UC San Diego
UC San Diego Previously Published Works

Title

One-Vote Veto: Semi-Supervised Learning for Low-Shot Glaucoma Diagnosis

Permalink

https://escholarship.org/uc/item/10x056pz

Journal

IEEE Transactions on Medical Imaging, 42(12)

ISSN

0278-0062

Authors

Fan, Rui
Bowd, Christopher
Brye, Nicole
et al.

Publication Date

2023-12-01

DOI

10.1109/tmi.2023.3307689

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

IEEE Trans Med Imaging. Author manuscript; available in PMC 2024 December 01.

Published in final edited form as:

IEEE Trans Med Imaging. 2023 December ; 42(12): 3764–3778. doi:10.1109/TMI.2023.3307689.

One-Vote Veto: Semi-Supervised Learning for Low-Shot Glaucoma Diagnosis

Rui Fan [Member, IEEE],

College of Electronics & Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China, and was with Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology, Shiley Eye Institute, and the Department of Computer Science & Engineering, the University of California San Diego, La Jolla, CA 92093, USA

Christopher Bowd,

Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, the University of California San Diego, La Jolla, CA 92093, USA

Nicole Brye,

Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, the University of California San Diego, La Jolla, CA 92093, USA

Mark Christopher,

Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, the University of California San Diego, La Jolla, CA 92093, USA

Robert N. Weinreb,

Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, the University of California San Diego, La Jolla, CA 92093, USA

David J. Kriegman [Fellow, IEEE],

Department of Computer Science &Engineering, the University of California San Diego, La Jolla, CA 92093, USA

Linda M. Zangwill

Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, the University of California San Diego, La Jolla, CA 92093, USA

Abstract

Convolutional neural networks (CNNs) are a promising technique for automated glaucoma diagnosis from images of the fundus, and these images are routinely acquired as part of an ophthalmic exam. Nevertheless, CNNs typically require a large amount of well-labeled data for training, which may not be available in many biomedical image classification applications, especially when diseases are rare and where labeling by experts is costly. This article makes two contributions to address this issue: (1) It extends the conventional Siamese network and

Corresponding author: Linda M. Zangwill lzangwill@health.ucsd.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

introduces a training method for low-shot learning when labeled data are limited and imbalanced, and (2) it introduces a novel semi-supervised learning strategy that uses additional unlabeled training data to achieve greater accuracy. Our proposed multi-task Siamese network (MTSN) can employ any back-bone CNN, and we demonstrate with four backbone CNNs that its accuracy with limited training data approaches the accuracy of backbone CNNs trained with a dataset that is 50 times larger. We also introduce One-Vote Veto (OVV) self-training, a semi-supervised learning strategy that is designed specifically for MTSNs. By taking both self-predictions and contrastive predictions of the unlabeled training data into account, OVV self-training provides additional pseudo labels for fine-tuning a pre-trained MTSN. Using a large (imbalanced) dataset with 66,715 fundus photographs acquired over 15 years, extensive experimental results demonstrate the effectiveness of low-shot learning with MTSN and semi-supervised learning with OVV self-training. Three additional, smaller clinical datasets of fundus images acquired under different conditions (cameras, instruments, locations, populations) are used to demonstrate the generalizability of the proposed methods.

Keywords

Convolutional neural networks; glaucoma diagnosis; low-shot learning; semi-supervised learning

I. Introduction

GLaucoma is a prevalent and debilitating disease that can lead to progressive and irreversible vision loss through optic nerve damage [1]. The global incidence of glaucoma was estimated at 64.3 million in 2013, and due to aging populations, this number is expected to rise to 111.8 million by 2040 [2]. Improvement in the management of glaucoma would have a major human and socio-economic impact [3]. Early identification and intervention would significantly reduce the economic burden of late-stage disease [4]. In addition, visual impairment in glaucoma patients has been associated with decreased physical activity and mental health [5], [6] and increased risk of motor vehicle accidents [7], [8].

With the recent advances in machine learning, convolutional neural networks (CNNs), trained via supervised learning, have shown promise in diagnosing glaucoma from fundus images (photographs of the back of eyes) [9]. However, this requires large amounts of empirical data for supervised training [10]. In this study, we use 66,715 fundus photographs from the **O**cular **H**ypertension **T**reatment Study (**OHTS**) [11]–[13], which is a 22-site multi-center, longitudinal (phase 1 and 2, 1994–2008) randomized clinical trial of 1,636 subjects (3,272 eyes). The primary goal of the OHTS was to determine if topical ocular hypotensive medications could delay or prevent the onset of glaucoma in eyes with high intraocular pressure [11]. Conversion to glaucoma was decided by a masked endpoint committee of three glaucoma specialists using fundus photographs and visual fields. Owing to its well-characterized ground-truth labels, the OHTS dataset provides us a basis to explore an effective way of training CNNs to diagnose glaucoma with low-shot learning when only a small quantity of labeled data is available, and/or semi-supervised learning when raw data is abundant, but labeling resources are scarce, costly, require strong expertise, or are just unavailable. However, as shown in Fig. 1, conventional semi-supervised learning approaches

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

typically require a reliable pre-trained CNN (using a small sample) as prior knowledge, which is often challenging due to the over-fitting problem. Moreover, there is also a strong motivation to design a feasible semi-supervised learning strategy capable of determining confident predictions and generating pseudo labels for unlabeled data. We focus specifically on fundus images and glaucoma diagnosis in this article because we have sufficient data to accurately characterize the effectiveness of our methods. The same techniques could also be applied to tasks where there is limited data, such as rare diseases or where limited labels are available (*e.g.*, asthma and diabetes prediction from fundus images). Therefore, this article aims to answer the following questions:

1. Can a CNN be developed to accurately diagnose glaucoma, compared to the expert graders of the OHTS? Will the model be generalizable to other datasets?
2. Is it necessary to train CNNs with thousands of labeled fundus images to diagnose glaucoma, or can diagnosis be achieved using only one image per patient (approximately 1.1K fundus images in the OHTS training set)?
3. Can the performance of a CNN trained using a small sample be improved further by fine-tuning it with additional unlabeled training data?

To answer these questions, we first evaluate the performance of state-of-the-art (SoTA) glaucoma diagnosis algorithms, including six supervised learning algorithms [3], [14]–[18], one low-shot learning algorithm [19], and two semi-supervised learning algorithms [20], [21], on the OHTS dataset. Their generalizabilities are further validated on three additional clinical datasets of fundus images: (a) ACRIMA (Spain) [22], (b) Large-Scale Attention-Based Glaucoma (LAG, China) [9], and (c) the UCSD-based Diagnostic Innovations in Glaucoma Study and African Descent and Glaucoma Evaluation Study (DIGS/ADAGES, US) [23].

Furthermore, we propose a novel extension to the conventional Siamese network, referred to as the **M**ulti-**T**ask **S**iamese Network (**MTSN**), as depicted in Fig. 2. By minimizing a novel **C**ombined **W**eighted **C**ross-**E**ntropy (**CWCE**) **Loss**, the MTSN can simultaneously perform two tasks: measuring the similarity of a given pair of images (primary task) and classifying them as healthy or glaucoma (secondary task). With a small training set of approximately 1.1K fundus images, we explore the feasibility of training an MTSN for glaucoma diagnosis. Although the MTSN may not provide complementary information, it effectively performs a type of "data augmentation" by generating $C(N, 2)$ pairs of fundus images for training instead of using N independent fundus images. The visual features learned from these two tasks prove to be more informative for glaucoma diagnosis when the training set is small. Our experimental results demonstrate that the MTSN greatly reduces over-fitting and achieves an accuracy on a small training set comparable to a large training set, which contains approximately 53K fundus images.

Moreover, we propose a novel semi-supervised learning strategy, referred to as **O**ne-**V**ote **V**eto (**OVV**) **Self-Training**, which generates reliable pseudo labels for the unlabeled training data and incorporates them into the labeled training data to fine-tune the MTSN for improved performance and generalizability. Our extensive experiments show that the MTSN fine-tuned with OVV self-training achieves similar performance to the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

corresponding backbone CNN trained via supervised learning on the OHTS dataset, and achieves higher area under the receiver operating characteristic curve (AUROC) scores on the additional fundus image datasets. The fine-tuned MTSN also outperforms SoTA semi-supervised glaucoma diagnosis approaches [20], [21], and in some cases, even outperforms SoTA supervised approaches. Additionally, we compare our proposed OVV self-training approach with four SoTA general-purpose semi-supervised learning methods, including FreeMatch [24], SoftMatch [25], FixMatch [26], and FlexMatch [27], all of which utilize vision Transformer [28] as their backbone network. The results demonstrate that our proposed OVV self-training approach outperforms these methods on the OHTS dataset and demonstrates better generalizability on three additional fundus image test sets.

We also conduct two additional few-shot biomedical image classification experiments (chest X-ray image classification [29], [30] and lung histopathological image classification [31]) to further validate the effectiveness of the MTSN on other types of image data. The promising results indicate that our proposed algorithms have the potential to solve a variety of biomedical image classification problems.

II. Related Works

Most SoTA glaucoma diagnosis algorithms are developed based on supervised fundus image classification. For example, Judy *et al.* [16] trained an AlexNet [32] to diagnose glaucoma. As VGG architectures [33] can learn more complicated image features than AlexNet, Gómez-Valverde *et al.* [15] employed a VGG-19 [33] model to diagnose glaucoma. Nevertheless, VGG architectures [33] consist of hundreds of millions of parameters, making them very memory-consuming. In contrast, GoogLeNet [34] and Inception-v3 [35] have lower computational complexities. Hence, Ahn *et al.* [36] and Li *et al.* [14] utilized transfer learning to re-train an Inception-v3 [35] model (pre-trained on the ImageNet [37] database) for glaucoma diagnosis, while Serener and Serte [17] re-trained a pre-trained GoogLeNet [34] model to diagnose glaucoma. However, with the increase of network depth, accuracy gets saturated and then degrades rapidly due to vanishing gradients [38]. To tackle this problem, the residual neural network (ResNet) [38] was developed. Due to its robustness, ResNet-50 [38] has been extensively used for biomedical image analysis, and it is a popular choice [3], [39]–[43] for fundus image classification. Additionally, developing low-cost and real-time embedded glaucoma diagnosis systems [18], [44], [45], *e.g.*, based on MobileNet-v2 [46], for mobile devices is also an emerging area.

Machine/deep learning has achieved compelling performance in data-intensive applications, but it is often challenging for these algorithms to yield comparable performance when only a limited amount of labeled training data is available [47]. Low-shot and semi-supervised learning can address these issues. Unfortunately, they are rarely discussed in the field of glaucoma diagnosis. To the best of our knowledge, [19] is the only published low/few-shot glaucoma diagnosis algorithm. This algorithm employs a conventional Siamese network to compare two groups of (negative and positive) fundus images. The Siamese network utilizes two identical CNNs to learn visual embeddings. A bi-directional long short-term memory [48] component is then trained over the CNN outputs for glaucoma diagnosis. However, the training process is complicated since different types of losses are minimized,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and the achieved glaucoma diagnosis results are unsatisfactory since each sub-network is only fed with one type of fundus images (either negative or positive). The lack of same-class comparisons leads to a performance bottleneck when compared to the MTSN proposed in this article.

A thorough search of the relevant literature yielded only two published studies on semi-supervised learning specifically for glaucoma diagnosis [20], [21]. Diaz-Pinto *et al.* [21] utilized a deep convolutional generative adversarial network (DCGAN) [49] for semi-supervised learning of glaucoma diagnosis, where the discriminator is trained to classify healthy and glaucomatous optic neuropathy (GON) fundus images, while also distinguishing between real and fake fundus images. The classifier for the former task is then employed for glaucoma diagnosis. On the other hand, Al Ghamdi *et al.* [20] developed a glaucoma diagnosis approach based on self-training [50], which is a typical semi-supervised learning approach that uses a pre-trained model (typically yielded via supervised learning) to produce pseudo labels of the unlabeled data. However, producing reliable pseudo labels is a significant challenge in self-training, and the pseudo labels generated by a single pre-trained CNN are usually not trustworthy enough for CNN fine-tuning [51]. Additionally, training a reliable pre-trained classifier with only a small amount of labeled data is notably demanding. In this article, we combine semi-supervised learning with low-shot learning to address these issues using glaucoma diagnosis as an example case. Specifically, our proposed OVV self-training strategy, as discussed in Sect. III-B, is inspired by the mechanism of *learning with external memory* (LwEM), used in low-shot learning [52], where the labels of unlabeled training data are predicted by a classifier trained via low-shot learning on a small collection of fundus images with ground-truth labels.

III. Methodology

A. Multi-Task Siamese Network

As illustrated in Fig. 1, conventional semi-supervised learning methods initialize a network by pre-training it with a small number of fundus images for glaucoma diagnosis. However, we observed that such approaches are highly sensitive to noise. As a result, we design a novel MTSN specifically for our semi-supervised learning approach, which requires not only predicting the label of a given fundus image but also determining the similarity between a pair of given fundus images to generate pseudo labels through a voting process.

Conventional Siamese networks have become a common choice for metric learning and few/low-shot image recognition tasks [53]. These networks comprise two identical sub-networks, as depicted in Fig. 2. Each pair of fundus images $\mathbf{x}_i$ and $\mathbf{x}_j$ are separately fed into these sub-networks, producing two 1D embeddings (features) $\mathbf{h}_i$ and $\mathbf{h}_j$, respectively. Another 1D embedding $\mathbf{h}_{i,j}$ is generated by $\Phi(\cdot)$. $\mathbf{h}_{i,j}$ is then passed through a fully connected (FC) layer to produce a scalar $q(\mathbf{x}_i, \mathbf{x}_j) \in [0,1]$ indicating the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are dissimilar, $q(\mathbf{x}_i, \mathbf{x}_j)$ approaches 1, and vice versa. The ground-truth labels of $\mathbf{x}_i$ and $\mathbf{x}_j$ are represented by $y_i \in \{0,1\}$ and $y_j \in \{0,1\}$, respectively, where 0 denotes a healthy image, and 1 denotes a GON image.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

However, a conventional Siamese network can only determine whether $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same category, rather than predicting their independent categories. A straightforward solution is to connect $\mathbf{h}_i$ and $\mathbf{h}_j$ to separate FC layers, producing two scalars $p(\mathbf{x}_i)$ and $\mathrm{p}(\mathbf{x}_j)$ indicating the probabilities that $\mathbf{x}_i$ and $\mathbf{x}_j$ are GON images, respectively. Refer to Fig. 2 and note that the two FC layers connected to $\mathbf{h}_i$ and $\mathbf{h}_j$ use the same weights. In this article, we refer to the network architecture in Fig. 2 as an MTSN, which can simultaneously measure the similarity of a given pair of fundus images and classify them as either healthy or GON. These two tasks are dependent yet not directly deducible from one another. A well-trained glaucoma diagnosis network can be employed to compare differences between given pairs of fundus images, but a well-trained fundus image similarity measurement network cannot directly output the category of a given fundus image.

In addition, the visual features learned from the primary and secondary tasks are distinct from one another. For the primary task, the network learns the visual features to classify same-class and different-class fundus image pairs. On the other hand, for the secondary task, the network learns the visual features to classify GON and healthy fundus images. Although this network architecture may not provide complementary information, it effectively performs a type of "data augmentation" by producing $C(N, 2)$ pairs of fundus images for training, rather than using N independent fundus images. The visual features learned from these two tasks prove to be more informative for glaucoma diagnosis when the training set is small. Furthermore, multi-task learning is effective because requiring an algorithm to perform well on a related task induces regularization, which can be superior to uniform complexity penalization for preventing over-fitting. This idea has been explored in many Siamese neural network works, such as [54]–[56].

In this article, we use n_0 and n_1 to respectively denote the numbers of healthy and GON fundus images used to train the MTSN, with $n = n_0 + n_1$. n_0 is usually much greater than n_1, because there are fewer patients with glaucomatous disease than healthy patients, resulting in a severely imbalanced dataset. Therefore, the MTSN is trained by minimizing a CWCE loss as follows:

$$\mathscr{L} = \mathscr{L}_{\mathrm{sim}} + \lambda \mathscr{L}_{\mathrm{cla}}, \tag{1}$$

where

$$\begin{aligned}\mathscr{L}_{\mathrm{sim}} = & -\frac{n_0(n_0-1)+n_1(n_1-1)}{n(n-1)}|y_i - y_j|\log(q(\mathbf{x}_i, \mathbf{x}_j)) \\ & -\frac{2n_0 n_1}{n(n-1)}(1-|y_i - y_j|)\log(1-q(\mathbf{x}_i, \mathbf{x}_j)),\end{aligned} \tag{2}$$

$$\begin{aligned}\mathscr{L}_{\mathrm{cla}} = & -\frac{1}{n}(n_0(y_i \ \log(p(\mathbf{x}_i)) + y_j \ \log(p(\mathbf{x}_j))) \\ & +n_1((1-y_i) \ \log(1-p(\mathbf{x}_i)) + (1-y_j) \ \log(1-p(\mathbf{x}_j)))),\end{aligned}$$

(3)

The hyper-parameter λ balances the primary task loss $\mathscr{L}_{sim}$ and the secondary task loss $\mathscr{L}_{cla}$. The choice of λ and $\Phi(\cdot)$ is discussed in Sect. IV-B. The motivations for using such a CWCE loss function instead of the commonly used triplet loss [57] or contrastive loss [58] to train the MTSN are:

1. Most datasets for rare disease diagnosis are imbalanced. As detailed in Sect. IV-A, the OHTS training set is severely imbalanced, with 50,208 healthy images and only 2,416 GON images for supervised learning, and 995 healthy images and 152 GON images for low-shot learning. Learning from such an imbalanced dataset without weights on different classes can result in many incorrect predictions, with most GON images likely to be predicted as healthy images. To address this issue, a higher weight should be used for the minority class to prevent the CNN from predicting all fundus images as the majority class.
2. In multi-task learning, weighing different types of losses, such as regression and classification, is typically challenging [59]. Assigning an incorrect weight may cause one task to perform poorly, even when other tasks converge to satisfactory results. Therefore, formulating $\mathscr{L}_{sim}$ as a weighted cross-entropy loss function is a simple but effective solution. However, due to the dataset imbalance problem, the cross-entropy losses have to be weighted.
3. As shown in Fig. 3, OVV self-training requires both labels and probabilities (of being GON images), predicted by a pre-trained model, to produce pseudo labels for unlabeled data. Such network architecture and training loss can efficiently and effectively provide both “self-predicted” and “contrastively-predicted” labels and probabilities, as described in Sect. III-B.

It should be noted here that using the fundus images from the same patient as an image pair for MTSN training is not necessary.

B. One-Vote Veto Self-Training

As discussed in Sect. I, self-training aims to improve the performance of a pre-trained model by incorporating reliable predictions of the unlabeled data to obtain useful additional information that can be used for model fine-tuning. A feasible strategy to determine such reliable predictions is, therefore, key to the success of self-training [60].

In conventional semi-supervised learning algorithms, a pre-trained image classification model (obtained through supervised learning) can be fine-tuned by assessing the reliability of unlabeled images. To determine the reliability of an unlabeled image, its probability distribution for the most likely class is compared to a pre-determined threshold. If the probability surpasses this threshold, the prediction is considered a pseudo label. Subsequently, the image and its corresponding pseudo label are utilized to fine-tune the pre-trained model.

However, relying solely on probability distributions to generate pseudo labels is often insufficient [50]. Drawing inspiration from LwEM [61], we introduce *One-Vote Veto*

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

self-training in this paper, as illustrated in Fig. 3[1]. Similar to LwEM [61], we use a collection of m reference (labeled) fundus images $\{\mathbf{x}_1^r, \ldots, \mathbf{x}_m^r\} \in \mathcal{X}^r$ to provide "*contrastive predictions*" to the target (unlabeled) fundus images $\{\mathbf{x}_1^t, \ldots, \mathbf{x}_m^t\} \in \mathcal{X}^t$. The contrastive predictions subsequently vote to veto the unreliable "*self-predictions*" $\{\tilde{y}_1^t, \ldots, \tilde{y}_m^t\}$ produced by the MTSN. Our OVV self-training is detailed in Algorithm 1, where the target model updates its parameters during self-training but the reference model does not.

When fine-tuning an MTSN pre-trained through low-shot learning, each mini-batch contains a discrete set of m reference

[1]The superscripts r and t denote "reference" and "target", respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Algorithm 1:

One-Vote Veto Self-Training Strategy.

Data: Reference fundus images $\mathscr{X}^r$ and their labels $\mathscr{Y}^r$, and target fundus images $\mathscr{X}^t$

while Training **do**

  Given a mini-batch consisting of $\{\mathbf{x}_1^r, \ldots, \mathbf{x}_m^r\} \in \mathscr{X}^r$, $\{y_1^r, \ldots, y_m^r\} \in \mathscr{Y}^r$ and $\{\mathbf{x}_1^t, \ldots, \mathbf{x}_m^t\} \in \mathscr{X}^t$;

  Initialize an empty set $\mathscr{P}$ to store reliable target fundus images and their pseudo labels ;

  for Each target fundus image $\mathbf{x}_k^t$ **do**

    if The self-prediction of $\mathbf{x}_k^t$ is reliable **then**

      for Any qualified reference image $\mathbf{x}_l^r$ **do**

        Compute the self-prediction and contrastive prediction;

      end

      if The criteria to generate pseudo labels are satisf **then**

        Update the set $\mathscr{P}$;

      end

    end

  end

  Fine-tune the target model using unique($\mathscr{P}$);

end

if The target model outperforms the reference model **then**

  Update the reference model parameters;

end

fundus images $\{\mathbf{x}_1^r, \ldots, \mathbf{x}_m^r\} \in \mathscr{X}^r$, their ground-truth labels $\{y_1^r, \ldots, y_m^r\} \in \mathscr{Y}^r$, and an equal number of m target fundus images $\{\mathbf{x}_1^t, \ldots, \mathbf{x}_m^t\} \in \mathscr{X}^t$ without labels. $\mathbf{h}_k^r$ and $\mathbf{h}_k^t$ represent the 1D

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

embeddings learned from $\mathbf{x}_k^r$ and $\mathbf{x}_k^t (k \in [1, m] \cap \mathbb{Z})$, respectively. Given a pair of reference and target fundus images, $\mathbf{x}_l^r$ and $\mathbf{x}_k^t$, the pre-trained MTSN can "self-predict":

- the scalars $p(\mathbf{x}_l^r)$ and $p(\mathbf{x}_k^t)$ which indicate the probabilities that $\mathbf{x}_l^r$ and $\mathbf{x}_k^t$ are GON images, respectively;
- their labels $\tilde{y}_l^r = \delta(p(\mathbf{x}_l^r))$ and $\tilde{y}_k^t = \delta(p(\mathbf{x}_k^t))$ using its fundus image classification functionality ($\delta(p) = 1$ when $p > 0.5$, and $\delta(p) = 0$ otherwise).

$p(\mathbf{x}_l^r)$ is then used to determine whether the reference fundus image $\mathbf{x}_l^r$ is qualified to veto unreliable predictions. If $|p(\mathbf{x}_l^r) - y_l^r| > \kappa_2$, its vote will be omitted, where κ_2 is a threshold used to select qualified reference fundus images (step 6 in Algorithm 1). In the meantime, the pre-trained MTSN can also "contrastively-predict" the scalar

$$p_{l,k}^{r \to t}(\mathbf{x}_l^r, \mathbf{x}_k^t) = \left|p(\mathbf{x}_l^r) - q(\mathbf{x}_l^r, \mathbf{x}_k^t)\right| \tag{4}$$

indicating the GON probability as well as the label

$$\tilde{y}_{l,k}^{r \to t}(\mathbf{x}_l^r, \mathbf{x}_k^t) = \left|\delta(p(\mathbf{x}_l^r)) - \delta(q(\mathbf{x}_l^r, \mathbf{x}_k^t))\right| \tag{5}$$

of $\mathbf{x}_k^t$ from $\mathbf{x}_l^r$ using its input similarity measurement functionality[2]. To determine the reliability of $\tilde{y}_k^t$ and whether it can be used as the pseudo label of $\mathbf{x}_k^t$, all the reference fundus images $\{\mathbf{x}_1^r, \ldots, \mathbf{x}_m^r\} \in \mathcal{X}^r$ in the mini-batch are used to provide additional judgements. Each pair of contrastively-predicted scalar (indicating GON probability) and label form a vote $(p_{l,k}^{r \to t}(\mathbf{x}_l^r, \mathbf{x}_k^t), \tilde{y}_{l,k}^{r \to t}(\mathbf{x}_l^r, \mathbf{x}_k^t))$. With all votes collected from the qualified reference fundus images, the OVV self-training algorithm determines whether $\tilde{y}_k^t$ should be used as the pseudo label for $\mathbf{x}_k^t$ based on the following criteria (step 9 in Algorithm 1):

- Identical to the process of determining qualified reference fundus images, if any $p_{l,k}^{r \to t}(\mathbf{x}_l^r, \mathbf{x}_k^t)(l \in [1, m] \cap \mathbb{Z}))$ or $p(\mathbf{x}_k^t)$ is not close to either 0 (healthy) or 1 (GON), as evaluated by the threshold κ_2, $\tilde{y}_k^t$ will not be assigned to $\mathbf{x}_k^t$.
- If a minority of more than κ_1 qualified reference fundus images disagree with the majority of the qualified reference fundus images, $\tilde{y}_k^t$ will not be assigned to $\mathbf{x}_k^t$.

As discussed in Sect. IV, $\kappa_1 = 0$ (all qualified reference images vote for the same category) achieves the best overall performance. Therefore, the aforementioned strategy is named "*One-Vote Veto*" in this paper. Since each target fundus image is required to be compared with all the reference fundus images in the same mini-batch, the proposed self-training strategy has a computational complexity of $\mathcal{O}(n^2)$, which is relatively memory-consuming. The reliable target fundus images and their pseudo labels are then included into the low-shot training data to fine-tune the pre-trained MTSN with supervised learning by minimizing a

[2]In rare cases, $\tilde{y}_{l,k}^{r \to t}(\mathbf{x}_l^r, \mathbf{x}_k^t)$ might not be equivalent to $\delta(p_{l,k}^{r \to t}(\mathbf{x}_l^r, \mathbf{x}_k^t))$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CWCE loss. The OVV self-training performance with respect to different κ_1, κ_2, and m values is discussed in Sect. IV.

IV. Experiments

A. Datasets and Experimental Setups

The datasets utilized in our experiments were collected by various clinicians in different countries using distinct fundus cameras. The ACRIMA [22] and LAG [9] datasets are publicly available, while the OHTS [11], [12] and DIGS/ADAGES [23] datasets are available upon request after appropriate data use agreements are initiated. Their details are as follows:

- The **OHTS** [11], [12] is the only multi-center longitudinal study that has precise information on the dates/timing of the development of glaucoma (the enrolled subjects did not have glaucoma at study entry) using standardized assessment criteria by an independent Optic Disc Reading Center and confirmed by three glaucoma specialist endpoint committee members. In our experiments, a square region centered on the optic nerve head was first extracted from each raw fundus image using a well-trained DeepLabv3+ [63] model. A small part of the raw data are stereoscopic fundus images, each of which was split to produce two individual fundus images. Through this image pre-processing approach, a total of 74,678 fundus images were obtained. Moreover, ENPOAGDISC (endpoint committee attributable to primary open angle glaucoma based on optic disc changes from photographs) [3] labels are used as the classification ground truth. The fundus images are divided into a training set (50,208 healthy images and 2,416 GON images), a validation set (7,188 healthy images and 426 GON images), and a test set (13,780 healthy images and 660 GON images) by participant. Splitting by participant (instead of by image) ensured that the validation and test sets did not contain images from any eyes or individuals used to train the model. More details on dataset preparation and baseline supervised learning experiments are provided in our recent publications [3], [10]. Additionally, we select one image (from only one eye) from each patient in the training set to create the low-shot training set (995 healthy images and 152 GON images). If both eyes of a patient do not convert to glaucoma in the study, the first captured fundus photograph is selected. If either eye of a patient converts to glaucoma in the study, the first glaucoma fundus photograph is selected.
- The **ACRIMA** [22] dataset consists of 309 healthy images and 396 GON images. It was collected as part of an initiative by the government of Spain. Classification was based on the review by a single experienced glaucoma expert. Images were excluded if they did not provide a clear view of the optic nerve head region [43].
- The **LAG** [9] dataset contains 3,143 healthy images and 1,711 GON images[3], collected by Beijing Tongren Hospital. Similar to the OHTS dataset, we also use

[3]The number of fundus images being published is fewer than what was reported in publication [9].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the well-trained DeepLabv3+ [63] model to extract a square region centered on the optic nerve head from each fundus image.

- The **DIGS** and **ADAGES** [23] are longitudinal studies designed to detect and monitor glaucoma based on optical imaging and visual function testing that, when combined, have generated tens of thousands of test results from over 4,000 healthy, glaucoma suspect, or glaucoma eyes. In our experiments, we utilize the DIGS/ADAGES test set (5,184 healthy images and 4,289 GON images) to evaluate the generalizability of our proposed methods.

Visualizations of the four test sets using t-SNE [62] are provided in Fig. 4. Since healthy and GON images are distributed similarly between the OHTS and LAG datasets, we expect models to perform similarly on these datasets. Dissimilar distributions in the ACRIMA and DIGS/ADAGES datasets led us to believe the performance of models on these datasets would be somewhat worse. Using these four datasets, we conduct three experiments:

1. **Supervised learning experiment:** We employ transfer learning [64] to train ResNet-50 [38], MobileNet-v2 [46], DenseNet [65], and EfficientNet [66] (pre-trained on the ImageNet database [37]), on the entire OHTS training set (including ~53K fundus images). The best-performing models are selected using the OHTS validation set. Their performance is subsequently evaluated on the OHTS test set, the ACRIMA dataset, the LAG dataset, and the DIGS/ADAGES test set.
2. **Low-shot learning experiment:** The four pre-trained models mentioned above are used as the MTSN backbones and trained on the OHTS low-shot training set (containing 1,147 images) to validate the effectiveness of our proposed low-shot glaucoma diagnosis algorithm. The validation and testing procedures are identical to those in the supervised learning experiment.
3. **Semi-supervised learning experiment:** The MTSNs trained on the low-shot training set are fine-tuned on the entire OHTS training set without using additional ground-truth labels. The fine-tuned MTSNs are referred to as MTSN+OVV. The validation and testing procedures are identical to those in the supervised learning experiment.

The fundus images are resized to 224 × 224 pixels. The initial learning rate is set to 0.001, which decays gradually after the 100th epoch. Due to the dataset imbalance problem, F1-score is utilized to select the best-performing models during the validation stage. Moreover, we adopt an early stopping mechanism during the validation stage to reduce over-fitting (the training will be terminated if the achieved F1-score has not increased for 10 epochs). We use three metrics: (1) accuracy, (2) F1-score, and (3) AUROC to quantify the performances of the trained models. While accuracy is generally reported in image classification papers, F1-score and AUROC are more comprehensive and reasonable evaluation metrics when the dataset is severely imbalanced.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

B. Ablation study and Threshold Selection

We set λ in (1) to 0.1, 0.2, 0.3, 0.4, and 0.5, respectively, and compare the MTSN performance when Φ computes the element-wise absolute difference (EWAD) and element-wise squared difference (EWSD), respectively. The comparisons in terms of F1-score and AUROC on the OHTS test set are provided in Fig. 5. It can be seen that the MTSN achieves the best overall performance when $\lambda = 0.3$. This is reasonable, as a higher λ weighs more on the image classification task, easily resulting in over-fitting. Additionally, the MTSN in which $\Phi(\cdot)$ computes the EWSD between $\mathbf{h}_i$ and $\mathbf{h}_j$ performs better when using ResNet-50 as the backbone CNN but slightly worse when using MobileNet-v2 as the backbone CNN. Therefore, we further evaluate their generalizability on three additional test sets, as shown in Table I. When $\Phi(\cdot)$ computes the EWAD, the MTSN generally performs better or very similarly on the additional test sets, especially when testing the MTSN assembled with ResNet-50 on the ACRIMA dataset. EWAD is, therefore, used in the following experiments. Furthermore, Table I provides the results of a baseline supervised learning experiment conducted on the low-shot training set. The results suggest that low-shot learning performs much better than supervised learning when the training size is small.

Furthermore, we discuss the selection of the thresholds κ_1 and κ_2 (used to select reliable "self-predictions" and "contrastive predictions" in our OVV self-training) as well as the impact of different mini-batch sizes $2m$ on OVV self-training (each mini-batch contains m pairs of reference and target fundus images). Table II shows the MTSN performances with respect to different κ_1, κ_2, and m. When evaluated on the OHTS test set, it can be seen that accuracy and F1-score increase slightly, but AUROC almost remains the same, with the decrease of m. Moreover, with the increase of κ_1 and κ_2, the standard to determine reliable predictions becomes lower, making the semi-supervised learning performance degrade. Based on this experiment, we believe OVV self-training benefits from smaller κ_1 and κ_2.

Additionally, MTSNs, trained under different m, are evaluated on the three additional test sets, as shown in Table II. It can be seen that the network trained with a larger typically shows better results. When m decreases, the generalizability of MTSN degrades dramatically, especially for F1-score (decreases by around 9–19%). Therefore, increasing the mini-batch size can improve the MTSN generalizability, as more reference fundus images are used to provide contrastive predictions for the target fundus images, which can veto more unreliable predictions on the unlabeled data. Hence, we to 30 to further improve OVV self-training when comparing it with other published SoTA algorithms, as shown in Sect. IV-D. Since our threshold selection experiments cover a very limited number of discrete sets of κ_1, κ_2, and m, we believe better performance can be achieved when more values are tested.

C. Comparison of Supervised, Low-Shot, and Semi-Supervised Glaucoma Diagnosis

Comparisons of supervised learning, low-shot learning, and semi-supervised learning (w.r.t. four backbone CNNs: ResNet-50, MobileNet-v2, DenseNet, and EfficientNet) for glaucoma diagnosis are provided in Table III. First, these results suggest that the MTSNs fine-tuned with OVV self-training that requires a small number of labeled fundus images perform

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

similarly (AUROC 95% CI overlaps considerably) and, in some cases, significantly better (AUROC 95% CI does not overlap) than the backbone CNNs trained with a large number of labeled fundus images (50 times larger) under full supervision.

Specifically, when using ResNet-50, MobileNet-v2, or DenseNet as the backbone CNN, semi-supervised learning performs similarly to supervised learning on the OHTS and DIGS/ADAGES test sets, and in most cases, significantly better than supervised learning on the ACRIMA and LAG datasets. Although EfficientNet trained through supervised learning performs unsatisfactorily on all four test sets, it shows considerable compatibility with MTSN in the low-shot and semi-supervised learning experiments. Second as expected, the AUROC scores achieved by low-shot learning are in most, but not all, cases slightly lower than those achieved by the backbone CNNs, when evaluated on the OHTS test set. However, low-shot learning shows better generalizability than supervised learning on the ACRIMA and LAG datasets. Moreover, since low-shot learning uses only a small amount of training data, training an MTSN is much faster than supervised learning. As MTSNs assembled with ResNet-50 and MobileNet-v2 typically demonstrate better performances than the ones assembled with DenseNet and EfficientNet, we only use the former two CNNs for the following experiments.

We also employ Grad-CAM++ [67] to explain the models' decision-making, as shown in Fig. 6. These results suggest that the optic nerve head areas impact model decisions most. The neuroretinal rim areas are identified as most important, and the periphery contributed comparatively little to model decisions for both healthy and GON eyes [42].

We also carry out a series of experiments with respect to different percentages of training data, as shown in Table IV, to further validate the effectiveness of our proposed low-shot and semi-supervised learning algorithms. The backbone CNNs trained via supervised learning on the small subsets generally perform worse than the MTSNs trained via low-shot and semi-supervised learning. With the labeled training data increase, the models' performance gets saturated. When using less labeled training data (0.5% and 1.0%), the MTSN performance degrades. However, its performance can still be greatly improved with OVV self-training (accuracy, F1-score, and AUROC can be improved by up to 6%, 11%, and 0.08, respectively). In addition, when using over 10% of the entire training data, the MTSN performance saturates, and the OVV self-training can bring very limited improvements on MTSNs.

D. Comparisons with other SoTA glaucoma diagnosis approaches

Table V provides comprehensive comparisons with nine SoTA glaucoma diagnosis algorithms[4]. The results suggest that (a) for low-shot learning, the MTSNs trained by minimizing our proposed CWCE loss perform significantly better than the SoTA low-shot glaucoma diagnosis approach [19] on all four datasets (accuracy, F1-score, and AUROC are up to 5%, 15%, and 0.08 higher, respectively), and (b) for semi-supervised learning, the MTSNs fine-tuned with OVV self-training also achieve the superior performances over another two SoTA semi-supervised glaucoma diagnosis approaches [20], [21] (accuracy,

[4]Our recent work [3] provides the baseline supervised learning results.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

F1-score, and AUROC are up to 6%, 11%, and 0.07 higher, respectively). Compared with the SoTA supervised approaches, the fine-tuned MTSNs demonstrate similar performance on the OHTS test set and better generalizability on three additional test sets. Therefore, we believe that MTSN with our proposed OVV self-training is an effective technique for semi-supervised glaucoma diagnosis.

E. Comparisons with SoTA general-purpose semi-supervised learning approaches

Table VI provides a comprehensive comparison of our proposed OVV self-training approach with four SoTA general-purpose semi-supervised learning methods: FreeMatch [24], SoftMatch [25], FixMatch [26], and FlexMatch [27], which all employ vision Transformer [28] as their backbone network. Our results demonstrate that the proposed OVV self-training approach outperforms these methods in terms of F1-score and AUROC on the OHTS dataset. Specifically, we observe improvements in F1-score ranging from approximately 11% to 13%, and improvements in AUROC ranging from around 9% to 20%. Furthermore, our method demonstrates better generalizability in terms of AUROC across three additional fundus image test sets. Although their results are inferior to ours, particularly in terms of AUROC, it may be unjust to compare them as they were not specifically designed for the diagnosis of glaucoma or other diseases.

F. MTSN and CWCE Loss for Few-Shot Multi-Class Biomedical Image Classification

We conduct two additional few-shot multi-class lung disease diagnosis experiments: (a) chest X-ray image classification for COVID-19 and viral pneumonia detection [29], [30], and (b) lung histopathological image classification for lung cancer diagnosis [31], to validate the effectiveness of our proposed MTSN and CWCE loss. The first experiment has three classes of images: (1) healthy, (2) viral pneumonia, and (3) COVID-19 (an example of each class is shown in Fig. 7), while the second experiment also has three classes of images: (1) benign tissue, (2) adenocarcinoma, and (3) squamous cell carcinoma (an example of each class is shown in Fig. 7). In these two experiments, we only select a few images from each class for MTSN training. The numbers of images used for training, validation, and testing are given in Table VII, where it can be observed that the training set is much smaller than the validation and test sets. Since (1) can only be used for binary image classification problems, we extend it here to tackle multi-class image classification problems.

Let us consider the chest X-ray image classification task as an example. Each image $\mathbf{x}$ is assigned a pair of two labels (r, s) with the following values:

- $r = 1$ and $s = 0$ when $\mathbf{x}$ is a healthy image;
- $r = 2$ and $s = 1$ when $\mathbf{x}$ is a viral pneumonia image;
- $r = 3$ and $s = 3$ when $\mathbf{x}$ is a COVID-19 image.

The numbers of healthy images (class 1), viral pneumonia images (class 2), and COVID-19 images (class 3) are denoted as n_1, n_2, and n_3, respectively. The total number of images is $N = n_1 + n_2 + n_3$. The weight $\omega_{\text{cla},e}$ used in the image classification loss $\mathscr{L}_{\text{cla}}$ w.r.t. class e is given by:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\omega_{\mathrm{cla},e} = \frac{N - n_e}{2N}. \tag{6}$$

$\sum_{e=1}^{3} \omega_{\mathrm{cla},e} = 1$. Therefore, $\mathscr{L}_{\mathrm{cla}}$ can be written as follows:

$$\mathscr{L}_{\mathrm{cla}} = -\sum_{e=1}^{3} \omega_{\mathrm{cla},e}\left(k_{e,i}\log\left(p_e(\mathbf{x}_i)\right) + k_{e,j}\log\left(p_e(\mathbf{x}_j)\right)\right), \tag{7}$$

where $p_e(\mathbf{x}) \in [0, 1]$ indicates the probability that $\mathbf{x}$ belongs to class e. $\sum_{e=1}^{3} p_e(\mathbf{x}) = 1$. $k_{e,i} = 1$ when $e = r_i$, and $k_{e,i} = 0$ otherwise. Given a pair of images $\mathbf{x}_i$ and $\mathbf{x}_j$ with ground-truth labels (r_i, s_i) and (r_j, s_j), respectively, there are four cases:

- case 1: $|s_i - s_j| = 1$ (healthy v.s. viral pneumonia);
- case 2: $|s_i - s_j| = 2$ (viral pneumonia v.s. COVID-19);
- case 3: $|s_i - s_j| = 3$ (COVID-19 v.s. healthy);
- case 4: $|s_i - s_j| = 0$ (two images are of the same class).

The weight $\omega_{\mathrm{sim},c}$ used in the image similarity comparison loss $\mathscr{L}_{\mathrm{sim}}$ with respect to case c is given by:

$$\omega_{\mathrm{sim},c} = \begin{cases} \frac{1}{3}\left(1 - \frac{2n_1n_2}{N(N-1)}\right) & \text{if } c = 1 \\ \frac{1}{3}\left(1 - \frac{2n_2n_3}{N(N-1)}\right) & \text{if } c = 2 \\ \frac{1}{3}\left(1 - \frac{2n_1n_3}{N(N-1)}\right) & \text{if } c = 3' \\ \frac{1}{3}\left(1 - \frac{\sum_{w=1}^{3} n_w(n_w - 1)}{N(N-1)}\right) & \text{if } c = 4 \end{cases} \tag{8}$$

$\sum_{c=1}^{4} \omega_{sim,c} = 1$. Therefore, $\mathscr{L}_{\mathrm{sim}}$ can be written as follows:

$$\mathscr{L}_{\mathrm{sim}} = -\sum_{c=1}^{4} \omega_{\mathrm{sim},c} h_{c,(i,j)} \log\left(q_c(\mathbf{x}_i, \mathbf{x}_j)\right), \tag{9}$$

where $q_c(\mathbf{x}_i, \mathbf{x}_j) \in [0, 1]$ indicates the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ under case c. $\sum_{c=1}^{4} q_c(\mathbf{x}_i, \mathbf{x}_j) = 1$. $h_{c,(i,j)} = 1$ when $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to case c, and $h_{c,(i,j)} = 0$ otherwise. The hyper-parameter λ is empirically set to 0.3.

The experimental results of these two multi-task biomedical image classification tasks are presented in Fig. 8 with two confusion matrices. These results demonstrate that our proposed

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

MTSN can be effectively trained with very few images to solve multi-class biomedical image classification problems. Specifically, the achieved accuracy values for chest X-ray image classification (~25-shot learning) and lung histopathological image classification (10-shot learning) are 93% and 90%, respectively. The chest X-ray image classification result compares favorably with the accuracy range of 82%−93% achieved by supervised methods (2,520 images for training, 840 images for validation, and 840 images for testing) using all available training data [68]. Although the accuracy achieved by MTSN for lung histopathological image classification is lower than the accuracy of over 97% reported in [69] by supervised approaches using the full training set (8,250 images for training, ~3,000 images for validation, and 3,744 images for testing), we believe that our proposed low-shot learning method can achieve comparable results when a small amount of additional images are incorporated for MTSN training.

V. Discussion

Extensive experiments demonstrate the effectiveness and efficiency of training an MTSN by minimizing our proposed CWCE loss. Such a low-shot learning approach significantly reduces over-fitting and achieves an accuracy on a small training set (1,147 fundus images) comparable to a large training set (approximately 53K fundus images). We also demonstrate its effectiveness on two additional multi-class few-shot biomedical image classification tasks. Additionally, the MTSNs fine-tuned with OVV self-training outperform the SoTA semi-supervised glaucoma diagnosis algorithms [20], [21] as well as general-purpose semi-supervised learning algorithms [24]–[27] trained for glaucoma diagnosis. They perform similarly, and in some cases, better than SoTA supervised algorithms. However, our proposed method has two limitations:

- In the OVV self-training, each target fundus image must be compared with all the reference fundus images in the same mini-batch, resulting in a computational complexity of $\mathcal{O}(n^2)$. As the mini-batch size increases, OVV self-training becomes relatively memory-consuming. The high computational complexity of OVV self-training may reduce the feasibility of this method in clinical practice for now. Therefore, we plan to improve the OVV self-training strategy in the future by adaptively selecting only a limited number of reference fundus photographs for semi-supervised glaucoma diagnosis, which can reduce the computational complexity and make the method more practical in clinical settings.
- Our proposed OVV self-training strategy is developed for binary image classification and may not be directly applicable to multi-class image classification problems. Therefore, we plan to extend the contrastive prediction procedure to handle multi-class image classification problems in future work. More hyper-parameter tuning can always be done, but it is so easy to over-fit with limited data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

VI. Conclusion

The main contributions of this paper include: (1) a multi-task Siamese network that can learn glaucoma diagnosis from very limited labeled training data; (2) an effective semi-supervised learning strategy, referred to as One-Vote Veto self-training, which can produce pseudo labels for the unlabeled data to fine-tune a pre-trained multi-task Siamese network. Extensive experiments conducted on four fundus image datasets demonstrated the effectiveness of these proposed techniques. The low-shot learning reduces over-fitting and achieves an accuracy on a small training set comparable to that of a large training set. Furthermore, with One-Vote Veto self-training, the multi-task Siamese networks perform similarly to their backbone CNNs (trained via supervised learning on the full training set) on the OHTS test set and show better generalizability on three additional test sets. The methods introduced in this paper can also be applied to other few-shot multi-class biomedical image classification problems, *e.g.*, COVID-19 and lung cancer diagnosis, and other diseases in which only a small quantity of ground-truth labels are available for network training.

Acknowledgments

This research was supported in part by the Fundamental Research Funds for the Central Universities, and was also supported in part by awards from the National Eye Institute (grants EY027510, R214278211, P30EY022589, K99EY030942, EY026574, EY11008, and EY19869), the National Center on Minority Health and Health Disparities, National Institutes of Health (grants EY09341 and EY09307), Horncrest Foundation, awards to the Department of Ophthalmology and Visual Sciences at Washington University, the NIH Vision Core Grant P30 EY 02687, Merck Research Laboratories, Pfizer, Inc., White House Station, New Jersey, and unrestricted grants from Research to Prevent Blindness, Inc., New York, NY.

References

[1]. Weinreb RN and Khaw PT, “Primary open-angle glaucoma,” The Lancet, vol. 363, no. 9422, pp. 1711–1720, 2004.

[2]. Tham Y-C et al. , “Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis,” Ophthalmology, vol. 121, no. 11, pp. 2081–2090, 2014. [PubMed: 24974815]

[3]. Fan R. et al. , “Detecting glaucoma in the ocular hypertension study using deep learning,” JAMA Ophthalmology, vol. 140, no. 4, pp. 383–391, 2022. [PubMed: 35297959]

[4]. Traverso C. et al. , “Direct costs of glaucoma and severity of the disease: a multinational long term study of resource utilisation in Europe,” British Journal of Ophthalmology, vol. 89, no. 10, pp. 1245–1249, 2005. [PubMed: 16170109]

[5]. Huang W. et al. , “The adverse impact of glaucoma on psychological function and daily physical activity,” Journal of Ophthalmology, vol. 2020, 2020.

[6]. Parrish RK et al. , “Visual function and quality of life among patients with glaucoma,” Archives of Ophthalmology, vol. 115, no. 11, pp. 1447–1455, 1997. [PubMed: 9366678]

[7]. Kwon M. et al. , “Association between glaucoma and at–fault motor vehicle collision involvement among older drivers: A population-based study,” Ophthalmology, vol. 123, no. 1, pp. 109–116, 2016. [PubMed: 26459997]

[8]. McGwin G Jr et al. , “Binocular visual field impairment in glaucoma and at-fault motor vehicle collisions,” Journal of Glaucoma, vol. 24, no. 2, pp. 138–143, 2015. [PubMed: 25642648]

[9]. Li L. et al., “Attention based glaucoma detection: A large-scale database and cnn model,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10571–10580.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[10]. Fan R. et al. , “Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization,” Ophthalmology Science, vol. 3, no. 1, p. 100233, 2023. [PubMed: 36545260]

[11]. Gordon MO and Kass MA, “The ocular hypertension treatment study: Design and baseline description of the participants,” Archives of Ophthalmology, vol. 117, no. 5, pp. 573–583, 1999. [PubMed: 10326953]

[12]. Kass MA et al. , “The ocular hypertension treatment study: A randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma,” Archives of Ophthalmology, vol. 120, no. 6, pp. 701–713, 2002. [PubMed: 12049574]

[13]. Gordon MO et al. , “Assessment of the impact of an endpoint committee in the ocular hypertension treatment study,” American Journal of Ophthalmology, vol. 199, pp. 193–199, 2019. [PubMed: 30471242]

[14]. Li Z. et al. , “Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs,” Ophthalmology, vol. 125, no. 8, pp. 1199–1206, 2018. [PubMed: 29506863]

[15]. Gómez-Valverde JJ et al. , “Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning,” Biomedical Optics Express, vol. 10, no. 2, pp. 892–913, 2019. [PubMed: 30800522]

[16]. Judy D. et al. , “Automated identification of glaucoma from fundus images using deep learning techniques,” European Journal of Molecular & Clinical Medicine, vol. 7, no. 2, pp. 5449–5458, 2020.

[17]. Serener A. and Serte S, “Transfer learning for early and advanced glaucoma detection with convolutional neural networks,” in 2019 Medical Technologies Congress (TIPTEKNO). IEEE, 2019, pp. 1–4.

[18]. Thakur A. et al. , “Predicting glaucoma before onset using deep learning,” Ophthalmology Glaucoma, vol. 3, no. 4, pp. 262–268, 2020. [PubMed: 33012331]

[19]. Kim M. et al., “Few-shot learning using a small-sized dataset of high-resolution fundus images for glaucoma diagnosis,” in Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care, 2017, pp. 89–92.

[20]. Al Ghamdi M. et al., “Semi-supervised transfer learning for convolutional neural networks for glaucoma detection,” in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3812–3816.

[21]. Díaz Pinto A. et al. , “Retinal image synthesis and semi-supervised learning for glaucoma assessment,” IEEE Transactions on Medical Imaging, vol. 38, no. 9, pp. 2211–2218, 2019. [PubMed: 30843823]

[22]. Díaz Pinto A. et al. , “CNNs for automatic glaucoma assessment using fundus images: An extensive validation,” Biomedical Engineering Online, vol. 18, pp. 1–19, 2019. [PubMed: 30602383]

[23]. Sample P. et al. , “The African descent and glaucoma evaluation study (ADAGES): Design and baseline data,” Archives of Ophthalmology, vol. 127, no. 9, pp. 1136–1145, 2009. [PubMed: 19752422]

[24]. Wang Y. et al., “Freematch: Self-adaptive thresholding for semi-supervised learning,” in the International Conference on Learning Representations (ICLR), 2023, in press.

[25]. Chen H. et al., “Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning,” in the International Conference on Learning Representations (ICLR), 2023, in press.

[26]. Sohn K. et al. , “Fixmatch: Simplifying semi-supervised learning with consistency and confidence,” Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 596–608, 2020.

[27]. Zhang B. et al. , “Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling,” Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 18408–18419, 2021.

[28]. Dosovitskiy A. et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in the International Conference on Learning Representations (ICLR).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[29]. Chowdhury ME et al. , “Can AI help in screening viral and COVID-19 pneumonia?” IEEE Access, vol. 8, pp. 132665–132676, 2020.

[30]. Rahman T. et al. , “Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images,” Computers in Biology and Medicine, vol. 132, p. 104319, 2021. [PubMed: 33799220]

[31]. Borkowski AA et al. , “Lung and colon cancer histopathological image dataset (LC25000),” CoRR, 2019.

[32]. Krizhevsky A, “One weird trick for parallelizing convolutional neural networks,” CoRR, 2014.

[33]. Simonyan K. and Zisserman A, “Very deep convolutional networks for large-scale image recognition,” in the International Conference on Learning Representations (ICLR), 2015, pp. 1–14.

[34]. Szegedy C. et al., “Going deeper with convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[35]. Szegedy C. et al., “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.

[36]. Ahn JM et al. , “A deep learning model for the detection of both advanced and early glaucoma using fundus photography,” PloS one, vol. 13, no. 11, p. e0207982, 2018. [PubMed: 30481205]

[37]. Deng J. et al., “ImageNet: A large-scale hierarchical image database,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.

[38]. He K. et al., “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[39]. Liu S. et al. , “A deep learning-based algorithm identifies glaucomatous discs using monoscopic fundus photographs,” Ophthalmology Glaucoma, vol. 1, no. 1, pp. 15–22, 2018. [PubMed: 32672627]

[40]. Ran AR et al. , “Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: A retrospective training and validation deep-learning analysis,” The Lancet Digital Health, vol. 1, no. 4, pp. e172–e182, 2019. [PubMed: 33323187]

[41]. Medeiros FA et al. , “Detection of progressive glaucomatous optic nerve damage on fundus photographs with deep learning,” Ophthalmology, vol. 128, no. 3, pp. 383–392, 2021. [PubMed: 32735906]

[42]. Christopher M. et al. , “Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs,” Scientific Reports, vol. 8, no. 1, pp. 1–13, 2018. [PubMed: 29311619]

[43]. Christopher M. et al. , “Effects of study population, labeling and training on glaucoma detection using deep learning algorithms,” Translational Vision Science & Technology, vol. 9, no. 2, pp. 27–27, 2020.

[44]. Jain D. et al. , “Open-source, ultra-low-cost smartphone attachment for non-mydriatic fundus photography-open indirect ophthalmoscope,” Investigative Ophthalmology & Visual Science, vol. 57, no. 12, pp. 1685–1685, 2016.

[45]. Matthew Lawson E. and Raskar R, “Smart phone administered fundus imaging without additional imaging optics,” Investigative Ophthalmology & Visual Science, vol. 55, no. 13, pp. 1609–1609, 2014.

[46]. Sandler M. et al., “MobileNetv2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.

[47]. Wang Y. et al. , “Generalizing from a few examples: A survey on few-shot learning,” ACM Computing Surveys (CSUR), vol. 53, no. 3, pp. 1–34, 2020.

[48]. Zhou P. et al., “Attention-based bidirectional long short-term memory networks for relation classification,” in Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 2: Short Papers), 2016, pp. 207–212.

[49]. Radford A. et al., “Unsupervised representation learning with deep convolutional generative adversarial networks,” in the International Conference on Learning Representations (ICLR), 2015.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[50]. Triguero I. et al. , “Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study,” Knowledge and Information Systems, vol. 42, no. 2, pp. 245–284, 2015.

[51]. Van Engelen JE and Hoos HH, “A survey on semi-supervised learning,” Machine Learning, vol. 109, no. 2, pp. 373–440, 2020.

[52]. Miller A. et al., “Key-value memory networks for directly reading documents,” in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 1400–1409.

[53]. Koch G. et al. , “Siamese neural networks for one-shot image recognition,” in ICML Deep Learning Workshop, vol. 2. Lille, 2015.

[54]. Jang J. and Kim CO, “Siamese network-based health representation learning and robust reference-based remaining useful life prediction,” IEEE Transactions on Industrial Informatics, vol. 18, no. 8, pp. 5264–5274, 2021.

[55]. Wang Q. et al., “Learning attentions: Residual attentional Siamese network for high performance online visual tracking,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4854–4863.

[56]. Wang W. et al., “Face recognition based on deep learning,” in 1st International Conference on Human Centered Computing (HCC). Springer, 2015, pp. 812–820.

[57]. Hoffer E. and Ailon N, “Deep metric learning using triplet network,” in International Workshop on Similarity-Based Pattern Recognition. Springer, 2015, pp. 84–92.

[58]. Khosla P. et al. , “Supervised contrastive learning,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 18661–18673.

[59]. Kendall A. et al., “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7482–7491.

[60]. McClosky D, Charniak E, and Johnson M, “Effective self-training for parsing,” in Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 2006, pp. 152–159.

[61]. Sukhbaatar S, Weston J, Fergus R. et al., “End-to-end memory networks,” in Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 2440–2448.

[62]. Maaten L. v. d. and Hinton G, “Visualizing data using t-SNE,” Journal of Machine Learning Research, vol. 9, no. 11, pp. 2579–2605, 2008.

[63]. Chen L-C et al. , “Rethinking atrous convolution for semantic image segmentation,” CoRR, 2017.

[64]. Tan C. et al., “A survey on deep transfer learning,” in International Conference on Artificial Neural Networks (ICANN). Springer, 2018, pp. 270–279.

[65]. Huang G. et al., “Densely connected convolutional networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.

[66]. Tan M. and Le Q, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in International Conference on Machine Learning (ICML). PMLR, 2019, pp. 6105–6114.

[67]. Chattopadhay A. et al., “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 839–847.

[68]. Jadon S, “COVID-19 detection from scarce chest x-ray image data using few-shot deep learning approach,” in Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications, vol. 11601. International Society for Optics and Photonics, 2021, p. 116010X.

[69]. Abbas MA et al. , “The histopathological diagnosis of adenocarcinoma & squamous cells carcinoma of lungs by artificial intelligence: A comparative study of convolutional neural networks,” MedRxiv, pp.2020–05, 2020.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

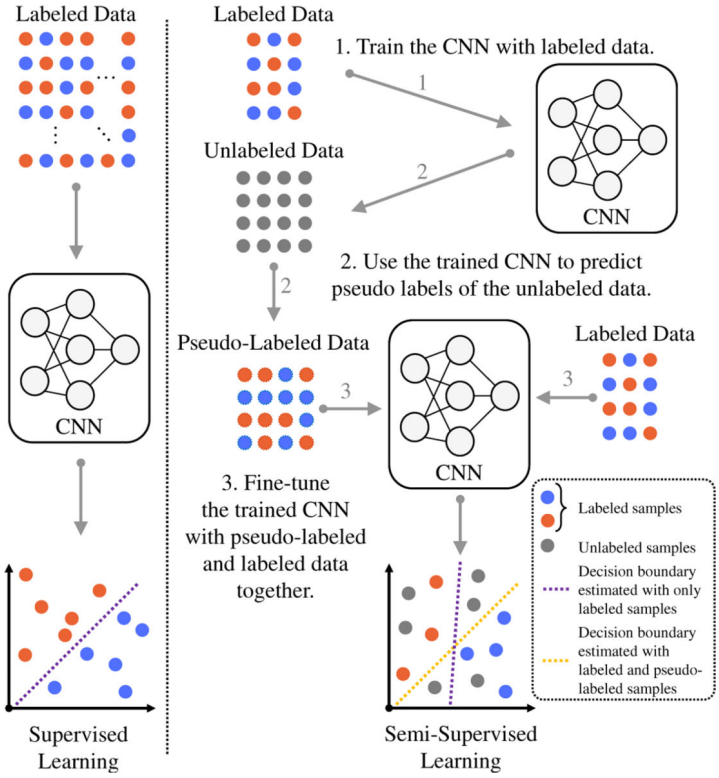


Fig. 1:
Supervised learning v.s. semi-supervised learning.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

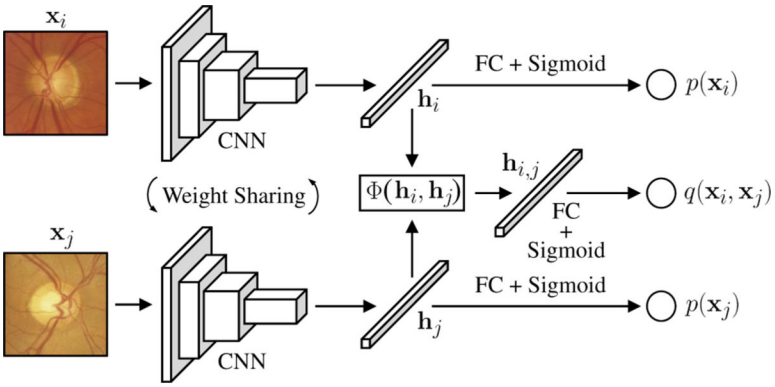


Fig. 2:

An illustration of our MTSN for joint learning of fundus image similarity measurement (primary task) and glaucoma diagnosis (secondary task) in a low-shot manner.

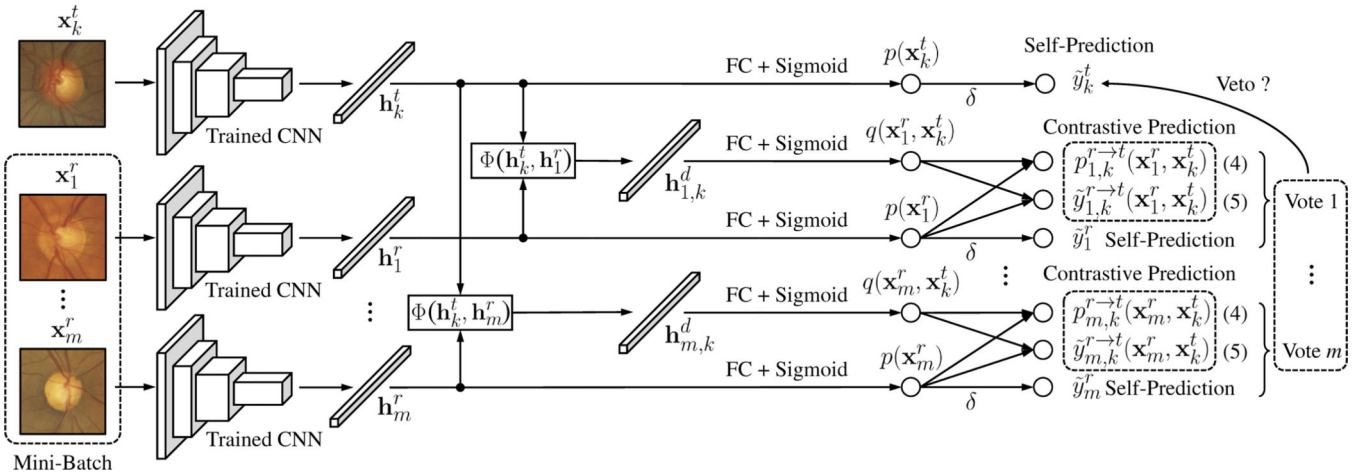


Fig. 3:

An illustration of our One-Vote Veto Self-Training strategy. $\mathbf{h}^d_{1,k}$ and $\mathbf{h}^d_{m,k}$ are two 1D embeddings, followed by an FC layer to produce scalars indicating the similarities between the given pairs of reference and target fundus images. The reference fundus images having ground-truth labels are used to train the MTSN by minimizing (1). The contrastive predictions are obtained using (4) and (5). The pseudo labels of the target fundus photographs are generated using One-Vote Veto Self-Training strategy, as detailed in Algorithm 1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

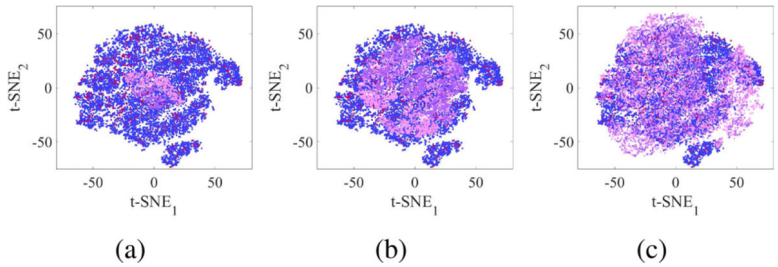
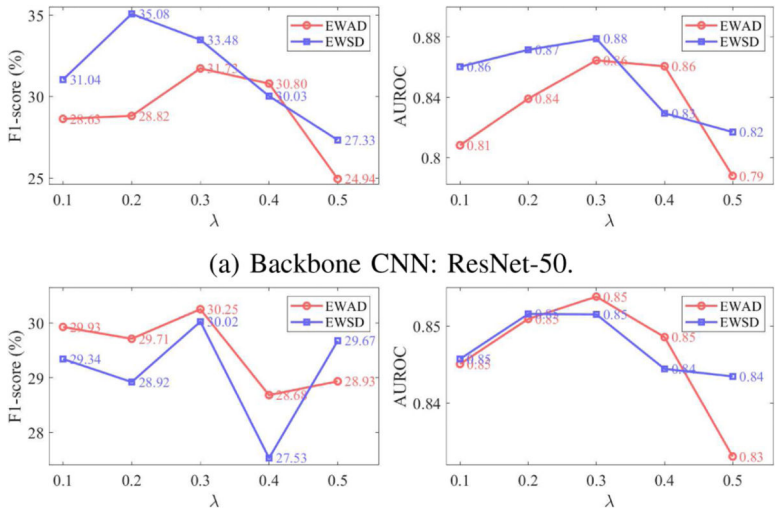


Fig. 4:

Comparisons of dataset visualizations produced by t-SNE, where • and • represent the healthy and GON images in the OHTS test set, respectively; • and • in (a) represent the healthy and GON images in the ACRIMA dataset, respectively; • and • in (b) represent the healthy and GON images in the LAG dataset, respectively; • and • in (c) represent the healthy and GON images in the DIGS/ADAGES test set, respectively.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript



(a) Backbone CNN: ResNet-50.

(b) Backbone CNN: MobileNet-v2.

Fig. 5:

MTSN performance on the OHTS test set with respect to different λ and $\Phi(\cdot)$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

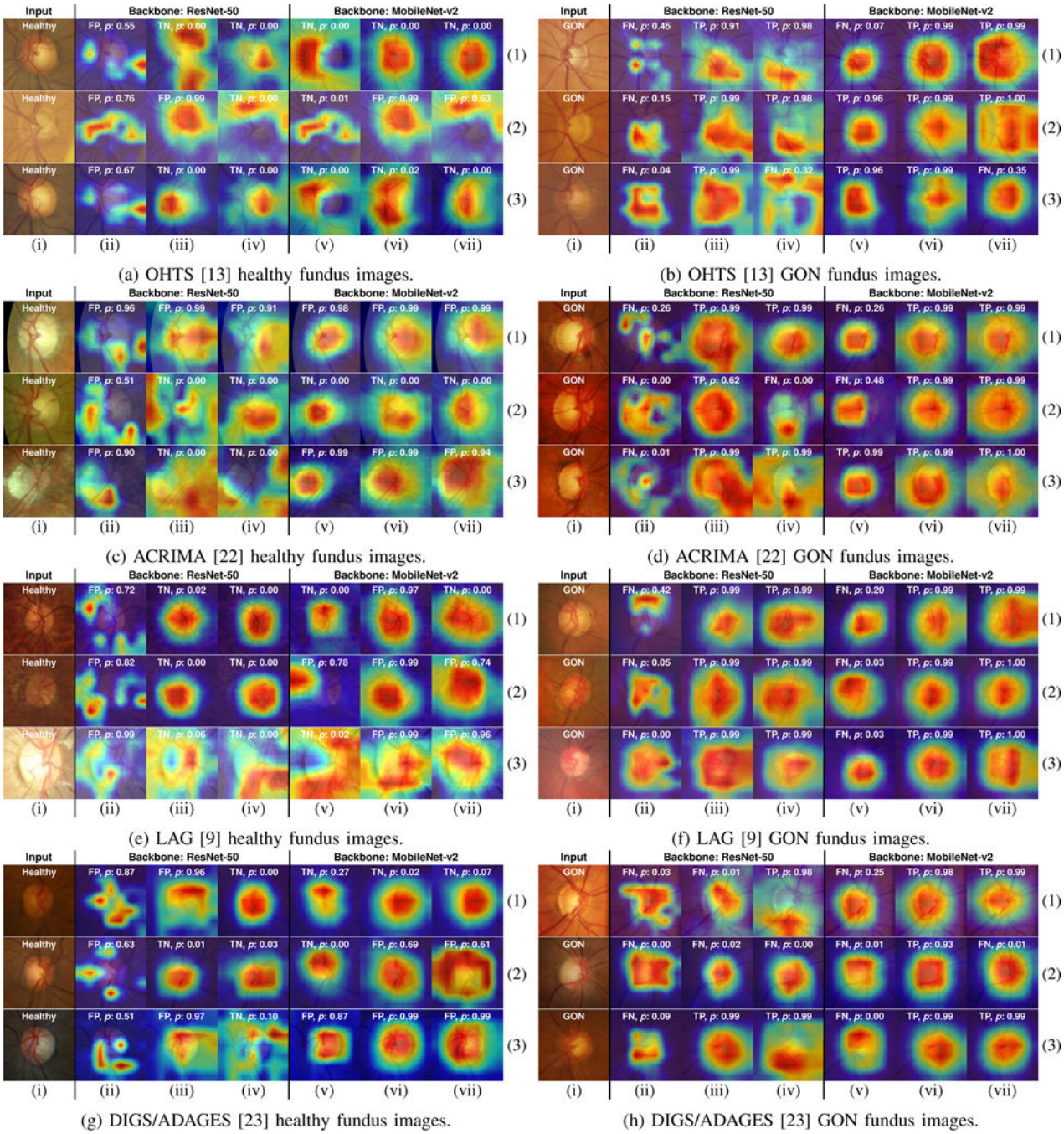


Fig. 6:

Examples of Grad-CAM++ [67] results: (i) fundus images; (ii) and (v) show the class activation maps of (i), obtained by the backbone CNNs trained through supervised learning on the entire training set (containing ~53K fundus images); (iii) and (vi) show the class activation maps of (i), obtained by MTSNs trained through low-shot learning on a small training set (containing 1,147 fundus images); (iv) and (vii) show the class activation maps of (i), obtained by MTSNs fine-tuned with our proposed OVV self-training on the entire training set (containing ~53K fundus images without ground-truth labels).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

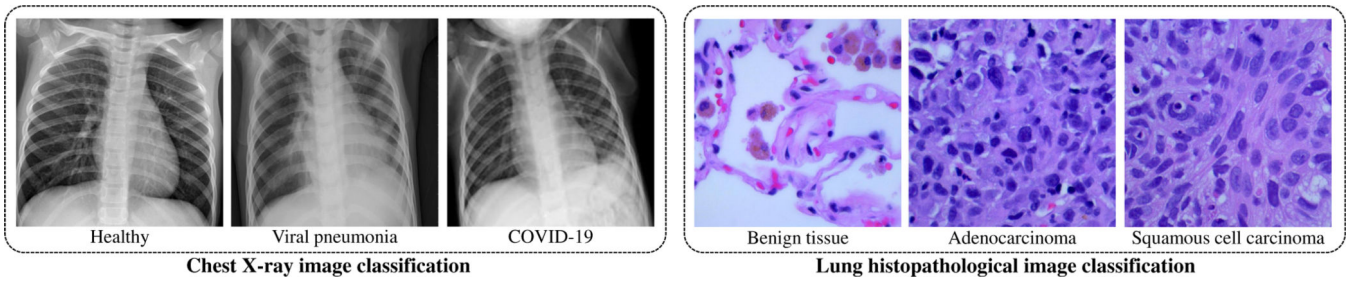


Fig. 7:
Examples of images used in two few-shot multi-class lung disease diagnosis tasks.

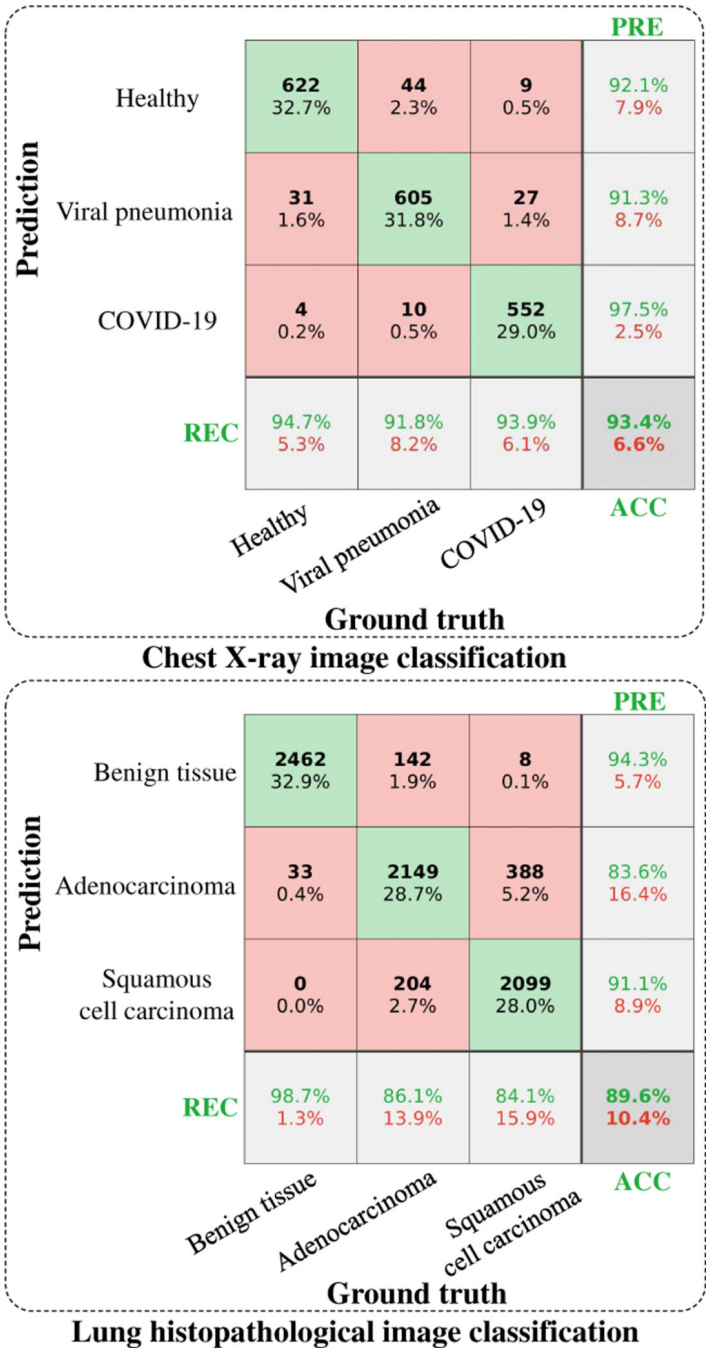


Fig. 8:

Experimental results of two few-shot lung disease diagnosis tasks. REC: recall; PRE: precision; ACC: accuracy.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE I:

Comparison between supervised and low-shot learning (both on the low-shot OHTS training set). λ is set to 0.3. The best results are shown in bold font.

Test set	Experiments	Training strategy	ResNet-50			MobileNet-v2		
		Accuracy (%)	F1-score (%)	AUROC	Accuracy (%)	F1-score (%)	AUROC	
ACRIMA [22]	Baseline	Supervised learning	57.163	56.734	0.625	**73.333**	**78.341**	0.779
	EWAD	Low-shot learning	**67.092**	**63.175**	**0.758**	70.355	67.797	0.820
	EWSD	Low-shot learning	49.504	25.523	0.437	66.241	59.107	**0.823**
LAG [9]	Baseline	Supervised learning	64.318	57.445	0.714	65.122	63.268	0.781
	EWAD	Low-shot learning	**79.028**	65.908	**0.841**	**79.007**	**69.482**	0.843
	EWSD	Low-shot learning	78.039	**68.179**	0.826	78.430	61.830	**0.846**
DIGS/ADAGES [23]	Baseline	Supervised learning	59.639	59.708	0.648	61.478	66.128	0.669
	EWAD	Low-shot learning	**67.745**	**60.754**	**0.743**	**69.176**	**68.145**	**0.748**
	EWSD	Low-shot learning	65.736	55.722	0.700	68.120	64.028	0.740

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE II:

Evaluation of our OVV self-training w.r.t. different κ_1, κ_2, and m. The best results are shown in bold font. ↑ indicates that semi-supervised learning outperforms low-shot learning.

Dataset	κ_1	κ_2	m	ResNet-50			MobileNet-v2		
				Accuracy (%)	F1-score (%)	AUROC	Accuracy (%)	F1-score (%)	AUROC
OHTS [11], [12]	0	0.01	20	91.415 ↑	41.148 ↑	0.898 ↑	90.609 ↑	36.960 ↑	0.887 ↑
	0	0.01	15	92.113 ↑	41.316 ↑	**0.898** ↑	93.470 ↑	36.976 ↑	**0.893** ↑
	0	0.01	10	**94.199** ↑	**43.759** ↑	0.890 ↑	**93.742** ↑	**37.779** ↑	0.863 ↑
	0	0.10	20	90.516 ↑	**38.139** ↑	**0.898** ↑	88.825 ↑	34.351 ↑	**0.878** ↑
	2	0.01	20	90.717 ↑	35.818 ↑	0.885 ↑	**93.463** ↑	**35.204** ↑	0.862 ↑
	2	0.10	20	92.656 ↑	32.017 ↑	0.851 ↓	90.772 ↑	32.616 ↑	0.858 ↑
	4	0.01	20	92.610 ↑	29.668 ↓	0.854 ↓	92.268 ↑	31.852 ↑	0.854 ↑
	4	0.10	20	**92.672** ↑	28.463 ↓	0.842 ↓	90.803 ↑	32.690 ↑	0.859 ↑
ACRIMA [22]	0	0.01	20	59.858 ↓	49.192 ↓	**0.775** ↑	**72.340** ↑	**70.229** ↑	**0.840** ↑
	0	0.01	15	**60.426** ↓	**49.365** ↓	0.751 ↓	63.404 ↓	54.895 ↓	0.814 ↓
	0	0.01	10	54.610 ↓	35.743 ↓	0.721 ↓	61.986 ↓	51.273 ↓	0.826 ↑
LAG [9]	0	0.01	20	**80.882** ↑	**66.304** ↑	**0.881** ↑	**79.625** ↑	**65.262** ↓	0.851 ↑
	0	0.01	15	76.864 ↓	56.252 ↓	0.825 ↓	76.670 ↓	55.906 ↓	0.841 ↓
	0	0.01	10	76.638 ↓	56.518 ↓	0.826 ↓	75.834 ↓	51.748 ↓	**0.866** ↑
DIGS/ADAGES [23]	0	0.01	20	**67.813** ↑	**58.315** ↓	**0.763** ↑	**69.653** ↑	**63.258** ↓	0.777 ↑
	0	0.01	15	63.045 ↓	44.727 ↓	0.753 ↑	66.383 ↓	52.888 ↓	0.773 ↑
	0	0.01	10	61.819 ↓	41.340 ↓	0.727 ↓	63.965 ↓	44.198 ↓	**0.789** ↑

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE III:

AUROC (shown along with 95% CI) and training time *t* (min) per epoch of supervised, low-shot, and semi-supervised glaucoma diagnosis.

Backbone	Method	Training strategy	OHTS [11], [12]	ACRIMA [22]	LAG [9]	DIGS/ADAGES [23]	*t* (min)
ResNet-50	Baseline	Supervised learning	0.904 (95% Cl: 0.865, 0.935)	0.736 (95% Cl: 0.698, 0.771)	0.794 (95% Cl: 0.780, 0.807)	0.744 (95% Cl: 0.696, 0.792)	52.1
	MTSN	Low-shot learning	0.869 (95% Cl: 0.833, 0.901)	0.758 (95% Cl: 0.723, 0.792)	0.841 (95% Cl: 0.829, 0.853)	0.743 (95% Cl: 0.683, 0.795)	1.8
	MTSN+OVV	Semi-supervised learning	0.898 (95% Cl: 0.857, 0.928)	0.775 (95% Cl: 0.741, 0.808)	0.881 (95% Cl: 0.870, 0.891)	0.763 (95% Cl: 0.695, 0.820)	203.7
MobileNet-v2	Baseline	Supervised learning	0.893 (95% Cl: 0.845, 0.932)	0.794 (95% Cl: 0.760, 0.825)	0.856 (95% Cl: 0.844, 0.867)	0.786 (95% Cl: 0.728, 0.835)	42.9
	MTSN	Low-shot learning	0.859 (95% Cl: 0.813, 0.896)	0.820 (95% Cl: 0.786, 0.850)	0.843 (95% Cl: 0.831, 0.855)	0.748 (95% Cl: 0.689, 0.802)	1.2
	MTSN+OVV	Semi-supervised learning	0.887 (95% Cl: 0.850, 0.920)	0.840 (95% Cl: 0.808, 0.867)	0.851 (95% Cl: 0.838, 0.862)	0.777 (95% Cl: 0.718, 0.826)	125.4
DenseNet	Baseline	Supervised learning	0.898 (95% Cl: 0.867, 0.927)	0.810 (95% Cl: 0.778, 0.841)	0.784 (95% Cl: 0.771, 0.798)	0.743 (95% Cl: 0.688, 0.789)	122.7
	MTSN	Low-shot learning	0.854 (95% Cl: 0.811, 0.894)	0.753 (95% Cl: 0.716, 0.786)	0.853 (95% Cl: 0.842, 0.865)	0.732 (95% Cl: 0.675, 0.785)	5.7
	MTSN+OVV	Semi-supervised learning	0.896 (95% Cl: 0.861, 0.926)	0.783 (95% Cl: 0.748, 0.817)	0.831 (95% Cl: 0.818, 0.843)	0.746 (95% Cl: 0.678, 0.800)	324.0
EfficientNet	Baseline	Supervised learning	0.768 (95% Cl: 0.684, 0.834)	0.633 (95% Cl: 0.590, 0.672)	0.650 (95% Cl: 0.634, 0.667)	0.658 (95% Cl: 0.611,0.702)	48.7
	MTSN	Low-shot learning	0.863 (95% Cl: 0.818, 0.899)	0.845 (95% Cl: 0.815, 0.873)	0.845 (95% Cl: 0.833, 0.856)	0.719 (95% Cl: 0.659, 0.774)	1.5
	MTSN+OVV	Semi-supervised learning	0.886 (95% Cl: 0.845, 0.918)	0.792 (95% Cl: 0.758, 0.824)	0.850 (95% Cl: 0.837, 0.861)	0.749 (95% Cl: 0.690, 0.800)	159.8

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE IV:

Comparisons of supervised learning, low-shot learning, and semi-supervised learning w.r.t. different percentages of labeled training data. In the baseline experiments (2.0% of training data), one fundus photograph is selected from each patient; In the experiments with 0.5% and 1.0% of training data, a subset is created by choosing patient IDs at regular intervals.

Percentage of training data	Supervised learning			Low-shot learning			Semi-supervised learning		
	Accuracy (%)	F1-score (%)	AUROC	Accuracy (%)	F1-score (%)	AUROC	Accuracy (%)	F1-score (%)	AUROC
0.5	80.402	18.066	0.720	84.769	20.228	0.759	88.856	25.351	0.797
1.0	79.557	21.573	0.799	88.538	25.654	0.806	88.453	32.472	0.857
2.0 (baseline)	84.141	26.743	0.838	87.150	31.726	0.865	91.415	41.148	0.898
10.0	85.651	32.288	0.890	89.988	38.612	0.891	91.446	39.760	0.899
50.0	92.950	40.188	0.907	94.223	40.826	0.887	92.067	38.187	0.889
90.0	89.556	37.582	0.905	92.897	41.878	0.887	93.331	43.421	0.898

(a) Backbone CNN: ResNet-50.

Percentage of training data	Supervised learning			Low-shot learning			Semi-supervised learning		
	Accuracy (%)	F1-score (%)	AUROC	Accuracy (%)	F1-score (%)	AUROC	Accuracy (%)	F1-score (%)	AUROC
0.5	83.033	19.035	0.720	81.954	18.035	0.745	87.918	29.246	0.826
1.0	67.659	16.458	0.748	91.764	20.628	0.772	89.407	31.357	0.841
2.0 (baseline)	78.144	23.547	0.840	86.018	30.252	0.854	90.609	36.960	0.887
10.0	90.429	34.873	0.884	89.562	36.808	0.897	89.624	36.887	0.888
50.0	92.438	43.361	0.906	91.043	39.050	0.888	93.230	40.653	0.889
90.0	93.754	42.401	0.908	93.750	37.519	0.890	91.857	41.602	0.896

(b) Backbone CNN: MobileNet-v2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE V:

Comparisons with other SoTA glaucoma diagnosis algorithms. The best results for each training strategy are shown in bold font.

Training strategy	Method	Accuracy (%)	F1-score (%)	AUROC
Supervised learning	Li *et al.* [14]	93.812	42.590	0.886
	Gómez-Valverde *et al.* [15]	**95.068**	**48.039**	0.903
	Judy *et al.* [16]	94.188	42.835	0.908
	Serener and Serte [17]	90.492	41.227	**0.912**
	Thakur *et al.* [18]	94.145	44.115	0.896
	Fan *et al.* [3] (Baseline Result)	93.261	43.016	0.904
Low-shot learning	Kim *et al.* [19]	84.203	23.161	0.786
	MTSN (Backbone: ResNet-50) (Ours)	**87.150**	**31.726**	**0.865**
	MTSN (Backbone: MobileNet-v2) (Ours)	86.018	30.252	0.854
Semi-supervised learning	Al Ghamdi *et al.* [20]	84.808	27.579	0.830
	Diaz-Pinto *et al.* [21]	76.619	20.721	0.748
	MTSN (Backbone: ResNet-50) + OVV Self-Training (Ours)	**90.244**	**38.454**	**0.899**
	MTSN (Backbone: MobileNet-v2) + OVV Self-Training (Ours)	89.360	36.599	0.891

(a) OHTS [12] test set.

Training strategy	Method	Accuracy (%)	F1-score (%)	AUROC
Supervised learning	Li *et al.* [14]	60.142	46.272	0.813
	Gómez-Valverde *et al.* [15]	63.546	53.358	**0.826**
	Judy *et al.* [16]	57.872	41.650	0.824
	Serener and Serte [17]	54.326	36.364	0.675
	Thakur *et al.* [18]	**65.106**	**58.020**	0.794
	Fan *et al.* [3] (Baseline Result)	53.333	31.601	0.736
Low-shot learning	Kim *et al.* [19]	64.965	58.627	**0.844**
	MTSN (Backbone: ResNet-50) (Ours)	67.092	63.175	0.758
	MTSN (Backbone: MobileNet-v2) (Ours)	**70.355**	**67.797**	0.820
Semi-supervised learning	Al Ghamdi *et al.* [20]	68.511	67.257	0.794

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

	Diaz-Pinto *et al.* [21]	59.149	44.828	0.818
	MTSN (Backbone: ResNet-50) + OVV Self-Training (Ours)	64.539	57.627	0.801
	MTSN (Backbone: MobileNet-v2) + OVV Self-Training (Ours)	**72.340**	**70.939**	**0.835**

(b) ACRIMA [22] dataset.

Training strategy	Method	Accuracy (%)	F1-score (%)	AUROC
A S	Li *et al.* [14]	76.535	53.756	0.855
Supervised learning	Gómez-Valverde *et al.* [15]	80.202	62.417	**0.883**
	Judy *et al.* [16]	78.348	60.174	0.860
	Serener and Serte [17]	77.379	**66.545**	0.806
	Thakur *et al.* [18]	**80.305**	65.882	0.856
	Fan *et al.* [3] (Baseline Result)	75.052	50.267	0.794
Low-shot learning	Kim *et al.* [19]	74.619	65.000	0.805
	MTSN (Backbone: ResNet-50) (Ours)	**79.028**	65.908	0.841
	MTSN (Backbone: MobileNet-v2) (Ours)	79.007	**69.482**	**0.843**
Semi-supervised learning	Al Ghamdi *et al.* [20]	79.028	**72.382**	0.860
	Diaz-Pinto *et al.* [21]	65.554	56.662	0.701
	MTSN (Backbone: ResNet-50) + OVV Self-Training (Ours)	**81.644**	69.929	**0.879**
	MTSN (Backbone: MobileNet-v2) + OVV Self-Training (Ours)	80.470	68.692	0.849

(c) LAG [9] dataset.

Training strategy	Method	Accuracy (%)	F1-score (%)	AUROC
Supervised learning	Li *et al.* [14]	65.293	46.955	0.780
	Gómez-Valverde *et al.* [15]	65.157	45.729	**0.795**
	Judy *et al.* [16]	63.556	41.144	0.760
	Serener and Serte [17]	69.108	57.478	0.757
	Thakur *et al.* [18]	**70.606**	**60.322**	0.786
	Fan *et al.* [3] (Baseline Result)	62.500	38.042	0.744
Low-shot learning	Kim *et al.* [19]	63.862	52.949	0.687
	MTSN (Backbone: ResNet-50) (Ours)	67.745	60.754	0.743

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

	MTSN (Backbone: MobileNet-v2) (Ours)	**69.176**	**68.145**	**0.748**
	Al Ghamdi *et al.* [20]	64.850	54.816	0.716
	Diaz-Pinto *et al.* [21]	64.441	59.908	0.677
	MTSN (Backbone: ResNet-50) + OVV Self-Training (Ours)	66.281	55.486	0.747
Semi-supervised learning	**MTSN (Backbone: MobileNet-v2) + OVV Self-Training (Ours)**	**70.402**	**66.015**	**0.776**

(d) DIGS/ADAGES [23] test set.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE VI:

Comparisons with SoTA general-purpose semi-supervised learning methods that use the vision Transformer [28] as the backbone network. The best results of each dataset are shown in bold font.

Test Set	Method	Accuracy (%)	F1-score (%)	AUROC
OHTS [11], [12]	FreeMatch [24]	86.510	27.745	0.784
	SoftMatch [25]	85.679	25.718	0.768
	FixMatch [26]	84.695	27.825	0.807
	FlexMatch [27]	**92.715**	26.331	0.696
	OVV Self-Training (Ours)	90.244	**38.454**	**0.899**
ACRIMA [22]	FreeMatch [24]	77.872	77.778	0.791
	SoftMatch [25]	78.582	79.115	0.795
	FixMatch [26]	**78.582**	**79.622**	0.792
	FlexMatch [27]	60.426	47.850	0.644
	OVV Self-Training (Ours)	64.539	57.627	**0.801**
LAG [9]	FreeMatch [24]	78.471	67.087	0.793
	SoftMatch [25]	79.110	67.972	0.798
	FixMatch [26]	81.314	**73.394**	0.827
	FlexMatch [27]	71.302	35.360	0.654
	OVV Self-Training (Ours)	**81.644**	69.929	**0.879**
DIGS/ADAGES [23]	FreeMatch [24]	65.429	56.716	0.685
	SoftMatch [25]	65.395	56.986	0.679
	FixMatch [26]	**68.835**	**65.090**	0.731
	FlexMatch [27]	57.050	26.558	0.594
	OVV Self-Training (Ours)	66.281	55.486	**0.747**

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE VII:

Training, validation, and test sample sizes in the two few-shot biomedical image classification experiments.

Class	Training	Validation	Test
Healthy	27	657	657
Viral pneumonia	27	659	659
COVID-19	24	588	588

(a) Chest X-ray image classification.

Class	Training	Validation	Test
Benign tissue	10	2,495	2,495
Adenocarcinoma	10	2,495	2,495
Squamous cell carcinoma	10	2,495	2,495

(b) Lung histopathological image classification.