Self Distillation for Improving the Generalizability of Retinal Disease Diagnosis Using Optical Coherence Tomography Images

Naveen Paluru, *Graduate Student Member, IEEE*, Hariharan Ravishankar, Sharat Hegde, and Phaneendra K. Yalavarthy, *Senior Member, IEEE*

***Abstract*—Optical coherence tomography (OCT) imaging has become a point-of-care imaging modality for the diagnosis of retinal diseases. Varying speckle noise in the OCT images across datasets and scanners worsens the performance of existing artificial intelligence (deep learning) models, that have been trained mostly with images having a particular noise level. The existing deep learning models for predicting retinal diseases are heavy, requires a sophisticated computing environment to train and deploy. Generalized lightweight deep learning models that can provide an automated diagnosis on an edge platform are highly appealing in the clinic. This work proposes a self distillation framework based on lightweight deep learning models for building generalizable deep models for retinal disease diagnosis. The proposed approach with three different baseline models ResNet18, MobileNetV2 and ShuffleNetV2, has been validated on simulated and real-time noisy OCT B-scans spanning a range of SNRs from four OCT datasets. The proposed method significantly outperforms the existing methods with improvement (as high as 14%) in precision, accuracy, and F1-score, to show that the self distillation framework can provide more generalizability for automated retinal diagnosis.**

***Index Terms*—Optical coherence tomography, lightweight CNNs, retinal diseases, knowledge distillation, regularization.**

I. Introduction

Optical Coherence Tomography (OCT) is the gold standard for diagnosing retinal diseases [1]. The most common retinal diseases include Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and DRUSEN. Fig. 1 shows the visually distinctive features of each retinal disease. A small retinal fluid formed near the retinal layer characterizes CNV, DME accounts for the formation of fluid-filled cysts, and Drusen results in irregular retinal boundaries. These distinct features makes computer-aided automated detection/classification of these abnormalities of vital interest, where the aim is to provide this diagnosis in real-time. A volumetric (three-dimensional, 3D) data capture in OCT imaging is the standard. However, interpretation and classification of these 3D OCT images is a tedious and time-consuming task, especially for screening. This diagnosis in an automated fashion at the image (two-dimensional, 2D) level is desirable for disease screening at the population level.

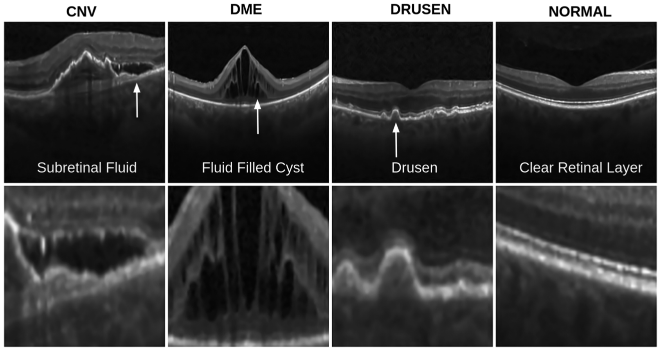


Fig. 1. Example OCT images (first row) from UCSD dataset [2] for each retinal disease given on top of each image correspondingly, with abnormal region shown by an arrow. Zoomed versions of these regions are shown in the second row correspondingly.

Several attempts have been made to fully automate this detection/classification without expert/clinician input. Hussain et al. [3] have proposed a classification algorithm (Random Forest) based on retinal features obtained from spectral-domain optical coherence tomography (SD-OCT) for identifying Age-related Macular Degeneration (AMD) and DME. Guillaume et al. have [4] deployed local binary patterns of SD-OCT images for separating the DME from the normal retinal condition. Albarrak et al. [5] have designed a pipeline that relies on Bayesian decision on the features projected by principal component analysis for age-related macular degeneration (AMD) identification. Fusion of histogram-of-oriented gradients (HoG) and local binary patterns within a multi-scale classification framework for identifying DME have shown exciting results [6]. In a similar study, Srinivasan et al. [7] have designed a support vector machine for discriminating the HoG features of AMD, DME, and the normal retinal condition. Venhuizen et al. [8] have developed

Manuscript received 21 September 2022; revised 21 December 2022 and 20 January 2023; accepted 26 January 2023. Date of publication 2 February 2023; date of current version 10 February 2023. This work was supported by the WIPRO-GE Collaborative Laboratory on Artificial Intelligence in Healthcare and Medical Imaging as well as Qatar National Research Fund (QNRF) under Grant NPRP12S-0228-190182. *(Corresponding author: Phaneendra K. Yalavarthy.)*

Naveen Paluru, Hariharan Ravishankar, and Phaneendra K. Yalavarthy are with the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore 560 012, India (e-mail: naveenp@iisc.ac.in; hariharanrav@iisc.ac.in; phaneendra.k.yalavarthy@gmail.com).

Sharat Hegde is with the Retina Services, Prasad Netralaya, Udipi 576101, India (e-mail: hegdesharath@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSTQE.2023.3240729.

Digital Object Identifier 10.1109/JSTQE.2023.3240729

an unsupervised technique based on bag of visual words and random forest classifier for identifying AMD. Lemaître et al. [4] have proposed an SVM-based classifier for DME identification; however, this algorithm did not report any results associated with AMD, which is known as the most dominant retinal disease in OCT images. 3D volume-based classifiers have also been designed [9] for retinal abnormality classification, but all volumes were constrained to have the same number of B-scans.

Convolutional Neural Networks (CNNs) based retinal disease diagnosis [2], [10], [11], [12], [13] and retinal image enhancements [14], [15], [16], [17] using OCT images have been explored in literature. Lu et al. [10] have proposed a deep neural network, specifically ResNet101 based automated retinal diagnosis using OCT images. Tan et al. [11] have explored deep learning driven fundus image based macular degeneration detection. Kermany et al. [2] have introduced a transfer learning based deep model for retinal diagnosis and a popular benchmarking OCT dataset with $\sim$ 109300 labeled B-scans. Li et al. [18] have proposed a VGG-16 [19] based deep network for retinal disease diagnosis from OCT images. The VGG-16 architecture has $\sim$ 138M parameters and is often known as one of the over-parameterized models for classical 1000 class Imagenet classification. A multi-scale ensemble of CNN's [20] has shown promising results in retinal diagnosis. This approach performs volume-based normalization, followed by ROI and VOI identification, and then an ensemble decision is made. Lesion Aware CNNs (LACNN) [21] with added segmentation and attention modules have improved the predictions of the deep model for classifying OCT B-scans. However, the cost of annotating the segmentation masks has been a bottleneck for re-training the model on another dataset. Further, the segmentation module adds more computation with increased model parameters, size, and inference time. Li et al. [22] have introduced an ensemble of deep residual networks for retinopathy classification. In short, the authors [22] have proposed an ensemble (in total four) of dilated convolutions based ResNet50 architecture for retinal disease detection. Again, ResNet50 is a heavy model with $\sim$ 23M parameters, making it unsuitable for point-of-care or resource-constrained settings. Semi-supervised methods [23], [24] have shown the way for annotation efficient deep learning methods for retinal disease diagnosis. A deep domain adaptation method [25] has improved the generalizability of retinal disease predictions on cross datasets. This work [25] have adapted the adversarial discriminative unsupervised domain adaptation approach [26] and added an entropy minimization module for building robust deep models for retinal diagnosis.

Most existing deep learning based retinal disease diagnosis methods are biased toward the training data distribution. For example, work of Luo et al. [25] focused on building generalizable deep models for retinal diagnosis. However, the authors [25] have used VGG-16 as a baseline model, which is not ideal for point-of-care or resource-constrained settings due to the model being heavy. One of the critical challenge for building such generalizable models via domain adaptation is the availability of the datasets from two different centers. Availing the data sources from multiple centers is a challenge, especially in medical image analysis, due to data privacy and security issues. To address this issue, one solution is to approach source-free domain adaptation techniques, where only the model (weights or parameters) trained on the source dataset is available along with an unannotated target dataset. This approach increases the complexity of the problem due to this additional step, and it is challenging to build a generalizable representation across two different distributions using only the model trained with the source dataset. The other approach is to go for federated deep learning [27] where the data privacy can be maintained, and a global model can be trained on the data from two different data centers. However, the standard federated learning methods assume that the data is labeled across two data centers, and often, the learned global model's generalizability can be poor compared to domain adaptation methods.

The source for the lack of generalizabilty in the retinal diagnosis using OCT images can be attributed to the speckle noise present in the OCT images [28], [29]. The noise modeling in OCT has been studied in detail in the literature. A stretched exponential distribution [30] has been proposed to model the distribution of intensities in OCT images. In a similar study, Amini et al. [31] proposed a mixture of normal-Laplace distribution model to enhance the OCT image contrast. The methods proposed in [32], [33], [34] used Gaussianization transforms to model the OCT image content for applications in denoising. Sudeep et al. [35] developed a Gamma prior for modeling the multi-frame OCT data and applied the designed priors to an iterative framework for denoising. Li et al. [36] designed the OCT image content from local statistics and proposed a maximum a-posteriori estimate for denoising. In a recent study, Tajmirriahi et al. [37] developed a stochastic differential equation based model and used it as a prior to denoise the OCT images.

The speckle noise is more dominant in SD-OCT, which is the most widely available OCT variant due to cost-effectiveness. As OCT imaging is making strides towards the point-of-care diagnosis of retinal diseases, highly generalized deep learning models that can provide an automated diagnosis on an edge platform will be highly appealing in the clinic, especially in resource-constrained settings. The potential of the lightweight models has been proven to a reasonable extent in deep learning based medical image analysis literature [13], [38], [39], [40], [41].

This work proposes noise regularized lightweight deep learning models trained via self distillation for improving the deployability and the generalizability of automated retinal diagnosis using OCT images. The high-speed OCT acquisition results in a poor signal-to-noise ratio (SNR) because of the induced speckle. To generate a high SNR OCT B-scan, multiple acquisitions are taken at the same axial position, and the final scan is an average of registered multiple acquisitions [42]. The multiple acquisitions result in increased scan time, and the increased scan time results in motion artifacts. This work presents a carefully engineered data-dependent noise regularized self-distilled loss function for deep learning based retinal diseases classification that can work on low SNR OCT images.

The proposed framework has distinct advantages compared to its counterparts and provides solution to three immediate

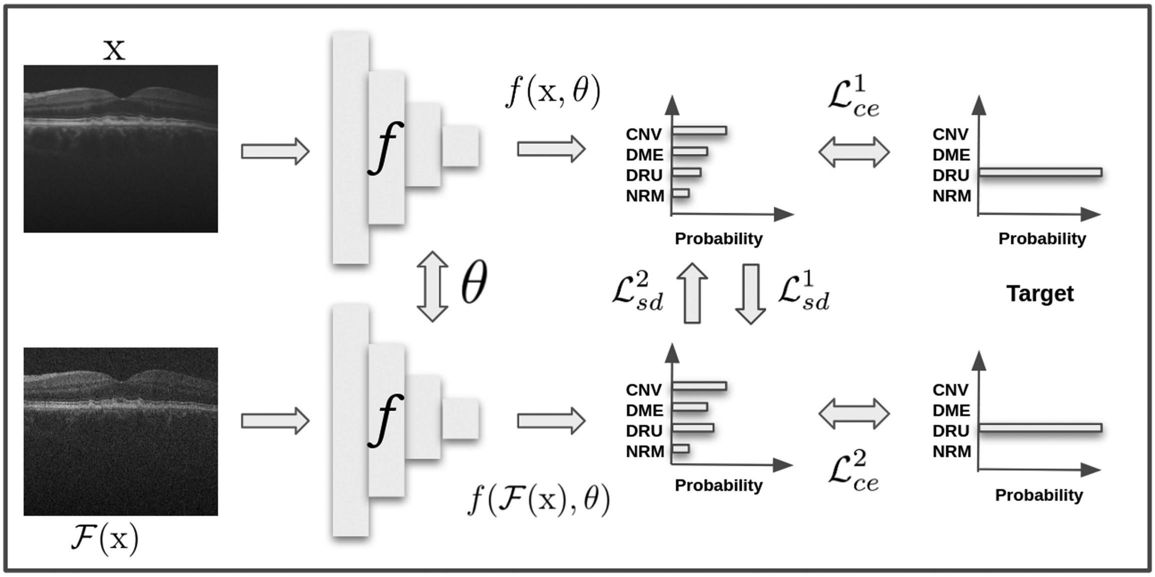


Fig. 2. Proposed NRSD (noise regularized self distillation) for robust retinal disease diagnosis. The lightweight model f is allowed to make predictions on x and the corresponding low SNR counterpart $\mathcal{F}(\mathrm{x})$. To have a consistent prediction irrespective of the SNR level, both these predictions are matched by minimizing KL divergence.

issues with deep learning based retinal diagnosis: (1) easy deployment (using existing lightweight models) for a point-of-care and resource constraint settings, (2) robustness towards variable SNR levels for retinal diagnosis during high-speed OCT acquisition, (3) The proposed method is agnostic to the noise models in OCT based retinal diagnosis and is computationally and quantitatively optimal than using the denoising algorithms followed by classification models for retinal disease diagnosis. The proposed method and the corresponding baseline models have been also systematically evaluated on four SD-OCT datasets. In short, the novelty of this work can be summarized as follows: Development of regularized lightweight models for robust retinal diagnosis via self distillation. Task tailored data-dependent noise regularized cost function has been proposed to provide better generalizability for the task at hand especially for building above mentioned lightweight models. This work also shows that irrespective of the SNR level, the proposed approach of building lightweight deep models for retinal diagnosis can accurately identify the region of interest for predicting the retinal abnormalities, providing better explainability of the proposed framework.

II. Methods

Hinton et al. [43] introduced the concept of knowledge distillation (KD) using soft targets for transferring knowledge between the models. The soft targets have been used as regularizers for building generalizable models [43], [44], [45]. The OCT images are corrupted by speckle noise arising out of interaction of multiple scattering sources during signal recording. SNR of the OCT image is one of the essential characteristic features that varies among the scans obtained from different OCT machines. This work proposes a data-dependent noise regularized self-distilled cost function to handle the SNR variability during the retinal diagnosis using OCT scans.

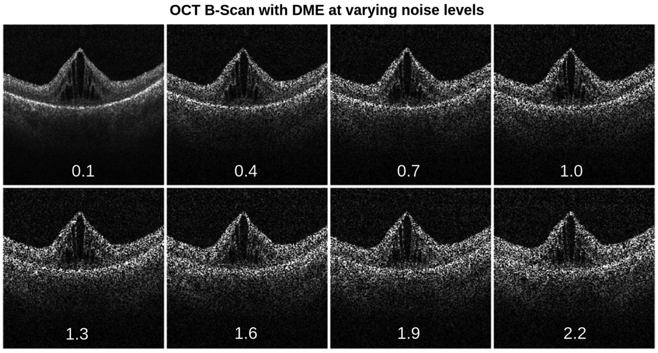


Fig. 3. Example OCT image with DME from UCSD dataset [2] at varying noise levels. The noise level (variance of the data-dependent/multiplicative noise) is shown at the bottom of the corresponding image. The SNR of the OCT image decreases with increase in the variance of the noise level.

The details of steps involved in the proposed approach are shown in Fig. 2. Given a mini-batch of high SNR (clean) B-scan images $\{\mathrm{x}_i\}_{i=1}^{N}$ along with corresponding expert annotations $\{y_i\}_{i=1}^{N}$ ($y_i \in \{1, 2, .., C\}$), a lightweight model f parameterized by θ is trained using the classical cross-entropy as shown in (2). Note that C is the total number of classes, y_{ij} is one hot representation for y_i and $\hat{\mathrm{y}}_{ij}^{c}$ denotes softmax probabilities (1) with f_j being the logit of the lightweight model for j^{th} class. Given a OCT B-scan x_i, its SNR is altered by adding a simulated data-dependent noise as shown in (3), where (a, b) denotes the spatial coordinates, η represents uniformly distributed random noise with 0 mean and variance v. Fig. 3 shows a sample OCT B-scan with varying noise levels. As shown in Fig. 2, the lightweight model f makes a prediction on x_i and the corresponding low SNR counterpart $\mathcal{F}(\mathrm{x}_i)$.

$$\hat{\mathrm{y}}_{ij}^{c} = \frac{\exp\left(f_j(\mathrm{x}_i, \theta)\right)}{\sum_{j=1}^{C} \exp\left(f_j(\mathrm{x}_i, \theta)\right)} \tag{1}$$

$$\mathcal{L}_{ce}^{1} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \mathrm{y}_{ij} \log \hat{\mathrm{y}}_{ij}^{c} \tag{2}$$

$$\mathcal{F}\left(\mathrm{x}_i(a,b)\right) = \mathrm{x}_i(a,b) + \eta(a,b) * \mathrm{x}_i(a,b) \tag{3}$$

$$\hat{\mathrm{y}}_{ij}^{n} = \frac{\exp\left(f_j(\mathcal{F}(\mathrm{x}_i), \theta)\right)}{\sum_{j=1}^{C} \exp\left(f_j(\mathcal{F}(\mathrm{x}_i), \theta)\right)} \tag{4}$$

$$\mathcal{L}_{ce}^{2} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \mathrm{y}_{ij} \log \hat{\mathrm{y}}_{ij}^{n} \tag{5}$$

The cross entropy loss has been enforced for the model predictions on noisy B-scans (5) as well, where $\hat{\mathrm{y}}_{ij}^{n}$ denotes softmax probabilities (4) for a noisy B-scan with f_j being the logit of the lightweight model for j^{th} class. To have a consistent prediction irrespective of the SNR level, the KL divergence between the soft predictions $\hat{\mathbf{y}}_{ij}^{c}$ (6) and $\hat{\mathbf{y}}_{ij}^{n}$ (7) is being minimized (8), (9) via deep mutual learning [46], [47]. This forms the critical step of self-distillation to provide better generalizability. As KL-divergence provides a statistical measure of information loss between distributions (original and noisy), the convergence in this case is much stronger in measures of information, thus leading to better self-distillation. Note that, (6) and (7) denote the temperature scaled soft predictions of f on the B-scan x_i and the noisy B-scan $\mathcal{F}(\mathrm{x}_i)$ with T being a temperature scaling parameter.

$$\hat{\mathbf{y}}_{ij}^{c} = \frac{\exp\left(f_j(\mathrm{x}_i, \theta)/T\right)}{\sum_{j=1}^{C} \exp\left(f_j(\mathrm{x}_i, \theta)/T\right)} \tag{6}$$

$$\hat{\mathbf{y}}_{ij}^{n} = \frac{\exp\left(f_j(\mathcal{F}(\mathrm{x}_i), \theta)/T\right)}{\sum_{j=1}^{C} \exp\left(f_j(\mathcal{F}(\mathrm{x}_i), \theta)/T\right)} \tag{7}$$

$$\mathcal{L}_{sd}^{1} = \frac{T^2}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \hat{\mathbf{y}}_{ij}^{c} \log \frac{\hat{\mathbf{y}}_{ij}^{c}}{\hat{\mathbf{y}}_{ij}^{n}} \tag{8}$$

$$\mathcal{L}_{sd}^{2} = \frac{T^2}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \hat{\mathbf{y}}_{ij}^{n} \log \frac{\hat{\mathbf{y}}_{ij}^{n}}{\hat{\mathbf{y}}_{ij}^{c}} \tag{9}$$

Finally, the total loss for optimizing the model parameters is defined in (10). The self-distilled cost function ensures that the lightweight model learns a generic and discriminative feature representation of OCT B-scans. $\mathcal{L}_{ce}$ is the classical cross-entropy ($\mathcal{L}_{ce}^{1} + \mathcal{L}_{ce}^{2}$), and $\mathcal{L}_{sd}$ is the data-dependent noise regularized cost function minimizing the KL divergence ($\mathcal{L}_{sd}^{1} + \mathcal{L}_{sd}^{2}$) enforcing a consistent prediction irrespective of the noise level. Note that λ is the weight for the data-dependent noise regularized terms in the cost function. The original knowledge distillation approach [43] matches the predictions on a single image from two networks. Yun et al. [44] proposed a self-distillation framework to match the predictions of two different images belonging to same class. In contrast, this work matches the predictions on a high SNR B-scan and low SNR B-scan from a single network, i.e., self-distillation. In short, the proposed approach can be written as noise regularized self distillation (NRSD).

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda\, \mathcal{L}_{sd} \tag{10}$$

TABLE I
DETAILS OF THE DATASETS UTILIZED IN THIS WORK

Dataset	Patients	CNV	AMD	DME	DRUSEN	NORMAL
UCSD	5420	37455	-	11598	8866	51390
DHU	45	-	723	1101	-	1407
NEH	148	-	1428	1086	-	1577

The numbers provided against each class are the total number of 2D images and the corresponding total number of patients in each dataset are listed in the second column.

III. EXPERIMENTS

A. Datasets

In this study, four publicly available datasets were considered to show the efficacy of the proposed approach. Specifically, the University of California San Diego (UCSD) dataset [2], the Noor Eye Hospital (NEH) dataset [20], the DHU dataset [7] and the noisy B-scans dataset [48], [49]. For the noisy B-scans dataset, the labels (AMD or Normal) were marked by an Ophthalmologist. The summary of these datasets was provided in Table I. The OCT B-scans from UCSD dataset [2], DHU dataset [7] and, NEH dataset [20] were captured using the Heidelberg Spectralis imaging system. The noisy OCT B-scans used from the dataset [48], [49] were captured using Bioptigen, Inc. imaging system. The UCSD dataset has ~ 109000 OCT B-scans captured from ~ 5420 subjects. The DHU dataset has 45 OCT volumes (15 each for AMD, DME, and, Normal). The axial resolution was 3.85μm; the lateral resolution ranged from 6-12 μm; the number of A-scans ranged from 512-1024, and 31-97 B-scans were acquired from different patients. The NEH database has 148 OCT volumes (48 for AMD, 50 for DME, and 50 for Normal). The axial resolution was 3.5μm. The lateral and azimuthal resolutions were not uniform across patients, and the number of A-scans ranged from 512-768, with 19-61 B-scans obtained from various patients. The dataset [48], [49] has 28 OCT B-scans collected from 28 subjects with and without AMD. The axial resolution was 4.5μm with 1000 A-scans per B-scan.

B. Lightweight Models

Three lightweight models namely ResNet18 [50], MobileNetV2 [51] and ShuffleNetV2 [52] were utilized for validating the proposed noise regularized self distillation approach for robust retinal diagnosis. The fully connected layers of these networks were modified to provide a four class classification for UCSD dataset and three class classification for NEH and DHU datasets respectively.

C. Baseline Models

For a fair evaluation of the proposed noise regularized self distillation method, the standard noise regularized (training lightweight models with noisy B-scans) results were also presented. In short, the baseline models were ResNet18 (denoted as R18); standard noise regularized R18 (denoted as R18+Noise); MobileNetV2 (denoted as MV2) and standard noise regularized MV2 (denoted as MV2+Noise); ShuffleNetV2 (denoted as SV2) and standard noise regularized SV2 (denoted as SV2+Noise). Note that the proposed lightweight networks were denoted as R18+NRSD, MV2+NRSD and SV2+NRSD.

D. Experimental Studies

1) Study 1: The OCT B-scans used in this study were taken from the UCSD dataset. The UCSD dataset consists of ~ 109000 B-scans. Table I shows the split among the retinal abnormalities. The dataset was split at the patient level into 70% training, 15% validation, and 15% testing (with $\sim$ 14500 OCT B-scans). As shown in (6), the B-scans with different data-dependent noise levels were simulated with variance $v \in [0.1, 2.2]$ with a step size of 0.3. Note that, during training of the proposed approach and the baseline (standard noise regularized) approaches, the noise levels considered were $[0.7, 1.0, 1.3]$. All lightweight models in this study were trained three times with different random seeds, and the quantitative metrics on the testing dataset were reported as mean ± standard deviation. For the retinal diagnosis of real time noisy OCT images (AMD or Normal, present in the dataset [48], [49]), the above mentioned lightweight models were fine tuned on NEH dataset (for the classes AMD, DME, and Normal). Finally, the proposed approach and equivalent baselines were evaluated on real time noisy B-scans (low SNR) and corresponding clean B-scans (high SNR) taken from an example dataset that has been widely used for bench-marking denoising algorithms of OCT images [48], [49].

2) Study 2: The OCT B-scans used in this study were taken from the NEH dataset. The NEH dataset consists of ~ 4000 B-scans captured from 148 patients. Table I shows the split (three categories) among the retinal abnormalities. The dataset was split at the patient level into 60% training, 20% validation, and 20% testing (with 995 B-scans). The pre-trained weights (till the fully connected layer) of the lightweight models trained in study 1 were used to initiate the transfer learning for lightweight models in this study. Note that, during training of the proposed approach and the baseline (standard noise regularized) approaches, the noise levels considered were $[0.09, 0.1, 0.11]$. All lightweight models in this study were also trained three times with different random seeds, and the quantitative metrics on the testing dataset were also reported as mean ± standard deviation. In this study, generalizability of the lightweight models was evaluated on unseen dataset (during training) by quantifying them on the DHU dataset.

3) Study 3: The OCT B-scans used in this study were taken from the DHU dataset. The DHU dataset consists of ~ 3200 B-scans captured from 45 patients. Table I shows the split (3 categories) among the retinal abnormalities. The dataset was split at the patient level into 60% training, 20% validation, and 20% testing (with 816 B-scans). The pre-trained weights (till the fully connected layer) of the lightweight models trained in study 1 were utilized to initiate the transfer learning for lightweight models in this study. Note that, during training of the proposed approach and the baseline (standard noise regularized) approaches, the noise levels considered were $[0.09, 0.1, 0.11]$. As with earlier studies, the lightweight models in this study were also trained three times with different random seeds, and the quantitative metrics on the testing dataset were reported as mean ± standard deviation. The generalizability of the lightweight models was evaluated on unseen dataset (during training) by quantifying them on the NEH dataset.

E. Implementation

The lightweight models considered in this work were trained using the PyTorch framework [53]. The model parameters in all studies were optimized using the Adam [54] optimizer along with a weight decay factor of $1e^{-4}$. In study 1, the initial learning rate was $5e^{-4}$ and was decayed by a factor of 0.1 once training completes for 50% and 75% of total epochs. The batch size in study 1 was 32. For the studies 2 and 3, the batch size was 64. The initial learning rate for these experiments was $5e^{-5}$ and was decayed by a factor of 0.1 once training completes for 50% and 75% of total epochs. The total number of epochs was fixed at 30 across all experiments. The values for λ and T were chosen as 1.0 and 4.0 respectively, and these values were consistent across all studies. To quantify the lightweight models performance, the figures of merit like F1-score, precision, and accuracy were computed in all experiments. All computations performed in this work, including training of lightweight models, utilized a Linux workstation with i9 9900X (CPU) with 128 GB RAM and two NVIDIA Quadro RTX 8000 GPU card having a memory capacity of 48˜GB. The metric that was used across all experiments was the F1-Score, which is the harmonic mean of precision and recall. It provides a better measure of the misdiagnosed cases than the precision and accuracy metrics.

IV. Results

The performance metrics from study 1 were detailed in Table II. Note that, during training of the proposed approach and the baseline (standard noise regularized) approaches, the noise levels considered were $[0.7, 1.0, 1.3]$. As shown in Table II, across all noise levels, the proposed methods R18+NRSD and MV2+NRSD outperformed the standard noise regularized models R18+Noise and MV2+Noise. Specifically, MV2+NRSD outperformed MV2+Noise with a margin of $\sim 3\%$ for the noise levels $[1.6, 1.9, 2.2]$. Similarly, R18+NRSD outperformed R18+Noise with a maximum margin of 6% for the noise levels $[1.3, 1.6, 1.9]$. The comparison between SV2+Noise and SV2+NRSD also followed a similar trend. The precision and accuracy scores were shown in Fig. 4. The proposed approach also showed a consistent improvement in terms of these metrics. Further, the proposed methods had minor deviation (refer to Table II) across multiple runs with different random seeds indicating the stability of lightweight models for varying noise levels of the input B-scans. The class activation maps were shown in Fig. 5. The columns show the overlayed gradients on the input images across different noise levels (column-wise),

TABLE II
AVERAGED F1-SCORES (MEAN ± STD) FROM THREE INDEPENDENT RUNS (WITH DIFFERENT RANDOM SEEDS) ON ~ 14500 TEST OCT B-SCANS FROM UCSD DATASET AT VARYING NOISE LEVELS

Noise Level	Original	(0.1)	(0.4)	(0.7)	(1.0)	(1.3)	(1.6)	(1.9)	(2.2)
Model | Metric	F1-score	F1-score	F1-score	F1-score	F1-score	F1-score	F1-score	F1-score	F1-score
R18	**0.95 ± 0.01**	0.92 ± 0.02	0.67 ± 0.10	0.57 ± 0.10	0.45 ± 0.08	0.38 ± 0.07	0.33 ± 0.06	0.29 ± 0.04	0.27 ± 0.04
R18+Noise	0.94 ± 0.01	0.93 ± 0.01	0.88 ± 0.02	0.84 ± 0.03	0.77 ± 0.04	0.70 ± 0.05	0.63 ± 0.06	0.56 ± 0.05	0.51 ± 0.06
R18+NRSD (Proposed)	0.94 ± 0.01	**0.94 ± 0.01**	**0.91 ± 0.01**	**0.89 ± 0.02**	**0.83 ± 0.02**	**0.76 ± 0.03**	**0.69 ± 0.05**	**0.63 ± 0.05**	**0.56 ± 0.04**
MV2	0.94 ± 0.01	0.92 ± 0.01	0.69 ± 0.04	0.59 ± 0.06	0.48 ± 0.06	0.40 ± 0.06	0.35 ± 0.06	0.30 ± 0.05	0.29 ± 0.05
MV2+Noise	**0.94 ± 0.01**	**0.94 ± 0.01**	0.89 ± 0.01	0.85 ± 0.02	0.79 ± 0.04	0.72 ± 0.06	0.66 ± 0.07	0.60 ± 0.08	0.55 ± 0.08
MV2+NRSD (Proposed)	0.94 ± 0.01	0.93 ± 0.01	**0.91 ± 0.01**	**0.88 ± 0.01**	**0.83 ± 0.02**	**0.76 ± 0.03**	**0.70 ± 0.04**	**0.63 ± 0.05**	**0.57 ± 0.05**
SV2	**0.93 ± 0.01**	0.89 ± 0.01	0.58 ± 0.06	0.51 ± 0.07	0.43 ± 0.07	0.39 ± 0.08	0.36 ± 0.08	0.34 ± 0.08	0.33 ± 0.08
SV2+Noise	0.91 ± 0.01	**0.91 ± 0.01**	0.85 ± 0.01	0.82 ± 0.01	0.77 ± 0.02	0.73 ± 0.02	0.69 ± 0.02	0.65 ± 0.02	0.61 ± 0.03
SV2+NRSD (Proposed)	0.91 ± 0.01	0.90 ± 0.01	**0.88 ± 0.01**	**0.86 ± 0.01**	**0.83 ± 0.01**	**0.79 ± 0.01**	**0.74 ± 0.01**	**0.70 ± 0.02**	**0.65 ± 0.03**

The study details were presented in sub-section III.D.1 (Study 1) with R18 : ResNet18, MV2 : MobileNetV2, SV2 : ShuffleNetV2 and NRSD : noise regularized self distillation (proposed approach). The best results are shown in bold.

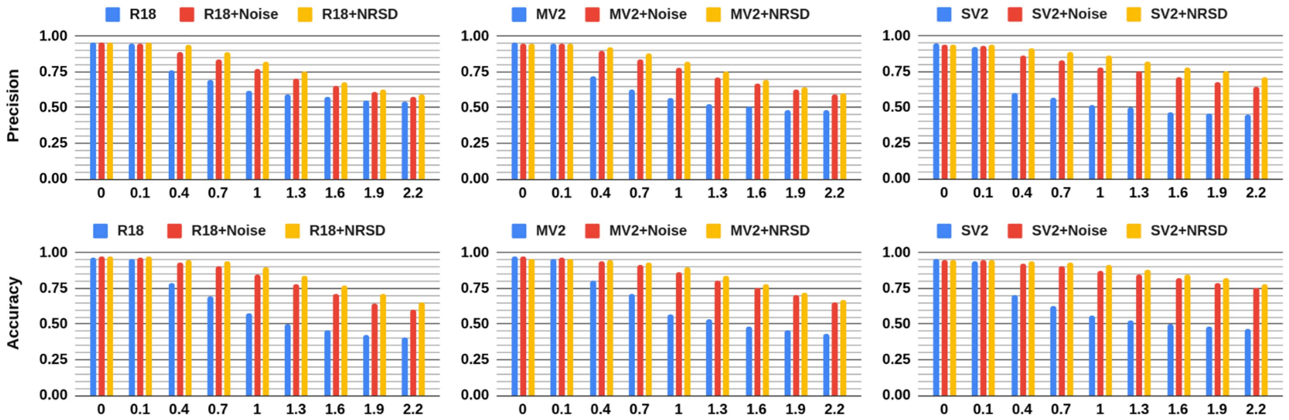


Fig. 4. Averaged precision and accuracy from three independent runs (with different random seeds) on ~ 14500 test OCT B-scans from the UCSD dataset as a function of noise level (X-axis). The study details were presented in Section III.D.1 (Study 1) with R18 : ResNet-18, MV2 : MobileNetV2, SV2 : ShuffleNetV2 and NRSD : noise regularized self distillation (proposed approach).

precisely the Grad-CAM [55] visualizations. The confidence score (softmax probability) of the lightweight models was given at the bottom of the corresponding image. The scores of failure cases (wrong predictions) were shown in the red color. As shown in Fig. 5, irrespective of the SNR level of the B-scan, the heatmaps of the proposed NRSD approach were more relevant to the region of interest (refer to DME and Drusen in Fig. 5) than the standard noise regularized counterparts. This can be attributed to the noise regularized self-distilled component of the loss function. The performance metrics on the real time noisy and corresponding clean B-scans were shown in Table III. The proposed method (deployed on several lightweight models) outperformed the equivalent baselines in predicting retinal diseases from real time noisy B-scans. The same was evident from the confidence scores detailed in Table III, also sample real time noisy B-scans and corresponding clean B-scans along with the confidence scores (softmax probabilities) were shown in Fig. 6. The SNR (11) shown in Fig. 6 is defined as the ratio of mean of the intensities in foreground region (μ_f) to the standard deviation of the intensities in foreground region (σ_f). Further, the final SNR was computed as an average of all SNRs from foreground region of interests selected in the OCT B-scan [16]. The foreground regions were illustrated using bounding boxes as shown in Fig. 6.

$$SNR = \frac{\mu_f}{\sigma_f} \tag{11}$$

TABLE III
ACCURACY AND CONFIDENCE SCORES ON REAL TIME NOISY (LOW SNR) AND CLEAN (HIGH SNR) B-SCANS FROM THE DATASET [48]

	Noisy B-scans (Low SNR)		Clean B-scans (High SNR)	
Model | Metric	Correct/Total	Confidence	Correct/Total	Confidence
R18	22/28	0.86 ± 0.17	22/28	0.84 ± 0.18
R18+Noise	24/28	0.87 ± 0.03	20/28	0.92 ± 0.10
R18+NRSD (Proposed)	25/28	0.88 ± 0.03	21/28	0.92 ± 0.05
MV2	18/28	0.99 ± 0.02	19/28	0.98 ± 0.02
MV2+Noise	25/28	0.71 ± 0.14	26/28	0.76 ± 0.14
MV2+NRSD (Proposed)	26/28	0.82 ± 0.17	25/28	0.83 ± 0.17
SV2	22/28	0.79 ± 0.14	23/28	0.81 ± 0.19
SV2+Noise	25/28	0.78 ± 0.03	23/28	0.87 ± 0.08
SV2+NRSD (Proposed)	24/28	0.80 ± 0.07	25/28	0.81 ± 0.09

The study details were presented in sub-section III.D.1 (Study 1) with R18 : ResNet18, MV2 : MobileNetV2, SV2 : ShuffleNetV2 and NRSD : noise regularized self distillation. Example real time noisy B-scans and the clean B-scans along with the confidence scores were shown in Fig. 6.

TABLE IV
QUANTITATIVE METRICS (MEAN ± STD) ON 995 TEST OCT B-SCANS FROM THE NEH DATASET AND 3231 OCT B - SCANS FROM THE DHU DATASET

Dataset	**NEH**			**DHU**		
Model | Metric	**Precision**	**F1-score**	**Accuracy**	**Precision**	**F1-score**	**Accuracy**
R18	**0.92 ± 0.01**	**0.91 ± 0.02**	**0.91 ± 0.01**	0.79 ± 0.02	0.67 ± 0.02	0.70 ± 0.02
R18+Noise	0.91 ± 0.01	**0.91 ± 0.01**	0.90 ± 0.01	0.80 ± 0.06	0.70 ± 0.09	0.73 ± 0.07
R18+NRSD (Proposed)	0.91 ± 0.01	**0.91 ± 0.01**	**0.91 ± 0.01**	**0.86 ± 0.05**	**0.84 ± 0.06**	**0.84 ± 0.06**
MV2	0.90 ± 0.01	0.90 ± 0.01	0.90 ± 0.01	0.84 ± 0.02	0.69 ± 0.07	0.71 ± 0.06
MV2+Noise	**0.91 ± 0.01**	**0.91 ± 0.01**	**0.91 ± 0.01**	0.84 ± 0.03	0.82 ± 0.04	0.82 ± 0.04
MV2+NRSD (Proposed)	**0.91 ± 0.02**	**0.91 ± 0.02**	0.90 ± 0.01	**0.85 ± 0.02**	**0.83 ± 0.01**	**0.84 ± 0.02**
SV2	0.88 ± 0.01	0.88 ± 0.01	0.88 ± 0.01	0.81 ± 0.07	0.73 ± 0.06	0.74 ± 0.05
SV2+Noise	**0.91 ± 0.02**	**0.91 ± 0.02**	**0.90 ± 0.02**	0.82 ± 0.02	0.77 ± 0.01	0.78 ± 0.00
SV2+NRSD (Proposed)	0.87 ± 0.01	0.87 ± 0.01	0.86 ± 0.01	**0.84 ± 0.01**	**0.81 ± 0.01**	**0.81 ± 0.01**

Models were trained using ~60% of OCT B-scans (across three independent runs with different random seeds) from the NEH dataset. The DHU dataset was completely unseen to the models under this experiment. The study details were presented in sub-section III.D.2 (Study 2) with R18 : ResNet18, MV2 : MobileNetV2, SV2 : ShuffleNetV2 and NRSD : noise regularized self distillation (proposed approach). The best results are shown in bold.

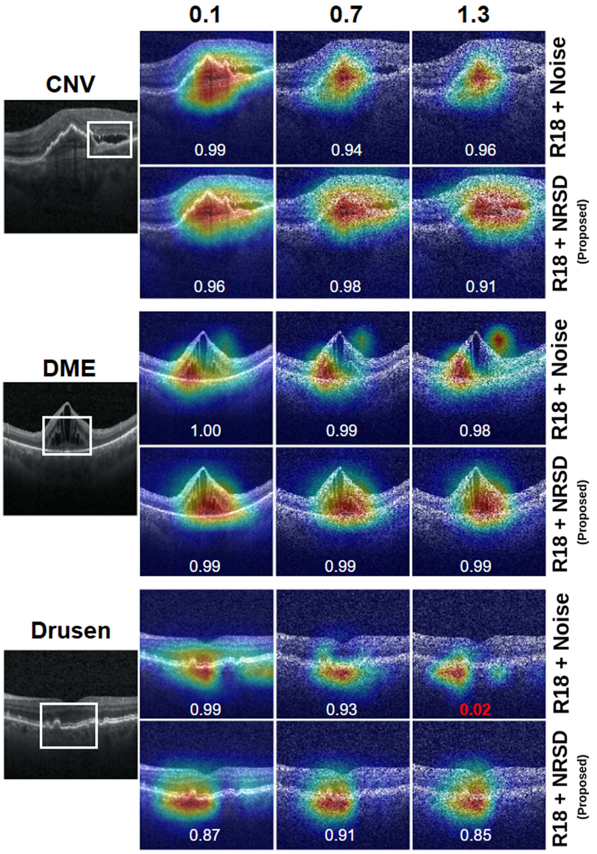
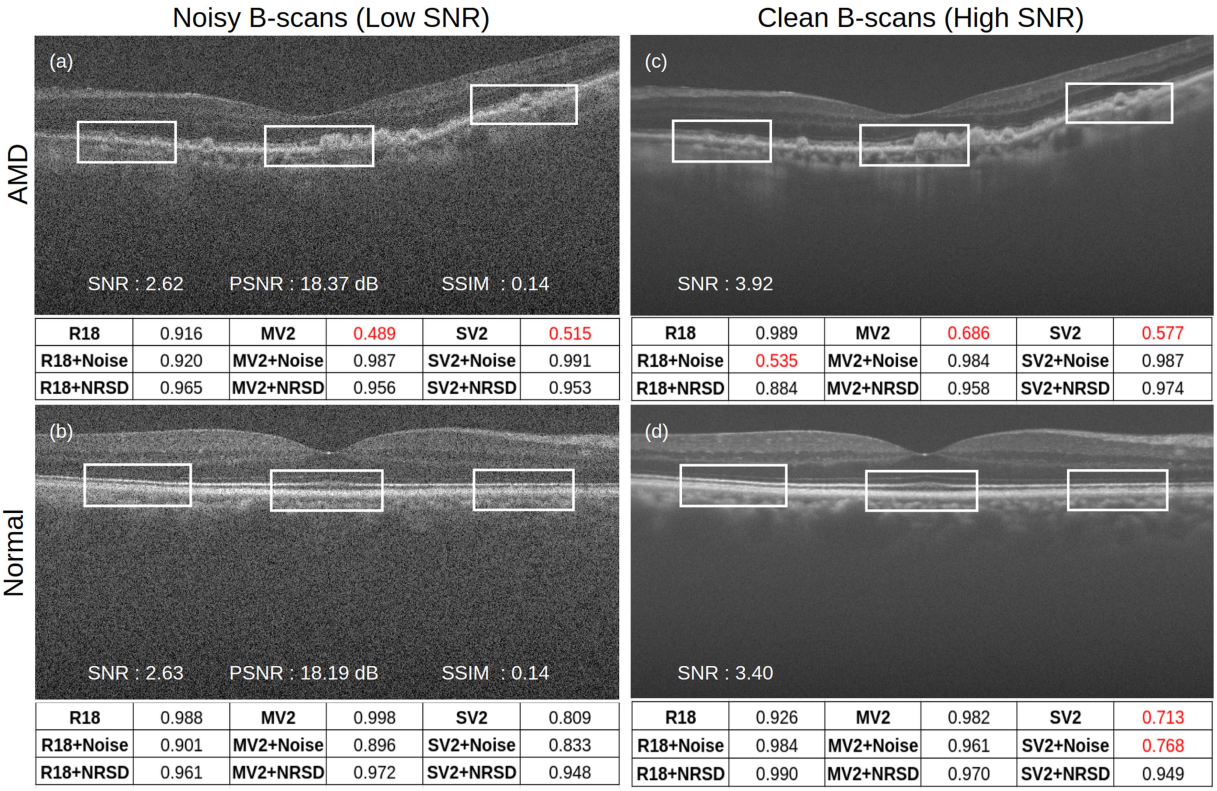


Fig. 5. Example OCT images (first column) from UCSD dataset for each class of retinal disease, with abnormal region marked by a bounding box. The subsequent columns show the overlayed gradients on the input images across different noise levels (column wise), precisely the Grad-CAM visualizations for the example ResNet18 model. The study details were presented in Sub-Section III.D.1 (Study 1). The corresponding confidence score of the utilized model was shown at the bottom of the image. The confidence scores of failure cases (wrong predictions) are shown in red color.

The quantified metrics for study 2 were detailed in Table IV. As mentioned earlier, the models were trained on a subset of the NEH dataset and were tested on unseen leftover B-scans from the NEH dataset. The models were also tested on the DHU dataset, which was entirely unseen during training and this dataset has been curated from a different center. As shown in Table IV, the performance of all lightweight models considered was comparable on the NEH dataset. However, while evaluating the generalizability of the lightweight models on the DHU dataset (different site), R18+NRSD outperformed R18+Noise with a margin of 6% in precision, 14% in F1-score, and 11% in terms of accuracy. Also, the deviation from multiple runs was minimal for the proposed NRSD approach. Similarly, with MV2, the proposed approach improved 1% in precision, 1% in F1-score, and 2% in accuracy with minimum deviation across multiple runs. A similar trend was observed for the models SV2+Noise and SV2+NRSD.

The quantified metrics for study 3 were presented in Table V. As mentioned earlier, the models were trained on a subset of the DHU dataset and were tested on unseen leftover B-scans from the DHU dataset. The models were also tested on the NEH dataset, which was unseen during training and was obtained from a different center. As shown in Table V, the performance of all lightweight models considered was comparable on the DHU dataset. However, while evaluating the generalizability of the lightweight models on the NEH dataset, R18+NRSD outperformed R18+Noise with a margin of 9% in precision, 9% in F1-score, and 7% in terms of accuracy. Also, the deviation from multiple runs was a bit high for the proposed NRSD approach. Similarly, with MV2, the proposed approach improved 5% in accuracy and 6% in F1-score with minimum deviation across multiple runs. The NEH dataset has poor SNR compared to the DHU dataset. As a result, the performance of deep models, including the proposed NRSD approach was low. However, the proposed NRSD approach was more consistent than the baseline models. In experimental studies 2 and 3, the SV2+Noise outperformed SV2+NRSD with a margin of 2%. However, SV2+NRSD was more competent in F1-score and accuracy. Further, in Table VI the proposed approach was compared with existing deep models for retinal diagnosis. The proposed method R18+NRSD outperformed Lu et al. [10] with 22%, Sunija et al. [13] with 7% and Li et al. [18] with 8% in terms of F1-score. The over-parameterized deep models (refer to Table VI) not only increased the computational burden, but



R18	0.916	MV2	0.489	SV2	0.515
R18+Noise	0.920	MV2+Noise	0.987	SV2+Noise	0.991
R18+NRSD	0.965	MV2+NRSD	0.956	SV2+NRSD	0.953

R18	0.989	MV2	0.686	SV2	0.577
R18+Noise	0.535	MV2+Noise	0.984	SV2+Noise	0.987
R18+NRSD	0.884	MV2+NRSD	0.958	SV2+NRSD	0.974

R18	0.988	MV2	0.998	SV2	0.809
R18+Noise	0.901	MV2+Noise	0.896	SV2+Noise	0.833
R18+NRSD	0.961	MV2+NRSD	0.972	SV2+NRSD	0.948

R18	0.926	MV2	0.982	SV2	0.713
R18+Noise	0.984	MV2+Noise	0.961	SV2+Noise	0.768
R18+NRSD	0.990	MV2+NRSD	0.970	SV2+NRSD	0.949

Fig. 6. Performance of the lightweight models on noisy (low SNR) OCT images (a,b) and corresponding high SNR (clean) OCT images (c,d) taken from the dataset [48], [49] with confidence scores (softmax probabilities) of the models shown at the bottom of the image. The confidence scores shown in red color (less than 0.8) indicate that the model performance is subpar. The study details were presented in Sub-Section III.D.1 (Study 1) with R18 : ResNet-18, MV2 : MobileNetV2, SV2 : ShuffleNetV2 and NRSD : noise regularized self distillation (proposed approach). The foreground regions were illustrated as white boxes and the metrics SNR, PSNR and, SSIM were reported for noisy B-scans and SNR was reported for corresponding clean B-scans.

TABLE V
QUANTITATIVE METRICS (MEAN ± STD) ON 816 TEST OCT B-SCANS FROM THE DHU DATASET AND 4091 OCT B - SCANS FROM THE NEH DATASET

Dataset	DHU			NEH		
Model | Metric	Precision	F1-score	Accuracy	Precision	F1-score	Accuracy
R18	**0.97 ± 0.01**	**0.96 ± 0.01**	**0.95 ± 0.02**	0.58 ± 0.03	0.52 ± 0.05	0.54 ± 0.04
R18+Noise	0.95 ± 0.01	0.93 ± 0.01	0.92 ± 0.01	0.57 ± 0.02	0.52 ± 0.01	0.54 ± 0.01
R18+NRSD (Proposed)	0.93 ± 0.01	0.89 ± 0.03	0.88 ± 0.03	**0.66 ± 0.07**	**0.61 ± 0.06**	**0.61 ± 0.06**
MV2	**0.97 ± 0.01**	**0.96 ± 0.01**	**0.95 ± 0.01**	0.62 ± 0.04	0.59 ± 0.04	0.59 ± 0.04
MV2+Noise	0.94 ± 0.01	0.92 ± 0.02	0.90 ± 0.01	**0.72 ± 0.03**	0.59 ± 0.06	0.60 ± 0.07
MV2+NRSD (Proposed)	0.95 ± 0.02	0.94 ± 0.02	0.93 ± 0.03	0.71 ± 0.03	**0.65 ± 0.03**	**0.65 ± 0.03**
SV2	**0.95 ± 0.02**	**0.94 ± 0.03**	**0.93 ± 0.03**	0.57 ± 0.03	0.53 ± 0.07	0.54 ± 0.05
SV2+Noise	0.92 ± 0.01	0.87 ± 0.03	0.86 ± 0.03	**0.68 ± 0.01**	0.56 ± 0.04	0.57 ± 0.03
SV2+NRSD (Proposed)	0.92 ± 0.01	0.86 ± 0.01	0.85 ± 0.01	0.65 ± 0.02	**0.59 ± 0.03**	**0.59 ± 0.03**

The models were trained using ~60% of OCT B-scans (across three independent runs with different random seeds) from the DHU dataset. The NEH dataset was completely unseen to the models under this experiment. The study details were presented in sub-section III.D.3 (Study 3) with R18 : ResNet18, MV2 : MobileNetV2, SV2 : ShuffleNetV2 and NRSD : noise regularized self distillation (proposed approach). The best results are shown in bold.

also had poor generalization for small-scale problems like retinal diagnosis.

V. Discussion

Automated retinal disease classification is of crucial interest in Ophthalmology, especially for screening of patients. Earlier techniques relied on the classical handcrafted features for building machine learning methods [3], [5], [8], [9] for retinal diagnosis. The technical advancements in automated feature extraction methods like CNN's have drawn significant attention to build deep learning based methods for retinal diagnosis. The deep learning based methods like Li et al. [18] used VGG-16 (∼ 134M parameters), and Kermany et al. [2] used Inception V3 (∼ 23M

TABLE VI
COMPARISON WITH THE EXISTING DEEP LEARNING BASED RETINAL DISEASE CLASSIFICATION METHODS

Metric|Model	Li et al. [18]	Lu et al. [10]	Sunija et al. [13]	R18+NRSD (Proposed)	MV2+NRSD (Proposed)	SV2+NRSD (Proposed)
Parameters	∼ 134 M	∼ 42.5 M	∼ 1.85 M	∼ 11 M	∼ 2.2 M	∼ 1.2 M
Size (MB)	512.22	162.13	7.07	42.25	8.50	4.79
MACs	∼ 3.95 G	∼ 7.77 G	∼ 1.16 G	∼ 1.74 G	∼ 0.31 G	∼ 0.14 G
Inference (s)	0.19, 0.2	0.18, 0.2	0.17, 0.2	0.17, 0.2	0.18, 0.2	0.19, 0.2
Precision	0.79 ± 0.06	0.74 ± 0.09	0.85 ±0.02	**0.86 ± 0.05**	0.85 ± 0.02	0.84 ± 0.01
F1-score	0.76 ± 0.06	0.62 ± 0.17	0.67 ± 0.11	**0.84 ± 0.06**	0.83 ± 0.01	0.81 ± 0.01
Accuracy	0.76 ± 0.06	0.64 ± 0.17	0.70 ± 0.08	**0.84 ± 0.06**	**0.84 ± 0.02**	0.81 ± 0.01
AUC	0.91 ± 0.01	0.90 ± 0.08	0.88 ± 0.03	**0.95 ± 0.01**	0.94 ± 0.02	0.92 ± 0.02
Precision	0.57 ± 0.03	0.57 ± 0.06	0.65 ± 0.05	0.66 ± 0.07	**0.71 ± 0.03**	0.65 ± 0.02
F1-score	0.50 ± 0.02	0.52 ± 0.05	0.60 ± 0.09	0.61 ± 0.06	**0.65 ± 0.03**	0.59 ± 0.03
Accuracy	0.53 ± 0.03	0.54 ± 0.05	0.61 ± 0.08	0.61 ± 0.06	**0.65 ± 0.03**	0.59 ± 0.03
AUC	0.83 ± 0.01	0.79 ± 0.04	0.80 ± 0.06	0.85 ± 0.03	**0.86 ± 0.02**	0.83 ± 0.01

The models were trained using ~60% of OCT B-scans from the NEH dataset and were tested on the 3231 OCT B-scans from the DHU dataset [Rows 6-9]. Similarly, the models were also trained using ~60% of OCT B-scans from the DHU dataset and were tested on the 4091 OCT B-scans from the NEH dataset [Rows 10-13]. Note that inference (s) reported were for (GPU, CPU). The best results are shown in bold.

parameters) [56] based models for retinal disease classification. These models provided promising classification performance, but at a heavy cost of computation. Due to advancements in lightweight models, the initial deep learning architectures like VGG are now recognized as overparameterized models for the 1000 class ImageNet classification problem. Given the small-scale classification problem of retinal diagnosis from OCT B-scans, using such heavy models is not necessary. Lightweight models (ResNet18 ∼ 11M parameters, MobileNetV2 ∼ 2.2M parameters, and ShuffleNetV2 ∼ 1.2M parameters), which were used in this study are useful for addressing overparameterization, model size, and heavy cost computation issues. The use of these lightweight networks, which are easy to train from scratch (training time of less than one hour) are beneficial especially for easy deployment in the low-resource settings, and the proposed methods are adequate for the immediate task of robust and efficient OCT image analysis.

OCT imaging systems such as Heidelberg Spectralis inherently apply the log-operator as a pre-processing step for B-mode images before they are displayed. The noise distribution in these B-mode OCT scans will be Gaussian. To analyze the performance of deep learning models on such type of images, an investigation was carried out involving OCT B-scans with Gaussian noise model. In this, we have added zero mean Gaussian noise with variance $\in$ [0.04, 0.1, 0.16] to B-scans from the UCSD dataset. The noisy B-scans were tested for retinal diagnosis using the baseline models and the proposed approach. R18 (ResNet18) and R18+Noise were the baseline models for this experiment, and R18+NRSD was the proposed approach. The results of the same were reported in Table VII. From the results it is clear that the performance of the proposed approach was superior to others in this case. This also asserts that the proposed method would be able to handle Gaussian noise models.

TABLE VII
AVERAGED F1-SCORES (MEAN ± STD) FROM THREE INDEPENDENT RUNS (WITH DIFFERENT RANDOM SEEDS) ON ∼ 14500 TEST OCT B-SCANS FROM UCSD DATASET AT VARYING NOISE LEVELS (ADDITIVE GAUSSIAN NOISE)

Noise Level	**(0.04)**	**(0.10)**	**(0.16)**
Model | Metric	**F1-score**	**F1-score**	**F1-score**
R18	0.87 ± 0.04	0.67 ± 0.04	0.50 ± 0.07
R18+Noise	0.92 ± 0.01	0.86 ± 0.01	0.78 ± 0.02
R18+NRSD (Proposed)	**0.93 ± 0.01**	**0.90 ± 0.01**	**0.86 ± 0.01**

The best results are shown in bold.

There are many OCT denoising algorithms for generating high SNR B-scans [16], [48], [57]. The effect of such denoising algorithms before the classification model has been investigated in this work. The popular OCT speckle denoising algorithms like BM3D [57], NLM [58], and SiameseGAN [16], were used to denoise the OCT B-scans. We also report the retinal disease diagnosis on B-scans denoised from the benchmarking UNet [59] architecture. The denoised B-scans obtained from above mentioned methods were sent to the classification model to predict the retinal diseases. These experiments were conducted in two stages: (a) This experiment considered noisy B-scans from the dataset [48], [49]. The baseline models were SiameseGAN+R18, i.e., SiameseGAN denoiser, followed by ResNet18 classifier and UNet+R18, i.e., UNet denoiser, followed by ResNet18 classifier. The SiameseGAN and UNet models were trained on subset of B-scans from the dataset [48], [49]. The quantitative metrics on the test cases for this experiment were shown in Table VIII. The proposed approach R18+NRSD outperforms SiameseGAN+R18 and UNet+R18 with a significant margin in terms of confidence scores. (b) This experiment considered test cases varying at noise levels [0.4, 1.0, 1.6] from the UCSD dataset. The baseline models were BM3D+R18, i.e., BM3D denoiser, followed by ResNet18

TABLE VIII
EFFECT OF B-SCANS DENOISING USING SIAMESEGAN AND UNET ON RETINAL DISEASE DIAGNOSIS

Model | Metric	Correct/Total	Confidence	Parameters	Inference (s)
R18	14/18	0.87 ± 0.15	∼ 11 M	0.2
SiameseGAN+R18	13/18	0.81 ± 0.24	∼ 30.6 M	1.5
UNet+R18	14/18	0.83 ± 0.20	∼ 41.1 M	1.5
R18+NRSD (Proposed)	**16/18**	**0.91 ± 0.05**	∼ 11 M	0.2

The confidence scores (mean ± std) from three independent runs (different random seeds) on 18 OCT B-scans from the noisy B-scans dataset [48], [49] were shown below. The best results are shown in bold.

TABLE IX
EFFECT OF B-SCANS DENOISING USING BM3D AND NLM ON RETINAL DISEASE DIAGNOSIS

Noise Level	(0.4)	(1.0)	(1.6)	
Model | Metric	F1-score	F1-score	F1-score	Inference (s)
R18	0.67 ± 0.10	0.45 ± 0.08	0.33 ± 0.06	0.2
BM3D+R18	0.59 ± 0.09	0.34 ± 0.07	0.22 ± 0.04	8.5
NLM+R18	0.62 ± 0.08	0.40 ± 0.08	0.29 ± 0.05	1.3
R18+NRSD (Proposed)	**0.91 ± 0.01**	**0.83 ± 0.01**	**0.69 ± 0.01**	0.2

Averaged F1-scores (mean ± std) from three independent runs (with different random seeds) on ∼14500 test OCT B-scans from UCSD dataset at three different noise levels were shown below. The best results are shown in bold.

TABLE X
ABLATION STUDY FOR DIFFERENT COMPONENTS OF THE LOSS FUNCTION

Method	$\mathcal{L}_{ce}^{1}$	$\mathcal{L}_{sd}^{1}$	$\mathcal{L}_{ce}^{2}$	$\mathcal{L}_{sd}^{2}$	F1-score
Ablation-1	✓	x	x	x	0.67 ± 0.02
Ablation-2	x	x	✓	x	0.70 ± 0.09
Ablation-3	✓	✓	x	x	0.78 ± 0.03
Ablation-4	x	x	✓	✓	0.79 ± 0.08
Proposed	✓	✓	✓	✓	**0.84 ± 0.06**

These experiments were performed using the ResNet18. The models were trained on the NEH dataset and were tested on the DHU dataset (Study 2). Averaged F1-scores were reported as the mean ± standard deviation.

classifier, and NLM+R18, i.e., NLM denoiser, followed by ResNet18 classifier. The quantitative metrics for this experiment were shown in Table IX. Retinal disease prediction from BM3D+R18 and NLM+R18 were sub-optimal compared to predictions by R18 because the SNR of the denoised B-scans is different from the SNR of B-scans from training data. Also, the denoisers will add additional compute time, resulting in delayed results/diagnosis."

OCT imaging protocols vary mainly depending on the geographical location and investigation needed [60]. Adding to this, the current emphasis for the OCT imaging systems has been on making it more affordable, accurate, and accessible [61], in short, moving towards point-of-care technology. Most of these point-of-care scanning protocols (including instruments) provide low SNR OCT images. The need for generalizable deep learning models that adapt to different scanning protocols is of utmost importance. The proposed method serves this purpose by carefully engineering SNR-dependent regularized constraints in the cost function. The SNR is an essential characteristic feature that often varies across the scanning protocols and scanners, and having such a regularized constraint enabled the proposed method to predict retinal diseases robustly. The same is evident from the generalizability studies 2 and 3 performed in this work. Using the ablation experiments, the significance of various loss components was also assessed. (refer to Table X), asserting that the proposed loss function provides the much needed improved performance. The existing generalizable deep models for retinal diagnosis are based on domain adaptation [25]. However, this assumes that the access to datasets from multiple centers is feasible, and in reality, this is challenging because of the data privacy and security. Domain adaptation methods have significantly improved generalizability of deep models in computer vision literature. Alongside, source-free domain adaptation included the vital criterion of data privacy. An interesting future direction for building beneficial deep models for retinal diagnosis is to embed lightweight deep models within a federated learning driven domain adaptation framework. The lightweight models will account for low cost computation, memory size, high-speed inference, etc.; federated learning will consider data privacy and security issues, and domain adaptation will build generalizable deep models.

The work presented here has a distinct advantage of working across datasets and varying SNR levels in OCT images providing same performance in terms of confidence score provided by the proposed model for retinal disease prediction (Fig. 6). Current artificial intelligence (AI) driven diagnosis of retinal diseases requires a denoising of these low SNR OCT images with several works in the literature [14], [15], [16], [17] highlighting the extra step to provide acceptable performance for patient screening and triaging. The proposed framework eliminates the denoising step without compromise in the performance, thus making AI based diagnosis more efficient along with much needed generalizability across datasets.

The proposed method has few limitations. The noise levels for building generalized models were chosen empirically. However, non reference based metrics like CNR or SNR might help to determine to what extent the model can provide a correct prediction irrespective of the SNR level. Also, developing methods to transform the B-scans during test time to match or lie within the SNR margin of the B-scans from training time might boost the performance. Embedding such methods via domain adaption techniques might improve the generalizability of deep learning based retinal diagnosis. Also, the proposed approach utilized KL-divergence loss minimization to provide better generalization, the other losses (including hybrid ones) which can take into account the physics of the problem were not explored in this work. The proposed framework has utilized cross entropy loss function, which is commonly used loss function for classification task. This frame work is capable of incorporation of noise regularization in other loss functions as well since it provides regularized gradients for better classification of OCT images. Even though the proposed framework is computationally efficient, it requires careful choice of hyperparameters regularization constant (λ) and temperature scaling (T), which were empirically chosen as 1 and 4 throughout the experiments. These are needed to be tuned or chosen empirically, if the underlying classification problem is different. These investigations will be taken up as a future work. Even though the proposed method is capable of handling the noisy B-scans (including various types of noise distributions), the method will not be able to handle the domain shift cases like the B-scans from the Swept Source (SS)-OCT data [37]. To handle such extreme variability, domain adaptation methods needs to be experimented and such investigations will carried as a future work. The developed codes were made publicly available for enthusiastic users at [62].

VI. Conclusion

This work introduced regularized lightweight models for robust retinal disease diagnosis using OCT B-scan images. Precisely, a novel regularizing term has been proposed for enforcing a consistent prediction irrespective of the SNR level of the B-scan. The proposed self-distillation approach provided highly generalizable models across scanning protocols and OCT acquisitions from different machines. It was shown through rigorous experimentation that the proposed NRSD can be easily integrated into popular deep/lightweight learning models to provide improved performance/generalization that is agnostic to SNR of the OCT images and datasets. Thus the proposed NRSD with lightweight models is an efficient and robust deep learning framework for the retinal diseases diagnosis using OCT images.

Acknowledgment

Naveen Paluru acknowledges Prime Ministers Research Fellowship (PMRF). The authors would also like to thank to the UCSD [2], DHU [7], NEH [20], and [48], [49] for building open source OCT datasets.

References

[1] P. A. Keane et al., “Evaluation of age-related macular degeneration with optical coherence tomography,” *Surv. Ophthalmol.*, vol. 57, no. 5, pp. 389–414, 2012.

[2] D. S. Kermany et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[3] M. A. Hussain et al., “Classification of healthy and diseased retina using SD-OCT imaging and random forest algorithm,” *PLoS One*, vol. 13, no. 6, 2018, Art. no. e0198281.

[4] G. Lemaître et al., “Classification of SD-OCT volumes using local binary patterns: Experimental validation for DME detection,” *J. Ophthalmol.*, vol. 2016, 2016, Art. no. 3298606.

[5] A. Albarrak et al., “Age-related macular degeneration identification in volumetric optical coherence tomography using decomposition and local feature extraction,” in *Proc. Int. Conf. Med. Image, Understanding Anal.*, 2013, pp. 59–64.

[6] K. Alsaih et al., “Machine learning techniques for diabetic macular edema (DME) classification on SD-OCT images,” *Biomed. Eng. Online*, vol. 16, no. 1, pp. 1–12, 2017.

[7] P.P. Srinivasan et al., “Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images,” *Biomed. Opt. Exp.*, vol. 5, no. 10, pp. 3568–3577, 2014.

[8] F. G. Venhuizen et al., “Automated age-related macular degeneration classification in OCT using unsupervised feature learning,” *Proc. SPIE*, vol. 9414, 2015, Art. no. 94141I.

[9] S. Apostolopoulos et al., “RetiNet: Automatic AMD identification in OCT volumetric data,” *Invest. Ophthalmol. Vis. Sci.*, vol. 58, no. 8, 2017, Art. no. 387.

[10] W. Lu et al., “Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images,” *Transl. Vis. Sci. Technol.*, vol. 7, no. 6, 2018, Art. no. 41.

[11] J. H. Tan et al., “Age-related macular degeneration detection using deep convolutional neural network,” *Future Gener. Comput. Syst.*, vol. 87, pp. 127–135, 2018.

[12] V. Das, E. Prabhakararao, S. Dandapat, and P. K. Bora, “B-scan attentive CNN for the classification of retinal optical coherence tomography volumes,” *IEEE Signal Process. Lett.*, vol. 27, pp. 1025–1029, 2020.

[13] A. Sunija et al., “OctNET: A lightweight CNN for retinal disease classification from optical coherence tomography images,” *Comput. Methods Programs Biomed.*, vol. 200, 2021, Art. no. 105877.

[14] Y. Huang et al., “Simultaneous denoising and super-resolution of optical coherence tomography images based on generative adversarial network,” *Opt. Exp.*, vol. 27, no. 9, pp. 12289–12307, 2019.

[15] Q. Hao et al., “High signal-to-noise ratio reconstruction of low bit-depth optical coherence tomography using deep learning,” *J. Biomed. Opt.*, vol. 25, no. 12, 2020, Art. no. 123702.

[16] N. A. Kande, R. Dakhane, A. Dukkipati, and P. K. Yalavarthy, “SiameseGAN: A generative model for denoising of spectral domain optical coherence tomography images,” *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 180–192, Jan. 2020.

[17] B. Qiu et al., “Comparative study of deep neural networks with unsupervised noise2noise strategy for noise reduction of optical coherence tomography images,” *J. Biophotonics*, vol. 14, no. 11, 2021, Art. no. e202100151.

[18] F. Li, H. Chen, Z. Liu, X. Zhang, and Z. Wu, “Fully automated detection of retinal disorders by image-based deep learning,” *Graefe’s Arch. Clin. Exp. Ophthalmol.*, vol. 257, no. 3, pp. 495–505, 2019.

[19] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[20] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, “Macular OCT classification using a multi-scale convolutional neural network ensemble,” *IEEE Trans. Med. Imag.*, vol. 37, no. 4, pp. 1024–1034, Apr. 2018.

[21] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, “Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification,” *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1959–1970, Aug. 2019.

[22] F. Li et al., “Deep learning-based automated detection of retinal diseases using optical coherence tomography images,” *Biomed. Opt. Exp.*, vol. 10, no. 12, pp. 6204–6226, 2019.

[23] Y. Luo et al., “Automatic detection of retinopathy with optical coherence tomography images via a semi-supervised deep learning method,” *Biomed. Opt. Exp.*, vol. 12, no. 5, pp. 2684–2702, 2021.

[24] J. P. Owen et al., “Student becomes teacher: Training faster deep learning lightweight networks for automated identification of optical coherence tomography B-scans of interest using a student-teacher framework,” *Biomed. Opt. Exp.*, vol. 12, no. 9, pp. 5387–5399, 2021.

[25] Y. Luo et al., “Cross-domain retinopathy classification with optical coherence tomography images via a novel deep domain adaptation method,” *J. Biophotonics*, vol. 14, no. 8, 2021, Art. no. e202100096.

[26] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.

[27] B. McMahan et al., “Communication-efficient learning of deep networks from decentralized data,” in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[28] J. M. Schmitt, S. Xiang, and K. M. Yung, “Speckle in optical coherence tomography,” *J. Biomed. Opt.*, vol. 4, no. 1, pp. 95–105, 1999.

[29] B. Karamata, K. Hassler, M. Laubscher, and T. Lasser, “Speckle statistics in optical coherence tomography,” *J. Opt. Soc. Amer. A*, vol. 22, no. 4, pp. 593–596, 2005.

[30] N. M. Grzywacz et al., “Statistics of optical coherence tomography data from human retina,” *IEEE Trans. Med. Imag.*, vol. 29, no. 6, pp. 1224–1237, Jun. 2010.

[31] Z. Amini and H. Rabbani, “Statistical modeling of retinal optical coherence tomography,” *IEEE Trans. Med. Imag.*, vol. 35, no. 6, pp. 1544–1554, Jun. 2016.

[32] Z. Amini and H. Rabbani, “Optical coherence tomography image denoising using Gaussianization transform,” *J. Biomed. Opt.*, vol. 22, no. 8, 2017, Art. no. 086011.

[33] Z. Amini, H. Rabbani, and I. Selesnick, “Sparse domain Gaussianization for multi-variate statistical modeling of retinal OCT images,” *IEEE Trans. Image Process.*, vol. 29, pp. 6873–6884, 2020.

[34] M. Samieinasab, Z. Amini, and H. Rabbani, “Multivariate statistical modeling of retinal optical coherence tomography,” *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3475–3487, Nov. 2020.

[35] P. Sudeep et al., “Enhancement and bias removal of optical coherence tomography images: An iterative approach with adaptive bilateral filtering,” *Comput. Biol. Med.*, vol. 71, pp. 97–107, 2016.

[36] M. Li, R. Idoughi, B. Choudhury, and W. Heidrich, “Statistical model for OCT image denoising,” *Biomed. Opt. Exp.*, vol. 8, no. 9, pp. 3903–3917, 2017.

[37] M. Tajmirriahi et al., “Modeling of retinal optical coherence tomography based on stochastic differential equations: Application to denoising,” *IEEE Trans. Med. Imag.*, vol. 40, no. 8, pp. 2129–2141, Aug. 2021.

[38] N. Paluru et al., “Anam-Net: Anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, pp. 932–946, Mar. 2021.

[39] N. Awasthi, A. Dayal, L. R. Cenkeramaddi, and P. K. Yalavarthy, “Mini-COVIDNet: Efficient lightweight deep neural network for ultrasound based point-of-care detection of COVID-19,” *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 68, no. 6, pp. 2023–2037, Jun. 2021.

[40] H. Peng et al., “Accurate brain age prediction with lightweight deep neural networks,” *Med. Image Anal.*, vol. 68, 2021, Art. no. 101871.

[41] Y. Li et al., “VolumeNet: A lightweight parallel network for super-resolution of MR and CT volumetric data,” *IEEE Trans. Image Process.*, vol. 30, pp. 4840–4854, 2021.

[42] A. W. Scott et al., “Imaging the infant retina with a hand-held spectral-domain optical coherence tomography device,” *Amer. J. Ophthalmol.*, vol. 147, no. 2, pp. 364–373, 2009.

[43] G. Hinton et al., “Distilling the knowledge in a neural network,” 2015, *arXiv:1503.02531*.

[44] S. Yun, J. Park, K. Lee, and J. Shin, “Regularizing class-wise predictions via self-knowledge distillation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13876–13885.

[45] L. Zhang et al., “Be your own teacher: Improve the performance of convolutional neural networks via self distillation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3713–3722.

[46] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, “Deep mutual learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4320–4328.

[47] T.-B. Xu and C.-L. Liu, “Data-distortion guided self-distillation for deep neural networks,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 01, pp. 5565–5572.

[48] L. Fang et al., “Sparsity based denoising of spectral domain optical coherence tomography images,” *Biomed. Opt. Exp.*, vol. 3, no. 5, pp. 927–942, 2012.

[49] L. Fang et al., “Fast acquisition and reconstruction of optical coherence tomography images via sparse representation,” *IEEE Trans. Med. Imag.*, vol. 32, no. 11, pp. 2034–2049, Nov. 2013.

[50] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[51] M. Sandler et al., “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[52] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “ShuffleNet V2: Practical guidelines for efficient CNN architecture design,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.

[53] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[54] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[55] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[57] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[58] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 60–65.

[59] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Springer, 2015, pp. 234–241.

[60] G. Montesano et al., “Optimizing OCT acquisition parameters for assessments of vitreous haze for application in uveitis,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–7, 2018.

[61] S. Kim et al., “Design and implementation of a low-cost, portable OCT system,” *Biomed. Opt. Exp.*, vol. 9, no. 3, pp. 1232–1243, 2018.

[62] “NRSD: Noise Regularized Self Distillation [Source Code],” [Online]. Available: https://github.com/NaveenPaluru/OCT-NRSD

Naveen Paluru (Graduate Student Member, IEEE) received the master’s by Research degree from the Department of Electrical Engineering, Indian Institute of Technology, Tirupati, India, in 2019. He is currently working toward the Ph.D. degree with the Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, India. His main research interests are machine learning and deep learning for automated medical image analysis. He was the recipient of the Prime Ministers Research Fellowship (PMRF) in 2020, for the Doctoral Fellowship.

Hariharan Ravishankar received bachelor’s degree in engineering from Anna University, Chennai, India, in 2008, and master’s degree in communications systems from the Indian Institute of Technology, Madras, India, in 2010. He is currently working toward the Ph.D. degree with the Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, India. He was a Senior Scientist with GE Healthcare, Bangalore, for the past four years. His main research interests include development of efficient deep learning methods for medical image analysis.

Sharat Hegde received the MBBS and Postgraduate degrees in ophthalmology from JJM Medical College, Davanagere, India, affiliated to Rajiv Gandhi University of Health Science in 2014. He is currently a consultant in vitreoretinal services with Prasad Netralaya, Udupi, India. He has various publications in prestigious international journals. His research interest include diabetic retinopathy, endophthalmitis, and artificial intelligence in retinal disorders. He was the recipient of the Fellowship in vitreoretinal super speciality from L V Prasad Eye Institute and was a consultant.

Phaneendra K. Yalavarthy (Senior Member, IEEE) received the M.Sc. degree in engineering from the Indian Institute of Science, Bangalore, India, in 2004, and the Ph.D. degree in biomedical computation from Dartmouth College, Hanover, NH, USA, in 2007. He is currently a Professor with the Department of Computational and Data Sciences, Indian Institute of Science, Bangalore. His research interests include medical image computing, medical image analysis, and biomedical optics. He is the Senior Member of SPIE and OSA. He was the Associate Editor of IEEE Transactions on Medical Imaging.