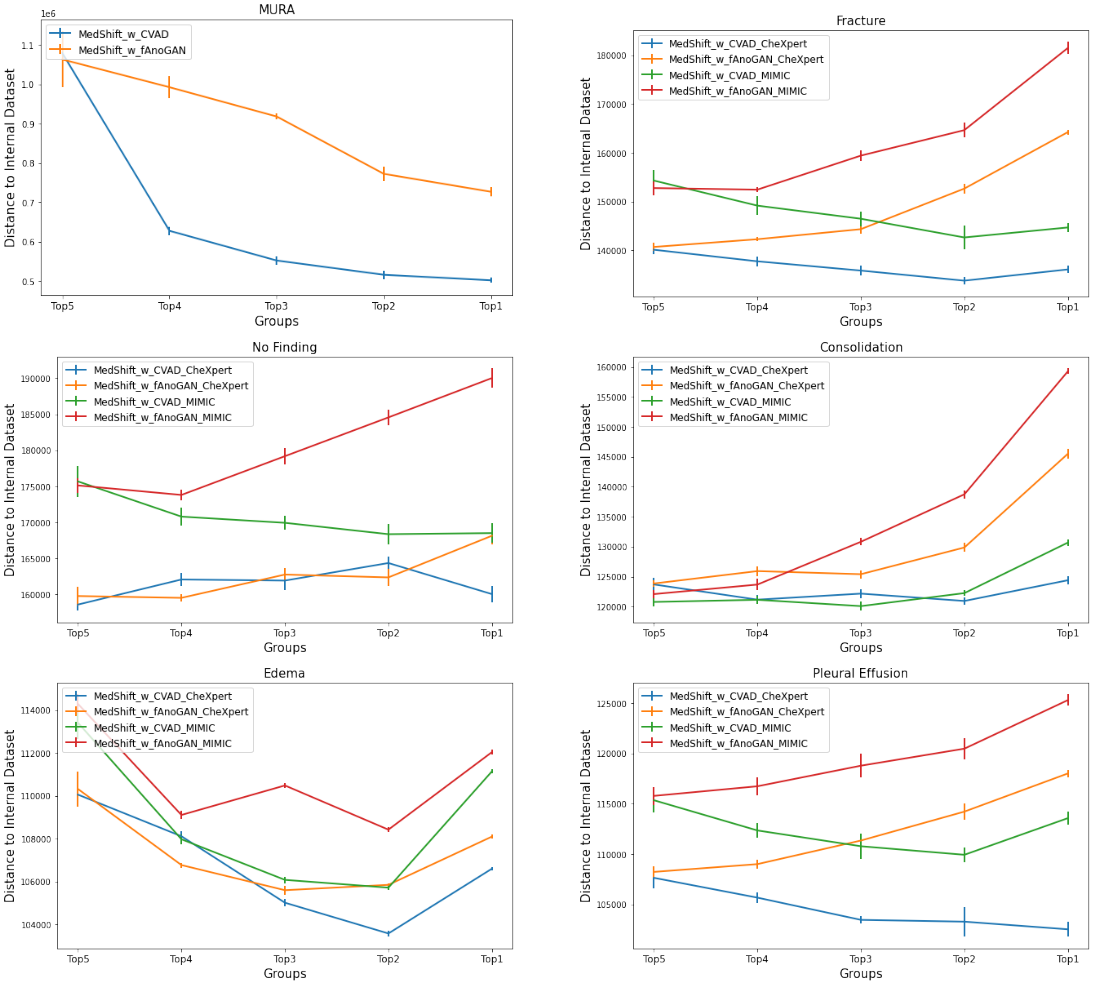


Figure 5.10: Dataset quality measurement results. From left to right, top to bottom, there are Stanford MURA whole dataset's quality, CheXpert and MIMIC *Fracture*, *No Finding*, *Consolidation*, *Edema* and *Pleural Effusion* class quality. X-axis values represent situations of the groups in use, and 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.

5.4 Conclusion

In this paper, 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 in an automatic way. The anomaly detection architecture involves two separate implementation phases - (1) internal training - time consuming and needs to trained for each targeted class labels, 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 any 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.

5.5 Discussions and future works

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 two different anomaly detector CVAD and fAnoGAN. Even though our CVAD version efficiently shortens the data quality matrix (OTDD) faster than f-AnoGAN and reaches convergence for the shift data removal by dropping lower number of cases from external data, the targeted final classification performance stays similar for both architectures.

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 represent the fact that the *MedShift* can identify relevant shift data that will degrade the performance of an in-domain model and able to reproduce the internal performance on an unseen external data. Moreover, the brief cluster exploration on the external dataset showed that higher anomaly cluster groups contains more variations in terms of image quality, positioning, noise, and the pipeline correctly identified the shift data. 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 matrix.

In it's current state, the proposed pipeline *MedShift* can server for domain-specific quality checks and derive powerful and actionable insights. The suggested workflow will be beneficial in future non-shareable healthcare collaboration where the *MedShift* pipeline will be setup as a browser-based service within the local firewall for automated dataset curation with multi-class labels.

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

Table 5.3: ChestXpert classification class-wise results.

Dataset		Metrics	No Finding	Enlarged Cardiomediastinum	Cardiomegaly	Lung Lesion	Lung Opacity	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural Other	Fracture	Support Devices	AVERAGE			
																	Micro	Macro	Weighted	Samples
Emory_CXR	**Test**	*#images*	7,936	522	1,256	397	2,141	830	151	439	2,315	150	711	98	177	9,994				
		Precision	0.753	0.082	0.385	0.090	0.531	0.416	0.107	0.153	0.577	0.066	0.516	0.066	0.058	0.93	0.489	0.338	0.705	0.633
		Recall	0.960	0.232	0.697	0.446	0.588	0.627	0.437	0.490	0.580	0.433	0.686	0.520	0.277	0.109	0.514	0.506	0.514	0.537
		F1-score	0.844	0.121	0.496	0.150	0.558	0.500	0.171	0.233	0.579	0.115	0.589	0.118	0.096	0.195	0.502	0.340	0.477	0.54
		AUC	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	0.710	0.698	0.719	0.661
TOP 5	**CheXpert**	*#images*	22,381	10,798	27,000	9,186	10,5581	52,246	14,783	6,039	33,376	19,448	86,187	3,523	9,040	116,001	212,273			
		Precision	0.356	0.098	0.319	0.081	0.581	0.462	0.122	0.055	0.196	0.118	0.623	0.033	0.091	0.670	0.343	0.272	0.479	0.351
		Recall	0.684	0.073	0.610	0.610	0.832	0.625	0.556	0.495	0.818	0.884	0.835	0.461	0.173	0.845	0.752	0.607	0.752	0.722
		F1-score	0.468	0.084	0.419	0.143	0.684	0.531	0.200	0.098	0.316	0.208	0.714	0.061	0.120	0.747	0.471	0.342	0.566	0.446
		AUC	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	0.726	0.664	0.724	0.709
	MIMIC	*#images*	143,352	10,042	64,346	10,801	76,423	36,564	14,675	26,222	65,047	14,257	76,957	3,460	7,605	84,073				
		Precision	0.674	0.040	0.412	0.060	0.318	0.358	0.096	0.133	0.315	0.084	0.564	0.028	0.036	0.502	0.315	0.259	0.433	0.4
		Recall	0.754	0.078	0.440	0.539	0.740	0.619	0.543	0.428	0.754	0.719	0.700	0.379	0.158	0.739	0.661	0.542	0.661	0.661
		F1-score	0.712	0.053	0.425	0.108	0.445	0.454	0.164	0.202	0.445	0.151	0.624	0.052	0.058	0.598	0.427	0.321	0.504	0.464
		AUC	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	0.731	0.667	0.727	0.712
TOP 4	**CheXpert**	*#images*	17,834 / 4,845	8,899 / 4,282	22,846 / 11,259	7,518 / 2,609	88,278 / 39,924	45,239 / 21,838	12,217 / 5,845	4,934 / 2,106	28,414 / 13,719	15,609 / 6,520	73,595 / 35,301	3,141 / 1,079	7,464 / 3,325	97,280 / 45,373	173,664 / 68,757			
		Precision	0.382 / 0.327	0.099 / 0.112	0.315 / 0.369	0.085 / 0.078	0.583 / 0.645	0.463 / 0.528	0.119 / 0.138	0.054 / 0.059	0.199 / 0.234	0.115 / 0.120	0.626 / 0.693	0.037 / 0.031	0.098 / 0.120	0.670 / 0.759	0.351 / 0.395	0.275 / 0.301	0.483 / 0.545	0.358 / 0.399
		Recall	0.640 / 0.638	0.071 / 0.064	0.624 / 0.631	0.602 / 0.594	0.848 / 0.853	0.647 / 0.660	0.553 / 0.562	0.505 / 0.519	0.833 / 0.848	0.884 / 0.902	0.855 / 0.858	0.469 / 0.450	0.157 / 0.146	0.864 / 0.865	0.765 / 0.772	0.611 / 0.614	0.765 / 0.772	0.737 / 0.750
		F1-score	0.478 / 0.433	0.083 / 0.082	0.419 / 0.466	0.150 / 0.138	0.691 / 0.735	0.540 / 0.587	0.196 / 0.222	0.098 / 0.106	0.321 / 0.366	0.204 / 0.212	0.723 / 0.767	0.069 / 0.059	0.121 / 0.132	0.755 / 0.808	0.481 / 0.522	0.346 / 0.365	0.573 / 0.618	0.457 / 0.498
		AUC	0.760 / 0.769	0.518 / 0.515	0.709 / 0.710	0.655 / 0.659	0.611 / 0.602	0.691 / 0.693	0.622 / 0.618	0.623 / 0.629	0.589 / 0.577	0.607 / 0.605	0.740 / 0.728	0.623 / 0.614	0.546 / 0.546	0.661 / 0.666	0.729 / 0.733	0.640 / 0.638	0.715 / 0.722	0.661 /0.656
	MIMIC	*#images*	115,457 / 25,060	8,503 / 4,053	53,833 / 24,040	9,033 / 2,629	63,674 / 26,173	30,636 / 14,797	12,159 / 6,146	21,621 / 8,548	54,894 / 25,083	11,529 / 4,943	64,695 / 30,863	2,915 / 1,192	6,223 / 2,236	70,786 / 34,346	290,657 / 89,546			
		Precision	0.675 / 0.634	0.042 / 0.057	0.413 / 0.503	0.062 / 0.062	0.322 / 0.386	0.359 / 0.422	0.097 / 0.135	0.133 / 0.163	0.323 / 0.407	0.084 / 0.107	0.565 / 0.653	0.029 / 0.039	0.036 / 0.048	0.506 / 0.643	0.319 / 0.374	0.260 / 0.304	0.433 / 0.476	0.399 / 0.421
		Recall	0.749 / 0.726	0.082 / 0.070	0.451 / 0.472	0.540 / 0.510	0.739 / 0.786	0.645 / 0.675	0.542 / 0.556	0.431 / 0.471	0.756 / 0.809	0.704 / 0.715	0.716 / 0.750	0.381 / 0.400	0.158 / 0.140	0.749 / 0.764	0.666 / 0.681	0.546 / 0.560	0.666 / 0.681	0.663 / 0.673
		F1-score	0.710 / 0.677	0.056 / 0.063	0.431 / 0.487	0.111 / 0.111	0.449 / 0.518	0.461 / 0.519	0.165 / 0.217	0.203 / 0.242	0.453 / 0.541	0.150 / 0.186	0.632 / 0.698	0.055 / 0.072	0.058 / 0.072	0.604 / 0.698	0.431 / 0.483	0.324 / 0.364	0.507 / 0.543	0.464 / 0.487
		AUC	0.755 / 0.782	0.513 / 0.508	0.653 / 0.650	0.639 / 0.638	0.651 / 0.635	0.754 / 0.746	0.661 / 0.646	0.602 / 0.608	0.694 / 0.675	0.693 / 0.683	0.779 / 0.770	0.627 / 0.634	0.532 / 0.535	0.757 / 0.750	0.727 / 0.726	0.665 / 0.661	0.724 / 0.720	0.710 / 0.702
TOP 3	**CheXpert**	*#images*	13,296 / 5,030	6,986 / 4,554	18,241 / 11,723	5,773 / 2,968	68,914 / 41,990	35,897 / 22,654	9,479 / 6,006	3,781 / 2,455	22,396 / 14,376	11,697 / 7,688	58,410 / 36,948	2,498 / 1,249	5,796 / 3,788	75,410 / 48,873	133,375 / 73,685			
		Precision	0.394 / 0.302	0.104 / 0.113	0.315 / 0.367	0.089 / 0.083	0.586 / 0.638	0.462 / 0.524	0.118 / 0.133	0.054 / 0.064	0.202 / 0.229	0.114 / 0.132	0.630 / 0.681	0.041 / 0.033	0.100 / 0.115	0.667 / 0.762	0.357 / 0.392	0.277 / 0.298	0.485 / 0.539	0.364 / 0.394
		Recall	0.607 / 0.623	0.075 / 0.067	0.638 / 0.628	0.582 / 0.597	0.856 / 0.853	0.669 / 0.660	0.548 / 0.552	0.515 / 0.514	0.842 / 0.844	0.876 / 0.899	0.866 / 0.851	0.467 / 0.441	0.130 / 0.132	0.875 / 0.861	0.772 / 0.768	0.610 / 0.609	0.772 / 0.768	0.745 / 0.745
		F1-score	0.478 / 0.407	0.087 / 0.084	0.421 / 0.463	0.154 / 0.146	0.696 / 0.730	0.547 / 0.584	0.194 / 0.214	0.097 / 0.114	0.326 / 0.360	0.202 / 0.231	0.729 / 0.756	0.075 / 0.061	0.113 / 0.123	0.757 / 0.809	0.488 / 0.519	0.348 / 0.363	0.577 / 0.612	0.464 / 0.493
		AUC	0.752 / 0.759	0.520 / 0.516	0.709 / 0.711	0.656 / 0.660	0.604 / 0.606	0.691 / 0.697	0.617 / 0.616	0.625 / 0.628	0.585 / 0.577	0.610 / 0.607	0.735 / 0.725	0.629 / 0.609	0.538 / 0.538	0.653 / 0.666	0.732 / 0.732	0.637 / 0.637	0.719 / 0.720	0.656 / 0.656
	MIMIC	*#images*	86,898 / 25,331	6,641 / 4,228	41,690 / 24,530	6,906 / 3,022	48,975 / 27,226	24,134 / 15,108	9,268 / 6,248	16,549 / 9,331	43,116 / 25,693	8,690 / 5,759	50,432 / 31,506	2,282 / 1,199	4,758 / 2,884	55,161 / 35,891	220,463 / 92,741			
		Precision	0.681 / 0.618	0.045 / 0.059	0.408 / 0.494	0.064 / 0.070	0.325 / 0.385	0.359 / 0.414	0.098 / 0.131	0.133 / 0.167	0.329 / 0.401	0.086 / 0.120	0.567 / 0.639	0.033 / 0.039	0.037 / 0.054	0.509 / 0.641	0.326 / 0.371	0.262 / 0.302	0.435 / 0.466	0.404 / 0.413
		Recall	0.741 / 0.712	0.090 / 0.070	0.472 / 0.465	0.528 / 0.522	0.739 / 0.785	0.677 / 0.681	0.541 / 0.556	0.432 / 0.461	0.766 / 0.813	0.686 / 0.719	0.740 / 0.750	0.380 / 0.405	0.149 / 0.128	0.764 / 0.767	0.673 / 0.678	0.550 / 0.560	0.673 / 0.678	0.668 / 0.668
		F1-score	0.710 / 0.662	0.060 / 0.064	0.437 / 0.479	0.114 / 0.123	0.452 / 0.516	0.469 / 0.515	0.166 / 0.212	0.204 / 0.245	0.461 / 0.537	0.153 / 0.206	0.642 / 0.691	0.060 / 0.070	0.059 / 0.076	0.611 / 0.698	0.439 / 0.480	0.328 / 0.364	0.511 / 0.535	0.469 / 0.480
		AUC	0.757 / 0.773	0.515 / 0.508	0.656 / 0.647	0.639 / 0.643	0.650 / 0.632	0.764 / 0.747	0.661 / 0.645	0.602 / 0.602	0.693 / 0.674	0.694 / 0.685	0.786 / 0.766	0.631 / 0.637	0.531 / 0.528	0.759 / 0.748	0.731 / 0.723	0.667 / 0.660	0.727 / 0.716	0.712 / 0.698
TOP 2	**CheXpert**	*#images*	8,808 / 5,029	4,911 / 4,668	13,019 / 11,951	3,927 / 4,193	47,978 / 43,264	25,529 / 22,804	6,498 / 6,155	2,585 / 2,639	15,707 / 14,707	7,841 / 8,602	41,341 / 37,688	1,754 / 1,700	4,022 / 3,803	51,977 / 50,012	91,428 / 76,520			
		Precision	0.406 / 0.270	0.109 / 0.098	0.313 / 0.360	0.092 / 0.111	0.590 / 0.632	0.459 / 0.517	0.115 / 0.130	0.053 / 0.067	0.204 / 0.227	0.115 / 0.142	0.634 / 0.676	0.044 / 0.044	0.106 / 0.104	0.665 / 0.755	0.364 / 0.389	0.279 / 0.295	0.487 / 0.529	0.371 / 0.390
		Recall	0.579 / 0.603	0.079 / 0.057	0.659 / 0.611	0.554 / 0.603	0.865 / 0.848	0.697 / 0.659	0.544 / 0.548	0.524 / 0.519	0.845 / 0.839	0.963 / 0.892	0.876 / 0.843	0.452 / 0.459	0.105 / 0.133	0.884 / 0.855	0.779 / 0.762	0.609 / 0.605	0.779 / 0.762	0.754 / 0.738
		F1-score	0.477 / 0.373	0.092 / 0.072	0.425 / 0.453	0.158 / 0.187	0.701 / 0.725	0.554 / 0.579	0.190 / 0.210	0.096 / 0.118	0.329 / 0.357	0.202 / 0.245	0.736 / 0.750	0.080 / 0.080	0.106 / 0.116	0.759 / 0.802	0.496 / 0.515	0.350 / 0.362	0.582 / 0.603	0.473 / 0.487
		AUC	0.744 / 0.744	0.521 / 0.512	0.710 / 0.705	0.655 / 0.661	0.600 / 0.603	0.690 / 0.699	0.612 / 0.613	0.626 / 0.630	0.581 / 0.579	0.619 / 0.605	0.729 / 0.725	0.630 / 0.616	0.532 / 0.536	0.648 / 0.666	0.736 / 0.729	0.636 / 0.635	0.723 / 0.717	0.653 / 0.655
	MIMIC	*#images*	57,859 / 25,256	4,581 / 4,386	28,853 / 24,588	4,648 / 3,590	33,249 / 27,724	17,193 / 15,187	6,303 / 6,311	11,188 / 10,120	30,295 / 25,939	5,799 / 6,271	34,920 / 31,486	1,563 / 1,502	3,256 / 3,008	38,400 / 36,113	147,980 / 94,532			
		Precision	0.692 / 0.593	0.046 / 0.049	0.399 / 0.491	0.070 / 0.079	0.327 / 0.385	0.359 / 0.413	0.101 / 0.130	0.134 / 0.174	0.336 / 0.396	0.093 / 0.129	0.568 / 0.631	0.037 / 0.045	0.038 / 0.050	0.513 / 0.634	0.338 / 0.366	0.265 / 0.300	0.438 / 0.457	0.414 / 0.405
		Recall	0.731 / 0.698	0.100 / 0.058	0.505 / 0.464	0.492 / 0.523	0.740 / 0.785	0.721 / 0.676	0.545 / 0.553	0.434 / 0.456	0.778 / 0.805	0.658 / 0.720	0.766 / 0.738	0.367 / 0.396	0.120 / 0.120	0.780 / 0.767	0.683 / 0.671	0.553 / 0.554	0.683 / 0.671	0.676 / 0.658
		F1-score	0.711 / 0.641	0.063 / 0.053	0.446 / 0.477	0.123 / 0.138	0.454 / 0.516	0.479 / 0.512	0.170 / 0.210	0.205 / 0.252	0.470 / 0.531	0.163 / 0.218	0.653 / 0.680	0.068 / 0.081	0.058 / 0.071	0.619 / 0.694	0.452 / 0.474	0.334 / 0.363	0.516 / 0.526	0.480 / 0.471
		AUC	0.761 / 0.762	0.517 / 0.502	0.660 / 0.647	0.640 / 0.642	0.650 / 0.632	0.776 / 0.746	0.664 / 0.644	0.603 / 0.598	0.691 / 0.670	0.699 / 0.687	0.793 / 0.761	0.633 / 0.630	0.526 / 0.523	0.761 / 0.747	0.738 / 0.719	0.670 / 0.656	0.732 / 0.711	0.715 / 0.694
TOP 1	**CheXpert**	*#images*	4,390 / 4,940	2,582 / 4,805	7,357 / 11,902	2,000 / 5,579	25,328 / 44,160	13,947 / 22,810	3,390 / 6,248	1,313 / 2,545	8,368 / 14,778	3,930 / 9,632	22,005 / 37,642	899 / 1,787	2,049 / 3,268	26,981 / 50,665	47,411 / 78,209			
		Precision	0.422 / 0.249	0.116 / 0.095	0.319 / 0.358	0.100 / 0.138	0.596 / 0.631	0.453 / 0.515	0.113 / 0.128	0.052 / 0.060	0.206 / 0.222	0.117 / 0.157	0.638 / 0.667	0.050 / 0.044	0.108 / 0.076	0.661 / 0.750	**0.374** / 0.386	0.282 / 0.292	0.489 / 0.523	0.381 / 0.386
		Recall	0.549 / 0.591	0.098 / 0.055	0.689 / 0.606	0.511 / 0.600	0.871 / 0.845	0.731 / 0.663	0.533 / 0.551	0.548 / 0.511	0.839 / 0.824	0.832 / 0.890	0.886 / 0.833	0.445 / 0.465	0.068 / 0.119	0.890 / 0.855	**0.786** / 0.759	0.606 / 0.600	0.786 / 0.759	0.761 / 0.735
		F1-score	0.477 / 0.351	0.106 / 0.069	0.436 / 0.450	0.168 / 0.224	0.707 / 0.722	0.559 / 0.580	0.187 / 0.207	0.095 / 0.108	0.330 / 0.350	0.205 / 0.266	0.741 / 0.741	0.089 / 0.080	0.084 / 0.092	0.759 / 0.799	**0.507** / 0.511	0.353 / 0.360	0.586 / 0.597	0.483 / 0.483
		AUC	0.736 / 0.735	0.527 / 0.510	0.709 / 0.705	0.655 / 0.656	0.596 / 0.602	0.681 / 0.703	0.606 / 0.612	0.632 / 0.621	0.572 / 0.576	0.632 / 0.608	0.725 / 0.723	0.640 / 0.613	0.521 / 0.528	0.644 / 0.665	**0.741** / 0.727	0.634 / 0.633	0.729 / 0.715	0.649 / 0.654
	MIMIC	*#images*	28,790 / 25,017	2,477 / 4,634	15,163 / 24,714	2,329 / 6,930	17,072 / 29,313	9,478 / 15,219	3,241 / 6,644	5,655 / 11,005	16,002 / 25,970	2,901 / 6,464	18,178 / 31,759	780 / 1,550	1,636 / 2,380	20,628 / 35,968	74,374 / 97,347			
		Precision	0.706 / 0.555	0.048 / 0.051	0.390 / 0.483	0.083 / 0.145	0.334 / 0.388	0.351 / 0.412	0.102 / 0.129	0.136 / 0.176	0.341 / 0.384	0.109 / 0.128	0.567 / 0.623	0.046 / 0.042	0.044 / 0.035	0.532 / 0.620	**0.357** / 0.359	0.271 / 0.298	0.442 / 0.442	0.434 / 0.393
		Recall	0.718 / 0.686	0.121 / 0.057	0.564 / 0.456	0.421 / 0.542	0.744 / 0.780	0.774 / 0.679	0.530 / 0.554	0.443 / 0.450	0.785 / 0.784	0.577 / 0.720	0.789 / 0.722	0.335 / 0.398	0.075 / 0.118	0.800 / 0.760	**0.694** / 0.661	0.548 / 0.550	0.694 / 0.661	0.686 / 0.648
		F1-score	0.712 / 0.614	0.068 / 0.054	0.461 / 0.470	0.139 / 0.229	0.461 / 0.518	0.483 / 0.513	0.171 / 0.209	0.208 / 0.253	0.476 / 0.516	0.183 / 0.217	0.660 / 0.669	0.080 / 0.077	0.055 / 0.054	0.639 / 0.683	**0.472** / 0.466	0.343 / 0.362	0.524 / 0.513	0.498 / 0.459
		AUC	0.765 / 0.748	0.519 / 0.502	0.669 / 0.645	0.636 / 0.649	0.651 / 0.625	0.783 / 0.750	0.659 / 0.640	0.606 / 0.591	0.685 / 0.663	0.693 / 0.685	0.797 / 0.755	0.630 / 0.626	0.519 / 0.519	0.765 / 0.747	**0.747** / 0.712	0.670 / 0.653	0.740 / 0.704	0.718 / 0.688

Chapter 6

Medical Image Retrieval with Limited Supervision

6.1 Outlier-sensitive radiography retrieval

With the widespread adoption of radiology in diagnosis and treatment planning, the amount of medical image data is rapidly increasing [67]. Fast and effective retrieval in large-scale medical image repositories has been demanding to support data management, research and clinical applications [141]. One common way to retrieval images is content-based, which has been widely researched and applied to the medical field [153, 42, 31, 29]. For a given query image, a content-based image retrieval (CBIR) system returns a ranked list of images from the database based on a similarity measure between the query and retrieved images [41, 121]. The core idea behind CBIR is to minimize the distance of an anchor image a to its positive counterparts ps and maximize the distance to the corresponding negative images ns in the feature space. Usually, the positive images are in the same class as the anchor image. How-

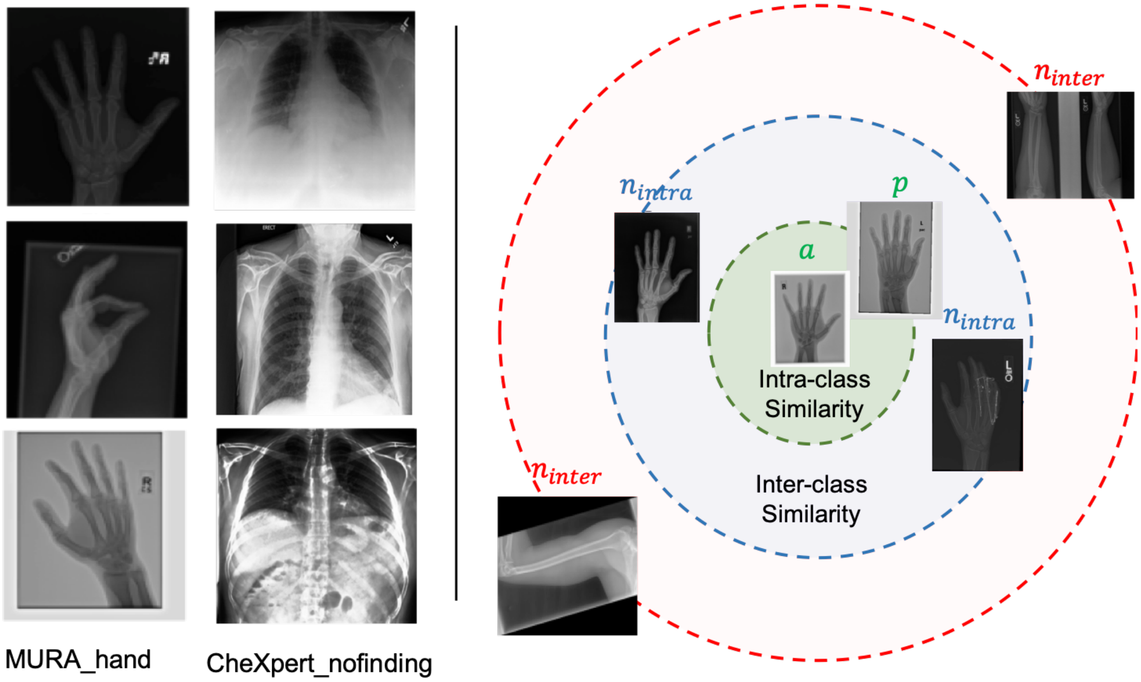


Figure 6.1: (Left) Examples of intra-class variations. 1st column shows samples from Stanford MURA *HAND* class and 2nd column presents CheXpert*No Finding* class data. (Right) Oscars learns the intra-class and inter-class similarity simultaneously. Images with intra-class similarity p should be closer to the given image a than the samples that show inter-class similarity n_{intra} in the feature space.

ever, adopting this strategy can be problematic as it only considers the inter-class variation. The assumption - as long as a and p are from the same class, they show similar visual features - is not realistic as samples from one class often exhibit certain intra-class variations. Noisy, under-represented data can exist, also called *outliers*. This phenomenon is more common in radiology as images are often acquired via different equipment from different sources and varies based on acquisition protocols. These variations, as shown in the left part of Figure 6.1, pose specific challenges in the consumer domain and need to be recognized in assessing image similarity [5].

In this paper, we focus on relevant radiograph image retrieval in external datasets which can contain lots of noisy data compared to the clean internal dataset. Such a system will help to *collect cleaner external image dataset with minimal human effort and accelerate AI evaluation.* To achieve the goal, we propose an **O**utlier-

Sensitive **C**ontent-based r**A**diologhy **R**etrieval **S**ystem (**OSCARS**), which takes both the intra-class and inter-class variations into consideration. To acquire the intra-class variation information, we adopt the unsupervised anomaly detectors trained on the internal dataset and utilize the assigned anomaly scores to the external dataset to split each class into several bins, with each bin in a certain range regarding the anomaly scores. Based on which, we construct the quadruplet data $(a, p, n_{intra}, n_{inter})$ with an anchor image a, a positive image p from the same class and same bin, an intra-class negative image n_{intra} from the same class but different bins, and an inter-class negative image n_{inter} that is from a different class.

With the proposed quadruplet sampling strategy, we incorporate the intra-class discriminative information into the training data and hence improve the retrieval of sensitivity outlier-related queries after model training. All the images in a quadruplet are fed into the feature extractor to learn their latent embeddings $(e_a, e_p, e_{n_{intra}}, e_{n_{inter}})$. As illustrated in the right of Figure 6.1, we then learn the intra-class embedding similarity to achieve $(Sim(e_a, e_p) > Sim(e_a, e_{n_{intra}}))$ with an intra-class triplet loss L_{intra} and the inter-class similarity for $(Sim(e_a, e_{n_{intra}}) > Sim(e_a, e_{n_{inter}}))$ with an inter-class triplet loss L_{inter} in a weighted way.

6.1.1 Contribution

Our summarized contributions are:

1. We introduce the task of outlier-sensitive image retrieval for noisy external medical image dataset and propose an effective image retrieval system **OSCARS** to enhance the relevance of outlier-related results.

2. We propose to acquire intra-class information of external datasets via anomaly detectors trained unsupervised. By training on clean internal datasets, the anomaly detectors assign each sample of the external dataset with a specific

anomaly score. Based on which, we split each class into several bins with different intra-class variations.

3. We sample both the intra-class and inter-class negative images to construct quadruplets for intra-class and inter-class similarity learning.

4. We demonstrate the model effectiveness with two public representative radiography datasets - Stanford Muscoloskeletal Radiography (MURA) [117] and CheXpert [68]. The code is available at `https://github.com/XiaoyuanGuo/oscars`.

Publication:

- Guo, Xiaoyuan, Jiali Duan, Saptarshi Purkayastha, Hari Trivedi, Judy Wawira Gichoya, and Imon Banerjee. 2022. "OSCARS: An Outlier-Sensitive Content-Based Radiography Retrieval System." In Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22). Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/3512527.3531425

6.1.2 Method

Given a clean internal dataset D_I and a noisy external dataset D_E, the external data of class c can contain outliers visually different from the internal class. Therefore, a conventional image retrieval system for the external dataset will be insufficient as it merely treats all the samples from one class as the same without considering the intra-class variations. Thus, the system will lack sensitivity to the outliers, undermining the retrieval accuracy. Our objective is to train an image retrieval model that will prioritize the images with both intra-class and inter-class dissimilarity during retrieval ranking. Figure 6.2 summarizes the whole framework of our model. There are mainly two steps involved. First, we design to learn intra-class information in an unsupervised way (introduced in Sec. 6.1.2). Second, we propose to sample training data that are with intra-class bin information and inter-class information (introduced in Sec. 6.1.2).

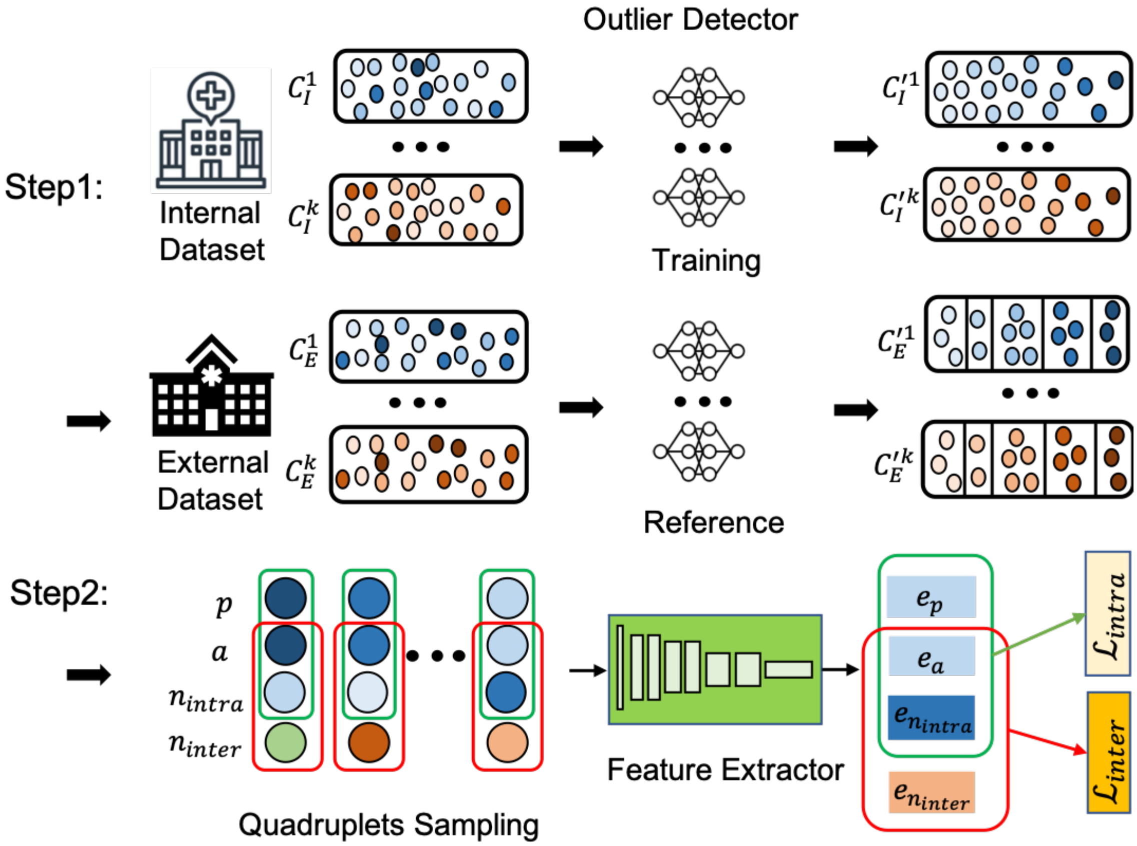


Figure 6.2: OSCARS architecture involves two main steps. Step1: train anomaly detectors on the internal dataset for each class C_I^i; learn clean in-distributions with anomaly scores assigned to $C_I'^i$; apply the trained anomaly detectors on each class C_E^i of the external dataset and split the data into several bins $C_E'^i$ according to the anomaly scores. (Dark colors mean more distribution shifts.) Step2: generate quadruplets $(a, p, n_{intra}, n_{inter})$ by sampling the intra-class positive, negative and inter-class negative simultaneously; learn the intra-class and inter-class similarity in feature space with the intra-class triplet loss L_{intra} and inter-class triplet loss L_{inter}.

With these steps, images with the same labels and similar contents are pulled together by maintaining intra-class similarity.

Learning intra-class information

Due to the difficulties of collecting annotated data with intra-class information provided in the medical domain, the outlier-sensitivity research on medical images has been delayed. To overcome the problem, we propose to generate intra-class labels automatically inspired by a recent work - MedShift [55]. Given a clean internal dataset D_I, MedShift has suggested an approach to identify outliers for noisy external dataset D_E. Following the same steps of MedShift, we first obtain the internal distribution information by training an unsupervised outlier detector named CVAD [54] for each class on the same internal datasets used in [55]. Then, the trained anomaly detectors are evaluated on the external datasets as they have learnt intra-class discriminative features. Thus, each external data has its anomaly score, based on which we split each class into B bins with the K-Means clustering techniques [95, 99]. B (5 in our paper) is determined by the Elbow method [147]. The resulting bins are in different anomaly score ranges. With the data from different bins, we get the intra-class labels. Given that both the intra-class and inter-class labels are available, for each image a, we randomly sample one intra-class positive image p, one intra-class negative sample n_{intra} and one inter-class negative sample n_{inter} accordingly, thus collecting the quadruplets $(a, p, n_{intra}, n_{inter})$ for training.

Balancing the inter- and intra-class influence

With the sampled quadruplets data, we feed each of the image to a CNN-based feature extractor to acquire latent embeddings $(e_a, e_p, e_{n_{intra}}, e_{n_{inter}})$. For simplicity, we adopt the ResNet18 [58] pre-trained on ImageNet [37] as the network backbone. OSCARS is designed to consider both the inter-class similarity and the intra-class similarity at the same time, which brings the model advantages of acquiring the sensitivity of intra-class outlier relevance during image retrieval. However, balancing the effect of the two parts is a challenging problem. Too much weight on intra-class information will distract the general retrieval accuracy of inter-class data. Therefore, we design an intra-class triplet margin loss and an inter-class triplet margin loss to optimize the model architecture. To balance the influence of intra-class and intra-class information on final ranking, we adopt a weighted loss formulated as:

$$\begin{aligned}\mathcal{L} &= \lambda\mathcal{L}_{intra}(e_a, e_p, e_{n_{intra}}) + (1-\lambda)\mathcal{L}_{inter}(e_a, e_{n_{intra}}, e_{n_{inter}}) \\ &= \lambda(max\{d(e_a, e_p) - d(e_a, e_{n_{intra}}) + \mathcal{M}_{intra}, 0\}) \\ &+ (1-\lambda)(max\{d(e_a, e_{n_{intra}}) - d(e_a, e_{n_{inter}}) + \mathcal{M}_{inter}, 0\})\end{aligned} \tag{6.1}$$

where $d(x_i, y_i) = \|x_i - y_i\|_2$. λ, $\mathcal{M}_{intra}$ and $\mathcal{M}_{inter}$ are set as 0.05, 1 and 2 in our experiments respectively.

When we have a query image unseen during training, we first acquire the query representation with the trained image feature backbone and then compute the cosine similarity between the representative features of the query image and dataset images. Images are ranked based on the similarity scores in the descending order.

6.1.3 Experiments

We have evaluated our approach on two publicly available large-scale radiograph image datasets. The first is Stanford MURA dataset, a large dataset of bone X-rays, which contains seven classes - *HAND*, *FORARM*, *FIGER*, *SHOULDER*, *ELBOW*, *WRIST*, *HUMERUS*. The second is CheXpert dataset, which in total has 14 classes - *No Finding*, *Enlarged Cardiomediastinum*, *Cardiomegaly*, *Lung Lesion*, *Lung Opacity*, *Edema*, *Consolidation*, *Pneumonia*, *Atelectasis*, *Pneumothorax*, *Pleural Effusion*, *Pleural Other*, *Fracture*, *Support Devices*. As the chest x-ray images are with two views - frontal and lateral. We here only use frontal view and leave the lateral view for future studies. See more details in the supplementary materials.

Evaluation Metrics: For the retrieval task, we report the retrieval *recall* at rank K ($R@K$, $K \in \{1, 5, 10, 50, 100\}$), *precision* at rank K ($P@K$, $K \in \{1, 5, 10, 50, 100\}$), outlier sensitivity ($S@K$, $K \in \{1, 5, 10, 50, 100\}$). The metric *recall* is the percentage of relevant images retrieved over the total number of retrieved images, defined as $recall = \frac{N_R}{K}$ where R represents the relevant images retrieved. The metric *precision* is assigned based on the existence of the same labels between the query image and the retrieved images. If $\delta(\cdot) \in \{0,1\}$ is an indicator function, the *precision* is defined as $precision = \frac{\sum_{i=1}^{K}\delta(R^i>0)}{K}$. Additionally, we evaluate the outlier sensitivity by calculating the anomaly score difference with $sensitivity = \sum_{i=1}^{N_R}\frac{|\mathcal{A}_R^i - \mathcal{A}_q|}{N_R}$, where $\mathcal{A}$ means anomaly score. We scale the anomaly scores of MURA dataset into [0,1] with the sigmoid function due to the large variations of its anomaly scores.

Implementation Details: The pipelines are developed using Pytorch 1.9.0, Python 3.7.3 and Cuda compilation tools V11.4 on a machine with 4 NVIDIA Quadro RTX A6000 GPUs with 48GB memory. The training for all the models is run for 50 epochs with a start learning rate 0.001 and a SGD optimizer.

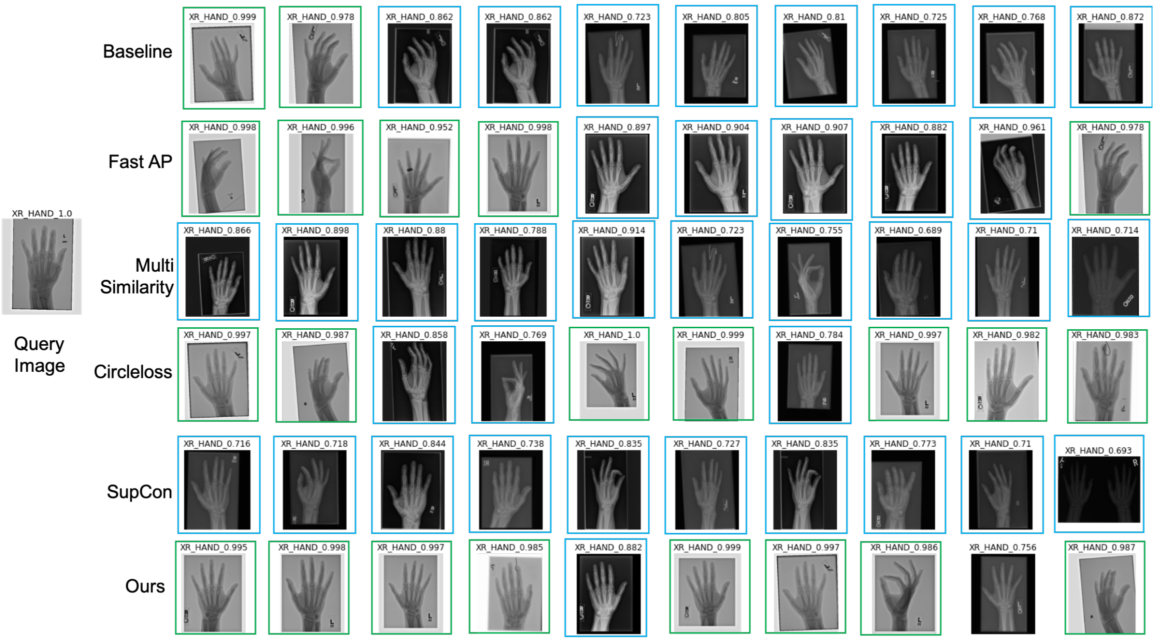


Figure 6.3: Hand results, left is the query image, right shows retrieval results. Green boxes mean both intra- and inter-class correct; blue boxes are for inter-class correct predictions. Each retrieval image has its label on top of itself. For correct predictions, we also put the anomaly scores on them. Closer anomaly scores mean more similarity.

Search Results As a representative image retrieval method with triplet data in training, we select DeepRank as our baseline. State-of-the-art CBIR approaches including FastAP, MultiSimilarity, CircleLoss and SupCon are used to compare the model performance. Notably, we keep the feature extractor consistent for all the

Table 6.1: Quantitative performance on Stanford MURA and CheXpert datasets . Bold indicates the best.

Method	MURA										CheXpert									
	R@1↑	R@5↑	R@10↑	R@50↑	R@100↑	P@1↑	P@5↑	P@10↑	P@50↑	P@100↑	R@1↑	R@5↑	R@10↑	R@50↑	R@100↑	P@1↑	P@5↑	P@10↑	P@50↑	P@100↑
DeepRank [153]	0.912	0.914	0.912	0.906	0.903	0.912	0.964	0.972	0.984	0.988	0.734	0.694	0.716	0.721	0.442	0.734	0.911	0.961	1.000	1.000
FastAP [20]	0.927	0.930	0.931	0.932	0.933	0.927	0.956	0.961	0.973	0.977	0.734	0.742	0.733	0.716	0.710	0.734	0.943	0.968	1.000	1.000
MultiSimilarity [157]	0.923	0.921	0.919	0.915	0.913	0.923	0.955	0.968	0.977	0.980	0.695	0.680	0.677	0.676	0.682	0.695	**0.957**	**0.975**	1.000	1.000
CircleLoss[144]	0.929	0.932	0.933	0.934	0.934	0.929	0.960	0.964	0.979	0.985	0.727	0.703	0.718	0.717	0.726	0.727	0.936	0.968	1.000	1.000
SupCon [71]	0.930	**0.933**	**0.936**	**0.938**	**0.937**	0.930	0.964	0.971	0.981	0.985	0.776	0.730	0.720	0.734	0.726	0.776	0.936	0.950	1.000	1.000
OSCARS *(ours)*	**0.931**	0.922	0.920	0.913	0.910	**0.931**	**0.965**	**0.974**	**0.986**	**0.991**	**0.787**	**0.763**	**0.747**	**0.745**	**0.743**	**0.787**	0.908	0.950	1.000	1.000

methods to ensure fair comparisons.

Quantitative Results: Table 6.1 presents the recall and precision performance for both Stanford MURA and CheXpert datasets respectively. Since the data in CheXpert can have multiple labels, we calculate the correct hit with the strategy - loose match, which means that for a query chest X-ray with multiple labels, a retrieval image is relevant as long as it has one label matched. Compared to the baseline DeepRank, Oscars can enhance the recall and precision performances on both datasets and achieve the best recall at 1 and precision at 1. In general, SupCon has the highest recall for MURA dataset. Nonetheless, Oscars achieves the best precision for MURA dataset and recall for CheXpert. Additionally, we report the sensitivity results in the supplementary materials.

Qualitative Results: Figure 6.3 shows an example of a *HAND* query image in MURA dataset. The corresponding retrieval results including ours are present in different rows. As can be seen, although many methods can achieve high recall and precision (see Table 6.1), they fail to distinguish the intra-class variations. Especially, MultiSimilarity and SupCon exhibit little sensitivity to the noisy query. Comparatively, our method can prioritize intra-class similarity and rank the images with similar anomaly semantics ahead. Please refer to the supplementary materials for more results.

Impact of Lambda: We also explore the impacts of applying different λ values to the loss function (Eqn. 6.1). A good balance between the intra-class and inter-class information will enable the retrieval system to acquire both accurate inter-class and outlier-sensitive intra-class results. Figure 6.4 illustrates the performance variations

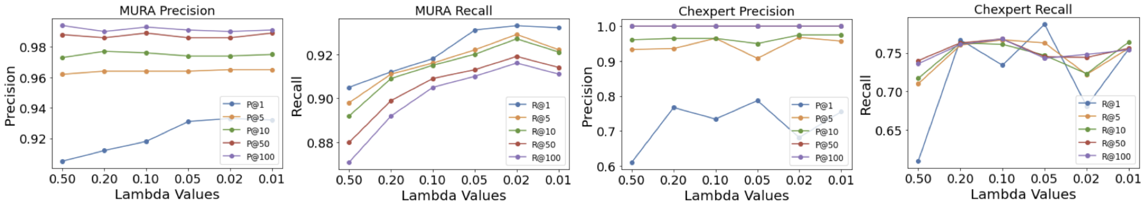


Figure 6.4: Effects with different lambda values on different datasets.

in different datasets under different settings. λ decides how the model learns to weight the intra-class and inter-class information simultaneously. We observe that too much weight on the intra-class similarity will degrade the inter-class similarity predictions. Experiments suggest 0.05 can work well.

6.1.4 Conclusion

In this work, we propose an outlier-sensitive radiography image retrieval system **OSCARS**, which goes beyond retrieving images with the most inter-class similarity but also inspects the intra-class similarity implicitly when query images show certain variations. Utilizing the automatic learning of clean internal distribution, the intra-class variations of external sources are captured and used to generate intra-class labels by splitting the class into several groups. Feeding the sampled quadruplets consisting of both the intra, inter-class positive and negative samples to the image feature learner, a weighted margin loss is adopted to optimize the retrieval network. The resulting retrieval system is sensitive to outlier-related queries as it has learnt to rank the retrieved results based on both intra-class and inter-class similarities. This outlier-sensitive image retrieval approach provides clinical users the access to receive more relevant medical images and allow radiologists to process and analyze radiography images more effectively.

6.2 Multi-label medical image retrieval

Multi-label image retrieval is a challenging problem in the medical area. First, compared to natural images, labels in the medical domain exhibit higher class-imbalance and much nuanced variations. Second, pair-based sampling for positives and negatives during similarity optimization are ambiguous in the multi-label setting, as samples with the same set of labels are limited. To address the aforementioned challenges, we propose a proxy-based multi-class similarity (**PMS**) framework, which compares and contrasts samples by comparing their similarities with the discovered proxies. In this way, samples of different sets of label attributes can be utilized and compared indirectly, without the need for complicated sampling. **PMS** learns a class-wise feature decomposition and maintains a memory bank for positive features from each class. The memory bank keeps track of the latest features, used to compute the class proxies. We compare samples based on their similarity distributions against the proxies, which provide a more stable mean against noise. We benchmark over 10 popular metric learning baselines on medical datasets and experiments show consistent stability of our approach under both exact and non-exact match settings.

In light of recent progress of AI applications in medical image domain, building large data collection for training and testing has become a necessity among institutions and hospitals. One commonly-desired application is content based image retrieval (CBIR). Given a query image, the goal is to retrieve a ranked list of images from the database based on a certain similarity measure [41]. An efficient image retrieval system can help accelerate image annotations, disease diagnosis and history queries, etc.

In this paper, we address the challenges by proposing a **P**roxy-based **M**ulti-**S**imilarity (PMS) framework and benchmark against representative metric learning approaches on two commonly-used medical datasets MIMIC-CXR [69] and CheXpert [68]. Figure 6.5 explains the core idea of our method. Instead of comparing sam-

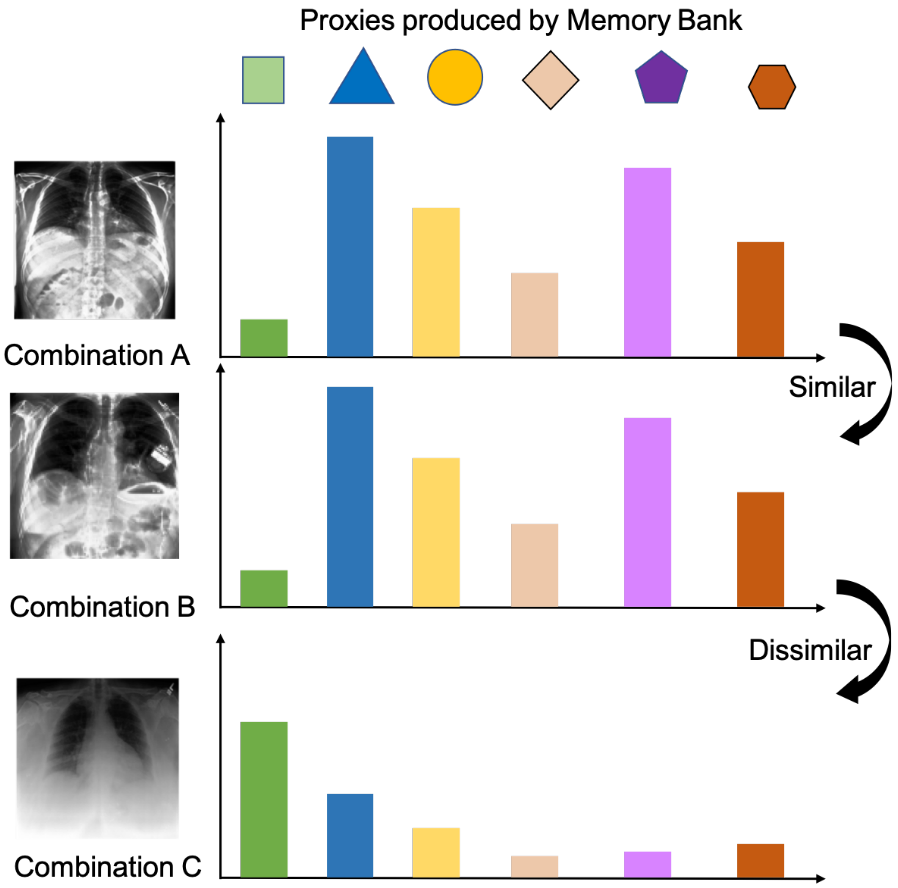


Figure 6.5: We propose to compare multi-label samples via their similarity distributions with respect to the proxies, produced by the memory bank. Each proxy is updated based on a set of latest activated-class features from each sample, designed to be robust to noise.

ples directly, where each sample exhibits a combination of multi-labels, we compare them indirectly by comparing their similarity distributions against the class proxies. In other words, we decompose multi-label representations into a linear combination of single-label representations via proxies. These proxies can be interpreted as class centroids and we enforce the decomposed feature representations corresponding to the class to be close to its centroid while far from other centroids. We dynamically adjust the class proxies based on the votings in the memory bank and update the memory bank in a First-In-First-Out (FIFO) manner. Our approach jointly optimizes the feature backbone, feature decomposition and feature-centroid metric in an unified framework as in Figure 6.6. During inference, we fix the memory bank while using the computed proxies to compute multi-class similarity.

We evaluate our approach against competitive baselines under full (exact-match: a hit requires matching all class labels) and partial (non-exact match: at least one label matches) settings on two benchmarks. Our approach shows a significant margin on the MIMIC_partial and CheXpert_full and is more stable compared to other methods.

6.2.1 Contribution

The core contributions of this work are summarized as follows,

1. We propose an unique solution for multi-label medical image retrieval, by addressing the complexity of label imbalance and class variations (Section 6.2.2).

2. We propose a proxy-based multi-class similarity metric (PMS), where multi-label samples are compared based on their similarity distributions against the proxies (Section 6.2.2). This obviates the need for complicated positive/negative sampling, as pairs with the same set of labels are limited.

3. We benchmark over 10 popular metric learning baselines, both real-valued and hash-based, on two common medical datasets (Section 6.2.3).

Publication:

- Guo, Xiaoyuan, Jiali Duan, Judy Gichoya, Hari Trivedi, Saptarshi Purkayastha, Ashish Sharma, and Imon Banerjee. "Multi-Label Medical Image Retrieval Via Learning Multi-Class Similarity." AI in Medicine, 2022, (under review).

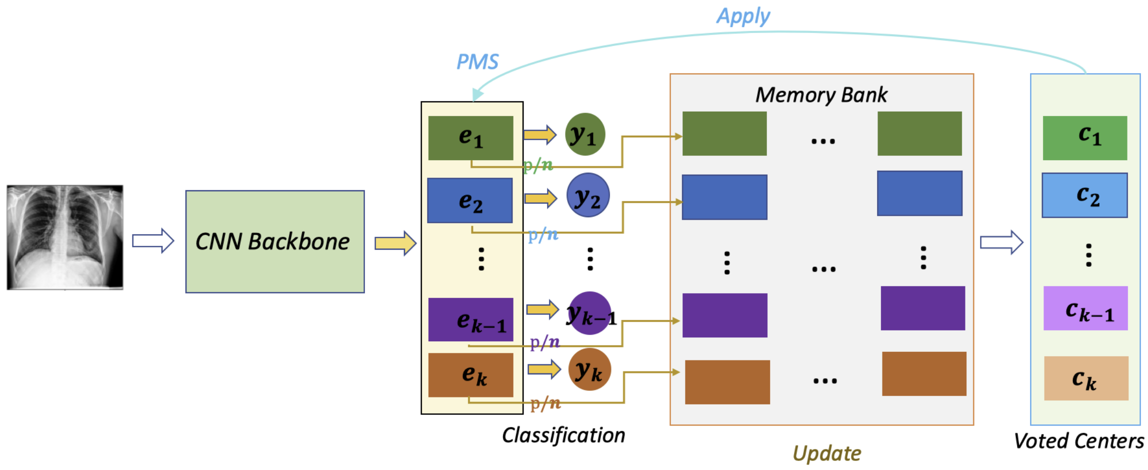


Figure 6.6: We train PMS via two stages. In the first stage, we warm up the memory bank by queuing in the decomposed features corresponding to the activated-classes from each sample (Section 6.2.2). The memory bank in turn votes class proxies (Section 6.2.2) which will be used in the second stage: proxy-based multi-class similarity learning (Section 6.2.2). During inference, the proxies are used to compute multi-class similarity between the image and candidates for retrieval (Section 6.2.2).

6.2.2 Method

Framework Overview

Figure 6.6 illustrates the architecture of PMS, which mainly consists of a feature backbone, a multi-branch feature decomposition and a memory bank. In experiments, we use ResNet18 [163] as our representation learning backbone for fair comparison with other baselines. However, the feature backbone is model agnostic, thus any deep learning architecture can be used for feature extraction. The proposed loss operates on features from the penultimate layer of the network and discriminates them using the class-specific classification heads. We train PMS into two phases - (i) We warm-up the model for several epochs by optimizing multi-label classification loss. In this stage, decomposed positive features will be used to queue up the memory bank for producing meaningful class proxies. Let $X = \{x_i\}_{i=1}^{N}$ and $Y = \{y_1, y_2, ..., y_k\}$ denote the images and the associated label space respectively, k is the total number of classes. Each x_i is associated with a subset of labels $y_i \in Y$. For decomposed class embeddings $e = \{e_1, e_2, ..., e_{k-1}, e_k\}$, only embeddings corresponding to positive classes (i.e., $y_i = 1$) will be stored into the memory bank. (ii) The proxies will be used to update the feature representations as introduced in Section 6.2.2.

Proxy Computation

Memory bank update: We keep a record of latest activated-class features from each sample in a memory bank $\mathcal{B}$. Suppose the dataset has k classes in total, for each class i, we cache M samples. Therefore, a memory bank is a $k \times M$ matrix. The memory bank is only updated during the first stage training in a First-In-First-Out (FIFO) manner. When the model is stable and the validation loss stops decreasing, the memory bank will stop updating and be applied to vote class-wise centers for usage in the second stage.

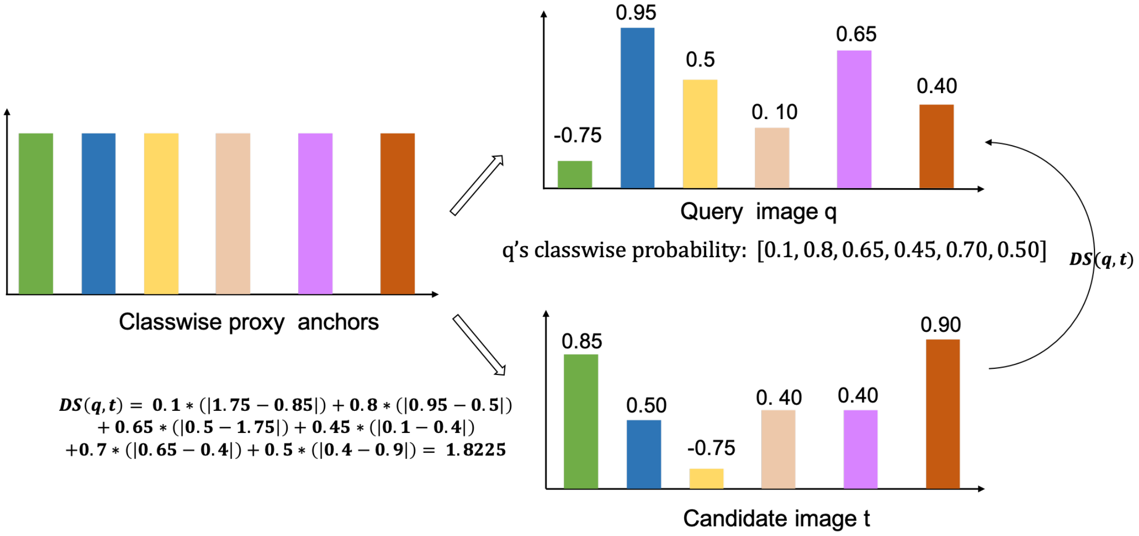


Figure 6.7: Multi-class similarity calculation.

Classwise voted center (proxy) computation: After the first stage, k center embeddings $\{c_1, c_2, ..., c_{k-1}, c_k\}$ are computed via calculating the mean of all the class-wise representations stored in the memory bank. For one class i, the voted center is calculated via $c_i = \frac{\sum_{j=1}^{M} \mathcal{B}_{i,j}}{M}$. The centers will be used as proxy anchors to optimize the classwise feature compactness.

Proxy-based Multi-class Similarity Learning

Although a multi-branch classifier is able to learn the optimal decision boundary in the feature space to classify inputs into different classes, the positive representation compactness of each class is not optimized. To enhance the discriminative ability of classwise negative samples, we propose to optimize the mutli-class similarity objective by maximizing the cosine similarity between the positive features and the corresponding classwise proxy, and simultaneously maximizing the dissimilarity between the negatives and the proxy with a certain margin. E_i^{pos} and E_i^{neg} are for the total classwise positives and negatives for class i. After the warm-up, we compute the latent class-wise representation learning via the proxy-based multi-class similarity objective defined as $L_{pms} = \frac{1}{k}(L_{pms}^{pos} + L_{pms}^{neg})$, where L_{pms}^{pos} and L_{pms}^{neg} are defined in Eqn. 6.2 and Eqn. 6.3 respectively.

$$L_{pms}^{pos} = \sum_{i=1}^{k} \frac{\sum(1-\tau_i)(1-\boldsymbol{COS}(c_i, E_i^{pos}))}{N_{E_i^{pos}}} \tag{6.2}$$

$$L_{pms}^{neg} = \sum_{i=1}^{k} \frac{\sum \tau_i |-1-\boldsymbol{COS}(c_i, E_i^{neg})|}{N_{E_i^{neg}}} \tag{6.3}$$

where $\boldsymbol{COS}(a,b) = a^T b/(||a||_2||b||_2)$ is the cosine similarity, and τ is the imbalance ratio vector, which records the imbalance ratio for each class. As introduced in Sec. 6.2.2, given a multi-label dataset which has N samples, the total number of class i's positives N_i^{pos} and negatives N_i^{neg} equals N, namely, $N = N_i^{pos} + N_i^{neg}$. The occurrence of different classes can be highly imbalanced. To handle the challenge, we weight the **PMS** learning for each class with the imbalance ratio $\tau_i = N_i^{pos}/N$. The reason that we use cosine similarity to optimize the representation learning instead of euclidean distance lies in the controllable value range. As cosine similarity naturally scales the output into [-1, 1], it can avoid dealing with negative values during the optimization. During the warm-up stage, we optimize the classification loss with the definition $L_{cls} = \sum_{i=1}^{k} BCE(\hat{Y_i}, Y_i)$; in the second stage, we incorporate the PMS loss and thus the overall loss is defined as $L_{overall} = \lambda L_{cls} + (1 - \lambda) L_{pms}$.

Multi-class Similarity for Inference

To get more relevant images, we suggest a multi-class similarity to rank the retrieved images. Figure 6.7 illustrates the computation process. Given a query image q, we first obtain its class-wise representations $E^q = \{e_1^q, e_2^q, ..., e_k^q\}$ and class-wise probability $P^q = \{p_1^q, p_2^q, ..., p_k^q\}$ predicted by the classifier. Based on the proxy anchors $C = \{c_1, c_2, ..., c_k\}$, we then calculate the classwise cosine similarity $S^q = \{s_1^q, s_2^q, ..., s_k^q\}$ following the Eqn. 6.4:

$$s_i^q = \boldsymbol{COS}(c_i, e_i^q) \tag{6.4}$$

Therefore, for a candidate image t in the database, we follow the same procedure to obtain its similarity vector S^t. With the similarity values of both the query image and candidate image to the same class center anchors available, the overall dissimilarity value is calculated via Eqn. 6.5:

$$DS = \sum_{i=1}^{k} p_i^q |s_i^q - s_i^t| \tag{6.5}$$

A low dissimilarity value indicates high similarity between the query image and the candidate image. For convenience, we use weighted_PMS to represent the foregoing defined DS in the following sections, and use unweighted_PMS for comparison by only computing the dissimilarity as $DS = \sum_{i=1}^{k} |s_i^q - s_i^t|$, without the class-wise probability involved.

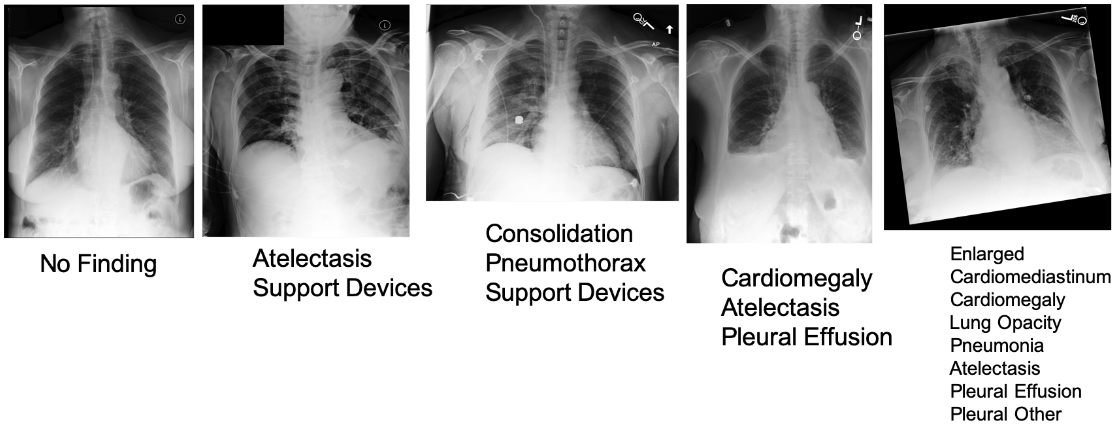


Figure 6.8: Chest X-ray examples from MIMIC dataset.

6.2.3 Experiments

Datasets

We experimented on the two public Chest Radiograph(CXR) dataset - MIMIC-CXR-2.0.0 [69] datasets (227,835 studies, MIMIC for short) and CheXpert [68] (114,526 frontal-view training images, 197 frontal-view test images). We split the MIMIC-CXR into training and validataion datasets with the ratio of 8:2. The chest X-ray data in both datasets has 14 classes - *No Finding* (NF), *Enlarged Cardiomediastinum* (EC), *Cardiomegaly* (CM), *Lung Lesion* (LL), *Lung Opacity* (LO), *Edema* (EM), *Consolidation* (CD), *Pneumonia* (PN), *Atelectasis* (AT), *Pneumothorax* (PX), *Pleural Effusion* (PE), *Pleural Other* (PO), *Fracture* (FT), *Support Devices* (SD). Examples from MIMIC dataset can be seen in Fig. 6.8. To unify the training pipelines, we resize all the chest X-rays into size of 224 × 224 × 3.

Comparative Baselines

For *classification*, we compare the performance with the standard binary cross entropy loss while keeping the same backbone architecture. For *image retrieval*, we compare our approach with 10 SOTA image retrieval methods - including 6 deep metric learning representatives - DeepRank [153], FastAP [20], MultiSimilarity [157], CircleLoss[144], ProxyAnchor [72], SupCon [71], and 4 deep hashing representatives - DPSH [82], DTSH [155], CSQ [166] and DBDH [177]. For fair comparison, we use the same backbone to extract features and encode the representations into 32-bits for deep hashing approaches.

Experimental Setup

We set d = 512, M = 1000 and fix λ to 0.25. For PMS model training, the initial learning rate is set as 1e-3, with the pretrained weights (on ImageNet [37]) to initialize the backbone, we warm up the multi-branch classifier with 5 epochs, and then optimize the classification objective and the discriminative representation learning objective with 10 epochs. We implement all the pipelines with Pytorch with 4 GPUs.

Image Retrieval Performance

The voted class-wise centers function as proxy anchors, and based on which, we perform content-based image retrieval via calculating the weighted total dissimilarity with Eqn. 6.4. A small distance indicates more closeness. By ranking the distances in ascending order, we get the retrieved results and calculate the metrics by considering top K results ($K \in \{1, 5, 10, 50\}$). For quantitative comparison, we report the precision at K (P@K). Although our paper focuses on exact match (a.k.a., full match) for multi-label image retrieval, we present both the full match and partial match performance in Tab. 6.2. Partial match means that for a query image with multiple labels, a retrieval image is relevant if it shares at least one same label as the query data.

Table 6.2 shows image retrieval performances on both MIMIC [69] and CheXpert [68] datasets, with MIMIC_full for full match results for MIMIC dataset. Others follow the same name style. We present the deep metric learning approaches first and then deep hashing algorithms. For MIMIC full match, deep hash approaches CSQ and DBDH can reach very high precision for top 1 retrieval and exceed other algorithms clearly. But the advantage fails to remain when considering more retrieval candidates and the performance of them drops significantly. In comparison, deep metric learning methods can output stable predictions for different numbers of retrieval candidates. Nonetheless, PMS (both the unweighted and weighted versions) outperforms others for P@5, P@10 and P@50. Interestingly, our model PMS maintains its advantages for MIMIC partial match and achieves the best over all the metrics. It is notable that full match performance on CheXpert is very low, the possible reason behind this is the noisy annotations. It also indicates that exact match retrieval for CheXpert is relatively challenging. Even so, PMS gets the best full match performance compared to all the other models while exhibiting satisfying partial match results.

To demonstrate the effectiveness of applying the class-wise probabilities to dissimilarity calculation, we present both the unweighted_PMS and weighted_PMS results at the bottom of Table 1. As can be observed, the weighted PMS generally surpasses the unweighted version on both the MIMIC and CheXpert dataset. The performance improvement shows the classification probability can be an indicator of the importance for the specific class, and eventually contributing to the retrieval accuracy.

Table 6.2: Image retrieval performance on MIMIC-CXR and CheXpert datasets . Bold indicates the best.

Method	MIMIC_full				MIMIC_partial				CheXpert_full				CheXpert_partial			
	P@1	P@5	P@10	P@50	P@1	P@5	P@10	P@50	P@1	P@5	P@10	P@50	P@1	P@5	P@10	P@50
DeepRank [153]	21.2	22.5	23.0	23.1	49.6	52.0	52.9	53.5	1.0	2.3	1.6	1.2	73.4	69.3	71.6	72.1
FastAP [20]	20.1	22.7	22.6	23.0	53.1	53.6	53.2	53.4	0.7	0.7	1.2	1.1	73.4	**74.2**	73.3	71.5
MultiSimilarity [157]	22.9	22.5	22.7	23.1	53.7	53.8	53.7	54.0	0.4	1.1	1.1	1.1	69.5	68.0	67.6	67.5
CircleLoss[144]	22.1	22.9	23.9	23.3	53.1	53.5	54.9	53.4	0.0	0.0	0.2	0.6	72.6	70.2	71.8	71.7
ProxyAnchor [72]	22.3	22.9	23.1	23.2	55.9	56.2	56.9	57.0	0.7	1.2	1.3	1.0	73.7	72.3	72.0	71.8
SupCon [71]	23.4	23.4	23.4	23.3	54.6	53.8	53.7	53.5	2.8	1.8	1.6	1.9	**77.6**	72.9	71.9	**73.4**
DPSH_32bits [82]	6.1	6.9	6.7	6.8	14.4	14.7	14.6	14.8	0.5	0.6	0.5	0.4	11.7	16.8	16.8	16.8
DTSH_32bits [155]	6.5	17.0	10.8	7.8	12.3	27.9	21.9	17.5	5.1	5.7	5.9	6.0	10.2	10.3	10.6	9.9
CSQ_32bits [166]	25.9	6.1	6.3	2.8	42.9	24.0	24.3	20.5	2.0	0.9	0.7	0.6	22.3	21.5	20.9	22.7
DBDH_32bits [177]	**32.2**	13.6	10.6	12.4	35.8	34.6	30.7	33.8	0.0	5.4	3.2	2.0	72.5	48.1	55.7	62.1
Unweighted_PMS*(ours)*	21.9	23.7	**24.3**	23.2	57.0	59.2	59.5	59.3	8.6	9.0	**9.1**	**8.5**	73.4	73.8	73.0	71.9
Weighted_PMS *(ours)*	25.8	**24.2**	**24.3**	**23.7**	**63.7**	**60.1**	**60.9**	**60.0**	**10.2**	**9.1**	9.0	8.2	73.1	74.0	**73.9**	72.8

Classification Performance

Since the metric learning part of PMS compacts the class-wise latent space and forces the negatives to be distant, jointly optimizing the classification objective with this metric can improve the classification to a certain extent. Thus, we compare the classification AUC score for each class considering with and without the metric learning. Table 6.3 reports the class-wise AUC score for the classification tasks on both MIMIC and CheXpert datasets. The class names are in short abbreviation to save space. As can be seen, the general accuracy for classifying the 14 distinct classes for both MIMIC and CheXpert has increased.

Table 6.3: Classification AUC performance for each class for MIMIC-CXR and CheXpert datasets.

Dataset	**Methods**	**NF**	**EC**	**CM**	**LL**	**LO**	**EM**	**CD**	**PN**	**AT**	**PX**	**PE**	**PO**	**FT**	**SD**	**Micro**	**Macro**	**Weighted**
MIMIC	Baseline	**77.5**	50.0	60.6	**50.2**	51.8	59.4	50.0	51.2	**64.6**	**57.9**	**81.3**	50.0	50.0	**80.0**	67.8	59.6	66.0
	PMS	77.2	50.0	**62.9**	50.0	**56.7**	**70.9**	50.0	**52.0**	58.6	51.5	77.9	50.0	50.0	79.9	**68.2**	**59.8**	**66.4**
CheXpert	Baseline	65.3	51.3	62.1	50.0	**69.7**	**75.3**	50.0	50.0	50.0	63.1	**82.1**	50.0	**54.6**	75.5	**71.9**	60.7	**67.7**
	PMS	**71.6**	**52.3**	**63.4**	50.0	61.8	69.2	**52.0**	**51.7**	**53.4**	**64.4**	81.9	50.0	51.9	**77.7**	70.5	**60.9**	67.5

6.2.4 Conclusion

Multi-label medical image retrieval upholds the promise to empower a variety of applications such as image annotations, disease diagnosis and history queries etc. The challenges reside in the gap between natural domain and medical domain, in terms of the imbalance of label combinations and the nuance of variations in medical datasets. In this paper, we bridge this gap by designing a proxy-based multi-class similarity metric, which compares and contrasts samples based on their similarity distributions with respect to the class proxies. We benchmark over 10 popular real-valued and hash-bashed metric learning methods on two medical datasets. Experiments show the effectiveness of our approach under both exact and non-exact settings across two common medical datasets.

6.3 Discussions and future works

In this chapter, we have investigated medical image retrieval under limited supervision and addressed outlier-sensitive image retrieval and multi-label radiology image retrieval problems. Nonetheless, our outlier-sensitive image retrieval mainly works on noisy external medical image datasets, requiring a clean internal dataset as the dataset quality standard. Moreover, our method can perform well for dataset that shows clear intra- and inter-class variations, and may fail when facing classes that share heavy similarities. Different from the outlier-sensitive image retrieval, our multi-label medical image retrieval approach focuses on learning the similarities of different samples by considering all the classes the dataset has. Although our model outputs decent results, there is still a big gap compared to the single-label image retrieval performance. We would like to design more effective models in the future that can capture multi-label similarities and retrieve the images efficiently.

Chapter 7

Publications

- **Guo, Xiaoyuan**, Hanyi Yu, Blair Rossetti, George Teodoro, Daniel Brat, and Jun Kong. "Clumped nuclei segmentation with adjacent point match and local shape-based intensity analysis in fluorescence microscopy images." In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3410-3413. IEEE, 2018.

- **Guo, Xiaoyuan**, Fusheng Wang, George Teodoro, Alton B. Farris, and Jun Kong. "Liver steatosis segmentation with deep learning methods." In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 24-27. IEEE, 2019.

- **Guo, Xiaoyuan**, Judy Wawira Gichoya, Hari Trivedi, W. Charles O'Neill, Rhakur Priya, Weijia Sun, Manisha Singh, Kathiravelu Pradeeban, Thomas J. Kim, Tianen Christopher Yang and Imon Banerjee. "Deeper Thinner UNet (DT-UNet) for Fine Vessel Segmentation of Breast Arterial Calcification (BAC)." CMIMI2020. `https://cdn.ymaws.com/siim.org/resource/resmgr/mimi20/abstracts/deeper_thinner_unet_guo.pdf`

- Saha, Monjoy, **Xiaoyuan Guo**, and Ashish Sharma. "TilGAN: GAN for Facilitating Tumor-Infiltrating Lymphocyte Pathology Image Synthesis With Improved Image Classification." IEEE Access 9 (2021): 79829-79840.

- **Guo, Xiaoyuan**, W. Charles O'Neill, Brianna Vey, Tianen Christopher Yang, Thomas J. Kim, Maryzeh Ghassemi, Ian Pan, Judy Wawira Gichoya, Hari Trivedi, and Imon Banerjee. "SCU-Net: A deep learning method for segmentation and quantification of breast arterial calcifications on mammograms." Medical physics 48, no. 10 (2021): 5851-5861.

- **Guo, Xiaoyuan**, Judy W. Gichoya, Saptarshi Purkayastha, and Imon Banerjee. "Margin-aware intraclass novelty identification for medical images." Journal of Medical Imaging 9, no. 1 (2022): 014004.

- **Guo, Xiaoyuan**, Judy Wawira Gichoya, Saptarshi Purkayastha, and Imon Banerjee. "CVAD-An unsupervised image anomaly detector." Software Impacts 11 (2022): 100195.

- **Guo, Xiaoyuan**, Judy Wawira Gichoya, Saptarshi Purkayastha, and Imon Banerjee. "CVAD: A generic medical anomaly detector based on Cascade VAE." MICCAI workshop (2022).

- **Guo, Xiaoyuan**, Judy Wawira Gichoya, Hari Trivedi, Saptarshi Purkayastha, and Imon Banerjee. "Shift data identification for external medical datasets." SIIM 2022.

- **Guo, Xiaoyuan**, Judy Wawira Gichoya, Hari Trivedi, Saptarshi Purkayastha, and Imon Banerjee. "MedShift: identifying shift data for medical dataset curation." Joural of Biomedical and Health Informatics 2022 (under 2nd round review).

- **Guo, Xiaoyuan**, Jiali Duan, Saptarshi Purkayastha, Hari Trivedi, Judy Wawira Gichoya, and Imon Banerjee. 2022. "OSCARS: An Outlier-Sensitive Content-Based Radiography Retrieval System." In Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22). Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/3512527.3531425

- **Guo, Xiaoyuan**, Jiali Duan, C-C. Jay Kuo, Judy Wawira Gichoya, and Imon Banerjee. "Augmenting Vision Language Pretraining by Learning Codebook with Visual Semantics." ICPR (2022).

- **Guo, Xiaoyuan**, Jiali Duan, Judy Gichoya, Hari Trivedi, Saptarshi Purkayastha, Ashish Sharma, and Imon Banerjee. "Multi-Label Medical Image Retrieval Via Learning Multi-Class Similarity." AI in Medicine, 2022, (under review).

Chapter 8

Conclusion and Future works

To summarize, this thesis has focused on designing and applying deep learning methods to solve various medical image tasks with limited supervision situation. Specifically, we have addressed problems under no supervision, weak supervision, limited supervision and collecting more supervisions, respectively. For no supervision, we have worked on nuclei segmentation and designed a classical segmentation method to avoid the necessity of annotations; and based on the segmentation algorithm, we have generated weak annotations for liver steatosis data to facilitate the application of deep learning methods. For weak supervision, there are two main tasks involved - BAC segmentation and outlier-sensitive image retrieval. Facing the weak and limited supervision, we have proposed a lightweight model for BAC segmentation and five quantification metrics to measure the relevance between predictions and ground-truth masks, which can quantify model performance more accurately instead of directly evaluating on pixel-level segmentation; for image retrieval, we have generated pseudo labels and assigned a small weight during optimization for learning intra-class variations, which enables the model to learn outlier-sensitive knowledge without affecting the normal image retrieval. For limited supervision, we have explored the OOD detection task. Due to the inaccessibility of all possible OOD categories, we designed unsupervised anomaly detection approaches following the one-vs-rest training procedure and verified the effectiveness on the several medical datasets. For more supervision data collection, we have proposed a unified automated pipeline to detect shift data among external datasets based on the self-supervised learning on a clean internal dataset. This pipeline can provide domain knowledge without sharing data across different institutions, which can be utilized widely. Generally speaking, the thesis conducts research on different medical tasks under different supervision scenarios and primarily addresses the problems from model design and algorithm efficiency perspective.

Nonetheless, there are still more complicated scenarios with different image modalities (e.g., Magnetic Resonance Imaging (MRI) images, 3D medical images, etc.) in real life, which are worthwhile more efforts to investigate. Specifically, the anomaly detection for MRI images, which is a more challenging task as the modality is much more noisy than the others. Besides, the algorithms we have adopted are mainly supervised and unsupervised learning. In the future, we can exploit semi-supervised learning, few-shot learning, meta learning techniques to address more challenging medical image tasks.

Bibliography

[1] Nada Abou-Hassan, Ekamol Tantisattamo, Ellen T D'Orsi, and W Charles O'neill. The clinical significance of medial arterial calcification in end-stage renal disease in women. *Kidney international*, 87(1):195–199, 2015.

[2] Jeongyoun Ahn, Myung Hee Lee, and Jung Ae Lee. Distance-based outlier detection for high dimension, low sample size data. *Journal of Applied Statistics*, 46(1):13–29, 2019.

[3] Muhammad Ahsan, Muhammad Mashuri, Heri Kuswanto, Dedy Dwi Prastyo, and Hidayatul Khusna. Outlier detection using pca mix based t 2 control chart for continuous and categorical data. *Communications in Statistics-Simulation and Computation*, 50(5):1496–1523, 2021.

[4] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018.

[5] Ceyhun Burak Akgül, Daniel L Rubin, Sandy Napel, Christopher F Beaulieu, Hayit Greenspan, and Burak Acar. Content-based image retrieval in radiology: current status and future directions. *Journal of digital imaging*, 24(2):208–222, 2011.

[6] Harish R Alappan, Gurleen Kaur, Shumila Manzoor, Jose Navarrete, and

W Charles O'Neill. Warfarin accelerates medial arterial calcification in humans. *Arteriosclerosis, thrombosis, and vascular biology*, 40(5):1413–1419, 2020.

[7] Harish R Alappan, Payaswini Vasanth, Shumila Manzoor, and W Charles O'Neill. Vascular calcification slows but does not regress after kidney transplantation. *Kidney International Reports*, 5(12):2212–2217, 2020.

[8] David Alvarez Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33, 2020.

[9] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.

[10] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.

[11] Thomas Atta-Fosu, Weihong Guo, Dana Jeter, Claudia M Mizutani, Nathan Stopczynski, and Rui Sousa-Neves. 3d clumped cell segmentation using curvature based seeded watershed. *Journal of imaging*, 2(4):31, 2016.

[12] Thomas Atta-Fosu, Weihong Guo, Dana Jeter, Claudia M Mizutani, Nathan Stopczynski, and Rui Sousa-Neves. 3d clumped cell segmentation using curvature based seeded watershed. *Journal of imaging*, 2(4):31, 2016.

[13] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[14] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017.

[15] Lena R Bartell, Lawrence J Bonassar, and Itai Cohen. A watershed-based algorithm to segment and classify cells in fluorescence microscopy images. *arXiv preprint arXiv:1706.00815*, 2017.

[16] Lena R Bartell, Lawrence J Bonassar, and Itai Cohen. A watershed-based algorithm to segment and classify cells in fluorescence microscopy images. *arXiv preprint arXiv:1706.00815*, 2017.

[17] Laura Beggel, Michael Pfeiffer, and Bernd Bischl. Robust anomaly detection in images using adversarial autoencoders. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 206–222. Springer, 2019.

[18] Gautam Bhattacharya, Koushik Ghosh, and Ananda S Chowdhury. Outlier detection using neighborhood rank difference. *Pattern Recognition Letters*, 60: 24–31, 2015.

[19] Ilker Bozcan and Erdal Kayacan. Uav-adnet: Unsupervised anomaly detection using deep neural networks for aerial surveillance. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1158–1164. IEEE, 2020.

[20] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1861–1870, 2019.

[21] Lisa Calvocoressi, Albert Sun, Stanislav V Kasl, Elizabeth B Claus, and Beth A Jones. Mammography screening of women in their 40s: impact of changes in screening guidelines. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 112(3):473–480, 2008.

[22] Tianshi Cao, Chinwei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

[23] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[24] Naga Chalasani, Zobair Younossi, Joel E Lavine, Anna Mae Diehl, Elizabeth M Brunt, Kenneth Cusi, Michael Charlton, and Arun J Sanyal. The diagnosis and management of non-alcoholic fatty liver disease: practice guideline by the american gastroenterological association, american association for the study of liver diseases, and american college of gastroenterology. *Gastroenterology*, 142 (7):1592–1609, 2012.

[25] Heang-Ping Chan, Ravi K. Samala, Lubomir M. Hadjiiski, and Chuan Zhou. Deep learning in medical image analysis. *Advances in Experimental Medicine and Biology*, 1213, 2020. doi: https://doi.org/10.1007/978-3-030-33128-3_1.

[26] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.

[27] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[29] Wei Chen, Yu Liu, Weiping Wang, Erwin Bakker, Theodoros Georgiou, Paul

Fieguth, Li Liu, and Michael S. Lew. Deep learning for instance retrieval: A survey, 2022.

[30] Won-Tak Choi, Kuang-Yu Jen, Dongliang Wang, Mehdi Tavakol, John P Roberts, and Ryan M Gill. Donor liver small droplet macrovesicular steatosis is associated with increased risk for recipient allograft rejection. *The American journal of surgical pathology*, 41(3):365–373, 2017.

[31] Manish Chowdhury, Samuel Rota Bulo, Rodrigo Moreno, Malay Kumar Kundu, and Örjan Smedby. An efficient radiographic image retrieval system using convolutional neural network. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3134–3139. IEEE, 2016.

[32] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

[33] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[34] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pages 3213–3223, 2016.

[35] Robert Jackson Williams David Rumelhart, Geoffrey Hinton. *Parallel distributed processing Explorations in the microstructure of cognition*, volume 1. MIT press Cambridge, MA, 1986.

[36] Erik Daxberger and José Miguel Hernández-Lobato. Bayesian variational autoencoders for unsupervised out-of-distribution detection. *arXiv preprint arXiv:1912.05651*, 2019.

[37] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[38] Robert Detrano, Alan D Guerci, J Jeffrey Carr, Diane E Bild, Gregory Burke, Aaron R Folsom, Kiang Liu, Steven Shea, Moyses Szklo, David A Bluemke, et al. Coronary calcium as a predictor of coronary events in four racial or ethnic groups. *New England Journal of Medicine*, 358(13):1336–1345, 2008.

[39] Federico Di Mattia, Paolo Galeone, Michele De Simoni, and Emanuele Ghelfi. A survey on gans for anomaly detection. *arXiv preprint arXiv:1906.11632*, 2019.

[40] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[41] Jiali Duan and C-C Jay Kuo. Bridging gap between image pixels and semantics via supervision: A survey. *arXiv preprint arXiv:2107.13757*, 2021.

[42] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[43] Valerie Duhn, Ellen T D'Orsi, Samuel Johnson, Carl J D'Orsi, Amy L Adams, and W Charles O'Neill. Breast arterial calcification: a marker of medial vascular calcification in chronic kidney disease. *Clinical Journal of the American Society of Nephrology*, 6(2):377–382, 2011.

[44] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *arXiv preprint arXiv:2012.02364*, 2020.

[45] R.N. Fiorini, J. Kirtz, and B. et al. Periyasamy. Development of an unbiased method for the estimation of liver steatosis. *Clin. Transplant*, 18(6):700–706, 2004.

[46] Shereen Fouad, Gabriel Landini, David Randell, and Antony Galton. Morphological separation of clustered nuclei in histological images. In *International Conference on Image Analysis and Recognition*, pages 599–607. Springer, 2016.

[47] Shereen Fouad, Gabriel Landini, David Randell, and Antony Galton. Morphological separation of clustered nuclei in histological images. In *International Conference on Image Analysis and Recognition*, pages 599–607. Springer, 2016.

[48] Guojun Gan and Michael Kwok-Po Ng. K-means clustering with outlier removal. *Pattern Recognition Letters*, 90:8–14, 2017.

[49] Mariana Garcia, Sharon L Mulvagh, C Noel Bairey Merz, Julie E Buring, and JoAnn E Manson. Cardiovascular disease in women: clinical perspectives. *Circulation research*, 118(8):1273–1293, 2016.

[50] M Jayesh George, D Anto Sahaya Dhas, et al. Preprocessing filters for mammogram images: A review. In *2017 Conference on Emerging Devices and Smart Systems (ICEDSS)*, pages 1–7. IEEE, 2017.

[51] Davide Giavarina. Understanding bland altman analysis. *Biochemia medica*, 25(2):141–151, 2015.

[52] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[53] Xiaoyuan Guo, Hanyi Yu, Blair Rossetti, George Teodoro, Daniel Brat, and Jun Kong. Clumped nuclei segmentation with adjacent point match and local shape based intensity analysis for overlapped nuclei in fluorescence in-situ hybridization images. *arXiv preprint arXiv:1808.04795*, 2018.

[54] Xiaoyuan Guo, Judy Wawira Gichoya, Saptarshi Purkayastha, and Imon Banerjee. Cvad: A generic medical anomaly detector based on cascade vae. *arXiv preprint arXiv:2110.15811*, 2021.

[55] Xiaoyuan Guo, Judy Wawira Gichoya, Hari Trivedi, Saptarshi Purkayastha, and Imon Banerjee. Medshift: identifying shift data for medical dataset curation. *arXiv preprint arXiv:2112.13885*, 2021.

[56] Xiaoyuan Guo, Judy Wawira Gichoya, Saptarshi Purkayastha, and Imon Banerjee. Cvad-an unsupervised image anomaly detector. *Software Impacts*, 11: 100195, 2022.

[57] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[59] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[60] Eva JE Hendriks, Pim A de Jong, Yolanda van der Graaf, P Th M Willem, Yvonne T van der Schouw, and Joline WJ Beulens. Breast arterial calcifications: a systematic review and meta-analysis of their determinants and their association with cardiovascular events. *Atherosclerosis*, 239(1):11–20, 2015.

[61] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

[62] André Homeyer, Patrik Nasr, Christiane Engel, Stergios Kechagias, Peter Lundberg, Mattias Ekstedt, Henning Kost, Nick Weiss, Tim Palmer, Horst Karl Hahn, et al. Automated quantification of steatosis: agreement with stereological point counting. *Diagnostic pathology*, 12(1):80, 2017.

[63] Md Shamim Hossain. Microc alcification segmentation using modified u-net segmentation network from mammogram images. *Journal of King Saud University-Computer and Information Sciences*, 2019.

[64] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[65] Huaibo Huang, Zhihang Li, Ran He, Zhenan Sun, and Tieniu Tan. Introvae: Introspective variational autoencoders for photographic image synthesis. *arXiv preprint arXiv:1807.06358*, 2018.

[66] Rebecca A Hubbard, Karla Kerlikowske, Chris I Flowers, Bonnie C Yankaskas, Weiwei Zhu, and Diana L Miglioretti. Cumulative probability of false-positive

recall or biopsy recommendation after 10 years of screening mammography: a cohort study. *Annals of internal medicine*, 155(8):481–492, 2011.

[67] Kyung Hoon Hwang, Haejun Lee, and Duckjoo Choi. Medical image retrieval: past and present. *Healthcare informatics research*, 18(1):3–9, 2012.

[68] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[69] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

[70] Shehroz S Khan and Michael G Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3): 345–374, 2014.

[71] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[72] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020.

[73] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[74] Nahum Kiryati and Yuval Landau. Dataset growth in medical image analysis research. *Journal of imaging*, 7(8):155, 2021.

[75] Jun Kong, Michael J Lee, Pelin Bagci, Puneet Sharma, Diego Martin, N Volkan Adsay, Joel H Saltz, and Alton B Farris. Computer-based image analysis of liver steatosis with large-scale microscopy imagery and correlation with magnetic resonance imaging lipid analysis. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pages 333–338. IEEE, 2011.

[76] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5, 2010.

[77] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.

[78] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

[79] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31:7167–7177, 2018.

[80] Michael J Lee, Pelin Bagci, Jun Kong, Miriam B Vos, Puneet Sharma, Bobby Kalb, Joel H Saltz, Diego R Martin, N Volkan Adsay, and Alton B Farris. Liver steatosis assessment: correlations among pathology, radiology, clinical data and automated image analysis software. *Pathology-Research and Practice*, 209(6): 371–379, 2013.

[81] Gen Li, Inyoung Yun, Jonghyun Kim, and Joongkyu Kim. Dabnet: Depth-wise

asymmetric bottleneck for real-time semantic segmentation. *arXiv preprint arXiv:1907.11357*, 2019.

[82] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. Feature learning based deep supervised hashing with pairwise labels. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1711–1717, 2016.

[83] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[84] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[85] Miao Liao, Yu-qian Zhao, Xiang-hua Li, Pei-shan Dai, Xiao-wen Xu, Jun-kai Zhang, and Bei-ji Zou. Automatic segmentation for cell images based on bottleneck detection and ellipse fitting. *Neurocomputing*, 173:615–622, 2016.

[86] Miao Liao, Yu-qian Zhao, Xiang-hua Li, Pei-shan Dai, Xiao-wen Xu, Jun-kai Zhang, and Bei-ji Zou. Automatic segmentation for cell images based on bottleneck detection and ellipse fitting. *Neurocomputing*, 173:615–622, 2016.

[87] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[88] G.E. Liquori, G. Calamita, D. Cascella, M. Mastrodonato, P. Portincasa, and D. Ferri. An innovative methodology for the automated morphometric and quantitative estimation of liver steatosis. *Histol. Histopathol.*, 24(1):49–60, 2009.

[89] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[90] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Deep supervised hashing for fast image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2064–2072, 2016.

[91] Mengyu Liu and Hujun Yin. Feature pyramid encoding network for real-time semantic segmentation. *arXiv preprint arXiv:1909.08599*, 2019.

[92] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.

[93] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[94] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760*, 2020.

[95] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[96] Shao-Yuan Lo, Hsueh-Ming Hang, Sheng-Wei Chan, and Jing-Jhih Lin. Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In *Proceedings of the ACM Multimedia Asia*, pages 1–6. 2019.

[97] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[98] Francesco Lupo. *Variational Autoencoder for unsupervised anomaly detection.* PhD thesis, Politecnico di Torino, 2019.

[99] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[100] Shumila Manzoor, Syed Ahmed, Arshad Ali, Kum Hyun Han, Ioannis Sechopoulos, Ansley O'Neill, Baowei Fei, and W Charles O'Neill. Progression of medial arterial calcification in ckd. *Kidney international reports*, 3(6):1328–1335, 2018.

[101] H. Marsman, T. Matsushita, and R. Dierkhising. Assessment of donor liver steatosis: pathologist or automated software? *Human pathology*, 35(4):430–435, 2004.

[102] Jason C Ni, Katie Shpanskaya, Michelle Han, Edward H Lee, Bao H Do, William T Kuo, Kristen W Yeom, and David S Wang. Deep learning for automated classification of inferior vena cava filter types on radiographs. *Journal of Vascular and Interventional Radiology*, 31(1):66–73, 2020.

[103] Jason C Ni, Katie Shpanskaya, Michelle Han, Edward H Lee, Bao H Do, William T Kuo, Kristen W Yeom, and David S Wang. Deep learning for automated classification of inferior vena cava filter types on radiographs. *Journal of Vascular and Interventional Radiology*, 31(1):66–73, 2020.

[104] Yuqi Ouyang and Victor Sanchez. Video anomaly detection by estimating likelihood of representations. *arXiv preprint arXiv:2012.01468*, 2020.

[105] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[106] Jialin Peng and Ye Wang. Medical image segmentation with limited supervision: A review of deep network models. *IEEE Access*, 2021.

[107] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2020.

[108] Pramuditha Perera and Vishal M Patel. Learning deep features for one-class classification. *IEEE Transactions on Image Processing*, 28(11):5450–5463, 2019.

[109] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2906, 2019.

[110] Adrian Alan Pol, Victor Berger, Cecile Germain, Gianluca Cerminara, and Maurizio Pierini. Anomaly detection with conditional variational autoencoders. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1651–1657. IEEE, 2019.

[111] Rudra PK Poudel, Ujwal Bonde, Stephan Liwicki, and Christopher Zach. Contextnet: Exploring context and detail for semantic segmentation in real-time. *arXiv preprint arXiv:1805.04554*, 2018.

[112] Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network. *arXiv preprint arXiv:1902.04502*, 2019.

[113] Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. Medical image retrieval using deep convolutional neural network. *Neurocomputing*, 266:8–20, 2017.

[114] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019.

[115] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning.* Mit Press, 2009.

[116] Stephan Rabanser, Stephan Günnemann, and Zachary C Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *arXiv preprint arXiv:1810.11953*, 2018.

[117] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017.

[118] Nikola Reljin, Marijeta Slavkovic-Ilic, Coya Tapia, Nikola Cihoric, and Srdjan Stankovic. Multifractal-based nuclei segmentation in fish images. *Biomedical microdevices*, 19(3):1–13, 2017.

[119] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14707–14718, 2019.

[120] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[121] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5107–5116, 2019.

[122] Josiane Rodrigues, Marco Cristo, and Juan G Colonna. Deep hashing for multi-label image retrieval: a survey. *Artificial Intelligence Review*, 53(7):5261–5307, 2020.

[123] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017.

[124] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[125] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):1–8, 2021.

[126] Mousumi Roy, Fusheng Wang, George Teodoro, Miriam B Vos, Alton Brad Farris, and Jun Kong. Segmentation of overlapped steatosis in whole-slide liver histopathology microscopy images. *arXiv preprint arXiv:1806.09090*, 2018.

[127] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[128] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402, 2018.

[129] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pages 4–11, 2014.

[130] Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pages 4–11, 2014.

[131] Patrick Schlachter, Yiwen Liao, and Bin Yang. Deep one-class classification using intra-class splitting. *arXiv preprint arXiv:1902.01194*, 2019.

[132] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

[133] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.

[134] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[135] August DuMont Schütte, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick Schwab. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ digital medicine*, 4, 2021.

[136] Vikash Sehwag, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 105–116, 2019.

[137] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[138] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[139] Liwei Song, Vikash Sehwag, Arjun Nitin Bhagoji, and Prateek Mittal. A critical evaluation of open-world machine learning. *arXiv preprint arXiv:2007.04391*, 2020.

[140] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[141] Camilo G Sotomayor, Marcelo Mendoza, Víctor Castañeda, Humberto Farías, Gabriel Molina, Gonzalo Pereira, Steffen Härtel, Mauricio Solar, and Mauricio Araya. Content-based medical image retrieval and intelligent interactive visual browser for medical education, research and care. *Diagnostics*, 11(8):1470, 2021.

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

[143] Jeremias Sulam, Rami Ben-Ari, and Pavel Kisilev. Maximizing auc with deep learning for classification of imbalanced mammogram datasets. In *VCBM*, pages 131–135, 2017.

[144] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020.

[145] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, 2020.

[146] Takaaki Tagawa, Yukihiro Tadokoro, and Takehisa Yairi. Structured denoising autoencoder for fault detection and analysis. In *Asian Conference on Machine Learning*, pages 96–111. PMLR, 2015.

[147] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[148] Gabriele Valvano, Gianmarco Santini, Nicola Martini, Andrea Ripoli, Chiara Iacconi, Dante Chiappino, and Daniele Della Latta. Convolutional neural networks for the segmentation of microcalcification in mammography imaging. *Journal of Healthcare Engineering*, 2019, 2019.

[149] Peter MA van Ooijen. Quality and curation of medical images and data. In *Artificial intelligence in medical imaging*, pages 247–255. Springer, 2019.

[150] Willem G Van Panhuis, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J Herbst, David Heymann, and Donald S Burke. A systematic review of barriers to data sharing in public health. *BMC public health*, 14(1):1–9, 2014.

[151] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[152] Hui Wang, Hong Zhang, and Nilanjan Ray. Clump splitting via bottleneck detection. In *2011 18th IEEE International Conference on Image Processing*, pages 61–64. IEEE, 2011.

[153] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393, 2014.

[154] Juan Wang, Huanjun Ding, Fatemeh Azamian Bidgoli, Brian Zhou, Carlos Iribarren, Sabee Molloi, and Pierre Baldi. Detecting cardiovascular disease from mammograms with deep learning. *IEEE transactions on medical imaging*, 36(5):1172–1181, 2017.

[155] Xiaofang Wang, Yi Shi, and Kris M Kitani. Deep supervised hashing with triplet labels. In *Asian conference on computer vision*, pages 70–84. Springer, 2016.

[156] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common

thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[157] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

[158] Yu Wang, Quan Zhou, Jian Xiong, Xiaofu Wu, and Xin Jin. Esnet: An efficient symmetric network for real-time semantic segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 41–52. Springer, 2019.

[159] Quan Wen, Hang Chang, and Bahram Parvin. A delaunay triangulation approach for segmenting clumps of nuclei. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 9–12. IEEE, 2009.

[160] Quan Wen, Hang Chang, and Bahram Parvin. A delaunay triangulation approach for segmenting clumps of nuclei. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 9–12. IEEE, 2009.

[161] Tianyi Wu, Sheng Tang, Rui Zhang, and Yongdong Zhang. Cgnet: A lightweight context guided network for semantic segmentation. *arXiv preprint arXiv:1811.08201*, 2018.

[162] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[163] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 1492–1500, 2017.

[164] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[165] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9518–9526, 2019.

[166] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3083–3092, 2020.

[167] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049, 2021.

[168] A.M. Zaitoun, H. Al Mardini, S. Awad, S. Ukabam, S. Makadisi, and C.O. Record. Quantitative assessment of fibrosis and steatosis in liver biopsies from patients with chronic hepatitis. *J. Clin. Pathol.*, 54(6):461–465, 2001.

[169] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018.

[170] Chao Zhang, Changming Sun, Ran Su, and Tuan D Pham. Segmentation of clustered nuclei based on curvature weighting. In *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*, pages 49–54, 2012.

[171] Chao Zhang, Changming Sun, Ran Su, and Tuan D Pham. Segmentation of clustered nuclei based on curvature weighting. In *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*, pages 49–54, 2012.

[172] Chao ZHANG, C Sun, and Tuan D Pham. Segmentation of clustered nuclei based on concave curve expansion. *Journal of microscopy*, 251(1):57–67, 2013.

[173] Chao ZHANG, C Sun, and Tuan D Pham. Segmentation of clustered nuclei based on concave curve expansion. *Journal of microscopy*, 251(1):57–67, 2013.

[174] Jianjing Zhang and Robert X Gao. Deep learning-driven data curation and model interpretation for smart manufacturing. *Chinese Journal of Mechanical Engineering*, 34(1):1–21, 2021.

[175] Pengyi Zhang, Yunxin Zhong, Yulin Deng, Xiaoying Tang, and Xiaoqiong Li. A survey on deep learning of small sample in biomedical image analysis. *arXiv preprint arXiv:1908.00473*, 2019.

[176] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. Deep semantic ranking based hashing for multi-label image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1556–1564, 2015.

[177] Xiangtao Zheng, Yichao Zhang, and Xiaoqiang Lu. Deep balanced discrete hashing for image retrieval. *Neurocomputing*, 403:224–236, 2020.

[178] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 665–674, 2017.

[179] David Zimmerer, Jens Petersen, and Klaus Maier-Hein. High-and low-level

image component decomposition using vaes for improved reconstruction and anomaly detection. *arXiv preprint arXiv:1911.12161*, 2019.