Aligning Small Datasets Using Domain Adversarial Learning: Applications in Automated in Vivo Oral Cancer Diagnosis

Kayla Caughlin, Elvis Duran-Sierra, Shuna Cheng, Rodrigo Cuenca, Beena Ahmed, *Member, IEEE*, Jim Ji, Mathias Martinez, Moustafa Al-Khalil, Hussain Al-Enazi, Yi-Shing Lisa Cheng, John Wright, Javier A. Jo, and Carlos Busso, *Senior Member, IEEE*

***Abstract*—Deep learning approaches for medical image analysis are limited by small data set size due to factors such as patient privacy and difficulties in obtaining expert labelling for each image. In medical imaging system development pipelines, phases for system development and classification algorithms often overlap with data collection, creating small disjoint data sets collected at numerous locations with differing protocols. In this setting, merging data from different data collection centers increases the amount of training data. However, a direct combination of datasets will likely fail due to domain shifts between imaging centers. In contrast to previous approaches that focus on a single data set, we add a domain adaptation module to a neural network and train using multiple data sets. Our approach encourages domain invariance between two multispectral autofluorescence imaging (maFLIM) data sets of in vivo oral lesions collected with an imaging system currently in development. The two data sets have differences in the sub-populations imaged and in the calibration procedures used during data collection. We mitigate these differences using a gradient reversal layer and domain classifier. Our final model trained with two data sets substantially increases performance, including a significant increase in specificity. We also achieve a significant increase in average performance over the best baseline model train with two domains ($p = 0.0341$). Our approach lays the foundation for faster development of computer-aided diagnostic systems and presents a feasible approach for creating a robust classifier that aligns images from multiple data centers in the presence of domain shifts.**

***Index Terms*—Automated oral cancer diagnosis, domain adaptation, gradient reversal, multispectral autofluorescence imaging, variance regularization.**

Manuscript received 25 April 2022; revised 26 August 2022 and 4 October 2022; accepted 14 October 2022. Date of publication 25 October 2022; date of current version 5 January 2023. This work was supported by NIH, under Grant R01:5R01CA218739-04. *(Corresponding author: Carlos Busso.)*

Kayla Caughlin and Carlos Busso are with the Department of Electrical and Computer Engineering, University of Texas, Dallas Richardson, TX 78712 USA (e-mail: busso@utdallas.edu).

Rodrigo Cuenca and Javier A. Jo are with the School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK 73019 USA.

Elvis Duran-Sierra and Shuna Cheng are with the Department of Biomedical Engineering, Texas A&M University, College Station, TX 77843 USA.

Beena Ahmed is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia.

Jim Ji is with the Department of Electrical and Computer Engineering, Texas A&M University, Doha, Qatar, and also with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA.

Mathias Martinez and Moustafa Al-Khalil are with the Department of Cranio-Maxillofacial Surgery, Hamad Medical Corporation, Doha, Qatar.

Hussain Al-Enazi is with the Department of Otorhinolaryngology Head and Neck Surgery, Hamad Medical Corporation, Doha, Qatar.

Yi-Shing Lisa Cheng and John Wright are with the College of Dentistry, Texas A&M University, Dallas, TX 75246 USA.

Digital Object Identifier 10.1109/JBHI.2022.3217015

I. Introduction

TRADITIONALLY, a large data set is required for training a successful deep learning model. Unfortunately, medical imaging data sets are typically small and heterogeneous, creating issues with overfitting and exhibiting lack of generalization to different domains and settings across data centers. Many strategies have avoided deep learning solutions altogether, instead, focusing on methods such as *support vector machines* (SVM) and *quadratic discriminant analysis* (QDA) that offer less flexibility, but are less prone to overfitting [1], [2], [3]. Other methods reduce problems associated with inter-patient variability by including images from each patient in the training set [4]. However, requiring an image from each patient in the training set is impractical for clinical translation, because the model needs to be re-trained with the addition of every new patient. Another alternative approach for modalities with some similarities to natural images (e.g., *magnetic resonance imaging* (MRI) and digital histology) is to adapt deep learning models pre-trained on large-scale data sets such as ImageNet [5], [6]. However, other medical imaging applications are highly specific and do not have a similar modality that can be pre-trained on an extensive data set (i.e., in vivo fluorescent lifetime images of oral lesions).

In the absence of a pre-trained model and to avoid using images from each test patient in the training set, the combination of small medical image data sets from different centers, including sets from slightly different domains, can increase the size of the data set, and, potentially, mitigate generalization problems. However, domain shifts between data centers caused by differences in imaging systems, data collection protocols, and sub-populations prevent direct combination [7]. In addition, mismatches between data sets from different imaging centers may occur when ground truth annotations contain ambiguity. Specifically in oral cancer identification, histology ground truth labels contain both inter- and intra- observer variability. Due to

center-specific variations, generalization problems persist even when a large data set has been collected at a single imaging location [8]. Classification challenges due to domain shift have been documented in multiple imaging modalities, including *computed tomography* (CT), *magnetic resonance imaging* (MRI), ultrasound, and histology images [8], [9], [10], [11], [12], [13], [14].

Previous studies on oral cancer classification using *multispectral autofluorescence imaging* (maFLIM) focused only on a single, small data set [1], [2], [3], [15], [16]. As a single large-scale maFLIM oral lesion data set is not currently accessible, we would like to combine images from various collection centers. Specifically, we aim to improve performance on a main dataset through the inclusion of auxiliary training data from a separate imaging center. Unfortunately, preliminary results indicate that the new training data can actually decrease performance on the main dataset. This result has been observed in other medical imaging domains [17], showing the need for domain adaptation techniques before combining databases. To mitigate domain shift problems, we propose the use of domain adaptation techniques in a deep learning framework to merge the data sets and create domain invariance feature representations. While domain adaptation techniques have been widely investigated for combination of domains in other applications, little work has been reported on domain adaptation in cancer diagnosis of oral lesions from in vivo maFLIM. Although we focus on the main data set performance, this work sets the foundations for a single, robust classifier that can be used for the diagnosis of oral lesions across multiple imaging systems and locations. The main contributions of this work are as follows:

- We quantify main performance decrease when training with multiple domains for automated oral cancer diagnosis from maFLIM.
- We propose a new domain adaptation model to mitigate the domain shift between data collection centers for maFLIM.
- We investigate failure modes of standard gradient reversal for domain adaptation in a small data setting.
- We propose a variance regularization technique to reduce mode collapse in gradient reversal.
- We explore model selection techniques to further improve the main domain performance.

We note that there is an important distinction between our formulation and other machine learning methods using domain adaptation. Our goal is to increase the size of the train set, which is different from the common objective of other domain adaptation studies aiming to improve the performance of a classifier on an unlabeled target domain. The domain adaptation helps us to reduce mismatches so we can combine the datasets in an effective way, which is crucial in medical problems characterized with limited data obtained with different protocols. Our approach borrows aspects from data augmentation, where the *new* data included in the train set corresponds to new samples from the auxiliary domain. This approach differs from traditional data augmentation methods which create transformations or modifications of existing samples to increase the train set. Our formulation represents an effective use of domain adaptation to reduce data mismatches so we can combine the datasets in an effective way.

This paper is organized as follows. Section II summarizes related studies in maFLIM cancer diagnosis and describes relevant background for domain adaptation. Section III details our proposed approach. Section IV presents the results of our method in comparison to the baselines and analyzes each component of our model. Section V summarizes the main contributions and results of this study.

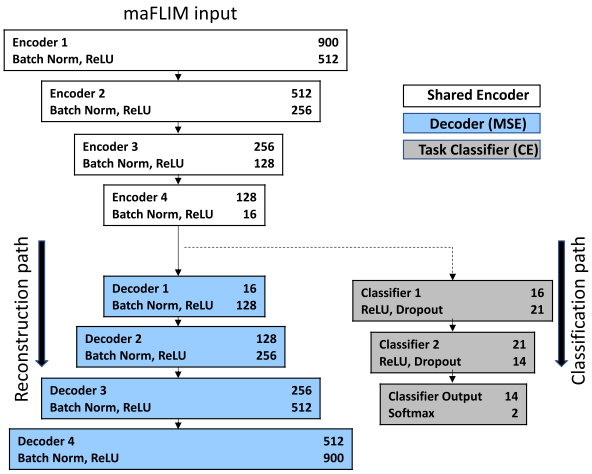


Fig. 1. Joint autoencoder and classifier neural network as presented in Caughlin et al. [15]. Our base model uses the same general structure of this neural network.

II. Related Work

A. In Vivo maFLIM Cancer Classification Methods

Although domain adaptation methods are commonly used in other deep learning applications, approaches for cancer diagnosis and margin delineation from in vivo maFLIM have been typically trained using a single domain [1], [2], [3], [15], [16], [18]. For example, Jo et al. [2] presented a solution based on a *quadratic discriminant analysis* (QDA) classifier for oral cancer diagnosis using a single data set. The inputs to the QDA classifier were hand-derived features based on signal intensity (spectral) and lifetime (temporal) features [2]. Marsden et al. [1] also used lifetime and intensity features to train a *support vector machine* (SVM) and a *random forest* (RF) classifier for margin detection in oral lesions from one imaging center. Chen et al. [19] used a SVM, but for microscopy images from a single collection site. Vasanthakumari et al. [3] used QDA and *linear discriminant analysis* (LDA) classifiers with hand-crafted features from phasor plots to classify skin lesions. Duran et al. [18] trained LDA, QDA, SVM and logistic regression classifiers on data from one imaging center and tests on data from another imaging center. However, the authors did not explore if an increase in performance might be obtained by using domain adaptation techniques. In addition, the authors report performance on the margin delineation task, while we focus on cancer diagnosis. Studies have only recently started to use deep learning solutions [1], [15]. Marsden et al. [1] explored the use of a *convolutional neural network* (CNN) for oral lesion margin delineation. However, the CNN-based model failed to outperform baselines implemented with traditional machine learning classifiers trained with hand-derived features [1]. In contrast, we previously reported a joint autoencoder and classifier structure (see Fig. 1) for oral cancer diagnosis using data-driven features that showed improved performance over the traditional baseline implemented with SVM in a single-domain setting [15].

While our previous work focused on automated oral cancer diagnosis using a single data set, our deep learning structure can accommodate domain adaptation techniques that have been successfully used in other domains. As our main data set is small, we

wish to improve performance on the main data set by including training data from another data set. Unfortunately, domain shifts between data sets can cause the model capacity to split between the domains, decreasing performance on the main data set. We hypothesize that the addition of domain adaptation methods to our joint autoencoder and classifier structure can create invariant representations such that auxiliary data from a different domain can act as augmented data for the main domain. In our method, we use an auxiliary data set to increase the training data set size, formulating the classification performance on the main data as our primary evaluation criterion. This formulation is radically different from other approaches used for in vivo oral cancer diagnosis from maFLIM.

B. Domain Adaptation Background

The goal of domain adaptation is to mitigate the mismatch between domains. Multiple domain adaptation techniques have been proposed, including using fine-tuning [10], *generative adversarial network* (GAN) [12], gradient reversal [20], transformation of data between domains [11], and augmentation methods [21].

Several studies have used a fine-tuning, transfer learning approach when at least some labels are present for both source and target data sets [5], [6], [10]. In these methods, a base classifier was trained using a source data set. Then, the model was minimally adjusted using the labeled target data to maximize performance on the target data. Fine-tuning differs from our objectives, as we wish to maximize performance on the main data set, using a second domain to augment the training set.

In contrast to fine-tuning, Zhang et al. [9] introduced a series of source-only data augmentation methods for MRI and ultrasound that minimized the performance gap between the source and target data in both small and large data settings. Though not in the medical imaging domain, Anaby et al. [22] similarly used a deep learning method to create synthetic examples for data augmentation in a small data setting. Our work takes inspiration from data augmentation and domain adaptation.

Ganin and Lepinsky [20] proposed the gradient reversal layer, which describes a popular plug-in general purpose technique for unsupervised domain adaptation. The model consisted of three blocks: a feature extractor, a domain classifier, and a task classifier. The domain classifier was trained to generate gradient updates in the direction of maximum domain separation. The gradients resulting from the domain classifier were then reversed during back-propagation to the feature extractor. Following training, the domain classifier was discarded and the feature extractor and task classifier were used for inference [20]. While Ganin et al. [20] focused on improving generalization to a new domain, our focus is on improving the main domain performance.

While the gradient reversal layer has been used in a variety of domain adaptation applications in both unsupervised and semi-supervised settings, some authors have noted issues retaining class-discriminative information [23], [24]. Specifically, Li et al. [23] asserted that gradient reversal does not ensure that class-discriminative information is preserved (referred to as *mode collapse*). Li et al. [23] used *joint adversarial domain adaptation* (JADA) to train two task classifiers using unlabeled target samples and labeled source samples. The model finds samples on the class boundary and optimizes the feature extractor to minimize differences in the predictions of the two classifiers. The authors hypothesize that since the two task classifiers are randomly initialized, if the two task classifiers disagree on the label of the target sample, it is likely near the decision boundary of the source. By training the feature extractor to minimize the discrepancy between the two task classifiers on the target data, the trained target representation will be near the source classes, improving task discrimination on the target data. Overall, the two task classifiers encouraged distinct class boundaries, while the domain classifier encouraged domain invariance [23]. Similarly, Kurmi et al. [24] mitigated mode collapse in domain adaptation by using an approach called *informative domain discriminator for domain adaptation* (IDDA). IDDA used a modified domain classifier and selection of source samples to train the domain classifier. The modified domain discriminator classified each sample as belonging to one of the source domain class labels or to a general label for the target domain [24]. Furthermore, source samples that were miss-classified in the task classifier were discarded from training the domain classifier [24]. However, the feature extractor only tried to make the domain classifier place the target sample in any of the source classification label bins, disregarding the target label class.

While JADA and IDDA mitigated mode collapse in the unsupervised domain adaptation framework, our framework is different. In our approach, rather than having a single moderate-to-large labeled source data set and an unlabeled target data set, both of our data sets contain labels. The size of the train set is also different. The JADA framework is implemented with thousands of examples [25]. Similarly, experiments with IDDA were run on multiple data sets, most with thousands of examples [24]. The smallest dataset contained 600 labeled examples, but used a network pre-trained on ImageNet. In contrast, we use only 113 images with no pre-trained model available since our "pixels" are time series signals. In our setting, task supervision helps prevent loss of discriminative information during domain adaptation. The stability problem in our formulation concerns the domain classifier and it is addressed with a variance regularization that is simpler than both JADA and IDDA approaches. In contrast to JADA, our variance regularization method does not rely on randomization in task classifier initialization to prevent mode collapse. Our work details a new method for reducing mode collapse in the domain discriminator that improves domain adaptation in a small data setting.

III. Methods

A. Motivation

Our deep learning framework builds on our previously reported neural network structure [15], shown in Fig. 1. Our original framework uses two blocks, an autoencoder and a classifier, that are simultaneously trained on a single data set. The autoencoder block provides regularization by reconstructing the input signal, but cannot ensure a task-discriminative bottleneck representation. Adding the task classifier provides supervision, ensuring the autoencoder generates task-discriminative representations in the bottleneck. While our original framework performed well when trained on a single data set, we observed a decrease in performance when training with multiple data sets. We extend our original framework to work in a multi-center setting by adding a domain adaptation module. This approach aims to create a feature representation that cannot distinguish between data collected from different imaging centers.

B. Domain Adaptation Using Gradient Reversal Layer

Fig. 2 describes our proposed approach, which combines the autoencoder-classifier framework with a domain classifier with gradient reversal. The model works at the pixel level, providing a

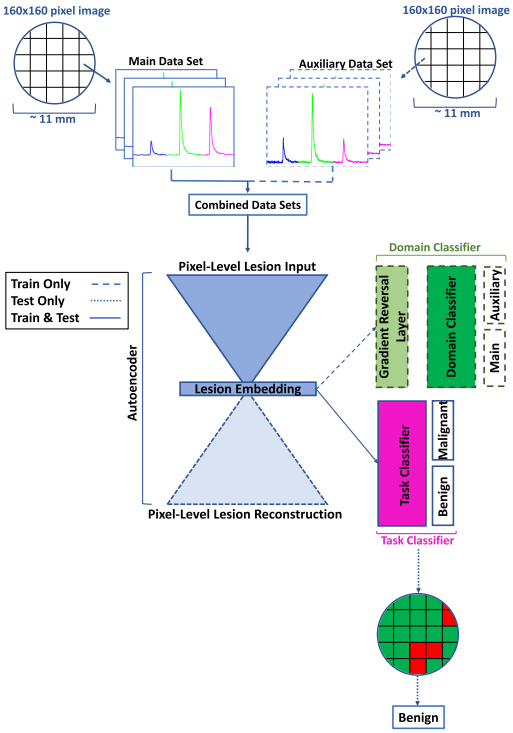


Fig. 2. Proposed architecture with autoencoder, task classifier and domain classifier implemented with gradient reversal layer. Two data sets from separate imaging centers are preprocessed to minimize center-specific differences and combined into a single data set before training the neural network with domain adaptation on a pixel level. At test time, the task classifier pixel-level diagnosis is aggregated for each image with a 50% threshold.

prediction for each pixel. Our architecture has three components: a feature extractor built with an autoencoder, a task classifier, and a domain classifier. In our architecture, the contracting path of the autoencoder, from the input to the bottleneck layer, plays the role of the feature extractor. Our domain classifier is trained to recognize center-specific differences in order to distinguish between the data sets (i.e., center one versus center two). The task classifier labels each pixel with a diagnosis (i.e., it classifies each pixel as benign or malignant). While margin delineation (classification of a pixel as a lesion or healthy) is another common task in oral maFLIM analysis, we only consider the diagnosis task in this work. The feature extractor is connected to the task and domain classifiers. The network is trained with an adversarial loss to maximize the performance of the task classifier while minimizing the performance of the domain classifier. We find a saddle point where the parameters of the feature extractor minimize the domain classifier performance. The key step in this formulation to minimize the performance of the domain classifier is to reverse the sign of the gradient updates generated by the domain classifier. When the performance of the domain classifier is at random level, the bottleneck layer generates feature representations that are indistinguishable between both domains, compensating for potential differences. Our network looks for the parameters at the saddle point:

$$(\hat{\theta}_{enc}, \hat{\theta}_t) = \text{argmin}\mathcal{L}_{total}(\theta_{enc}, \theta_t, \hat{\theta}_d) \quad (1)$$

$$\hat{\theta}_d = \text{argmax}\mathcal{L}_{total}(\hat{\theta}_{enc}, \hat{\theta}_t, \theta_d) \quad (2)$$

where $\hat{\theta}_{enc}$, $\hat{\theta}_t$, and $\hat{\theta}_d$ refer to the optimal parameters for the encoder, task classifier, and domain classifier, respectively. Equation 1 states that we find the parameters for the encoder and the task classifiers that minimize the task classifier loss, while (2) states that we want the parameters for the domain classifier that minimizes the domain classifier loss.

The trade-off between the domain classifier and the task classifier is controlled by the hyperparameter λ. The total loss ($\mathcal{L}_{total}$) is given by (3),

$$\mathcal{L}_{total} = \alpha L_{task} + \beta L_{AE} + \mathcal{L}_{VR} - \lambda \mathcal{L}_{domain} \quad (3)$$

where $\mathcal{L}_{task}$ is the cross entropy loss for the task classifier, $\mathcal{L}_{AE}$ is the reconstruction loss for the autoencoder, $\mathcal{L}_{VR}$ is a variance regularization penalty discussed in Section III-C, and $\mathcal{L}_{domain}$ is the domain classifier loss. The negative sign on the term $\lambda\mathcal{L}_{domain}$ reflects the reversal of the gradient for the domain classifier. As in Ganin and Lepinsky [20], the hyperparameter λ gradually increases during training, allowing the domain discriminator to distinguish the classes before applying large updates to merge the domains.

Since the relative strength of the domain adaptation increases during training as λ increases, the model takes some time to achieve domain invariance. Ideally, the domain classifier should approach 50% accuracy on both data sets. This scenario indicates that the domain classifier can no longer tell the difference between the domains (the domain classifier has converged). If we choose the best model before the domain classifier converges, good performance on only one domain may dominate the task classifier loss, while the other domain may have poor classification performance. Thus, we select the best model only after the validation set domain performance falls near 50% for both the main and auxiliary data. Then, we select the best model as determined by the task classifier performance on the validation set.

After training, the network predicts a label for each pixel in the test images. The predicted diagnosis for an image is generated by aggregating the pixel-level predictions using a 50% threshold. Although better performance may be achieved by determining the optimal threshold, we use the majority class of an image's pixels as the final image label for simplicity and to avoid adding an extra hyperparameter that may lead to overfitting given the size of the corpus. The long dashed borders and arrows in Fig. 2 indicate that the auxiliary data is only used during training. Similarly, the decoder and the domain classifier are discarded after training. The small dashed arrows indicate that the image-level aggregation only applies during testing.

C. Strategies to Increase Regularization

Since we are dealing with small data sets, we implement several approaches to increase the regularization of the architecture, improving the generalization of the model to avoid overfitting. Using an autoencoder is our first regularization approach. The autoencoder includes an unsupervised loss to reconstruct the input features. While we do not have many images, we do have over 2.5 million pixels to train this model.

Our second approach to increase regularization is the use of dropout in the classifier layers. Dropout randomly removes nodes during training, resulting in a model that approximates an average over a collection of sub-models [26]. Dropout discourages the model from relying on specific combinations of nodes (referred to as "co-adaptation") and reduces overfitting [26]. Each layer in the encoder and decoder used batch normalization. Batch normalization was introduced by Ioffe and Szegedy [27]

to minimize internal covariate shift and was placed before the activation layer. However, experiments by Santurkar et al. [28] show that internal covariate shift is not always reduced by batch normalization. Instead, the likely performance improvements may be due to a smoother loss surface during the optimization [28]. The results of batch normalization can also regularize the network by encouraging favorable model initialization and increasing generalization [28]. While Ioffe and Szegedy [27] asserted that batch normalization can reduce the need for dropout, Chen et al. [29] found improvements in combining batch normalization with dropout and using these layers after the activation. In our framework we do not use batch normalization and dropout in the same layer. Instead, we use batch normalization in the autoencoder layers and dropout in the task classifier layers. However, we acknowledge that performance increases may be obtained by exploring alternatives such as the combination of dropout and batch normalization in the same layer.

A technical contribution in this study is the use of variance regularization for adversarial domain adaptation. Although we experimented with multiple schedules for the hyperparameter λ, we were unable to stabilize the domain classifier with hyperparameter tuning alone. The domain classifier was unstable, reaching the expected 50% accuracy by collapsing to the predictions of a single domain and switching between which domain was predicted. During preliminary experiments, we found switching was reduced by changing the activation function to a sigmoid at the bottleneck layer (see original activations in Fig. 1). However, the sigmoid activation did not entirely stabilize the training of the model, with the domain classifier still frequently collapsing. One way domain collapsing can happen is when the gradient updates cause the model to have low diversity in the embedding representation that feeds directly the domain classifier. In this case, the domain discriminator will collapse to predict a single class. This phenomenon is similar to the well-studied mode collapse problem in GANs, where the generator will only produce samples of one or few classes. The variance regularization penalizes the model when the variance in the embedding drops below a threshold to reduce the likelihood of our model to converge to this poor solution. We implement the variance regularization proposed by Chong et al. [30], shown in (4),

$$\mathcal{L}_{\mathcal{VR}} = \max\left\{0, \gamma - \frac{1}{p(n-1)} \sum_{q=1}^{p} \sum_{i=1}^{n} \left(\phi_{i,q} - \frac{1}{n} \sum_{k=1}^{n} \phi_{k,q}\right)^2\right\} \tag{4}$$

where $\mathcal{L}_{\mathcal{VR}}$ is the variance regularization penalty, γ is the threshold, p is the size of the bottleneck (16 in our implementation), n is the number of samples in the mini-batch (256 in our implementation), and ϕ is the value of the feature. The only change from the approach presented by Chong et al. [30] is that our threshold γ is a single value that does not change as the training progresses. We apply the regularization to the bottleneck layer, adding a penalty to the loss when the variance within a mini-batch falls below a threshold. With this approach, the input to the domain classifier is more varied and reduces the likelihood that the domain classifier collapses to predict every sample belonging to one domain.

D. Imaging System

The basis for fluorescence imaging of oral cancer is that changes in the levels of endogenous fluorophores may reflect changes in tissue structure and metabolism. Collagen, reduced *nicotinamide adenine dinucleotide* (NADH), and *flavin adenine dinucleotide* (FAD) are endogenous fluorophores of interest in cancer diagnosis and margin delineation [1], [2], [3], [31]. Several morphological and biochemical changes within the two primary layers of the human oral mucosa have been shown to accompany the development of *squamous cell carcinoma* (SCC). We can investigate these changes by studying the autofluorescence properties of the two primary layers of the human oral mucosa (the stratified squamous epithelium and lamina propria or connective tissue). Morphological changes include thickening of the epithelium and extracellular matrix remodeling in the lamina propria, which lowers the measured collagen autofluorescence [32].

Several publications show that the development of SCC causes changes in the concentration of the reduced form of NADH and FAD in the epithelial tissue [33], [34], [35]. Autofluorescence properties are typically quantified by two values: the intensity and the lifetime. The ratio of the intensity between these two endogenous fluorophores is referred to as the optical redox ratio and can indicate changes in the cells metabolic processes [36], [37], [38], [39]. A decrease in this ratio is associated with an increase in the metabolic rate of cells, a hallmark of neoplastic cell transformation [40]. In addition to the intensity, changes in protein binding affect the lifetime of NADH and FAD. Skala et al. [41] showed that these changes in lifetime accompany carcinogenesis [42]. The excitation and emission bands of the endoscope used in our study are designed to measure the autofluorescence properties of collagen, NADH and FAD [31].

All images ($\sim$11 mm circular field of view, $\sim$100 um lateral resolution) were acquired using the maFLIM system described in Cheng et al. [31]. Using this system, the tissue autofluorescence was excited at 355 nm with a pulse width of 1 ns. During the acquisition time, which was less than 3 s, 2.8 mJ were deposited to the tissue (*maximum permissible exposure* (MPE) = 29.8 mJ [43]). The emission bands imaged collagen (390$\pm$20 nm), NADH (452$\pm$22.5 nm), and FAD (>500 nm). Following imaging of the biopsy region, an incisional tissue biopsy was performed following standard clinical protocols. The histopathological diagnosis for each lesion biopsy was used as the ground truth for training and evaluating the proposed classifier. This process results in an image where each pixel contains a three-channel fluorescence decay. Fig. 3 shows an example of this three-channel fluorescence decay. As shown in the figure, the raw images provide no direct visual information on cancer diagnosis or lesion appearance. Feature representations need be estimated from the data.

E. Data: Main and Auxiliary Sets

Data was collected from two imaging centers. Each imaging center used two separate prototype versions of the maFLIM endoscope system previously reported in Cheng et al. [31]. At both centers, each patient in the study had a clinically suspicious oral lesion. Each lesion was imaged in vivo with the maFLIM endoscope, followed by surgical resection and histopathological diagnosis. Each lesion was imaged only once following our protocol (i.e., we do not have multiple images from each lesion). All images contained 160 $\times$ 160 pixels, with three channel decays per pixel corresponding to collagen, reduced NADH, and FAD, respectively. In both data sets, we use the raw signal instead of iterative deconvolution to avoid signal approximations. We rely on our neural network to adjust for differences in the impulse response of the imaging systems.

The main data set contains 67 oral lesions from nine anatomical locations, as shown in Table I. The data was collected in Doha, Qatar. The *institutional review board* (IRB) for the main data collection and processing was approved by the Hamad

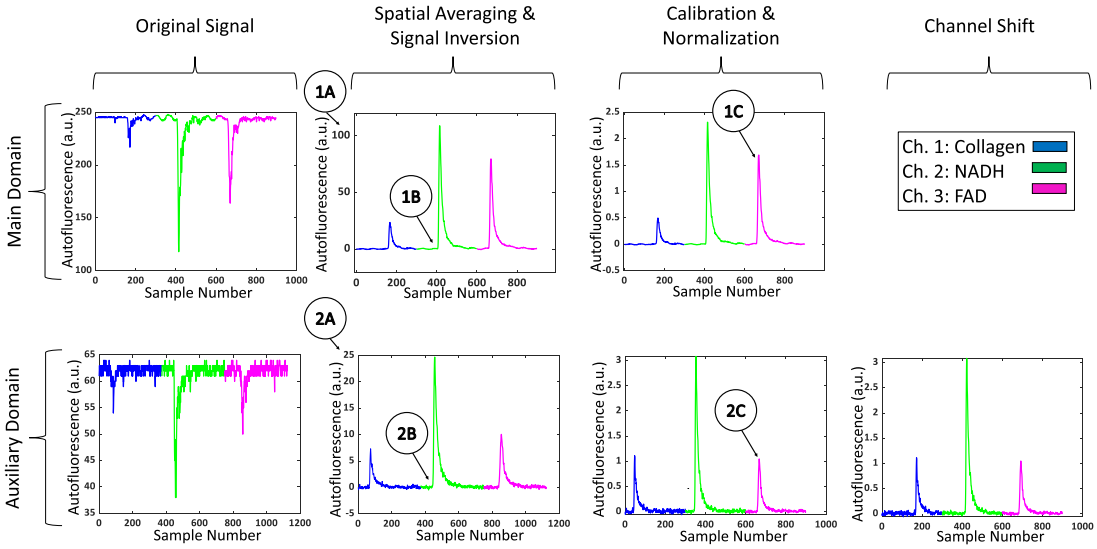


Fig. 3. Visualization of the preprocessing steps for main and auxiliary domains. Each domain was separately preprocessed to reduce variation from gain and peak locations. Peak value trends by channel remained different between domains, even among images from the same class. The arrows labeled 1A-2 C denote points of interest discussed in Section III-G. Top row: main domain. Bottom row: auxiliary domain.

TABLE I
ANATOMICAL DISTRIBUTION OF THE LESIONS IN THE MAIN DATA SET COLLECTED IN DOHA, QATAR

Location	Benign	Dysplasia	SCC
Mucosa	10	3	9
Floor of Mouth	2	0	1
Gingiva	0	2	3
Lip	10	0	2
Mandible	0	0	1
Palate	1	0	0
Maxilla	0	0	1
Retromolar	1	0	0
Tongue	9	0	12
Total	33	5	29

TABLE II
ANATOMICAL DISTRIBUTION OF THE LESIONS IN THE AUXILIARY DATA SET COLLECTED IN DALLAS, USA

Location	Benign	Dysplasia	SCC
Mucosa	8	1	2
Floor of Mouth	1	0	0
Gingiva	5	1	5
Mandible	1	0	1
Palate	1	0	0
Tongue	7	7	6
Total	23	9	14

Medical Corporation in Doha, Qatar (with study number HMC-IRB 16332/16, and approval date of January 29, 2019). The histopathological diagnosis revealed that 33 lesions were benign, 5 lesions were dysplasia, and 29 lesions were *squamous cell carcinoma* (SCC).

The full auxiliary data set contained 84 oral lesions, including 9 dysplasia, 14 SCC, and 61 benign lesions. The data was collected in Dallas, USA. To avoid introducing highly imbalanced data, we used only 23 of the 61 benign lesions from the auxiliary data set. Our preliminary results show that using different sets of randomly selected benign images led to minimal variations in the average between sensitivity and specificity. Therefore, we used the same 23 benign lesions for all experiments reported in our paper. Therefore, the auxiliary data used in this study contains 46 oral lesions from 6 anatomical locations (see Table II). IRB approval for the auxiliary data set experimental protocol was obtained from Texas A&M University (with protocol number IRB2010-0177D, and approval date of October 22, 2014). Notably, the auxiliary data set contains a larger portion of dysplasia cases compared to the main data set distribution. Dysplasia cases from both data sets were considered as malignant in our problem formulation. In our binary classification task, the positive class consists of dysplasia and SCC lesions, and the negative class consists of benign lesions.

In addition to different anatomical locations and dysplasia distributions, the experimental protocol was not consistent between imaging centers. Specifically, the main data was collected after calibrating the imaging system before each image acquisition. In contrast, the auxiliary data set was initially calibrated, but not subsequently adjusted throughout the experiment. The lack of precise calibration factors corrupts the relative intensity values between channels, especially without patient normalization. While an image from the normal contralateral side was collected for each patient, this study only uses the lesion information, resulting in greater challenges due to calibration differences.

TABLE III
MODEL STRUCTURE OF OUR ARCHITECTURE. THE DESCRIPTION INCLUDES THE ENCODER, DECODER, TASK CLASSIFIER AND DOMAIN CLASSIFIER

Block	Layer	Size In	Size Out	Activation
Encoder	1	900	512	ReLU
	2	512	256	ReLU
	3	256	128	ReLU
	4	128	16	Sigmoid
Decoder	1	16	128	ReLU
	2	128	256	ReLU
	3	256	512	ReLU
	4	512	900	ReLU
Task Classifier	1	16	21	ReLU
	2	21	14	ReLU
	3	14	2	Softmax
Domain Classifier	GR	16	16	None
	1	16	8	Sigmoid
	2	8	2	Softmax

F. Implementation

Table III shows the sizes and activations of the layers for the autoencoder and classifier. Each layer is fully conncected. The autoencoder accounts for most of the parameters, with 1,261,500 trainable parameters. The task classifier adds 695 parameters,

while the domain classifier used for gradient reversal adds only 154 parameters. The total number of *floating point operations per second* (FLOP) (as calculated using the keras-flops library) is 0.00253 G, which is significantly lower than conventional image-based networks. Though the domain classifier slightly increases the model complexity during training, the domain classifier is not used during inference. Once the model and data is loaded, our implementation (AMD Ryzen 9 3900X 12-Core Processor, NVIDIA GeForce RTX 3090) generates a pixel prediction in an average of 0.02 ms (i.e., approximately 512 ms per image). Images from the separate data sets are separately preprocessed to reduce correctable variations between data sets. We describe the preprocessing steps in Section III-G. After preprocessing of each pixel, the data sets are combined and the classifier is trained on a pixel level.

Our training auxiliary data set is smaller than that of the main data set (Sec. III-E). Therefore, we use sample weighing on the domain classifier output to prevent the model from collapsing to only predicting the domain with more samples. The values for the sample weights are given by the *sci-kit learn* toolkit based on the training domain distribution as detailed in Pedregosa et al. [44]. Similarly, we reduce class bias by using the same strategy on the task classifier, generating the sample weights from the class distribution of the training data.

The dropout rates are set to $p = 0.5$ and $p = 0.25$ for the task and domain classifiers, respectively. We use Adam [45] as the optimizer with a learning rate of 10^{-5}, training all the models for 25 epochs. All models are trained using Keras with Tensorflow [46]. The hyperparameter for the gradient reversal layer (λ) is initialized to zero and incremented by 0.025 for five epochs. After five epochs, we reduced the increment to 0.015 for the remainder of the training. The threshold γ for the variance regularization is set to 0.05 for all experiments, based on the stability of the domain classifier observed on the validation set.

Our training strategy for the main data relies on cross-validation given the limited size of our data sets. We split the data into 10 partitions, with each partition as class-balanced as possible. One partition is reserved as the test set, which is only used to evaluate the performance of the system. From the remaining nine partitions, two are randomly selected as the validation set, and the other seven partitions as the train set. We denote the experiments in a specific division of the data into train, validation and test sets as a *run*. The data is partitioned by patient, where all the pixels of a single image belong only to the train, validation, or test set within a specific run. With the cross-validation, every partition is eventually used as the test set. We build the system using the train set, maximizing the performance on the validation set. The final model is then evaluated on the test set. We refer to this cross-validation process as a *trial*. Since the partitions of the data set can affect the performance of the system, we repeat this process 10 times, creating different partitions for each trial. We report the average results across the 10 trials. In addition to the train and validation sets for the main data, we append train and validation partitions from the auxiliary data set. The same auxiliary data is added to the training and validation sets in each run. We use 34 lesions in the auxiliary training set and 12 lesions in the auxiliary validation set.

Across all experimental conditions, we use consistent data splits to minimize sources of variation across conditions due to variations in the pairing of train, validation, and test sets. Therefore, the comparisons across baselines are consistent since they are using the same partitions.

TABLE IV
BENIGN VERSUS MALIGNANT CLASSIFICATION RESULTS FOR THE PROPOSED APPROACH AND THE BASELINES

Model	Sens.	Spec.	Avg.	Prec.	F1	Acc.
AE [15]	87.50	67.58	77.54	76.25	79.80	77.62
AE Joint	70.08	81.17	75.62	75.68	70.80	75.74
AE (modified)	85.08	68.42	76.75	74.74	78.04	76.93
AE (modified) Joint	77.33	74.92	76.13	74.20	73.63	76.21
SVM SFS [15]	81.00	67.25	74.13	73.82	74.92	74.05
SVM SFS Joint	82.00	50.83	66.42	65.41	70.78	66.57
SVM L1 [15]	79.17	73.33	76.25	78.10	75.92	76.36
SVM L1 Joint	88.00	63.50	75.75	73.48	78.32	75.74
Proposed Approach	80.83	75.75	78.29	79.09	77.73	78.05

Note: sensitivity equals recall.

G. Preprocessing

Both the main and auxiliary data sets were preprocessed with signal inversion, spatial averaging, and SNR masking to reduce noise and reject pixels with low or saturated SNR. Since the spatial averaging used a sliding window, the original number of pixels in each image is maintained. We did not average all pixels from the image into a single decay. Each channel was chopped or zero padded to a consistent length. In addition, each data set had individual preprocessing steps to reduce variations between the domains. Fig. 3 shows an example pixel from each domain and the initial preprocessing steps.

The preprocessing for each data set is identical through the second column in Fig. 3, which shows the fluorescence decay following spatial averaging and signal inversion. Arrows 1 A and 2 A in the figure highlight the first dataset-dependent variation, where the different gain used in image acquisition changes the peak value of the channels by 100 units. In the next preprocessing step, the decays are calibrated and normalized to sum to 100 to put the peak values on a similar scale between data sets. In addition, each main data set channel had a temporal resolution of 0.25 ns and length of 300 samples after zero padding. The auxiliary data set had a temporal resolution of 0.16 ns and length of 375 samples after zero padding. The auxiliary data set was interpolated to match the length and sample rate of the main data set. The peak of each channel in the auxiliary data set was also shifted to approximately match the peak of each channel in the main data set (see Arrows 1B and 2B in Fig. 3). Finally, Arrows 1 C and 2 C show an unmitigated difference between images. Both pixels represented in the figure are from benign lesions. However, the ratios between channels peaks are different, with the main domain pixel showing a much higher channel 3 peak than that of the auxiliary domain. While this could partially be due to inter-patient variability, the difference highlights the difficulty in merging images from heterogeneous, small data sets. The relationship between the pre-processed data and the overall image is noted in Fig. 2, where each pixel in the 160x160 image contains a time series signal similar to the one shown in Fig. 3.

IV. EXPERIMENTS

All experimental results are reported using the main domain on the cancer diagnosis task (classification of benign versus malignant). While we report a total of six metrics, we focus our analysis on sensitivity, specificity, and the average between sensitivity and specificity.

A. Comparison With Baselines

We report baseline results from a variety of methods, as shown in Table IV. We use SVM and different feature selection

methods with standard features as input, following the work of Jo et al. [2] and Marsden et al. [1]. The standard feature set includes 21 features derived from signal lifetime, intensity, and bi-exponential decay parameters. The lifetime (τ_k) and intensity (I_k) of a single channel are given by (5) and 6 below:

$$\tau_k = \frac{\int t h_k(t) \partial t}{\int h_k(t) \partial t} \tag{5}$$

$$I_k = \int h_k(t) \partial t \tag{6}$$

where t specifies the sample time point and h_k denotes the deconvolved fluorescence decay for channel k. The deconvolved fluorescence decay was obtained by iterative least-squares deconvolution using a bi-exponential decay model, as shown in (7):

$$h_k = \alpha_{fast,k} e^{-t/\tau_{fast,k}} + \alpha_{slow,k} e^{-t/\tau_{slow,k}} \tag{7}$$

where $\alpha_{fast,k}$ and $\alpha_{slow,k}$ are the coefficients of the two exponential terms and $\tau_{slow,k}$ and $\tau_{fast,k}$ determine the decay rates of the exponential terms. We also use two different methods to focus the model on the best features: *sequential feature selection* (SFS) and L1 regularization. SFS sequentially chooses the best features, while L1 regularization produces a sparse solution, where coefficients corresponding to less discriminative features approach zero during training. We report results for SVM classifiers for single-domain training, as well as the results for training with the main and auxiliary sets (i.e., training and validation sets), denoted as *joint* in Table IV.

In addition to the SVM classifiers with standard features, we list the single-domain results using the autoencoder and classifier neural net shown in Fig. 1 and reported by Caughlin et al. [15]. This neural network structure uses the pre-processed fluorescence decays without iterative deconvolution and generates data-driven features. As detailed in Section III, the autoencoder for our approach is modified with a sigmoid activation at the bottleneck. We report the single-domain training performance using the modified autoencoder in Table IV to ensure that the improved results with our method are not due to an improved autoencoder structure. While the change in activation helps domain stability with gradient reversal, the sigmoid activation may lose some performance benefits resulting from sparsity in the bottleneck representation when using the *rectified linear unit* (ReLU) activation.

Table IV shows the results. Joint training reduces the average main domain performance in all baseline models. For the modified autoencoder method, joint training reduces the average performance by 0.62% compared with the main only model. The performance of the joint SVM model with L1 regularization dropped an average of 0.5% compared with the main only model. Joint training with SVM and SFS completely failed, with an average performance of 7.71% below the single domain SVM SFS model. The decrease in average performance highlights the problem of domain shift, which causes the main performance to decrease when the model is trained with both main and auxiliary sets, even in a fully-supervised setting. Furthermore, a similar performance drop in models trained using hand-crafted features from the deconvolved signal (SVM models) and the pre-processed signal without deconvolution (autoencoder models) shows that the domain shift problem cannot be solely attributed to differences in the imaging system impulse response.

In contrast to all other baseline models, our full neural network with domain adaptation and variance regularization improves the main domain performance over all other methods including the main-only training settings. Our full model significantly increases the average performance over the best SVM multi-domain baseline by 2.54% ($p = 0.0341$ using a one-tailed, paired t-test). Similarly, our full model shows no significant change in specificity and significantly increases sensitivity over the single-domain training using the same base structure (i.e., modified autoencoder trained on the main data only) by 7.33% ($p = 0.0023$ using a two-tailed, paired t-test). The improved results show that domain shift between data centers can be reduced using gradient reversal.

TABLE V
CONTRIBUTIONS OF THE MODEL'S COMPONENTS. THE BEST MODEL USES BOTH DOMAINS IN TRAINING (JOINT), GRADIENT REVERSAL (GR), AND VARIANCE REGULARIZATION (VR)

Joint	GR	VR	Sens.	Spec.	Avg.	Prec.	F1	Acc.
		✓	85.33	64.67	75.00	73.89	77.27	75.19
✓		✓	79.75	72.83	76.29	76.23	75.34	76.26
✓			77.33	74.92	76.13	74.20	73.63	76.21
✓	✓		80.00	71.67	75.83	74.97	75.06	75.67
✓	✓	✓	80.83	75.75	78.29	79.09	77.73	78.05

Note: sensitivity equals recall.

B. Contributions of the Model's Components

To evaluate the contributions of each component in our domain adaptation framework, we present a breakdown of each component in Table V. *Joint* refers to training with both main and auxiliary data in the training and validation sets. *GR* refers to joint training with gradient reversal. *VR* refers to the addition of variance regularization to the bottleneck representation (discussed in Section III). Comparing rows two (Joint) and three (Joint + GR) from Table V, we find that gradient reversal failed to correct the performance drop when the variance regularization is not used. This setting produces similar main performance to the model trained without any domain adaptation. After analyzing the domain performance during adaptation, we found evidence of domain collapse where the domain classifier always predicts a single class. When this happens, the domain adaptation module does not achieve its goal of reducing the mismatch between main and auxiliary sets. We provide further analysis on domain collapse and the effect on domain adaptation in Section IV-D. The domain collapse problem is corrected with the variance regularization. Since domain adaptation failed until the addition of variance regularization, we confirm that the increased performance of the full model is due to an appropriate implementation of domain adaptation, rather than an unforeseen effect of variance regularization by training our main-only autoencoder model with variance regularization at the bottleneck (row 1 in Table V). As expected, the addition of variance regularization outside of domain adaptation reduces average performance by 1.75% compared with our base autoencoder. When training the joint model with variance regularization but not gradient reversal, the average performance is within 0.16% of the joint only model. This result shows that the variance regularization is not effective without gradient reversal. The full model using joint training, gradient reversal, and variance regularization successfully merged the domains, with an increase of 3.50% sensitivity and 0.83% specificity compared to joint training without domain adaptation. Furthermore, the full domain adaptation model outperforms single-domain classifier performance using the same autoencoder by an average of 1.54%. Taken together, Table V shows that all three components of our model (multi-center training, gradient reversal, and domain adaptation) are required to successfully train using multiple data sets.

TABLE VI
EFFECT OF DOMAIN STABILITY IN MODEL SELECTION. JOINT + GR: JOINT MODEL TRAINED WITH THE MAIN AND AUXILIARY DATASETS IMPLEMENTED WITH GRADIENT REVERSAL

Model	MS	Sens.	Spec.	Avg.	Prec.	F1	Acc.
Joint+GR	–	77.17	72.33	74.75	73.31	73.12	74.70
	✓	80.00	71.67	75.83	74.97	75.06	75.67
Joint+GR+VR	–	81.50	74.25	77.87	77.84	77.45	78.21
	✓	80.83	75.75	78.29	79.09	77.73	78.05

Joint + GR + VR: joint model with gradient reversal and variance regularization. MS: model selection.
Note: sensitivity equals recall.

TABLE VII
EFFECT OF AUXILIARY DATA IN MODEL SELECTION. USING THE AUXILIARY DATA IN THE VALIDATION SET (AUX. DEV.) ONLY IMPROVES PERFORMANCE WHEN THE DOMAINS ARE MERGED USING GRADIENT REVERSAL (GR) AND VARIANCE REGULARIZATION (VR)

Model	AD	Sens.	Spec.	Avg.	Prec.	F1	Acc.
Joint	–	78.33	75.33	76.83	72.91	73.71	76.86
	✓	77.33	74.92	76.13	74.20	73.63	76.21
Joint+GR	–	79.83	74.75	77.29	74.89	75.22	77.31
	✓	80.00	71.67	75.83	74.97	75.06	75.67
Joint+GR+VR	–	80.92	72.25	76.58	77.10	76.46	76.81
	✓	80.83	75.75	78.29	79.09	77.73	78.05

Joint: model trained with the main and auxiliary datasets. Joint + GR: joint model implemented with gradient reversal. Joint + GR + VR: joint model with gradient reversal and variance regularization. AD: auxiliary in development set.
Note: sensitivity equals recall.

C. Model Selection Constraints

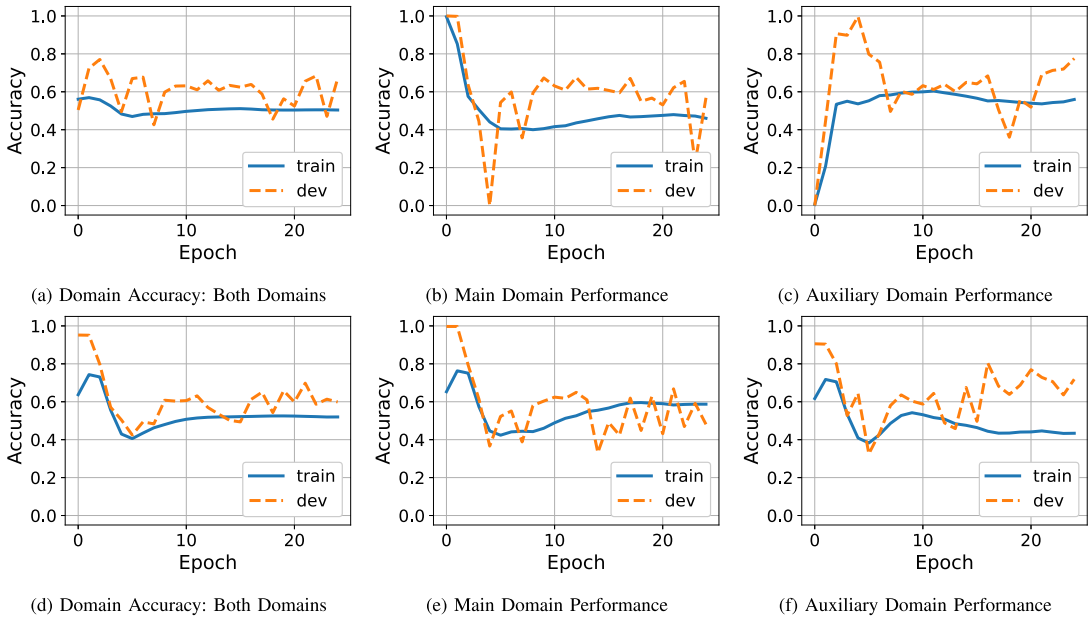
In addition to the use of domain adaptation, the model performance is also affected by the model selection criteria. In general, the best model from the training process is chosen by the performance on the validation set. However, in our domain adaptation framework, the task and domain classifiers are trained at the same time. However, the trade-off hyperparameter for gradient reversal increases over training. If the best model is selected based on the minimum validation set performance without considering domain convergence, the model may be selected before the domains have merged. To prevent sub-optimal model performance, we used a model selection constraint in our full model, as described in Section III-B. We verify the efficacy of our model selection constraint for two settings. One setting uses joint training and gradient reversal, and the other setting uses joint training, gradient reversal and variance regularization. Table VI shows the results when we either select the model with the best results on the validation set, even if the domain adaptation has not fulfilled its role, or we select the best results in the validation set after waiting for the domain adaptation to converge (denoted with the symbol ✓). In both settings, adding the model selection constraint improves average performance by 1.08% and 0.42%, respectively. The greater need for model selection criteria without variance regularization may be due to increased domain instability.

We hypothesized that when the domains have merged, the inclusion of the auxiliary data in the validation set will increase the main performance. The validation set is used to identify the best model to be used in the test set. If the domains have not merged, the validation performance may be skewed by higher classification performance on the auxiliary data set, affecting the performance of the main domain. We test this hypothesis for all three joint training scenarios: joint training without domain adaptation (Joint), joint training with gradient reversal (Joint + GR), and joint training, gradient reversal and variance regularization (Joint + GR + VR). Notice that in the three conditions, the auxiliary set is used on the training set. Table VII shows the results, denoting with the symbol ✓ when the auxiliary data is included in the validation set. When the variance regularization is not used and the system ineffectively merges the domains (Joint and Joint + GR models), removing the auxiliary data from the model selection (i.e., validation set) resulted in better performance on the main domains. Specifically, the joint model without domain adaptation dropped 0.70% on average. The joint model with domain adaptation dropped 1.46%. The decrease in performance with the auxiliary validation data indicates that any advantages from a larger, more diverse data set were offset by the domain shift. In contrast, our full model with variance regularization and gradient reversal benefits from the addition of auxiliary data in the validation set by 1.55% on average. Therefore, Table VII further confirms our hypothesis that the main domain performance can be increased by the inclusion of data from another domain when domain shift has been corrected with domain adaptation.

D. Domain Adaptation Verification

We plot the domain accuracy to monitor the effects of gradient reversal. Fig. 4 shows the performance of the train and validation set. We set the gradient reversal hyperparameter $\lambda = -0.25$ for the first epoch and increment it by 0.025 for five epochs and 0.015 thereafter. Note that this setting only applies to this section to verify domain adaptation. For the rest of the evaluation, we begin with $\lambda = 0$ and gradually increment it. The negative sign allows the domain accuracy to initially increase, showing that the domain classifier is working properly. Once the value of λ increases and domain adaptation progresses, the domain accuracy should reach near random performance, indicating domain invariance. The top row of Fig. 4 shows this process for the model without variance regularization. The validation accuracy reaches above 75.0% and then decreases as expected (Fig. 4(a), epoch 3). However, the way the model achieves 50% domain accuracy matters. For example, Kim and Kim [47] show that semi-supervised domain adaptation does not ensure that the representations are well-ordered. Rather than creating a compact and homogeneous mix of the domains, the main domain can encompass distinct clusters of the target domain and prevent improvements in target performance [47]. Similarly, we find that the domain classifier can reach a 50% accuracy by always predicting a single domain. In the top row of Fig. 4, the initial domain performance is split between the main and auxiliary data sets, with the main accuracy at 100% (Fig. 4(b), epoch 1) and the auxiliary performance at 0% (Fig. 4(c), epoch 1). Furthermore, the domain performance again collapses at epoch 4, when the main domain performance is 0% (Fig. 4(b), epoch 4) and the auxiliary domain performance is 100% (Fig. 4(c), epoch 4). When the domain classifier predicts a single domain, the overall domain accuracy calculated using the validation sets from both domains reaches ∼50%, but incorrectly indicates domain invariance. In contrast, the model with variance regularization reaches approximately 95% initial accuracy (Fig. 4(d), epoch 2), with similar domain accuracy for both domains by the end of training (Figs. 4(e) and 4(f), epoch 25). The overall performance of the domain classifier is near random by the end of training. Additionally, the collapse at epoch 4 is not present, showing that the model with variance regularization had greater stability.

We expect the changes in domain invariance to appear in *t-distributed stochastic neighbor embedding* (T-SNE) plots as

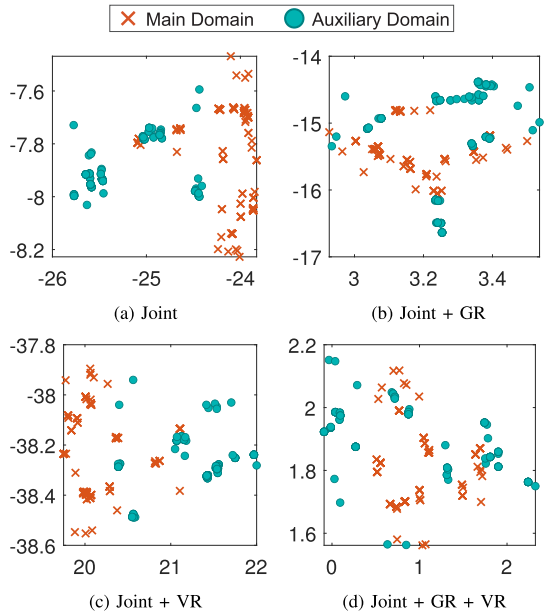


(a) Domain Accuracy: Both Domains (b) Main Domain Performance (c) Auxiliary Domain Performance

(d) Domain Accuracy: Both Domains (e) Main Domain Performance (f) Auxiliary Domain Performance

Fig. 4. Importance of domain classifier performance for each domain (Solid lines: training set. Dashed lines: validation set). Top row: Domain classifier performance for the model without variance regularization. Bottom row: Domain classifier performance for the model with variance regularization. Left column: Domain accuracy for the validation set using both domains. Center column: Domain accuracy for the validation set samples from the main data set. Right column: Domain accuracy for the validation set samples from the auxiliary data set.

improved domain invariance. T-SNE plots use distribution matching to create a lower dimensional representation in which points that are close in the original distribution remain nearby in the reduced representation [48]. We randomly sample 10 data points from each image in the validation set using the 16-dimensional bottleneck representation to create the T-SNE plots. We use the MATLAB's implementation of T-SNE from the Statistics and Machine Learning Toolbox. We use the sampling of points to prevent the visualization from showing local structure associated with the similarity of data from each lesion rather than the more global structure associated with differences in the domains (Kobak and Berens [49] describe the trade-off between local and global structure visualization in medical data cases where data points easily cluster by subgroups in the data). The resulting T-SNE plots for the models trained with and without variance regularization are shown in Fig. 5. Rows 1 and 2 show the T-SNE embedding of the same points before and after gradient reversal without variance regularization and with variance regularization, respectively. In contrast, Fig. 5(d) shows that the model at the end of domain adaptation has the more distributed T-SNE plots, with the data from both domains mixed together. Taken together, the T-SNE plots show a two-dimensional visualization that indicates improved domain invariance following domain adaptation.



(a) Joint (b) Joint + GR

(c) Joint + VR (d) Joint + GR + VR

Fig. 5. Visualization of domain invariance using T-SNE plots. (main domain: x's. Auxiliary domain: circles). The axes represent the values of the projected embedding created by T-SNE. Top row: T-SNE representation of the bottleneck embedding for the Joint, and Joint + GR models without variance regularization. Bottom row: T-SNE representation of the bottleneck embedding for the Joint + VR and Joint + GR + VR models. Left column: Bottleneck representation before GR. Right column: Bottleneck representation after GR. Domain invariance improves during training with variance regulation.

E. Heartbeat Classification

To show the flexibility of our domain adaptation approach, we apply our method to another computer aided diagnosis problem using three publicly available arrhythmia databases from PhysioNet [50]: MITDB, SVDB, and INCART databases. The task is to use a single-lead *electrocardiogram* (ECG), as is commonly found in personal health monitoring devices, and our neural network model to detect three types of irregular heart beats. The MITDB, SVDB, and INCART databases are commonly used to evaluate domain adaptation algorithms [51], [52]. We chose to apply our method to heartbeat classification due to the similarities in the shape of the model input (i.e., a time series rather than an entire image such as *computed tomography* (CT) or MRI). Like our three biexponential-like FLIM input, we use three beats as the input to our model and classify the third

TABLE VIII
CLASSIFICATION RESULTS ON HEARTBEAT CLASSIFICATION TASK, REPORTED AS F1 SCORE

Main	Auxiliary	GR	N	VEB	SVEB	F
MITDB	-	-	0.91	0.50	0.17	0.00
MITDB	SVDB	-	0.94	0.85	0.43	0.04
MITDB	SVDB	✓	0.95	0.87	0.42	0.06
MITDB	-	-	0.91	0.50	0.17	0.00
MITDB	INCART	-	0.92	0.57	0.15	0.00
MITDB	INCART	✓	0.93	0.67	0.11	0.04

N: normal, VEB: ventricular ectopic beat, SVEB: supraventricular ectopic beat, F: fusion beat.

beat as *normal* (N), *ventricular ectopic* (VEB), *supraventricular ectopic* (SVEB), or *fusion* (F).

We use the same set up used for our oral cancer detection task, with a main domain (the MITDB corpus) and an auxiliary domain with a domain shift (either SVDB or INCART corpus). We define the main domain as the MITDB data set, splitting the samples into a train, validation, and test sets. We use the entire SVDB or INCART data set as the auxiliary data. We use the same method discussed in Section III-F for our oral cancer detection system. Since the input dimension of each heartbeat is 128 and we use three beats, the total input length is 384. We adjust the size of the autoencoder to account for the smaller input. Our adjusted encoder layers have input size 384, 256, 128, and 16 (bottleneck). The encoder and decoder are symmetric. In addition, we increase the size of the task classifier output to accommodate all four classes. Table VIII shows the result. The baseline model is improved by training on two domains without domain adaptation, indicating that the two datasets likely show less domain shift than the fluorescence lifetime images. However, our network with gradient reversal further improves results on the MITDB test set for N, VEB and F. We note that our method does not include features such as the pre-RR interval included in other networks, relying only on features extracted from the bottleneck in the autoencoder. The pre-RR interval provides the network with a summary of the patient's typical heart rate. Exploring a way to add this information into our network is likely to improve the results.

V. CONCLUSION

This study presented a neural network framework for merging data from multiple maFLIM collection sites using gradient reversal. This model is valuable for clinical practice because current oral cancer diagnosis requires a biopsy. Models such as ours will enable clinicians to immediately and non-invasively diagnose oral cancer. Our experiments showed that joint training data from two centers without considering the domain shift decreases classifier performance on an individual data set. However, on small data sets, gradient reversal suffers from domain collapse even in the supervised setting. We presented an architecture adjustment and a variance regularization method to stabilize the training process and successfully merge the two domains. In future work, we plan to use a similar method to create a margin delineation classifier (using data from multiple sites) that can help ensure the margin is fully removed during surgical resection. Effective margin removal is clinically challenging, but essential to prevent recurrence.

In addition to oral cancer diagnosis and delineation, our methodology can have relevance for the development of data-driven tools for other medical applications, as a common requirement is the access to comprehensive health data collected from different institutions, using diverse instrumentation. For example, our group is also developing a FLIM dermoscopy system for early detection and margin assessment of malignant skin lesions [53]. Our methodology can also be applied in very different medical diagnosis tasks that share a similar data type. Although arrhythmia detection is a different task than cancer detection, due to the type of data, we can also apply our method to heartbeat classification as demonstrated in our evaluation.

In future work, we plan to extend our approach to the semi-supervised setting, using task labels for the main domain only and introducing the images from the auxiliary domain as unlabeled data. Labeling the images is one of the difficult tasks since it involves surgical resection and histopathological diagnosis. Therefore, we expect that this semi-supervised approach will allow us to leverage more images. Likewise, we are interested in evaluating domain adaptation in related problems, where the domain shift can be expected to be larger. For example, the proposed approach can be useful in detection of skin and oral lesions. We expect domain shift from differences in sub-populations and imaging centers, as discussed here, as well as in the underlying tissue and pathology characteristics. Extension to domain adaptation for multiple imaging centers and lesion types will aid the validation of a single robust classifier that can be used by the same endoscope for multiple applications. In addition, we would like to explore the generation of domain invariant representations to new domains unseen during training. Domain adaptation is often used when two or more domains are used during training. Data from some domains may be unlabeled, but available during training. The setting where a model is trained on multiple domains and tested on a third domain not used in training is called domain generalization. In our paper, we focus on the domain adaptation problem. The unique formulation of our domain adaptation technique is that the auxiliary data is only used to augment the size of the train set. Unfortunately, we currently lack a third data set to further test our model in the domain generalization setting. However, we very recently began collecting data with an improved version of the imaging system at multiple locations. We are interested in exploring the domain generalization problem when we have collected enough data with the new system. When extending our approach to more than two domains, we expect regularization strategies (such as those we present) to become increasingly important. Studies have used domain adaptation techniques for multi target domains [54]. Therefore, we are confident that a variation of our approach can be effective in these settings.

REFERENCES

[1] M. Marsden et al., "Intraoperative margin assessment in oral and oropharyngeal cancer using label-free fluorescence lifetime imaging and machine learning," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 3, pp. 857–868, Mar. 2021.

[2] J. Jo et al., "Endogenous fluorescence lifetime imaging (FLIM) endoscopy for early detection of oral cancer and dysplasia," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Honolulu, HI, USA, 2018, pp. 3009–3012.

[3] P. Vasanthakumari et al., "Classification of skin-cancer lesions based on fluorescence lifetime imaging," *Proc. SPIE* vol. 11317, w2020, pp. 245–253.

[4] W. Li, S. Liao, Q. Feng, W. Chen, and D. Shen, "Learning image context for segmentation of the prostate in CT-guided radiotherapy," *Phys. Med. Biol.*, vol. 57, no. 5, 2012, Art. no. 1283.

[5] A. Das, I. Tashev, and S. Mohammed, "Ultrasound based gesture recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, New Orleans, LA, USA, 2017, pp. 406–410.

[6] N. M. Khan, N. Abraham, and M. Hon, "Transfer learning with intelligent training data selection for prediction of Alzheimer's disease," *IEEE Access*, vol. 7, pp. 72726–72735, 2019.

[7] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022.

[8] E. H. Pooch, P. Ballester, and R. C. Barros, "Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification," in *Proc. Int. Workshop Thoracic Image Anal.*, Lima, Peru: 2020, pp. 74–83.

[9] L. Zhang et al., "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2531–2540, Jul. 2020.

[10] M. Saiz-Vivó, A. Colomer, C. Fonfría, L. Martí-Bonmatí, and V. Naranjo, "Supervised domain adaptation for automated semantic segmentation of the atrial cavity," *Entropy*, vol. 23, no. 7, 2021, Art. no. 898.

[11] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, 2019, vol. 33, pp. 865–872.

[12] L. Diao, H. Guo, Y. Zhou, and Y. He, "Bridging the gap between outputs: Domain adaptation for lung cancer IHC segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 6–10.

[13] H. Zhang, J. Liu, P. Wang, Z. Yu, W. Liu, and H. Chen, "Cross-boosted multi-target domain adaptation for multi-modality histopathology image translation and segmentation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 7, pp. 3197–3208, Jul. 2022.

[14] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "Measuring domain shift for deep learning in histopathology," *IEEE J. Biomed. health Informat.*, vol. 25, no. 2, pp. 325–336, Feb. 2021.

[15] K. Caughlin et al., "End-to-end neural network for feature extraction and cancer diagnosis of in vivo fluorescence lifetime images of oral lesions," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Guadalajara, Mexico, 2021, pp. 3894–3897.

[16] R. Romano, R. T. Rosa, A. Salvio, J. Jo, and C. Kurachi, "Multispectral autofluorescence dermoscope for skin lesion assessment," *Photodiagnosis Photodyn. Ther.*, vol. 30, Jun. 2020, Art. no. 101704.

[17] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," *NPJ Digit. Med.*, vol. 5, no. 1, pp. 1–8, 2022.

[18] E. Duran-Sierra et al., "Machine-learning assisted discrimination of precancerous and cancerous from healthy oral tissue based on multispectral autofluorescence lifetime imaging endoscopy," *Cancers*, vol. 13, no. 19, 2021, Art. no. 4751.

[19] B. Chen et al., "Support vector machine classification of nonmelanoma skin lesions based on fluorescence lifetime imaging microscopy," *Anal. Chem.*, vol. 91, no. 16, pp. 10640–10647, Aug. 2019.

[20] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 1180–1189.

[21] X. Zhang, C. Broun, R. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP J. Adv. Signal Process.*, vol. 1, pp. 1228–1247, Jan. 2002.

[22] A. Anaby-Tavor et al., "Do not have enough data? deep learning to the rescue!" in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 7383–7390.

[23] S. Li, C. H. Liu, B. Xie, L. Su, Z. Ding, and G. Huang, "Joint adversarial domain adaptation," in *Proc. 27th ACM Int. Conf. Multimedia*, Nice, France, 2019, pp. 729–737.

[24] V. K. Kurmi, V. K. Subramanian, and V. P. Namboodiri, "Informative discriminator for domain adaptation," *Image Vis. Comput.*, vol. 111, Jul. 2021, Art. no. 104180.

[25] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brighton, U.K., 2019, pp. 6675–6679.

[26] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[28] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2488–2498.

[29] G. Chen, P. Chen, Y. Shi, C.-Y. Hsieh, B. Liao, and S. Zhang, "Rethinking the usage of batch normalization and dropout in the training of deep neural networks," 2019, *arXiv:1905.05928*.

[30] P. Chong, L. Ruff, M. Kloft, and A. Binder, "Simple and effective prevention of mode collapse in deep one-class classification," in *Proc. Int. Joint Conf. Neural Netw.*, Glasgow, U.K., 2020, pp. 1–9.

[31] S. Cheng et al., "Handheld multispectral fluorescence lifetime imaging system for in vivo applications," *Biomed. Opt. Exp.*, vol. 5, no. 3, pp. 921–931, Mar. 2014.

[32] I. Pavlova, M. Williams, A. El-Naggar, R. Richards-Kortum, and A. Gillenwater, "Understanding the biological basis of autofluorescence imaging for oral cancer detection: High-resolution fluorescence microscopy in viable tissue," *Clin. Cancer Res.*, vol. 14, no. 8, pp. 2396–2404, 2008.

[33] M. G. Müller et al., "Spectroscopic detection and evaluation of morphologic and biochemical changes in early human oral carcinoma," *Cancer: Interdiscipl. Int. J. Amer. Cancer Soc.*, vol. 97, no. 7, pp. 1681–1692, 2003.

[34] A. T. Shah, M. D. Beckler, A. J. Walsh, W. P. Jones, P. R. Pohlmann, and M. C. Skala, "Optical metabolic imaging of treatment response in human head and neck squamous cell carcinoma," *PLoS One*, vol. 9, no. 3, 2014, Art. no. e90746.

[35] J. A. Jo et al., "In Vivo simultaneous morphological and biochemical optical imaging of oral epithelial cancer," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 10, pp. 2596–2599, Oct. 2010.

[36] C. Gulledge and M. Dewhirst, "Tumor oxygenation: A matter of supply and demand," *Anticancer Res.*, vol. 16, no. 2, pp. 741–749, 1996.

[37] R. Drezek et al., "Autofluorescence microscopy of fresh cervical-tissue sections reveals alterations in tissue biochemistry with dysplasia," *Photochemistry Photobiol.*, vol. 73, no. 6, pp. 636–641, 2001.

[38] N. Ramanujam, R. Richards-Kortum, S. Thomsen, A. Mahadevan-Jansen, M. Follen, and B. Chance, "Low temperature fluorescence imaging of freeze-trapped human cervical tissues," *Opt. Exp.*, vol. 8, no. 6, pp. 335–343, 2001.

[39] B. Chance, B. Schoener, R. Oshino, F. Itshak, and Y. Nakase, "Oxidation-reduction ratio studies of mitochondria in freeze-trapped samples. NADH and flavoprotein fluorescence signals," *J. Biol. Chem.*, vol. 254, no. 11, pp. 4764–4771, 1979.

[40] Z. Zhang et al., "Redox ratio of mitochondria as an indicator for the response of photodynamic therapy," *J. Biomed. Opt.*, vol. 9, no. 4, pp. 772–778, 2004.

[41] A. T. Shah, T. M. Heaster, and M. C. Skala, "Metabolic imaging of head and neck cancer organoids," *PLoS One*, vol. 12, no. 1, 2017, Art. no. e0170415.

[42] M. C. Skala et al., "In Vivo multiphoton microscopy of NADH and FAD redox states, fluorescence lifetimes, and cellular morphology in precancerous epithelia," *Proc. Nat. Acad. Sci.*, vol. 104, no. 49, pp. 19494–19499, Dec. 2007.

[43] L. America, "Safe use of lasers, ansi z136. 1–2007," *Amer. Nat. Standards Inst.*, 2014.

[44] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[45] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, pp. 1–13.

[46] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. Symp. Operating Syst. Des. Implementation*, Savannah, GA, USA, 2016, pp. 265–283.

[47] T. Kim and C. Kim, "Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 591–607.

[48] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[49] D. Kobak and P. Berens, "The art of using t-SNE for single-cell transcriptomics," *Nature Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[50] A. L. Goldberger et al., "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[51] G. Wang, M. Chen, Z. Ding, J. Li, H. Yang, and P. Zhang, "Inter-patient ECG arrhythmia heartbeat classification based on unsupervised domain adaptation," *Neurocomputing*, vol. 454, pp. 339–349, 2021.

[52] J. Li, G. Wang, M. Chen, Z. Ding, and H. Yang, "Mixup asymmetric tri-training for heartbeat classification under domain shift," *IEEE Signal Process. Lett.*, vol. 28, pp. 718–722, 2021.

[53] P. Vasanthakumari et al., "Discrimination of cancerous from benign pigmented skin lesions based on multispectral autofluorescence lifetime imaging dermoscopy and machine learning," *J. Biomed. Opt.*, vol. 27, no. 6, 2022, Art. no. 066002.

[54] J. Gideon, M. G. McInnis, and E. Mower Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1055–1068, Oct.–Dec. 2021.