Multi-Task and Few-Shot Learning-Based Fully Automatic Deep Learning Platform for Mobile Diagnosis of Skin Diseases

Kyungsu Lee, T. C. Cavalcanti, Sewoong Kim, Hah Min Lew, Dae Hun Suh, Dong Hun Lee, and Jae Youn Hwang, *Member, IEEE*

***Abstract*—Fluorescence imaging-based diagnostic systems have been widely used to diagnose skin diseases due to their ability to provide detailed information related to the molecular composition of the skin compared to conventional RGB imaging. In addition, recent advances in smartphones have made them suitable for application in biomedical imaging, and therefore various smartphone-based optical imaging systems have been developed for mobile healthcare. However, an advanced analysis algorithm is required to improve the diagnosis of skin diseases. Various deep learning-based algorithms have recently been developed for this purpose. However, deep learning-based algorithms using only white-light reflectance RGB images have exhibited limited diagnostic performance. In this study, we developed an auxiliary deep learning network called fluorescence-aided amplifying network (FAA-Net) to diagnose skin diseases using a developed multi-modal smartphone imaging system that offers RGB and fluorescence images. FAA-Net is equipped with a meta-learning-based algorithm to solve problems that may occur due to the insufficient number of images acquired by the developed system. In addition, we devised a new attention-based module that can learn the location of skin diseases by itself and emphasize potential disease regions, and incorporated it into FAA-Net. We conducted a clinical trial in a hospital to evaluate the performance of FAA-Net and to compare various evaluation metrics of our developed model and other state-of-the-art models for the diagnosis of skin diseases using our multi-modal system. Experimental results demonstrated that our developed model exhibited an 8.61% and 9.83% improvement in mean accuracy and area under the curve in classifying skin diseases, respectively, compared with other advanced models.**

***Index Terms*—Deep learning, few-shot learning, multi-modal system, skin diagnosis, fluorescence imaging.**

I. Introduction

ROSACEA and dermatitis are common skin diseases in modern society. Since both are inflammatory skin disorders with similar clinical phenotypes characterized by erythema, papules, edema, or telangiectasia, their symptoms are highly similar [1]. Therefore, precise diagnosis is essential to apply appropriate treatment methods for these diseases. In general, dermatitis is treated with topical corticosteroids, whereas rosacea is treated with once-daily brimonidine, a topical alpha-adrenergic receptor agonist, metronidazole, or ivermectin, depending on the type of rosacea. However, since they exhibit similar properties, misdiagnosis of these diseases often occurs clinically. Note that the long-term use of corticosteroids for rosacea leads to its aggravation. Therefore, for better treatment, precise diagnosis is highly important.

Recently, fluorescence imaging has been widely used to diagnose rosacea and dermatitis. Since fluorescence imaging can offer important diagnostic and localized information non-invasively without contrast agents and is a suitable diagnostic modality for skin disease, fluorescence systems have been widely studied for many applications [2]–[5]. However, because the conventional fluorescence imaging system was primarily designed for an expert in a clinical setting, it is of limited use as a healthcare device for an inexpert user at home. Furthermore, it is essential to use white-light with fluorescence in diagnosing skin disease rather than the independent utilization of fluorescence imaging [6].

To this end, we developed a multimodal skin disease diagnostic system using white-light and fluorescence imaging modalities. To acquire multimodal skin images, including fluorescence and white-light reflectance RGB images, and to achieve an accurate mobile diagnosis of various skin diseases at low cost at home, we constructed a smartphone-based fluorescence imaging system. Moreover, we developed deep learning-based software to diagnose skin diseases using images captured by the system. To the best of our knowledge, multimodal imaging systems

Manuscript received 10 September 2021; revised 13 February 2022 and 18 July 2022; accepted 19 July 2022. Date of publication 25 July 2022; date of current version 5 January 2023. This work was supported in part by the National Research Foundation of Korea (NRF) under Grant NRF-2020R1A2B5B01002786 and in part by the Bio & Medical Technology Development Program of the National Research Foundation (NRF), the Korean Government (MSIT) under Grant 2017M3A9G8084463. *(Corresponding authors: Jae Youn Hwang; Dong Hun Lee.)*

Kyungsu Lee, T. C. Cavalcanti, Sewoong Kim, and Hah Min Lew are with the Department of Electrical Engineering and Computer Science, DGIST, Daegu 42988, Korea (e-mail: ks_lee@dgist.ac.kr; thiago@dgist.ac.kr; dion.kim@dgist.ac.kr; rpm0130@dgist.ac.kr).

Dae Hun Suh and Dong Hun Lee are with the Department of Dermatology, Seoul National University College of Medicine, Institute of Human-Environment Interface Biology, Seoul National University, Seoul 03080, South Korea (e-mail: daehun@snu.ac.kr; ivymed27@snu.ac.kr).

Jae Youn Hwang is with the Department of Electrical Engineering and Computer Science, The Interdisciplinary Studies of Artificial Intelligence, DGIST, Daegu 42988, Korea (e-mail: jyhwang@dgist.ac.kr).

Digital Object Identifier 10.1109/JBHI.2022.3193685

using white-light and fluorescence imaging modalities for skin disease have rarely been studied; thus, datasets of multimodal images have not been constructed to fully optimize deep learning models, compared to other skin disease datasets containing a large number of skin disease images [7], [8]. Therefore, since using a limited number of images for training typically degrades deep learning performance, we devised a meta-learning-based deep learning model, called a fluorescence-aided amplifying network (FAA-Net) [9], [10]. FAA-Net was developed based on the few-shot learning algorithm due to the small number of images captured by our system. Furthermore, since the fluorescence imaging modality enables the localization of skin disease, a detection task for skin disease regions is included in FAA-Net alongside the classification of skin diseases [11], [12]. Therefore, FAA-Net is designed based on multi-task learning (classification and detection tasks) to use the localization property of the fluorescence imaging modality. The main contributions of this study are summarized as follows:

- We developed an end-to-end deep learning network based on multitask and few-shot learning using multimodal images. The algorithm outperforms other modern deep learning-based algorithms for the multiclass classification of skin diseases, even with a small number of images.
- We devised novel supporting modules such as recyclable attention collection (RAC) and amplifying focused similarity (AFS) blocks to take full advantage of the benefits of attention mechanisms, few-shot learning, and multitask learning.
- The RAC and AFS-Blocks enhanced the performance of the deep learning network, thus resulting in an 8.61% improvement in accuracy for skin diagnosis and classification compared to the other state-of-the-art models.
- A smartphone-based fluorescence imaging system that offers fluorescence and white-light reflectance RGB images was developed, and FAA-Net was incorporated into the system for multiclass classification of skin diseases.

II. Related Works

A. Fluorescence Imaging

The potential of fluorescence imaging has been studied. Several studies used the fluorescence imaging modality to classify target diseases, and verified that fluorescence imaging could localize the targets, and thus improve the diagnostic performance [4], [5]. Fluorescence imaging modalities have also been studied to treat skin diseases. Farkas et al. have examined the application of fluorescence imaging and found that it is a safe and effective tool for the diagnosis of pseudoxanthoma elasticum [13]. Furthermore, dermatologists have used fluorescence imaging of the skin regions to diagnose pigmentary disorders, cutaneous infections, and porphyria [14]. Previous studies have therefore established that fluorescence imaging was able to localize the targets and thus improve diagnostic performance [15]–[17].

B. Deep Learning-Based Skin Diagnostic Modality

Over the decades, deep learning-based algorithms for skin diseases have been widely studied owing to their precise diagnostic performance. Yap et al. used multi-modal skin images and patient metadata to perform deep learning-based skin disease diagnosis [18]. Yang *et al.* developed a deep learning model for both the categorization and segmentation of skin diseases, and the developed architecture improved skin disease diagnostic performance by compensating for each task [19]. Rahman *et al.* applied ensemble learning combined with deep learning to classify skin diseases, and they showed that mathematical analysis combined with a deep learning model could be used to classify various categories of skin diseases [20]. Furthermore, deep learning models have been developed for use in mobile-based environments owing to their effective diagnosis and low cost [21], [22]. However, current deep learning models still have room for improvement to exhibit superior diagnostic performance.

C. Meta-Learning-Based Diagnostic Modality

Recently, advanced algorithms have been developed in several studies based on meta-learning, including multi-task and few-shot learning. Lanchantin *et al.* developed a multi-task learning-based classification model using transformers [23]. Wertheimer *et al.* reformulated and significantly improved few-shot classification for various benchmarks using a large number of images [24]. Zhu *et al.* combined multi-task learning with the few-shot modality and concluded that multi-modality and few-shot learning could be superior tools for classification [25]. Deep learning-based methodologies have been used in healthcare due to their superior performance. Mahajan *et al.* proposed a meta-learning-based few-shot learning algorithm to identify skin diseases [26]. Mathews *et al.* developed a range of simple to advanced deep learning methods, including dictionary learning and representation learning, using sensor signal representations for classification tasks [27]–[30]. Furthermore, unsupervised representation learning has achieved state-of-the-art performance for rare disease classification [31]. Therefore, it has been demonstrated that the advantages of meta-learning-based deep learning methodologies significantly outweigh their drawbacks in classification tasks. Hence, in this study, a few-shot and multi-task learning-based deep learning model was devised to address the aforementioned problems.

III. Smartphone-Based Fluorescence Imaging System and Dataset

This section illustrates the structure of the smartphone-based fluorescence imaging system and the dataset obtained using it. In addition, since the number of images obtained via clinical trials was insufficient, the long-tail problem of the constructed dataset is discussed. Fig. 1 illustrates the graphical description of the developed system.

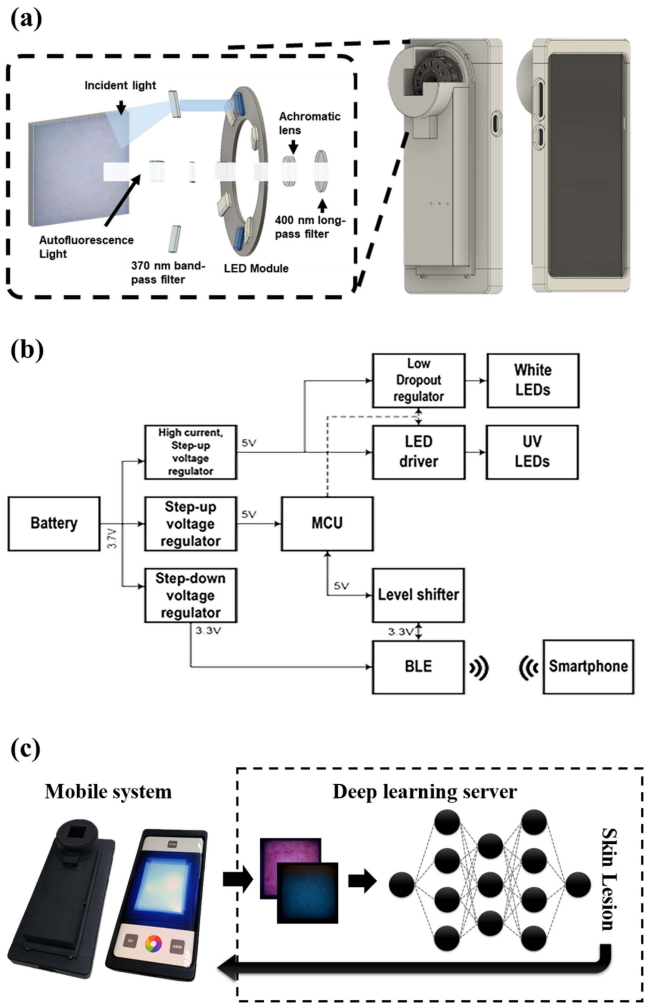

Fig. 1. Smartphone-based fluorescence imaging system: (a) Schematics of optical components for the developed system, (b) Circuit diagram of the interface circuit, (c) Orthographic image of the integrated system using a smartphone-based imaging system and a deep learning-based diagnosis system. Smartphone transfers captured images and receives the analysis results through wireless communications.

A. Smartphone-Based Fluorescence Imaging System

We developed a smartphone-based fluorescence imaging system for the low-cost quantitative mobile diagnosis of skin diseases. It comprises a fluorescence imaging module optimized to obtain an fluorescence image of the skin with an excitation LED array, various optical components, and an interface circuit to synchronize the LED array and a smartphone camera through Bluetooth Low Energy (BLE) technology, and a deep learning-based analysis platform linked to a smartphone (SM-G950 N, SAMSUNG) [Fig. 1(c)].

The illumination source of the system comprises four white-light LEDs (Iws-351-white, ITSWELL) and four UV LEDs (LTPL-c034uvh365, Liteon Optoelectronics). The white-light LEDs are used to search the skin regions of interest, and the UV LEDs are used to obtain fluorescence images of the skin regions. The UV light passes through optical bandpass filters (86-982, EDMUND OPTICS) in front of each UV LED, and then is delivered to excite the skin. The light emitted from the skin is collected by an achromatic lens (63-729, EDMUND OPTICS), passes through a long-pass filter (62-974, EDMUND OPTICS), and then is recorded in the 12 M pixel color CMOS sensor on the smartphone [Fig. 1(a)].

An interface circuit was built to synchronize the CMOS camera and LED modules. This consists of a microcontroller unit (ATMEGA128, Atmel), a Bluetooth low-energy module (RN4871, Microchip Technology), an LED sink driver (STP04CM05, ST Microelectronics) supplying constant current to the UV LEDs, a low dropout regulator (LT3085, Linear Technology) driving the white LEDs (CLM1C-WKW-CWbXb233, Cree Inc.), a level shifter (BSS138BKS, Nexperia), and three voltage regulators, one with a 3.3 V output (TPS73033, Texas Instruments) supplying power to the Bluetooth module, another with a 5 V output powering the microcontroller unit, and the last with a 5 V output supplying a higher current to the UV LEDs. The microcontroller unit receives a command signal from an Android application via Bluetooth and then turns a group of LEDs on or off to select the light wavelengths. The level shifter adjusts the voltage level to achieve reliable communication between the microprocessor unit and Bluetooth module. Finally, a 3.7-V lithium-polymer battery is used to supply electric power to the interface circuit [Fig. 1(b)].

B. Dataset and Pre-Processing

Using the developed system, we obtained skin disease images from 31 patients and 20 healthy subjects. Among the patients, 16 were diagnosed with rosacea and 15 were diagnosed with dermatitis by dermatologists. Three to six white-light RGB and fluorescence images were obtained from each patient's cheeks, forehead, and arms. Thus, 123 rosacea, 104 dermatitis, and 200 normal white-light RGB and fluorescence images were obtained, for a total of 427 white-light RGB and 427 fluorescence images. The training, validation, and test sets were divided by 3:1:1 for 5-fold cross-validation, and thus each fold includes three to four rosacea patients, three dermatitis patients, and four normal subjects. Here, three, one, and one folds were used for the training, validation, and test sets, respectively. The images were collected through a tertiary referral hospital under the approval of the Institutional Review Board (IRB No. 1908-161-1059) and were obtained with the consent of the subjects according to the principles of the Declaration of Helsinki.

To understand the properties of the acquired data, the white-light RGB data were compared with the ISIC2020 dataset, a public dataset on skin diseases [7]. Fig. 2 shows the density of data corresponding to each PCA characteristic after converting our data and the ISIC2020 combined data into a normalized PCA distribution. As presented in Fig. 2, our dataset did not belong to the head class representing the general characteristics of the data but was distributed in the tail class with unusual characteristics. This implies that the conventional methodology cannot easily analyze the characteristics of the dataset [32]. In addition, this indicated that there were too few images, and therefore, it

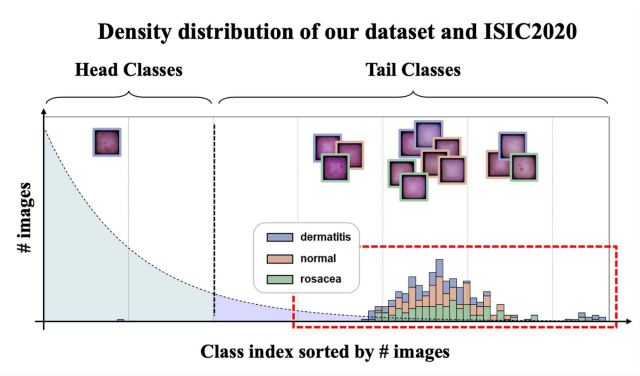

Fig. 2. Class distribution of our dataset and ISIC dataset. After the normalized PCA projections, images are grouped into different classes wiht a certain interval (0.01), and the class indexes are sorted by the number of images in each class. The images in our dataset were projected in the long-tail of ISIC2020 dataset.

was necessary to apply other methodologies such as few-shot learning and multi-task learning or meta-learning [32]–[34].

Therefore, our dataset has the advantage of capturing white-light RGB images and acquiring fluorescence data simultaneously. However, it is difficult to confirm whether the number of images is sufficient for deep learning. In addition, the features extracted from the acquired images were more extensive than typically used for this purpose. For this reason, we did not use universal classification to develop a deep learning algorithm using our data, but developed a novel algorithm based on few-shot learning and multi-task learning suitable for our data characteristics.

IV. Methods

A. Overview of Diagnostic Algorithm

The deep learning-based diagnostic algorithm was designed to improve the constraint that only a limited number of images were collected using our developed system. FAA-Net integrates multi-task learning and few-shot learning, which are effective in improving the issues caused by a limited dataset [33]; (1) In terms of multi-task learning, FAA-Net has two parallel training pipelines (Training phase II in Fig. 3). The training pipeline for the supplementary task (1st training task in Fig. 3) compensates that of the main task (2nd Training Task in Fig. 3), which helps improve classification accuracy. (2) To enhance the cognition of deep learning algorithms, despite the small number of training images, new few-shot learning-based components called recyclable attention collections (RAC) and amplifying focused similarity (AFS) blocks are proposed. A RAC is a collection of pre-extracted features that indicate highlighted regions by insight into a deep learning network by means of a class activation map (CAM) [35]. Subsequently, AFS-Blocks exploit the RAC in a detection task to localize suspicious areas similar to the extracted disease in the RAC. FAA-Net was thus optimized to ascertain the location of the target diseases, and

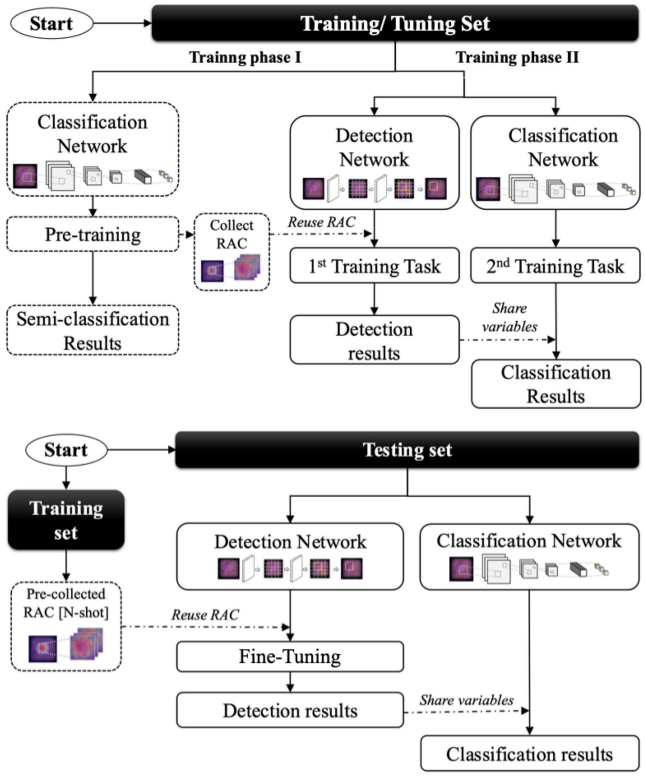

Fig. 3. Comprehensive workflow diagram of the proposed network in training and testing phases.

was therefore able to demonstrate superior classification performance. Additionally, the RAC and AFS block were designed to ensure consistency in visual diagnosis.

B. Model Construction

Since FAA-Net has individual training pipelines for detection and classification tasks, as shown in Fig. 4(a), DarkNet [36], which is widely used in both tasks, was adopted as a baseline architecture for FAA-Net. In addition, to compensate for the different formats of individual outputs of the classification network and a detection network, distinct layers were connected to each network: a fully connected (FC) layer was connected to the penultimate layers of the classification network, and a global average pooling (GAP) layer was placed at the end of the classification network, as shown in Fig. 4(a). Here, the CAM was generated by multiplying the generated feature map by the GAP result and the weights of a SoftMax layer, as shown in Fig. 4(b) [35].

To detect disease-suspicious regions and store the detected regions in the RAC, CAMs were generated as a by-product of the classification network in the pre-training phase (Training phase I in Fig. 3). Here, a CAM represents an insight into classifying diseases using a deep learning network. To maximize the insight of the network, only the top 15% of disease-like regions were rectified and emphasized with normalization and standardization methodologies, as shown in Fig. 4(b). The highlighted regions were then cropped and stored in the RAC for future use.

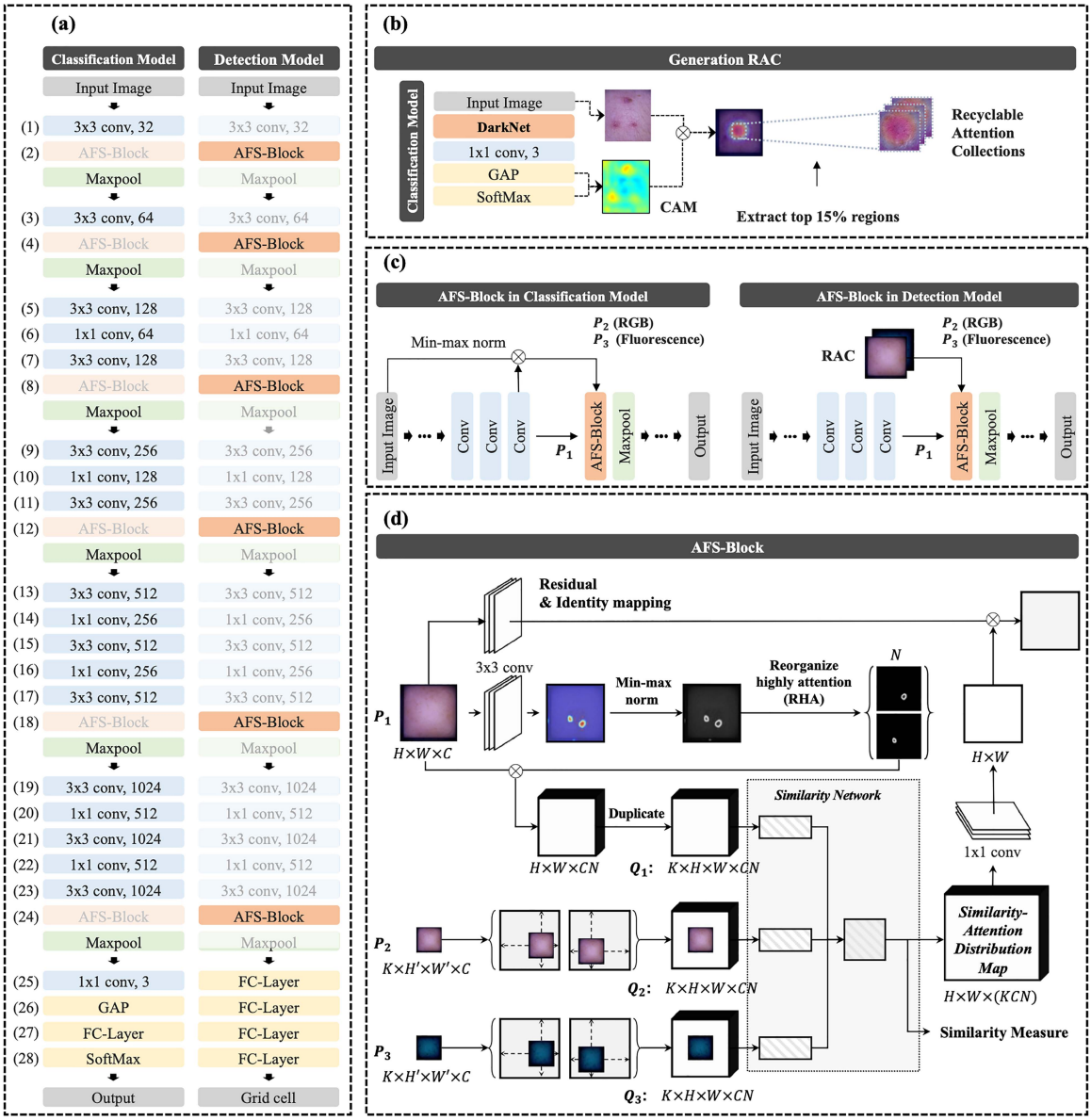

Fig. 4. (a) Detailed architecture of FAA-Net. The translucent and transparent layers indicate the trained and skipped-in-training layers, respectively. (b) Generation of a class activation map (CAM) by using the classification network of FAA-Net. (c) Inputs of AFS-Blocks. (d) Detailed architecture of AFS-Blocks. SiameseNetwork and triplet are used as baseline architecture. Here, stored features in RAC (P_2 and P_3) are zero-padded after the center alignment with the RHAs.

Subsequently, the normalized and standardized features in the RAC affected the AFS-Blocks of the detection network during the main training phase (Training phase II in Fig. 3) and the testing phase (Pre-collected RAC in Fig. 3). The pre-extracted features, which were stored in the RAC, were reused to identify disease-like regions in the detection task for feature extraction by convolution operations in AFS-Blocks.

AFS-Blocks were designed to amplify similarity in the inputs. Thus, AFS-Blocks were used in FAA-Net to highlight disease-like regions in the input images using disease-like features stored in the RAC. Thus far, AFS-Blocks were structured to redeem one output, which shows the magnified similarity distribution of three inputs: P_1 is the feature map of the previous convolution operation, and P_2 and P_3 respectively are the RGB and fluorescence features in the RAC. However, P_2 and P_3 come from different feature maps depending on the model, as shown in Fig. 4(c). In the classification task, the multiplication of the input image and feature maps from a previous convolution layer were used as P_2 and P_3. In contrast, in the detection task, the feature maps stored in the RAC were used as P_2 and P_3. Note that P_2 and P_3 should be disease-like regions that localize the diseases in the input image. However, because the RAC was used only in the detection task, the multiplied feature maps indicating the highlighted disease-like probability distribution

by the previous convolution layers were used in the classification task instead of the RAC. Therefore, since AFS-Blocks require a feature-extracted and normalized matrix, which is generally produced by a convolution operation [P_1 in Fig. 4(c)], the AFS-Blocks were placed right before every max-pooling operation [Fig. 4(a)].

C. Amplifying Focused Similarity (AFS)-Block

The detailed architecture of the AFS-Blocks is shown in Fig. 4(d). The structure is designed to extract disease-like features from an input [P_1 in Fig. 4(c) and (d)] using the pre-extracted disease-like features [P_2 and P_3 in Fig. 4(c) and (d)], which are stored in the RAC. To this end, AFS-Blocks were developed based on the Siamese network [33] and triplet [37], which are the basic networks for few-shot learning [Similarity Network in Fig. 4(d)]. Three inputs, which were a target feature map to be evaluated for skin diseases, a white-light RGB-based feature map, and a fluorescence-based feature map indicating disease features, are fed into the similarity network. The similarity network detects a similar region from the target feature map using the disease-like features from the other two feature maps from white-light RGB and fluorescence images. The similarity network then generates two outputs: (1) a similarity measurement score, which is the general output of a similarity measurement task [33], and (2) a similarity attention distribution map (SADM), which is a CAM-like feature map. Here, the similarity network is pre-trained and further optimized in the main training steps while making the similarity measurement score equal to 1, which indicates that the three inputs become similar. The similarity network creates an optimized SADM indicating disease-like regions in the target feature map using a pre-highlighted map with the normalization and reorganization of features from the previous step. In addition, since the AFS-Blocks have a deep architecture and diverse layers, residual and identical mappings were placed to compensate for the vanishing gradient. Note that AFS-Blocks were further trained and fine-tuned in the testing phase to improve the classification performance of FAA-Net. The detailed optimization process of FAA-Net is explained in the next section.

D. Mathematical Modeling

To illustrate the training process of FAA-Net, several functions and notations are defined, and the corresponding mathematical symbols are defined in Table I. $I_a^{(i)}$ indicates an output feature map of the i$^{\text{th}}$ layer in a-task network, and $\theta_a^{(i)}$ indicates a corresponding variable as an input to the i$^{\text{th}}$ layer in a-task network. Here, a can be 0 for a classification task and 1 for a detection task. For instance, $\theta_c^{(19)}$ indicates a 3×3 $conv, 1024$ layer in the classification model [Fig. 4(a)], and $I_0^{(19)}$ indicates the output from the layer of $\theta_0^{(19)}$. The height, width, and number of channels of the matrix are respectively denoted as (h, w, c). Here, the functions of resizing ($f_{\text{R};(H,W)}$) and padding ($f_{\text{P};(H,W)}$) are $\mathbb{R}^3 \rightarrow \mathbb{R}^3$ mappings that accept an input matrix with shape $(H,' W,' C)$ and generate a feature map with shape (H, W, C). The quick-selection function ($f_{\text{qs};k}$) with a value of k selects the top k% percentile from an input image and then generates a feature map for which the values corresponding to the top k% are maintained intact, and all other values are rectified to 0.

TABLE I
MATHEMATICAL SYMBOLS AND CORRESPONDING DEFINITIONS

Symbol	Definition
$I_a^{(i)}$	Output from the i$^{\text{th}}$ layer in the a-task network [See Fig. 4(a)]
$\theta_a^{(i)}$	Variable corresponding to $I_a^{(i)}$
$[I_a^{(i)}]_{h,w,c}$	(h, w, c) element in $I_a^{(i)}$
$f_{\text{R};(H,W)}(X)$	Function for resizing an input shape as $H \times W$
$f_{\text{P};(H,W)}(X)$	Function for padding to X to be size of $H \times W$
$f_{\text{qs};k}(X)$	Function for quick-selecting the top k% elements from X
$f_{\text{mm}}(X)$	Min-max normalization of X
$f_{\text{cc}}(X)$	Function for cropping an image after a contour extraction algorithm
$f_{\text{RHA}}(X)$	Function for reorganizing highly attended areas (See Eq. 2)
$f_{\text{CAM}}(I)$	$f_{\text{CAM}}(I) = C_{am}$ is a CAM-feature-map of an input image (I).

Furthermore, the min-max normalization is used here to standardize the output feature maps generated by other operations. The function for reorganizing highly attended (RHA) regions is $\mathbb{R}^3 \rightarrow \mathbb{R}^3$ mappings for feature maps of F and F' of sizes $(H \times W \times C)$ and $(H \times W \times CN)$, respectively, as follows:

$$f_{\text{RHA}} : F \rightarrow F'$$

$$\sum_{n}^{N} [f_{\text{P};(H,W)}(F')]_{h,w,cn} = [F]_{h,w,c} \tag{1}$$

That is, the RHA function splits arbitrary areas, which are defined as closures in terms of topology, to N different feature maps of the same size as the input of the RHA function.

The feature map indicating a CAM ($C_{am} = f_{\text{CAM}}(I_a^{(i)})$) is defined as follows:

$$[C_{am}]_{h,w} = \sum_{c}^{C} \left[f_{\text{R};(H,W)} \left(I_0^{(26)} \right) \right]_{h,w,c} \cdot \left[\theta_0^{(27)} \right]_c \tag{2}$$

Equation (2) demonstrates that a CAM [35] is generated by multiplication of the output from the GAP layer and the parameters of FC layers as illustrated in Fig. 4(a) and (b). In addition, as shown in Fig. 4(b), the highlighted feature map after a CAM is stored in the RAC. The following equation illustrates the elements stored in the RAC:

$$RAC = \{x | x = f_{cc}([f_{\text{RHA}}(f_{\text{qs};15}(f_{mm}(\quad f_{\text{CAM}}(I_a^{(i)})))) \cdot I]_i), \\ i = 1, 2, \ldots, N\} \tag{3}$$

That is, the normalized CAM is rectified and combined with an input image (I) by the multiplication operation. The top 15% of the regions are then extracted and stored in the RAC. Here, we consider the top 15% of areas as disease-like regions.

Three inputs (P_1, P_2, and P_3) are fed into the AFS-Blocks as shown in Fig. 4(c) and (d). As illustrated in Fig. 4(d), the three

inputs are processed as follows:

$$[Q_1]_{k,h,w,c} = [(f_{\mathrm{R};(H,W)} \circ f_{\mathrm{RHA}} \circ f_{\mathrm{qs};15} \circ f_{mm})(WP_1)]_{h,w,c}$$
$$[P_2^k]_{h,w,c} = [P_2]_{k,h,w,c} \Rightarrow [Q_2]_{k,h,w,c} = [f_{\mathrm{P};(H,W)}(P_2^k)]_{h,w,c}$$
$$[P_3^k]_{h,w,c} = [P_3]_{k,h,w,c} \Rightarrow [Q_3]_{k,h,w,c} = [f_{\mathrm{P};(H,W)}(P_3^k)]_{h,w,c} \quad (4)$$

where W is a convolution parameter, and $k = 1, 2, \ldots, K$ is the number of images in the RAC for each category of few-shot learning. If $K = 3$, each category of disease includes three images. Here, we considered $K = 3, 5, 7$ shots for few-shot learning. The processed feature maps Q_1, Q_2, Q_3 from P_1, P_2, and P_3 are fed into the Siamese network [33], and the two outputs generated by the Siamese network are the similarity measure, which is the loss value of the Siamese network, and the SADM between inputs. The similarity measure is used to optimize FAA-Net, and the SADM, as the output of the AFS block, passes over a 1×1 convolution layer. Note that AFS-Blocks are designed to highlight the disease-like areas of inputs, and thus the skin regions similar to diseases are highlighted by the attention mechanism using the similarity between inputs and the pre-extracted disease-like features stored in the RAC.

E. Optimization and Inference

FAA-Net includes three training phases: (1) training a classification network, (2) training a detection network, and (3) training a Siamese network in an AFS block. In these phases, the loss functions are denoted as $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$, respectively. Here, the optimization of a classification network is realized by supervised learning, which decreases the disparity between the prediction by the classification network and the pre-labeled ground truth indicating the disease class. In contrast, optimization of the detection network is realized by semi-supervised learning, which uses a CAM generated by the classification network as the ground truth indicating disease-like regions. The loss function for the classification network is configured as follows:

$$\mathcal{L}_1 = \sum_{c}^{3} [G]_c \cdot \left[I_0^{(28)} \right]_c \quad (5)$$

where $I_0^{(28)}$ is the output from the SoftMax layer for an input image in the classification network, G is the one-shot labelled ground truth corresponding to the input image, and c is a class label. The maximum value of c was determined to be 3 because there are three classes in the classification task. In contrast, the loss function for a detection network was designed based on the loss function introduced in YOLO [36]. The ground truths for a detection task were not manufactured by medical experts, but the CAMs generated by the classification network were used as ground truths. The following equation demonstrates the loss function for the detection network:

$$\mathcal{L}_2 = \mathcal{L}_{\mathrm{YOLO}}(I, f_{\mathrm{qs};15}(f_{\mathrm{CAM}}(I))) \quad (6)$$

where the first and the second parameters of $\mathcal{L}_{\mathrm{YOLO}}$ are an input image and a ground truth, respectively. Here, the points of objects are generated using the bounding boxes of an arbitrary topological closure, which is revealed by a contour extraction algorithm, and the class of the bounding box is always regarded as a disease category. Furthermore, the optimization of the AFS block is configured as follows:

$$\mathcal{L}_3 = ||SN(Q_1) - SN(Q_2)||_2 + ||SN(Q_1) - SN(Q_3)||_2 \quad (7)$$

where $SN(X)$ indicates the Siamese network. The loss function is designed based on a triplet of three inputs. However, since the feature maps of $SN(Q_1)$, $SN(Q_2)$, and $SN(Q_3)$) generated by the Siamese network could be the same, the L2 loss function is used to construct the Siamese loss function ($\mathcal{L}_3$). Therefore, the following loss function is used to optimize all parameters of FAA-Net:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \alpha \sum^{Q} \sum^{K} \mathcal{L}_3 \quad (8)$$

where Q indicates input feature maps of AFS-Blocks, and K indicates the number of images for each class in the RAC, which corresponds to K-shot learning. In addition, α is an unchangeable parameter to compensate for the large scale of $\mathcal{L}_3$. The value of α is determined to be 0.1, since $\mathcal{L}_3$ is much larger than $\mathcal{L}_1$ and $\mathcal{L}_2$.

The inference steps using FAA-Net are as follows. Using input images, a classification network generates provisional predictions, as well as their corresponding CAMs. Using these, the AFS-Blocks in a detection network are fine-tuned. $\mathcal{L}_2$ and $\mathcal{L}_3$ are used to fine-tune the network. The parameters of the AFS-Blocks in the detection network are shared with the classification network. This step represents the fine-tuning stage of FAA-Net. Finally, the fine-tuned classification network classifies the input images into the skin disease categories. Furthermore, the diagnosis pipelines using our developed system are to acquire skin images from patients or subjects using the system whereas the further steps are to diagnose skin diseases using the FAA-Net. Note that the diagnostic phase using FAA-Net involves a fine-tuning phase that is distinct from other deep learning models.

V. Experiments

The following experiments were performed to evaluate the performance of FAA-Net: 1) An ablation study was performed to investigate the effect of each module on FAA-Net, as well as the effect of fluorescence images with white-light RGB images being supplied as an input to FAA-NET. 2) A quantitative analysis was conducted to evaluate the classification performance of FAA-Net for the diagnosis of skin diseases compared with other state-of-the-art networks. Algorithm performance was analyzed using a receiver operating characteristic (ROC) curve and confusion matrix.

A. Experimental Setup

In the experiments, FAA-Net was trained and tested using the constructed dataset illustrated in Section III-B. White-light RGB and fluorescence images captured by the proposed system are shown in Fig. 5. A total of 51 patients and 427 images were split into training, tuning, and testing sets, and 5-fold for cross-validation was applied [38]. Since images acquired from the

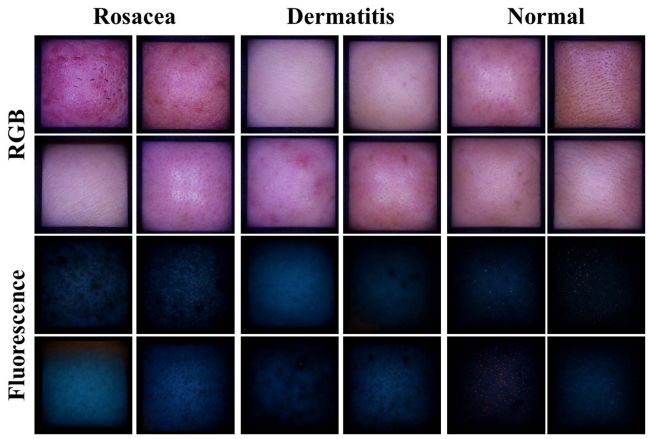

Fig. 5. RGB and fluorescence sample images captured by our system from different patients.

TABLE II
DETAILED DESCRIPTION OF DATASET SPLITS

# persons	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Total
Rosacea	4	3	3	3	3	16
Dermatitis	3	3	3	3	3	15
Normal	4	4	4	4	4	20
Total	11	10	10	10	10	51
# images	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Total
Rosacea	24	24	25	26	24	123
Dermatitis	20	19	22	21	22	104
Normal	40	40	40	40	40	200
Total	83	83	87	87	87	427

same patient tend to have similar characteristics, the dataset was split into patients rather than images. Table II provides a detailed description of the dataset splits.

Furthermore, data augmentation, such as rotation and flip, were applied to the dataset [39]. One image was augmented to eight images by applying left, right, and vertical reversals and 90°, 180°, and 270° rotations simultaneously. Since the field of view was fixed in the dataset, augmentations involving zoom or scale changes were not applied.

The experiments were conducted using the public platform TensorFlow version 1. All deep learning networks were implemented using TensorFlow, and the implemented deep learning models were trained using an Adam optimizer with a stochastic gradient descent (SGD) method and a mini-batch size of 5. Hyperparameters were applied to all deep learning networks with the same values. For instance, the decay rates (β_1 and β_2) for the Adam optimizer were set to 0.9 and 0.999, respectively, and ϵ was 1e-8. In addition, the learning rate was initially set to 1e-3 and then divided by two every 50 epochs. Group normalization was applied instead of batch normalization since the batch size was too small. The G value for group normalization was 16, and the initial values of group normalization were set to exhibit a Gaussian distribution. In the experiments, the few-shot learning models (those of Zhu, Mahajan, and Wertheimer) were trained in the manner illustrated in their own studies. In contrast, other deep learning models applied early stopping by checking the loss value every 50 epochs. Therefore, FAA-Net and the models of Yauney and Yap were trained for 250 epochs and then stopped early. In contrast, the models of Pan and Lanchamtin were trained for 300 epochs, and the others were trained for 350 epochs. The TensorFlow server consisted of two-way CPUs, 32 GB of RAM, and four-way Titan XP (11 GB) GPUs.

TABLE III
RESULTS OF ABLATION STUDY

Normal	Sensitivity	Specificity	F1	Accuracy	AUC
With F, AFS-7	63.81%	**89.11%**	**72.49%**	**77.21%**	**0.766**
With F, AFS-5	**64.50%**	84.94%	71.10%	75.32%	0.743
With F, AFS-3	62.87%	81.89%	68.62%	72.94%	0.726
No F, AFS-7	43.62%	77.78%	51.74%	61.71%	0.602
No F, AFS-5	49.94%	77.28%	56.91%	64.41%	0.635
No F, AFS-3	43.69%	69.39%	49.05%	57.29%	0.562
With F, No AFS	46.38%	75.33%	53.27%	61.71%	0.605
No F, No AFS	37.81%	72.22%	44.73%	56.03%	0.555
Dermatitis	Sensitivity	Specificity	F1	Accuracy	AUC
With F, AFS-7	**75.22%**	83.36%	**67.94%**	**81.21%**	**0.798**
With F, AFS-5	68.22%	**84.80%**	64.84%	80.41%	0.762
With F, AFS-3	62.44%	79.80%	57.14%	75.21%	0.711
No F, AFS-7	57.22%	71.44%	48.38%	67.68%	0.643
No F, AFS-5	53.67%	73.20%	47.05%	68.03%	0.647
No F, AFS-3	44.44%	76.04%	42.13%	67.68%	0.595
With F, No AFS	52.00%	73.76%	46.25%	68.00%	0.638
No F, No AFS	45.00%	69.72%	39.28%	63.18%	0.581
Rosacea	Sensitivity	Specificity	F1	Accuracy	AUC
With F, AFS-7	**76.33%**	**83.88%**	**69.05%**	**81.88%**	**0.804**
With F, AFS-5	71.00%	81.44%	63.80%	78.68%	0.763
With F, AFS-3	61.22%	82.00%	57.97%	76.50%	0.713
No F, AFS-7	50.00%	75.08%	45.62%	68.44%	0.623
No F, AFS-5	52.78%	77.44%	48.99%	70.91%	0.647
No F, AFS-3	48.00%	71.24%	42.13%	65.09%	0.594
With F, No AFS	48.44%	73.84%	43.82%	67.12%	0.611
No F, No AFS	43.56%	70.36%	38.56%	63.26%	0.569

The Highest Value is Highlighted in **Bold**, and the Second Highest Value is Underlined

B. Ablation Study

We performed an ablation study to evaluate the effect of AFS-Blocks on the performance of FAA-Net and analyze the effect of the use of fluorescence images. In addition, to search for the optimal number of k for few-shot learning, different values of k were investigated, specifically 3, 5, and 7. Table III presents the results of the ablation study for FAA-Net. In the table, "*With F*" indicates that FAA-Net uses fluorescence images with RGB images, and "*AFS-k*" indicates the value of k for few-shot learning. Note that k refers to the number of images collected in the RAC for each skin disease category. As illustrated in Table III, the ablation study demonstrated that the performance of FAA-Net was improved by using fluorescence images, AFS-Blocks, and a large number of images in the RAC. By using the fluorescence images as inputs, the performance of FAA-Net can be improved in most cases, especially when the number of images (k) stored in the RAC increases. Image stored in the RAC represent a disease-like region, and with an increasing number of images, the disease-like region extracted in advance could help improve the attention of a deep learning model when feature extraction was carried out in the prediction phase.

Fig. 6 illustrates a confusion matrix obtained using different versions of FAA-Net shown in the ablation study. The dark-blue color indicates a higher portion of predicted results, and

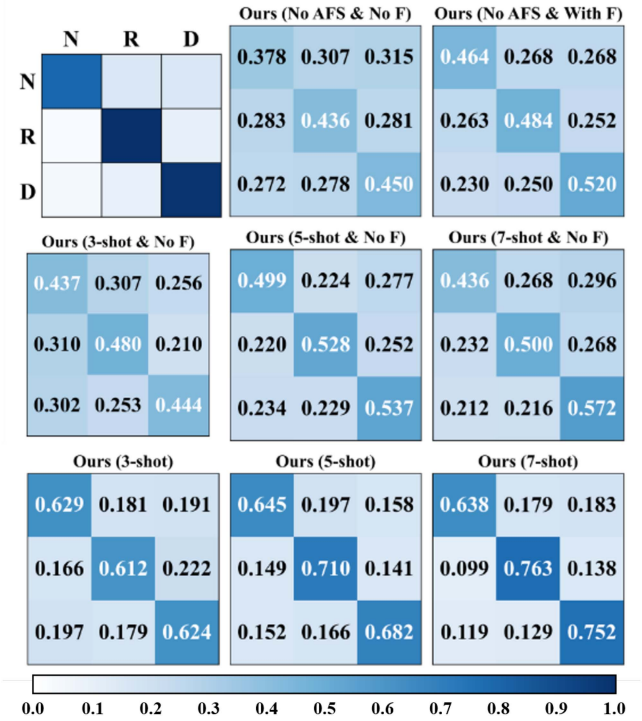

Fig. 6. Confusion matrix for different versions of FAA-Net. N, R, and D indicate Normal, Rosacea, and Dermatitis classes, respectively.

the white color indicates the spare portion. The diagonal cells in the matrix indicate correct predictions, and the other cells indicate mispredictions. Therefore, high contrast in the diagonal direction of the confusion matrix suggests that the predictive performance is better. In Fig. 6, FAA-Net using fluorescence images as inputs showed higher contrast at the diagonal direction in the predictions than FAA-Net without fluorescence images. Thus, the use of fluorescence images enhances FAA-Net for skin diagnosis. In addition, Fig. 6 demonstrates that the performance of FAA-Net improves when FAA-Net uses fluorescence images and AFS-Blocks with a larger number of images stored in the RAC.

C. Quantitative Comparison Analysis

To verify the superior performance of FAA-Net compared with other state-of-the-art networks for the multicategorical classification of skin diseases, the deep learning networks proposed by the following authors were used: Yauney [4], Pan [5], Yap [18], Zhu [25], Yang [19], Lanchamtin [23], Mahajan [26], Wertheimer [24], and Rahman [20]. While FAA-Net uses multimodal autofluorescence and white-light RGB images, other deep learning networks use only one type of image to diagnose skin diseases. Therefore, the inputs to other state-of-the-art networks were constructed by concatenating white-light RGB and fluorescence images; thus, the channels of the inputs were six; of which three from white-light RGB images and another three from fluorescence images.

Table IV summarizes the quantitative analysis results for FAA-Net and the other models. FAA-Net outperformed other

TABLE IV
QUANTITATIVE COMPARISON OF FAA-NET AND OTHER STATE-OF-THE-ART MODELS

NORMAL	Sensitivity	Specificity	F1	Accuracy	AUC
Ours (7-shot)	63.81%	89.11%	72.49%	77.21%	0.766
Ours (5-shot)	64.50%	84.94%	71.10%	75.32%	0.743
Ours (3-shot)	62.87%	81.89%	68.62%	72.94%	0.726
Ours (only F)	46.38%	75.33%	53.27%	61.71%	0.605
G.Yauney	60.06%	81.50%	66.41%	71.41%	0.713
S.Pan	61.31%	80.50%	66.92%	71.47%	0.716
J.Yap	64.31%	81.50%	69.48%	73.41%	0.734
Y.Zhu	61.19%	81.89%	67.40%	72.15%	0.718
X.Yang	63.63%	82.00%	69.20%	73.35%	0.727
J.Lanchamtin	60.63%	81.00%	66.62%	71.41%	0.699
K.Mahajan	49.44%	75.33%	55.80%	63.15%	0.628
D.Wertheimer	43.00%	76.00%	50.59%	60.47%	0.599
Z.Rahman	58.75%	79.67%	64.69%	69.82%	0.692
Improvement	21.50%	13.78%	21.90%	16.74%	16.65%
DERMATITIS	Sensitivity	Specificity	F1	Accuracy	AUC
Ours (7-shot)	75.22%	83.36%	67.94%	81.21%	0.798
Ours (5-shot)	68.22%	84.80%	64.84%	80.41%	0.762
Ours (3-shot)	62.44%	79.80%	57.14%	75.21%	0.711
Ours (only F)	52.00%	73.76%	46.25%	68.00%	0.638
G.Yauney	62.67%	80.88%	58.08%	76.06%	0.721
S.Pan	62.44%	80.76%	57.85%	75.91%	0.715
J.Yap	61.22%	81.96%	57.94%	76.47%	0.719
Y.Zhu	62.67%	81.20%	58.32%	76.29%	0.730
X.Yang	65.89%	81.72%	60.82%	77.53%	0.740
J.Lanchamtin	58.89%	81.24%	55.82%	75.32%	0.702
K.Mahajan	50.56%	74.28%	45.55%	68.00%	0.618
D.Wertheimer	51.78%	70.80%	44.47%	65.76%	0.616
Z.Rahman	57.56%	77.72%	52.46%	72.38%	0.679
Improvement	24.67%	14.00%	23.47%	15.44%	18.26%
ROSACEA	Sensitivity	Specificity	F1	Accuracy	AUC
Ours (7-shot)	76.33%	83.88%	69.05%	81.88%	0.804
Ours (5-shot)	71.00%	81.44%	63.80%	78.68%	0.763
Ours (3-shot)	61.22%	82.00%	57.97%	76.50%	0.713
Ours (only F)	48.44%	73.84%	43.82%	67.12%	0.611
G.Yauney	62.00%	79.76%	56.82%	75.06%	0.709
S.Pan	60.56%	80.80%	56.62%	75.44%	0.711
J.Yap	63.56%	81.44%	59.09%	76.71%	0.731
Y.Zhu	65.89%	81.28%	60.48%	77.21%	0.739
X.Yang	65.11%	83.12%	61.43%	78.35%	0.744
J.Lanchamtin	64.22%	79.56%	58.12%	75.50%	0.722
K.Mahajan	53.00%	76.40%	48.50%	70.21%	0.644
D.Wertheimer	51.11%	75.04%	46.37%	68.71%	0.638
Z.Rahman	59.78%	80.76%	56.07%	75.21%	0.706
Improvement	25.22%	8.84%	22.67%	13.18%	16.53%

models in most cases. In particular, the accuracy values of FAA-Net trained with 7-shot learning increased to 16.74%, 15.44%, and 13.18% for the NORMAL, DERMATITIS, and ROSACEA classes, respectively. Similarly, the area under the curve (AUC) values of FAA-Net trained with 7-shot learning were 16.65%, 18.26%, and 16.53% higher than those of other state-of-the-art models for the NORMAL, DERMATITIS, and ROSACEA classes, respectively. These results confirm that FAA-Net with fluorescence images and AFS-Blocks achieves improved multiclass classification performance. Furthermore, FAA-Net demonstrated significantly improved sensitivity compared with the other models in the skin diseases classification.

Fig. 7 presents the confusion matrix for FAA-Net and other networks. FAA-Net with AFS-Blocks trained by seven-shot fluorescence and white-light RGB images exhibited higher contrast in the diagonal of the confusion matrix, thus demonstrating that it outperformed the other networks in the classification tasks. Furthermore, Fig. 8 presents the ROC curves for FAA-Net and the other networks. The quantitative comparisons demonstrate

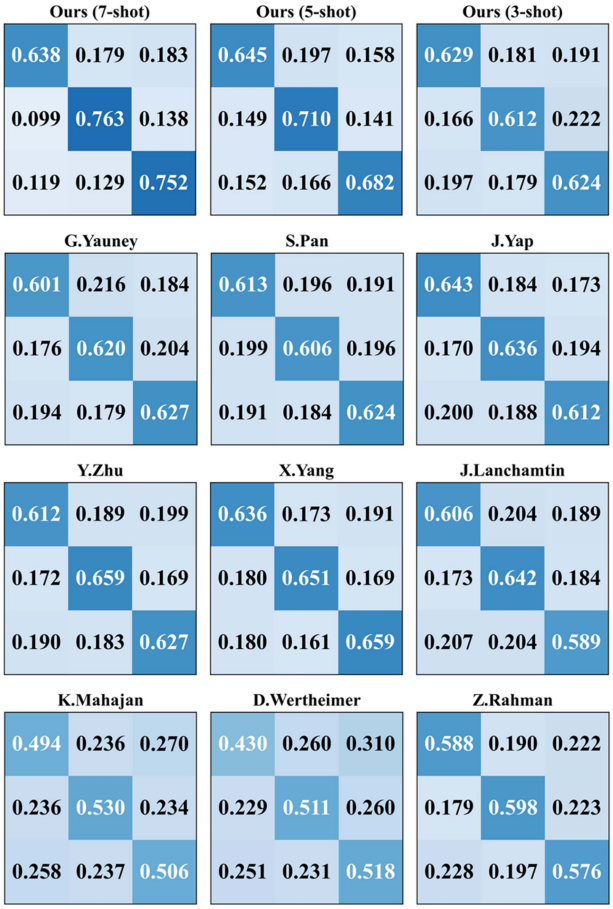

Fig. 7. Confusion matrix by each deep learning network. Dark blue color indicates a higher portion, and a white color refers to a sparse portion.

the outstanding performance of FAA-Net compared to the other state-of-the-art deep learning models.

D. Qualitative Analysis of Explainability

FAA-Net can be regarded as unsupervised learning by using the activation map as the ground truth to optimize detection networks. Therefore, it is necessary to ensure that the generated activation map truly represents the disease areas [1]. Fig. 9 shows the input image and the corresponding activation maps: CAM [2], LIFT-CAM [3], RISE [4], and Grad-CAMs [5]. In addition, a board-certified dermatologist examined the CAM from all images of the patients. The results showed that the generated activation representations were in good agreement with the results obtained by a dermatologist, indicating that FAA-Net exhibited good diagnostic performance, even in terms of unsupervised learning. Note that, since rosacea not only appears on a wider range of skin than dermatitis, but also affects skin color more than dermatitis, the activation representations of rosacea exhibit a wider distribution than those of dermatitis.

VI. Discussion

The experiment confirmed that the performance of FAA-Net significantly improved as the value of k increased. When the

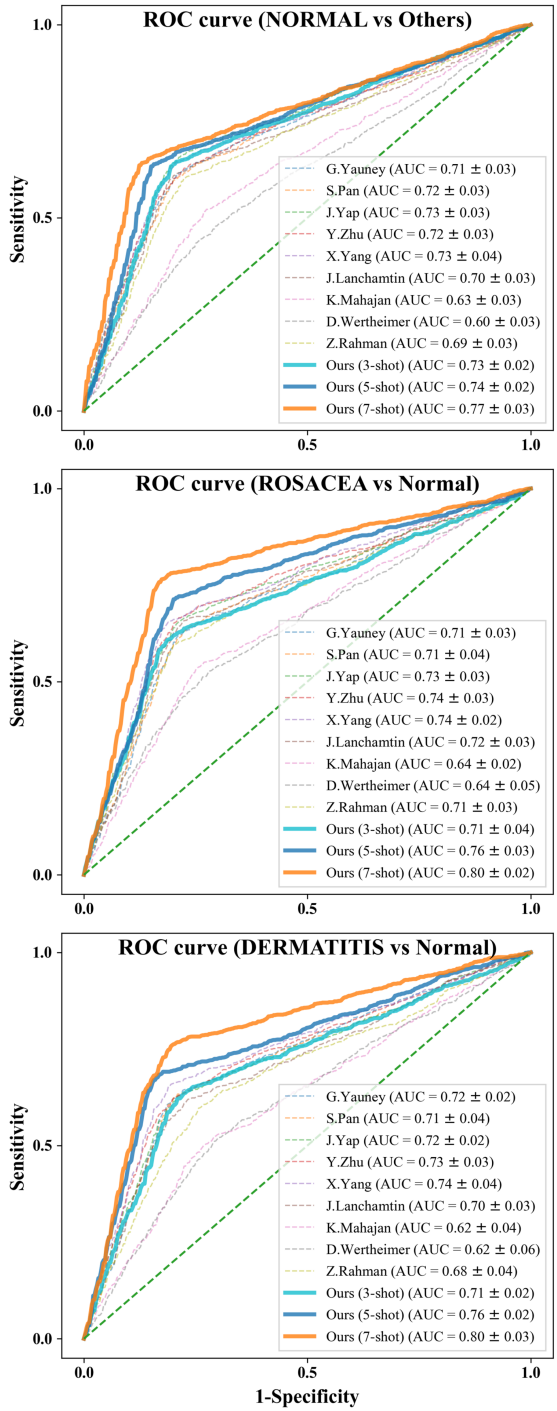

Fig. 8. Receiving operating characteristics (ROC) curve by deep learning networks. The numbers in the legend field indicate the average AUC and the standard deviation.

value of k increases, the number of images stored in the RAC could be increased. Therefore, the input images can be compared with more stored feature maps that are likely to be diseases from the point of view of few-shot learning comparing similarity, which may enhance the performance of FAA-Net. However, FAA-Net was designed considering the possibility of real-time applications. The structure of AFS-Blocks demonstrated that if the value of k increases, the number of channels to be calculated

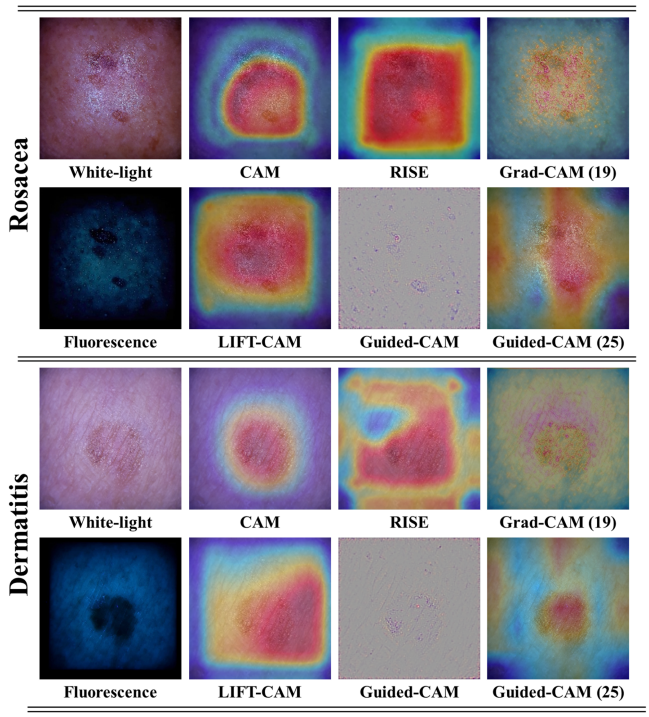

Fig. 9. Explainability representations of the FAA-Net. *Grad-CAM* (n) indicates the Grad-CAM of n^{th} layer in Fig. 4 (a). The red color indicates the highlighted area whereas the blue color indicates the region that is not importantly considered by the FAA-Net.

also increases, resulting in an increase in the computation time of FAA-Net, which is not suitable for real-time applications. Therefore, the value of k was determined to be at most seven to make the calculation time of FAA-Net suitable for real-time applications.

Another limitation is that the number of images acquired by the developed system was small. However, we found that the proposed algorithm, constructed using a limited dataset, achieved excellent performance in classifying multiple skin diseases. In addition, the deep learning-based algorithm with fluorescence and white-light RGB imaging further enhanced the performance even with a small number of images, implying that it can be employed to address unmet clinical needs in limited environments for the data acquisition of skin diseases.

In addition, if we suppose that the baseline deep learning model costs $O(N)$ time complexity and $O(M)$ space complexity (where M is the number of parameters), and that the classification and detection models share the same parameters without FC layers, as illustrated in Fig. 4(a), FAA-Net has the same space complexity as the baseline model. However, as illustrated in Fig. 3(b), prediction is performed twice in the detection network and the classification network, and further fine-tuning is performed after the prediction by the detection network. Therefore, the time complexity of FAA-Net is $O(kN)$ where $k \geq 2$, which implies that FAA-Net incurs a heavy computational cost. To embed the deep learning models in a mobile environment, which contains less computational power than a PC environment, the heavy computation of the proposed model should be improved in future work.

Furthermore, although DarkNet is an old-fashioned architecture, it was used as the baseline architecture of FAA-Net owing to its feasible use for both classification and detection tasks. However, a more advanced network for classification and detection tasks with better performance could be applied as a baseline model for FAA-Net. This remains a topic for future research. In this study, we emphasize the potential of the novel FAA-Net incorporated with a smartphone-based fluorescence imaging system for the mobile diagnosis of multiple skin diseases.

VII. Conclusion

In this study, we developed a novel deep learning model called FAA-Net with AFS-Blocks integrated with a smartphone-based fluorescence imaging system to classify multicategorical skin diseases. The developed fluorescence imaging system allows users to acquire white-light RGB and fluorescence images. Simultaneously, FAA-Net, which receives white-light RGB and fluorescence images as inputs, can improve diagnostic performance by adopting AFS-Blocks. These enable the network to store disease-like features and reuse those features in the diagnostic phase. In addition, FAA-Net is designed based on both few-shot learning and multi-task learning to guarantee high performance even in limited environments with few training images. The results shown in this study demonstrate that FAA-Net outperformed other state-of-the-art models for the classification of multicategorical skin diseases. FAA-Net's mean accuracy and AUC values for a skin disease classification task were 80.10% and 78.93%, respectively, which were 15.12% and 17.17% higher than those of the other models. The ablation study demonstrated that AFS-Blocks in combination with fluorescence and white-light RGB images significantly enhanced the accuracy of FAA-Net. Furthermore, it was confirmed that the fluorescence images obtained using the developed imaging system are important for improving the performance of FAA-Net for the diagnosis of skin diseases. Finally, the proposed system and diagnostic deep learning model can also be applied to the multiclass classification of various skin diseases in addition to those illustrated.

References

[1] D. H. Lee, K. Li, and D. H. Suh, "Pimecrolimus 1% cream for the treatment of steroid-induced rosacea: An 8-week split-face clinical trial," *Brit. J. Dermatol.*, vol. 158, no. 5, pp. 1069–1076, 2008.

[2] T. A. Valdez et al., "Multiwavelength fluorescence otoscope for video-rate chemical imaging of middle ear pathology," *Anal. Chem.*, vol. 86, no. 20, pp. 10454–10460, 2014.

[3] N. Le, H. M. Subhash, L. Kilpatrick-Liverman, and R. K. Wang, "Non-invasive multimodal imaging by integrating optical coherence tomography with fluorescence imaging for dental applications," *J. Biophotonics*, vol. 13, no. 7, 2020, Art. no. e202000026.

[4] G. Yauney et al., "Convolutional neural network for combined classification of fluorescent biomarkers and expert annotations using white light images," in *Proc. IEEE 17th Int. Conf. Bioinf. Bioeng. (BIBE)*, 2017, pp. 303–309.

[5] S. Pan et al., "Multi-task learning-based immunofluorescence classification of kidney disease," *Int. J. Environ. Res. Public Health*, vol. 18, no. 20, 2021, Art. no. 10798.

[6] J. P. Celli et al., "Imaging and photodynamic therapy: Mechanisms, monitoring, and optimization," *Chem Rev.*, vol. 110, no. 5, pp. 2795–2838, 2010, doi: 10.1021/cr900300p.

[7] V. Rotemberg et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, 2021, Art. no. 34. https://doi.org/10.1038/s41597-021-00815-z
[8] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.
[9] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.,* 2017, pp. 1126–1135.
[10] T. Hospedales et al., "Meta-learning in neural networks: A survey," 2020, *arXiv:2004.05439*.
[11] H. Yaguchi, R. Tsuboi, R. Ueki, and H. Ogawa, "Immunohistochemical localization of basic fibroblast growth factor in skin diseases," *Acta Derm Venereol*. vol. 73, no. 2, pp. 81–83, Apr. 1983. doi: 10.2340/00015555738183.
[12] E.A. Shirshin et al., "Two-photon fluorescence lifetime imaging of human skin papillary dermis in vivo: Assessment of blood capillaries and structural proteins localization," *Sci. Rep.* vol. 7, 2017, Art. no. 1171. Online. [Available]: https://doi.org/10.1038/s41598-017-01238-w
[13] K. Farkas et al., "Fluorescence imaging of the skin is an objective non-invasive technique for diagnosing pseudoxanthoma elasticum," *Diagnostics*, vol. 11, no. 2, 2021, Art. no. 260.
[14] Y. Wang, E. Gutierrez-Herrera, A. Ortega-Martinez, R. R. Anderson, and W. Franco, "UV fluorescence excitation imaging of healing of wounds in skin: Evaluation of wound closure in organ culture model," *Lasers Surg. Med.*, vol. 48, no. 7, pp. 678–685, 2016.
[15] S. M. Odeh, F. de Toro, I. Rojas, and M. J. Saéz-Lara, "Evaluating fluorescence illumination techniques for skin disease diagnosis," *Appl. Artif. Intell.*, vol. 26, no. 7, pp. 696–713, 2012.
[16] S. M. Odeh, E. Ros, I. Rojas, and J. M. Palomares, "Skin disease diagnosis using fluorescence images," in *Proc. Int. Conf. Image Anal. Recognit.*, 2006, pp. 648–659.
[17] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin diseases characterization," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 5, pp. 721–733, 2009.
[18] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin disease classification using deep learning," *Exp. Dermatol.*, vol. 27, no. 11, pp. 1261–1267, 2018.
[19] X. Yang, Z. Zeng, S. Y. Yeo, C. Tan, H. L. Tey, and Y. Su, "A novel multi-task deep learning model for skin disease segmentation and classification," 2017, *arXiv:1703. 01025*.
[20] Z. Rahman, M. S. Hossain, M. R. Islam, M. M. Hasan, and R. A. Hridhee, "An approach for multiclass skin disease classification based on ensemble learning," *Informat. Med. Unlocked*, vol. 25, 2021, Art. no. 100659.
[21] F. X. Marin-Gomez, J. Vidal-Alaball, P. R. Poch, C. J. Sariola, R. T. Ferrer, and J. M. Pena, "Diagnosis of skin lesions using photographs taken with a mobile phone: An online survey of primary care physicians," *J. Primary Care Community Health*. vol. 11, Jan. 2020, Art. no. 2150132720937831.
[22] C. A. Hartanto and A. Wibowo, "Development of mobile skin cancer detection using faster R-CNN and MobileNet v2 model," in *Proc. 7th Int. Conf. Inf. Technol., Computer, Elect. Eng. (ICITACEE)*, 2020, pp. 58–63, doi: 10.1109/ICITACEE50144.2020.9239197.
[23] J. Lanchantin et al., "General multi-label image classification with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16478–16488.
[24] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8012–8021.
[25] Y. Zhu, C. Liu, and S. Jiang, "Multi-attention meta learning for few-shot fine-grained image recognition," in *Proc. IJCAI*, 2020, pp. 1090–1096.
[26] K. Mahajan, M. Sharma, and L. Vig, "Meta-DermDiagnosis: Few-shot skin disease identification using meta-learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 730–731.
[27] S. M. Mathews, C. Kambhamettu, and Kenneth E. Barner, "A novel application of deep learning for single-lead ECG classification," *Comput. Biol. Med.*, vol. 99, pp. 53–62, 2018.
[28] S. M. Mathews, C. Kambhamettu, and K. E. Barner, "Centralized class specific dictionary learning for wearable sensors based physical activity recognition," in *Proc. 51st Annu. Conf. Inf. Sci. Syst.*, 2017, pp. 1–6, doi: 10.1109/CISS.2017.7926100.
[29] S. M. Mathews, C. Kambhamettu, and K. E. Barner, "Maximum correntropy based dictionary learning framework for physical activity recognition using wearable sensors," in *Proc. Int. Symp. Vis. Comput.*, 2016, pp. 123–132.
[30] S. M. Mathews, "Dictionary and deep learning algorithms with applications to remote health monitoring systems," in *ProQuest Diss. & Theses Glob.,* 2017.
[31] J. Sun et al., "Unsupervised representation learning meets pseudo-label supervised self-distillation: A new approach to rare disease classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 519–529.
[32] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2537–2546.
[33] G. Koch, R. Zemel, and R. Salakhutdinov, and Others, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, 2015, pp. 17–21.
[34] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, to be published, doi: 10.1109/TKDE.2021.3070203.
[35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
[36] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[37] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.
[38] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Roy. Stat. Soc.: Ser. B. (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.
[39] S. M. Piryonesi and T. E. El-Diraby, "Data analytics in asset management: Cost-effective prediction of the pavement condition index," *J. Infrastructure Syst.*, vol. 26, no. 1, 2020, Art. no. 04019036.
[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representative*, 2015.
[41] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186.
[42] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[44] S. M. Mary, "Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review," in *Proc. Intell. Comput.-Proc. Comput. Conf.*, 2019, pp. 1269–1292.
[45] H. Jung and Y. Oh, "Towards better explanations of class activation mapping," in *Proc. IEEE/CVE Int. Conf. Comput. Vis.*, 2021.
[46] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.
[47] 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.