IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 8, 2021, accepted January 20, 2021, date of publication January 25, 2021, date of current version February 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3054403

Robust Skin Disease Classification by Distilling Deep Neural Network Ensemble for the Mobile Diagnosis of Herpes Zoster

SEUNGHYEOK BACK[1], (Graduate Student Member, IEEE), SEONGJU LEE[1], SUNGHO SHIN[1], YEONGUK YU[1], TAEKYEONG YUK[1,2], SAEPOMI JONG[2], SEUNGJUN RYU[3], AND KYOOBIN LEE[1], (Member, IEEE)

[1]School of Integrated Technology, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea
[2]Research Department, Unitria Inc., Gwangju 61177, South Korea
[3]Department of Neurosurgery, Yonsei University Health System, Yonsei University College of Medicine, Seoul 03722, South Korea

Corresponding authors: Kyoobin Lee (kyoobinlee@gist.ac.kr) and Seungjun Ryu (pedisj@yuhs.ac)

This work was supported in part by the Institute for Information & Communications Technology Promotion (IITP) Grant funded by Korea Government (MSIT) (No.2019-0-01335, Development of AI technology to generate and validate the task plan for assembling furniture in the real and virtual environment by understanding the unstructured multi-modal information from the assembly manual).

ABSTRACT Herpes zoster (HZ) is a common cutaneous disease affecting one out of five people; hence, early diagnosis of HZ is crucial as it can progress to chronic pain syndrome if antiviral treatment is not provided within 72 hr. Mobile diagnosis of HZ with the assistance of artificial intelligence can prevent neuropathic pain while reducing clinicians' fatigue and diagnosis cost. However, the clinical images captured from daily mobile devices likely contain visual corruptions, such as motion blur and noise, which can easily mislead the automated system. Hence, this paper aims to train a robust and mobile deep neural network (DNN) that can distinguish HZ from other skin diseases using user-submitted images. To enhance robustness while retaining low computational cost, we propose a knowledge distillation from ensemble via curriculum training (KDE-CT) wherein a student network learns from a stronger teacher network progressively. We established skin diseases dataset for HZ diagnosis and evaluated the robustness against 75 types of corruption. A total of 13 different DNNs was evaluated on both clean and corrupted images. The experiment result shows that the proposed KDE-CT significantly improves corruption robustness when compared with other methods. Our trained MobileNetV3-Small achieved more robust performance (93.5% overall accuracy, 67.6 mean corruption error) than the DNN ensemble with smaller computation (549x fewer multiply-and-accumulate operations), which makes it suitable for mobile skin lesion analysis.

INDEX TERMS Biomedical image processing, convolutional neural networks, deep learning, dermatology.

I. INTRODUCTION

Herpes zoster (HZ) is a virus-induced skin disease characterized by a painful rash accompanied by blisters. The occurrence of HZ in a lifetime is 10%–30% [1], [2]. However, if proper antiviral treatments are not provided within seventy-two hours after the onset of a rash, it can progress to chronic disease with severe pain [3]. Thus, early diagnosis of HZ is crucial for a complete recovery; otherwise, it leads to severe complications. Despite its frequent occurrence and severity, most people are unaware of their risk for HZ [4]. The onset of HZ is accompanied by mild symptoms such as fever, itching, and chills and it can progress to persistent neuropathic pain. HZ decreases the quality of life, which severely affects an individual's sleep and social activities [5].

Despite its clinical importance, there has been little focus on the automated system that can diagnose HZ using only clinical images. To date, most studies related to automated skin lesion diagnosis have focused on melanoma [6], and early approaches adopted machine learning methods [7], [8]. Recently, convolutional neural network (CNN) has been utilized with a large amount of data and it has achieved superior performance comparable to that of actual dermatologists [9], [10]. However, these studies only focused on dermatoscopic

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

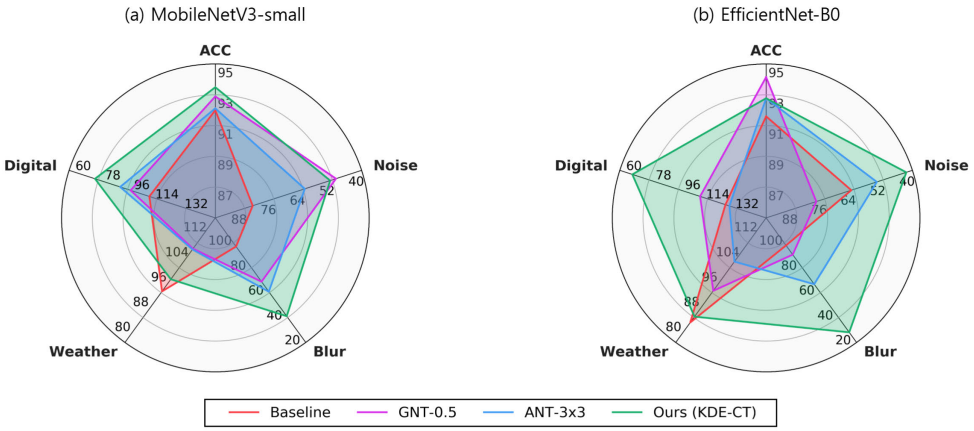

FIGURE 1. Summary of key results. Our knowledge distillation from ensemble via curriculum training (KDE-CT) enhanced the corruption robustness for both (a) MobileNetV3-Small and (b) EfficientNet-B0 compared to the standard training (baseline), Gaussian noise training (GNT), and adversarial noise training (ANT). ACC is the overall accuracy, and the rest are the corruption errors for noise, blur, weather, and digital corruptions.

images [6], which are collected under controlled environments using a fixed camera configuration and a dermatoscope. For easy and accessible diagnosis of HZ, an automated skin analysis system can be used on clinical images in mobile environments. If artificial intelligence (AI) can diagnose HZ from user-submitted skin images captured using a mobile camera, users can quickly submit their skin images and thus increase their chances of having the skin disease examined before symptoms are noticed. Considering the desire of those who do not wish for their private skin images to be exposed, it is not rational to diagnose the skin images on the server. Instead, the user is recommended to self-diagnose the skin condition offline using their own mobile devices. This requires a mobile model that can accurately and robustly diagnose skin diseases in low-quality images, which may contain noise and involve various illumination conditions.

In recent years, some studies have considered clinical images to classify skin lesion [11], [12]. Liu *et al.* [11] collected a large number of skin images from 16,114 cases, including both dermatoscopic and clinical images for the differential diagnosis of 26 skin conditions using deep learning. Han *et al.* [12] trained a DNN using 220,680 clinical images to suggest the treatment options for 134 skin disorders. Although the aforementioned models achieved promising performance and were extensively applied in the clinical domain, it is difficult to incorporate these models directly into mobile devices owing to their high computational costs. Specifically, [11] utilizes a large neural network consisting of one to six Inception-V4 networks [13], and [12] uses an ensemble of four different convolutional networks. To reduce the network size in skin lesion analysis, neural architecture search (NAS) [14] can be applied, but it requires a convoluted optimization process based on trial and error, depending on the task and datasets. However, our model utilizes knowledge distillation (KD), which makes it highly compatible with various models, datasets, and NAS, while reducing the computational cost and ensuring high accuracy.

For mobile diagnosis of skin diseases, the robustness of DNNs against possible visual corruptions should also be considered, as the skin disease classification performance of DNNs deteriorates in the case of low-quality images [12]. Mishra *et al.* [15] evaluated the robustness of DNNs against expected user noise and reported a significant reduction in the overall performance. However, only two types of visual corruption (i.e., shot noise and Gaussian blur) were examined. By contrast, we conducted an in-depth investigation of the effect of 75 types of visual corruptions on the network performance, following the protocol in [16]; such an investigation has not been conducted thus far in the context of skin lesion classification.

In this study, we aimed to develop an accurate, robust, and mobile model that can diagnose HZ from clinical images. To the best of our knowledge, this is the first study to focus on skin disease classification from clinical images using KD to train a mobile network effectively. The main contributions of this study are summarized as follows:

- Benchmarking of various DNNs, ranging from small mobile models to large ensemble models, for diagnosis of HZ from clinical skin images.
- Assessing the corruption robustness of DNNs by introducing the mean corruption error (mCE) in skin disease classification with 75 types of visual corruptions.
- Proposing an effective training strategy, KD from ensemble via curriculum training (KDE-CT), to enhance the robustness of DNNs against visual corruptions as well as the performance on clean images via progressive teacher selection.
- Achieving robust performance with an accuracy of 93.5% and mCE of 67.6 using KDE-CT on MobileNetV3-Small (Fig. 1), which can be applied to AI-based mobile prescreening for HZ diagnosis.

II. RELATED WORK

A. COMPUTER-AIDED DIAGNOSIS OF SKIN LESION

Computer-aided diagnosis (CAD) of skin lesions can reduce the diagnostic cost and improve the reliability of diagnosis by assisting dermatologists or physicians in decision-making [17]. Many CAD methods have been proposed to diagnose skin lesions [6], focusing mainly on the detection of melanoma from dermatoscopic images [18], [19]. Recently, a deep neural network (DNN) was utilized, and Esteva *et al.* [9] showed that a CNN can classify skin cancer as effectively as dermatologists can. Haenssle *et al.* [10] compared the performances of CNNs and 58 dermatologists in terms of melanoma detection using dermatoscopic images and reported that the CNN outperformed most dermatologists. For the analysis of skin lesions, an ensemble of neural networks is considered as a state-of-the-art strategy as it can enhance the predictive performance of the neural networks by reducing inductive biases [20]. The top-scoring methods from the 2017–2019 ISIC Skin Lesion Classification Challenges [21]–[23] utilized the ensemble of DNN by aggregating multiple deep neural networks. Specifically, Gessert *et al.* [23] achieved the best performance in the 2019 ISIC Skin Lesion Classification Challenge by using an ensemble of multiple efficient networks [24] and SENet [25]. Perez. *et al.* conducted extensive experiments for melanoma classification [26] with 135 models, and they showed that the ensembles of neural networks perform significantly better than a single model, even when the models used for the ensemble are randomly selected. However, the ensemble of the neural networks leads to a heavy and larger model, which is infeasible for mobile on-device inference and mobile diagnostic services. In this study, we jointly utilized the ensemble techniques with KD, obtaining an overall performance gain for skin lesion classification with low computational cost.

B. SKIN LESION DIAGNOSIS FROM CLINICAL IMAGES

Clinical images can be a more convenient and economical option for tele-dermatology and mobile applications, as they do not require additional equipment and can be easily obtained by users by photographing skin lesions using smartphones. However, these user-submitted images are more challenging to analyze than dermoscopic images. Unlike dermoscopic images captured under a controlled environment and camera configuration using a digital dermatoscope, smartphone-captured images can contain an illumination variation, defocus, and motion blur [27]–[29]; This can alter the visual appearance of skin lesions, inreducing the effectiveness of segmentation and classification [30]. A typical method for enhancing the generalization ability of DNNs for input variations is to increase the amount of training data by collecting extensive data [11], [12] or employing various data augmentation techniques [31]. Liu *et al.* [11] classified 26 common skin conditions at a level comparable to that of board-certified dermatologists by using deep learning systems on large-scale datasets, including 16,114 clinical cases. Perez *et al.* [31] examined the proper data augmentation pipeline for melanoma classification and outperformed the top-ranked method in the 2017 ISIC Challenge without additional external data.

However, a skin lesion diagnosis system based on DNNs is still prone to the visual artifacts induced by the user when images are captured with noise and blur, and these input shifts can easily degrade the performance of neural networks [12], [15]. Han *et al.* [12] reported that although DNNs are trained with 220,680 clinical photographs, tend to misclassify when the input images are of low quality with blur or shade. This instability can reduce the reliability of the diagnosis system and make it impractical for using in mobile clinical diagnosis. Despite its importance, however, there has been little interest in the robustness of skin lesion diagnosis. Only Mishra *et al.* [15] evaluated the robustness of DNN against user noise with imitated conditions and showed a significant decrease in the overall performance. However, in-depth studies have not yet been conducted to assess the robustness against input visual corruption in the domain of skin lesion diagnosis. While [15] examined the robustness against only shot noise and Gaussian blur, we followed the standard method proposed in [16] to assess the robustness of DNN for skin lesion diagnosis against 75 types of input corruptions.

C. EFFICIENCY AND ROBUSTNESS OF DNNs

In mobile applications of DNNs, computational efficiency is crucial for satisfying the computational resources and latency requirements. Several studies have focused on model compression and acceleration [32]; KD [33] is a promising method for these tasks as it is easily extensible to other tasks. In KD, the knowledge of large teacher networks can be transferred to a small student network by minimizing the difference between the student and teacher networks in terms of logits or feature levels. Thus, KD can significantly reduce the number of parameters and the inference time by utilizing the mobile models as a large-scale substitute model. The use of KD has been extended to distillation of neural-network ensembles [34], [35]. In addition, it has been demonstrated that more informative and richer knowledge can be distilled from multiple teacher networks. In this study, we utilized various training strategies for KD of the DNN ensemble, which have not yet been applied in the domain of skin lesion classification.

Robustness is also significant for deep learning on mobile devices, especially for diagnosis of skin diseases wherein the decision has to be trustworthy and safe. A naive approach for increasing the robustness against input corruption is to apply data augmentation in the training phases. However, the pipeline of data augmentation should be appropriately determined after careful consideration, as data augmentation for a certain corruption type can degrade the robustness against other corruption types [36]. Several methods have been proposed to enhance robustness against corruptions; these include mixed data augmentation [37], adversarial noise

FIGURE 2. Examples images in SD-HZ dataset: (a) Acne (b) Herpes Zoster (c) Tinea (d) Other Disease.

training [38], and assembling CNN techniques with KD [39]. However, in the skin lesion analysis domain, wherein random nuisances are observed, the correlation between the network and metrics is difficult to analyze from the ImageNet perspective [26]. Furthermore, methods that improve corruption robustness have not yet been extensively tested for a skin lesion analysis. Unlike [40], which focuses on mitigating adversarial noises in dermoscopic images, we considered various types of corruption in clinical images, including noise, blur, weather, and digital corruptions, to enhance the robustness of DNN against noise that is likely to be generated by the user.

III. METHODS

A. DATASET

1) SD-HZ: SKIN DISEASE DATASET FOR THE HERPES ZOSTER

For the visual diagnosis of HZ from clinical skin images, we established the SD-HZ dataset based on two public datasets (i.e., SD-198 [28] and SD-260 [41]) and our custom dataset (i.e., HZ-W). Table 1 presents the number of images corresponding to each class of disease in each dataset used in the experiment. SD-198 [28] and SD-256 [41] are currently the most extensive public datasets for images of clinical skin diseases, including eczema, acne, and various malignant conditions. We selected SD-198 and SD-256 datasets in this study as the images from these two datasets are captured from mobile phones and digital cameras under various illuminations, and camera configurations for various skin types. Example images from these datasets are depicted in Fig. 2, which is well suited for our purpose to evaluate performances using user-submitted images under mobile skin diagnosis settings.

SD-198 contains 6,548 skin disease images from 198 classes, whereas SD-260 is an extended version of SD-198, which has more diversity and imbalance, consisting of 20,600 images and 260 classes. However, they contain only 24 and 12 HZ images, respectively, which are insufficient for a DNN to learn richer and comprehensible features. Thus, we constructed a custom dataset (i.e., HZ-W) for the diagnosis of HZ based on web crawling. We initially collected 746 images from Google and Bing using the keyword ''herpes zoster'', and then we manually removed the duplicate images using an automatic tool. Then, a clinician (S. Ryu) prudentially examined the entire images twice and filtered out 369 non-herpes images. As a result, we collected 377 clinical images of HZ under various light conditions and camera configurations, as shown in Fig. 3, which finely imitate the condition for the mobile diagnosis of HZ.

TABLE 1. Summary of the SD-HZ dataset used in this study: Number of images and the disease names corresponding to the class.

Class	Disease	SD-198	SD-260	HZ-W	SD-HZ (total)
Acne	Acne Keloidalis Nuchae, Acne Vulgaris, Steroid Acne, Pomade Acne	155	937	-	1,092
HZ	Herpes Zoster	24	12	377	413
Tinea	Tinea Manus, Tinea Versico, Tinea Cruris, Tinea Faciale, Tinea Corporis, Tinea Pedis	660	468	-	793
Other Diseases	187 diseases including Eczema, Ulcer, and Malignant Melanoma	6,074	-	-	6,074

2) SD-HZ-C: CORRUPTED SKIN DISEASE DATASET

In the mobile diagnosis of skin diseases, user-submitted images are likely to be exposed to various corruptions, such as noise and blur. Thus, the automated diagnosis system should be robust against user noise for reliable and consistent diagnostic results. To measure the robustness of the neural network to the input image corruptions, Hendricks *et al.* [16] proposed an mCE metric and ImageNet-C benchmark with standard corruptions. Following the protocol of [16], we created an SD-HZ-C dataset for the fair evaluation of the robustness of the automated skin diagnosis system. We applied a total of 75 corruptions on the test split of the SD-HZ dataset including 15 corruption types with five severity levels for each type. The four categories of corruptions applied to SD-HZ-C are noise (i.e., Gaussian noise, shot noise, impulse noise), blur (defocus blur, frosted glass blur, motion blur, zoom blur), weather (snow, frost, fog), and digital (brightness, contrast, elastic, pixelation, JPEG). Example images of SD-HZ-C are shown in Fig. 4, which mimics well the various realistic corruptions in the user-submitted images for skin disease diagnosis.

B. DEEP NEURAL NETWORKS

For the classification of clinical skin images, we evaluated 13 different DNN models belonging to three categories.

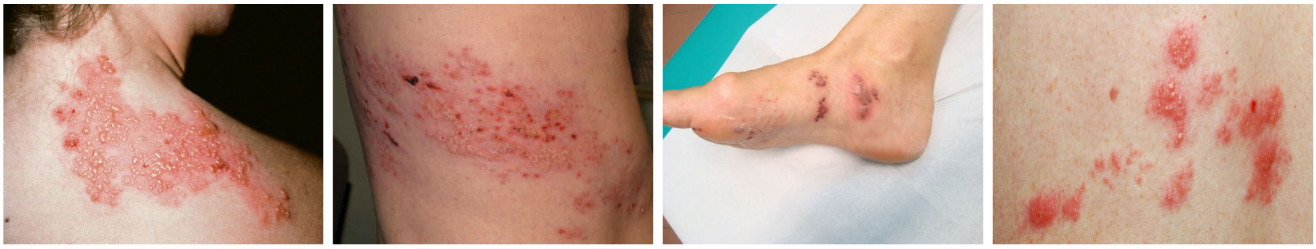

FIGURE 3. Examples of HZ image in HZ-W where light condition and camera configurations are varied (from close-up to long shot).

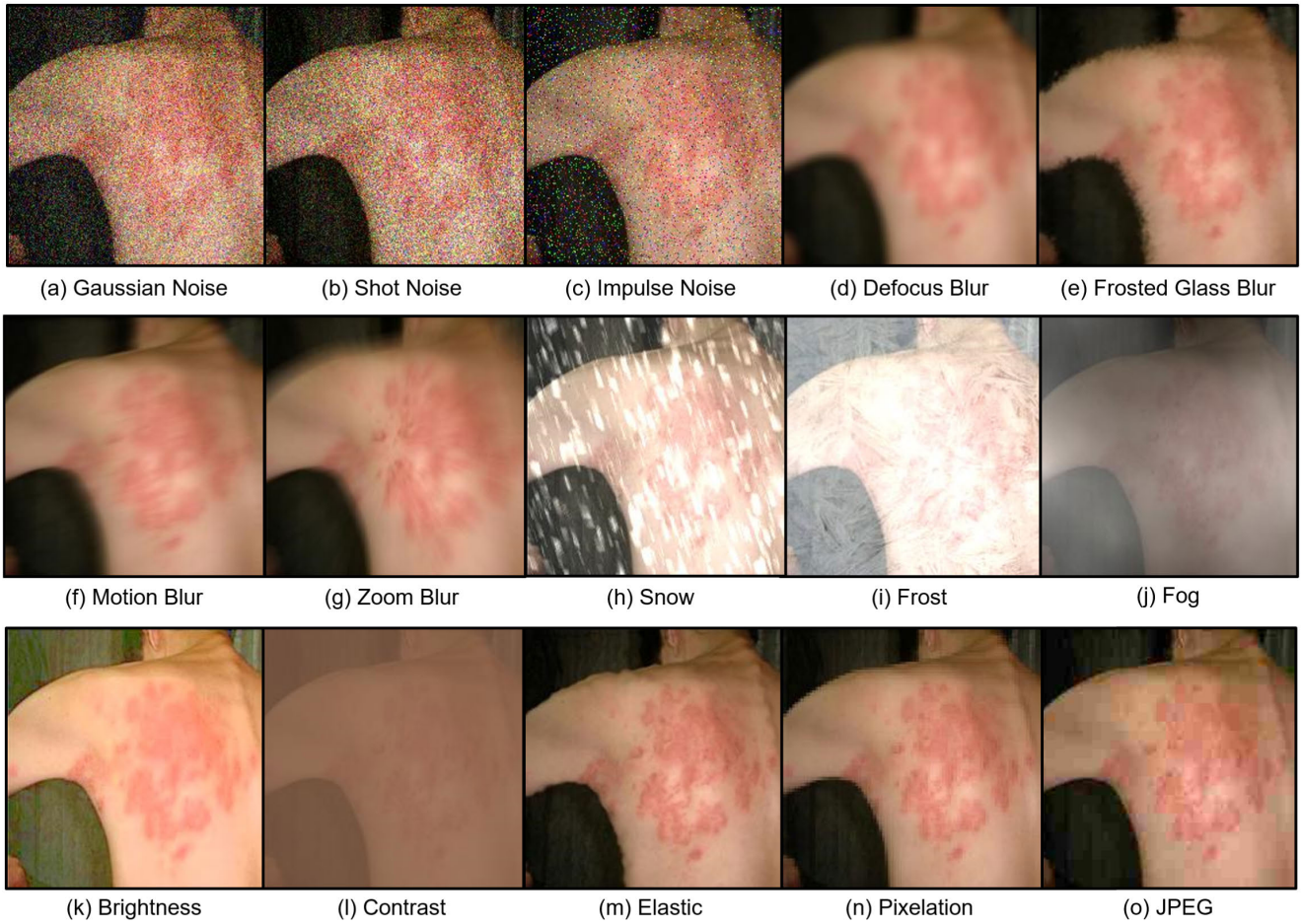

FIGURE 4. Examples of corrupted images in SD-HZ-C owing to (a)-(c) noise, (d)-(g) blur, (h)-(j) weather, and (k)-(o) digital categories with severity level of 3.

The first category pertains to *basic models*: CNN architectures that are widely used in computer vision and have shown considerable performance for image classification. The second category pertains to *mobile models*: mobile neural networks optimized for mobile settings with fewer parameters and faster computations. The third category is the DNN ensemble (*DNN ensemble*), which has outperformed single models in various skin lesion analysis tasks, including those presented in ISIC 2019. The details of each model are described in the following subsections.

1) BASIC MODELS

1) **AlexNet** [42] is an early type of CNN with five convolutional layers, ReLU activation function, and max-pooling layers.
2) **Vgg-16** [43] consists of 16 convolutional layers with a small kernel size of 3 x 3.
3) **InceptionV3** [44] has Inception modules, which produce outputs from convolutional kernels with different shape and auxiliary classifiers to improve the convergence of the deep network.

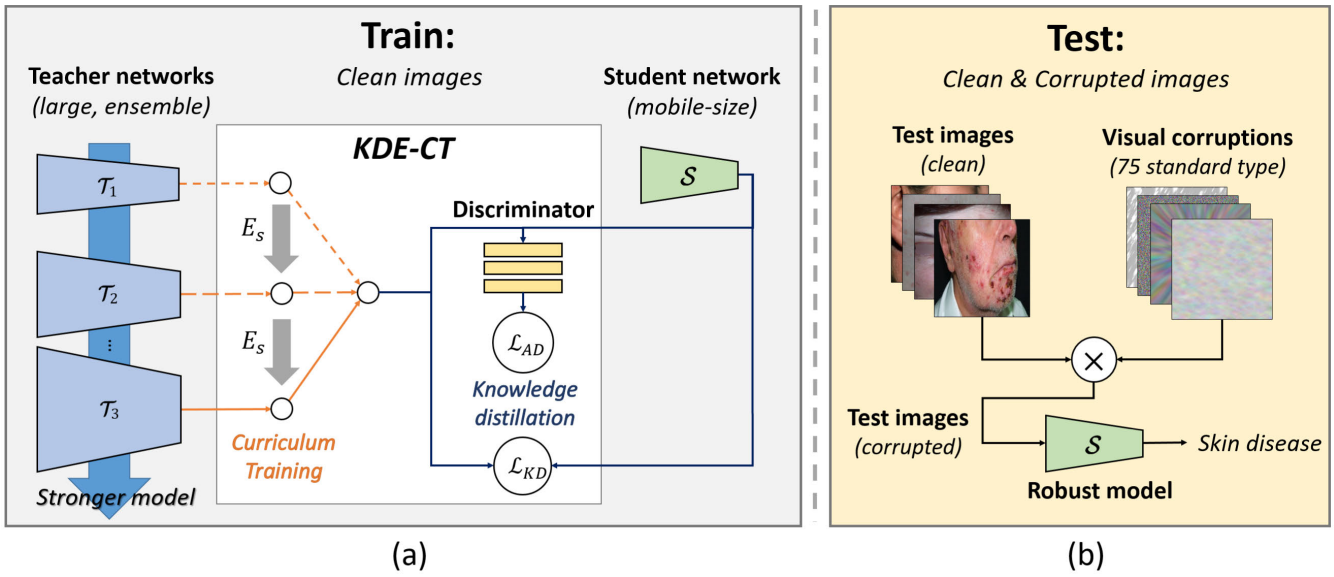

FIGURE 5. Flowchart of proposed KDE-CT. For the diagnosis of HZ from a clinical image, (a) we train a mobile neural network ($\mathcal{S}$) on clean images, and then (b) test this model on both clean and corrupted images, where the corrupted images are generated with 75 standard visual corruptions (e.g., Gaussian noise, zoom blur, fog, and pixelation). The student network ($\mathcal{S}$) learns from an ensemble of large teacher networks ($\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N$) using KD from ensemble via curriculum training (KDE-CT) by minimizing the sum of distillation loss and adversarial loss. The teacher network for training is progressively updated for every E_s epoch so that the student network can learn more robust and richer features from multiple teacher networks under a better curriculum.

4) **ResNet-50** [45] is a deep CNN of 50 layers that utilize the shortcut connection between two residual blocks to alleviate the degradation problem of a deep network.
5) **DenseNet-121** [46] consists of four dense blocks, and connections from each block to every other block are applied to alleviate the vanishing-gradient problem and strengthen feature propagation.
6) **ResNext-101** [47] is constructed by repeating a building block that aggregates a set of transformations with the same topology. This method exposes a new dimension, cardinality, as an essential factor in addition to the dimensions of depth and width. In the experiments, we use hyper parameters of 32 groups and 8 widths per group for the building blocks.
7) **SEResNext-101** [25] is the adapted version of ResNext-101 model with a squeeze-and-excitation block to focus on the channel relationship by adaptively recalibrating the channel-wise feature. We utilize the hyper parameters of 32 groups and 4 widths per group for the building blocks.

2) MOBILE MODELS

1) **EfficientNet-B0** [24] Its network depth, width, and resolution are balanced by utilizing the compound coefficient of width, depth, and resolution. A base architecture is searched by a neural architecture that optimizes both accuracy and efficiency. We utilized 1.0 for the channel multiplier and depth multiplier and 224 for the resolution as the hyper parameter of the compound scaling method.
2) **MNasNet-B1** [48] is a CNN designed via mobile NAS, which incorporates the model latency and the accuracy into the main objective of the search, with a channel multiplier of 1.0.
3) **MobileNetV2** [49] is based on an inverted residual structure where the shortcut connections are between the thin bottleneck layers and intermediate expansion layer for reducing the computational cost. Channel multiplier of 1.0 was used in this paper.
4) **MobileNetV3-Small** [50] is a CNN for low resources used on the mobile phone CPUs and optimized via platform-aware network architecture search and NetAdapt algorithm.

3) DNN ENSEMBLE

1) **Ensemble-M** An ensemble model that averages the outputs of MobileNetV3-Small, DenseNet-121, ResNext-101, and SEResNext-101.
2) **Ensemble-E** An ensemble model that averages the outputs of EfficientNet-B0, DenseNet-121, ResNext-101, and SEResNext-101.

C. KNOWLEDGE DISTILLATION FROM DNN ENSEMBLE

1) KNOWLEDGE DISTILLATION WITH ADVERSARIAL LEARNING

To enhance the performance of a single network, we used a KD from an ensemble (*KDE*), as shown in Fig. 6. In KDE, the knowledge of the teacher network ensemble is distilled into a single student network by learning the probability distribution of the output logits of multiple teacher network.

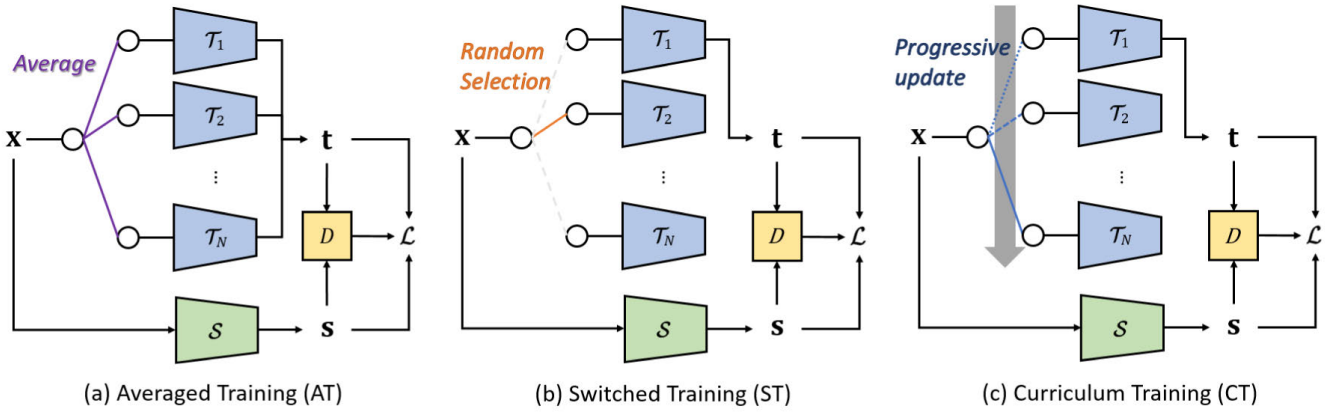

FIGURE 6. Illustration of training strategies for the knowledge distillation from DNN ensemble. $\mathcal{S}$ is student network and $\mathcal{T}$ is teacher network. To distill the knowledge of multiple teacher networks, (a) the output of all teacher models is averaged in AT, (b) a random teacher for each batch is selected in ST, and (c) a teacher is progressively updated for every (e_c) epoch in CT.

Using KDE, a student network can learn more robust and representative features from a DNN ensemble while maintaining a low computational cost. Thus, it can be especially useful for skin disease classification where ensembles of networks are likely to outperform single models [26]. We trained a student network with the softened probability distribution of teacher model outputs, which was proposed by Hinton *et al.* [33]. Given the input data $\mathbf{x}$, the knowledge of a teacher network can be transferred to the student network by minimizing the distillation loss $\mathcal{L}_{KD}$, which is the difference between the output logits of student network $\mathbf{s}$ and that of teacher network $\mathbf{t}$ with temperature scaling:

$$\mathcal{L}_{KD} = \mathcal{L}_{KL}\left(\sigma\left(\frac{\mathbf{s}}{T}\right), \sigma\left(\frac{\mathbf{t}}{T}\right)\right), \tag{1}$$

where σ is the softmax function, T is a temperature value that softens the output logits, and $\mathcal{L}_{KL}$ is the Kullback–Leibler divergence loss. Furthermore, $\mathbf{s}$ and $\mathbf{t}$ refer to the output logits of the teacher networks (i.e., target logits) and student network, respectively. Here, we use $T = 2$, which is the best value reported in [33]. After the teacher network is trained, its parameters are fixed during the training of the student network.

Additionally, we used adversarial learning [51] in the KD, which can ensure better convergence of the student model from multiple teacher networks [35]. In adversarial learning, the discriminator D, consisting of three fully connected layers attempts to distinguish whether the output logits $\mathbf{o}$ is from a teacher or student, by maximizing the following equation:

$$\max_{D}\left[\mathbb{E}_{\mathbf{o}\sim\mathbf{t}}\log D(\mathbf{o}) + \mathbb{E}_{\mathbf{o}\sim\mathbf{s}}\log(1 - D(\mathbf{o}))\right], \tag{2}$$

where $\mathbf{o} \sim \mathbf{t}$ and $\mathbf{o} \sim \mathbf{s}$ are the output logits from the teacher and student networks, respectively. Furthermore, $\mathbf{o}$ is the concatenation of $\mathbf{t}$ and $\mathbf{s}$. Thus, the loss function for the adversarial learning is

$$\mathcal{L}_{AD} = -\left[\mathbb{E}_{\mathbf{o}\sim\mathbf{t}}\log D(\mathbf{o}) + \mathbb{E}_{\mathbf{o}\sim\mathbf{s}}\log(1 - D(\mathbf{o}))\right], \tag{3}$$

Overall, the training proceeds with simultaneous minimization of the overall loss, L, which is the sum of the distilling loss and adversarial loss:

$$\mathcal{L} = \mathcal{L}_{KD} + \mathcal{L}_{AD}, \tag{4}$$

2) TEACHER SELECTION FOR DNN ENSEMBLE DISTILLATION

For KD from the ensemble of neural networks, the student network performance can vary depending on the teacher selection strategy in training. We proposed curriculum training for KDE (KDE-CT) to distill more effectively the ensemble teachers via an adaptive learning strategy. Using KDE-CT the student network can learn from a stronger teacher network as the training proceeds so that more comprehensive and robust features can be learned progressively from the ensemble of teacher networks under a better curriculum. The standard methods for the teacher selections are KDE via averaged training (KDE-AT) [34] and KDE via switched training (KDE-ST) [52], and the three training strategies are shown in Fig. 6, and the details are as follows:

1) **KDE via Averaged Training (KDE-AT)** This method is a typical DNN ensemble distillation method that uses the average of the outputs logits of all the teacher networks (*i.e.,* output logits of traditional ensemble) as the target logits [34], [52]. Given the input data $\mathbf{x}$ and a set of teacher networks $\mathcal{S}_{\mathcal{T}} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N\}$ sorted by the overall accuracy, the target logits $\mathbf{t}$ can be expressed by the following equation:

$$\mathbf{t} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{T}_n(\mathbf{x}), \tag{5}$$

where N is the number of teacher networks.

2) **KDE via Switched Training (KDE-ST)** It was proposed in [52] and [35] and showed beneficial effects on the performance of DNN ensemble distillation. This method randomly selects a teacher network from a set of teacher networks $\mathcal{S}_{\mathcal{T}}$ at every mini batch during the

training step. The target logits $\mathbf{t}$ with the input $\mathbf{x}$ can be formulated as the following equation:

$$\mathbf{t} = \mathcal{RSM}(\{\mathcal{T}_1(\mathbf{x}), \mathcal{T}_2(\mathbf{x}), \ldots, \mathcal{T}_N(\mathbf{x})\}), \tag{6}$$

where $\mathcal{RSM}(\mathcal{S})$ is the random selection module in a given set $\mathcal{S}$.

3) **KDE via Curriculum Training (KDE-CT)** We proposed a KDE strategy (KDE-CT) to train a more accurate and robust student network via progressive teacher network updates. Motivated from [53], which changes the training dataset from a low-level-knowledge to a high-level knowledge dataset, we sequentially updated the teacher network used for KD from a low-performance to a high-performance network. We hypothesize that this curriculum training strategy can help in finding better local minima for the student network as the student network can gradually learn more robust and richer features from multiple teacher networks under a better curriculum. The performance of the teacher network can be evaluated using various metrics; we used the overall accuracy of the teacher networks in this study. As illustrated in Figures 5-(a) and 6-(c), the teacher network is progressively updated for every E_s epoch in a given teacher set $\mathcal{S}_\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N\}$. Specifically, given the sequence of teacher networks $\mathcal{S}_\mathcal{T}$ sorted by a certain metric (i.e. accuracy, and F1), the target logits can be defined using the following equation:

$$\mathbf{t} = \mathcal{T}_{\min(\lceil E_c/E_s \rceil, N)}(\mathbf{x}), \tag{7}$$

where $\lceil\ \rceil$ is the ceiling function, E_c is the number of the current training epoch, and E_s is the step number of the training epoch that changes the teacher network. Additionally, the student network was trained by the ensemble of teacher networks for last 100 training epochs. In our experiments, we used 100 for E_s and 4 for N.

IV. EXPERIMENTS AND RESULTS

We compared the performances of 13 different models on both clean images (SD-HZ dataset) and corrupted images (SD-HZ-c dataset). Then, we conducted experiments to verify the effectiveness of our KD method to improve the robustness against input visual corruption in the context of skin disease classification.

A. EXPERIMENT SETUP AND EVALUATION METRICS

The performance for skin disease classification was assessed using the following criteria: overall accuracy (*ACC*), and the macro average of F1-score (*F1*), precision (*PR*), recall (*RE*), and area under the curve of the receiver operating characteristic (*AUROC*). Additionally, we used the F1 score for the HZ (*F1-HZ*) and accuracy for the HZ (ACC-HZ) as a metric to compare the performance of HZ diagnosis. Cohen's kappa coefficient (Kappa) [54] was measured to assess the agreement between the prediction of the neural network and experts. In addition, we measured the computational efficiency for each model using *MACs*, *Latency*, and *Params*. *MACs* is the number of multiply-and-accumulate operations and represents computational cost of the neural network. *Latency* is defined as the test time required for a model to process a single image. *Params.* is the storage cost, defined as the total number of parameters of the DNN. These metrics are calculated using FP32 on an Intel i7-9700K CPU with a batch size of 1, considering that users are likely to submit a single image per request to a mobile diagnosis application.

For mobile diagnosis of skin diseases from a user-submitted image, robustness against user noise is another important factor. To assess the robustness of DNNs against input visual corruptions, we used corruption errors (*CE*) and *mCE* [16]. For a classifier f, the *CE* against the input visual corruption type c at the level of severity s is defined as

$$CE_c^f = \Big(\sum_{s=1}^{5} E_{s,c}^f\Big) \Big/ \Big(\sum_{s=1}^{5} E_{s,c}^{AlexNet}\Big), \tag{8}$$

where E is the top-one error rate on the clean dataset. Meanwhile, mCE is the average of *CE* for all 15 corruption types and is computed as

$$mCE_c^f = \Big(\sum_{c \in C} CE_c^f\Big) \Big/ |C|, \tag{9}$$

where C = { Gaussian noise, shot noise, $\cdots$, and JPEG}. Thus, *mCE* measures the error of a classifier with respect to AlexNet [42] against all 75 corruption errors standardized by [16].

We used only the SD-HZ dataset for training the networks in all the experiments. The model trained on SD-HZ was tested on both clean images (test split of SD-HZ) and 75 types of corrupted image sets (SD-HZ-C). The SD-HZ dataset was randomly split into training (70%), validation (10%), and test (20%) datasets using stratified sampling. The detailed sample distribution of the training, validation, and test datasets is shown in Table 3. In training, all models were trained on SD-HZ after modifying the last fully connected layer of the model pre-trained on ImageNet [55] to a 4-ary classification layer. During training, the validation loss was monitored on every epoch. We stopped the training if no improvement was observed in the validation loss after the 10th epoch. For data augmentation, a similar pipeline proposed by [31] was utilized using the Albumentation library [56]. Images were resized to 224×224, and we applied the horizontal and vertical flip, random brightness, random contrast, random hue, saturation value, and an affine transformation (translate, scale, rotate) with the probability of 0.5 for each operation in every batch of training.

When training the DNNs in Section III-B, RAdam optimizer [57] was aplied with the batch size of 64, learning rate of 2×10^{-4}, and weight decay of 2×10^{-5} using Pytorch. We used the same train-val-test split and data augmentation pipelines for the KD models; however, the hyper parameter and optimizer slightly differed as we followed

TABLE 2. Performances of 13 different models on the SD-HZ dataset (Latency: millisecond, Params.: million, MACs: million).

Category	Model	ACC	F1	PR	RE	F1-HZ	AUROC	Kappa	Latency	Params.	MACs
Basic models	AlexNet	0.93	0.84	0.87	0.82	0.80	0.946	0.794	21.0	57.0	710
	Vgg-16	0.93	0.85	0.87	0.86	0.79	0.967	0.779	140.3	134.3	1,550
	InceptionV3	0.92	0.83	0.85	0.82	0.78	0.960	0.766	73.2	21.8	285
	ResNet-50	0.94	0.87	0.87	0.88	0.81	0.979	0.797	93.3	23.5	412
	ResNext-101	0.93	0.87	0.87	0.89	0.82	0.975	0.809	320.7	86.8	1,651
	SEResNext-101	0.94	0.88	0.89	0.87	**0.85**	0.981	0.846	226.8	46.9	805
	DenseNet-121	0.93	0.85	0.87	0.84	0.80	0.969	0.791	328.6	7.0	285
Mobile models	EfficientNet-b0	0.92	0.81	0.82	0.83	0.68	0.954	0.658	55.5	4.0	40
	MNasNet	0.92	0.84	0.85	0.84	0.76	0.947	0.749	38.6	3.1	33
	MobileNetV2	0.90	0.79	0.80	0.83	0.61	0.955	0.586	39.5	2.2	32
	MobileNetV3-Small	0.92	0.84	0.85	0.83	0.82	0.927	0.811	**20.6**	**1.5**	**5**
DNN Ensemble	Ensemble-M	0.95	0.89	0.92	0.88	0.82	0.979	0.880	666.7	142.2	2,746
	Ensemble-E	**0.97**	**0.93**	**0.94**	**0.92**	0.78	**0.992**	**0.910**	699.5	144.6	2,781

TABLE 3. Sample distribution for training, validation, and test.

	SD-HZ			SD-HZ-C
Class	Train	Val	Test (clean)	Test (corrupted)
Acne	764	109	219	219 × 75 corruptions
HZ	289	41	82	82 × 75 corruptions
Tinea	555	79	159	159 × 75 corruptions
Other Diseases	4252	607	1215	1215 × 75 corruptions
Total	5860	836	1676	1676 × 75 corruptions

the same setting for KDE methods in [35]. We used the model trained on the SD-HZ dataset as a teacher network and then distilled the knowledge from the teacher network using stochastic gradient descent (SGD) optimizer [58] with the batch size of 128. The learning rates for $\mathcal{L}_{KD}$ and $\mathcal{L}_{D}$ were 0.1 and 0.001, respectively. The step weight decay was applied at the 150-th and 250-th training epoch with the decay rate of 0.1. We used a weight decay constant of 10^{-5}, and a maximum training epoch of 500 for the convergence of the student network. The evaluation of KD was conducted under the same condition as in [35].

B. PERFORMANCE OF SKIN DISEASE CLASSIFICATION

We first evaluate the 13 models described in Section III-B on the SD-HZ dataset for the diagnosis of HZ and other skin diseases. The results of 13 different models for skin disease classification on the SD-HZ dataset are presented in Table 2. The ensemble of DNNs showed the best performance among the different models, where the accuracy, F1, and Kappa are 0.95, 0.89, and 0.880, respectively, for Ensemble-M and 0.97, 0.93, and 0.910, respectively, for Ensemble-E. Although the DNN ensembles outperformed all other models, the computational costs of the ensemble models are high where the latency is 73.94 ms, the number of parameters is 144.6 M, and MACs are 2786 M, which could be infeasible for mobile on-device diagnosis.

The performances of basic models were similar to each other, where the accuracy and F1 were ranged between 0.92–0.94 and 0.84–0.87, respectively. Furthermore, the accuracy and F1 of mobile models were similar to those of the basic models, with a small performance drop, where the accuracy was between 0.90 and 0.92, and the F1 was between 0.79 and 0.84. However, F1-HZ and Kappa of the mobile models dropped drastically, especially for EfficientNet-b0 (0.68, 0.658) and MobileNetV2 (0.61, 0.586). Interestingly, MobileNetV3-Small achieved fine performance, even when compared with the higher performance of InceptionV3, with F1 of 0.84 (+0.01), F1-HZ of 0.82 (+0.04), and Kappa of 0.811 (+0.045) with much low computational cost (5x faster inference, 43x smaller Params, x57 smaller MAC). Thus, using MobileNetV3-Small could be an acceptable option for diagnosing skin diseases, especially HZ.

C. ROBUSTNESS IMPROVEMENT VIA ENSEMBLE KD

To assess the robustness against common visual corruptions, the 13 models in Section III-B were evaluated on the SD-HZ-C dataset. Table 4 shows the performance of 13 models against visual corruption in SD-HZ-C. For the single models, no significant correlation was observed between the performance on the clean images and corrupted images. The mCE does not decrease, as the overall accuracy and architecture are improved, which is different from the tendency observed for ImageNet. For example, InceptionV3 achieved the best mCE performance among the basic models, but it exhibited a low accuracy (79.2 mCE, 91.8 ACC). It agrees with the observation in [26] ImageNet may not be a safe proxy for skin lesion analysis. Moreover, this emphasizes the need for careful consideration when training the DNN for the skin disease classification from clean images. Meanwhile, the ensemble strategy leads to improvement in diagnosis using both clean images and corrupted images. The corruption robustness of MobileNetV3-Small and EfficientNet-B0 was enhanced via ensemble, as the mCE of Ensemble-M

TABLE 4. Robustness of 13 different models on corrupted images (SD-HZ-C), where the models are trained using only clean images (SD-HZ). Here, ACC is the overall accuracy (higher is better). The values in Noise, Blur, Weather, and Digital columns are corruption errors, and mCE is the mean of these values (lower is better).

	(%)		Noise			Blur				Weather				Digital			
Model	**ACC**	**mCE**	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
AlexNet	92.8	100.0	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Vgg-16	93.2	124.1	44	116	169	116	148	125	149	89	130	116	94	101	152	174	140
InceptionV3	91.8	**79.2**	31	77	74	42	63	105	49	108	96	57	93	47	124	125	96
ResNet-50	93.7	106.6	50	142	128	97	117	124	140	85	123	92	90	46	185	91	90
ResNext-101	93.4	132.0	99	248	251	111	107	187	143	93	125	149	70	71	119	106	101
SEResNext-101	**94.1**	87.2	34	96	80	51	76	100	106	86	94	91	83	73	119	130	88
DenseNet-121	93.4	116.6	49	143	148	116	121	107	129	84	85	115	87	89	170	161	146
EfficientNet-B0	91.6	**96.2**	33	83	79	75	86	104	127	82	98	79	97	44	149	190	117
MNasNet-B1	**92.4**	100.3	59	132	129	35	69	78	92	104	133	70	124	60	146	158	115
MobileNetV2	90.0	142.2	74	191	184	97	144	161	91	105	113	84	149	80	225	273	163
MobileNetV3-Small	92.0	98.6	43	102	109	67	62	149	110	107	115	67	110	78	125	125	110
Ensemble-M	95.0	81.1	34	89	85	70	79	90	115	87	73	84	64	77	101	92	75
Ensemble-E	**96.5**	**69.7**	27	83	63	66	74	65	96	70	62	87	48	65	91	84	65

and Ensemble-E was improved with a large margin 69.7 (−26.5) and 81.1 (−17.5), respectively. Additionally, the mCE of the ensembles surpassed the value of the single models used for the ensemble, where the mCE of MobileNetV3-Small, EfficientNet-B0, DenseNet-121, ResNext-101, SEResNext-101 are 98.6, 96.2, 116.6, 132.0, and 87.2, respectively.

To improve corruption robustness, we applied a KD from an ensemble (KDE) for MobileNetV3-Small and EfficientNet-B0 with three different training strategies, namely, averaged training (KDE-AT), switched training (KDE-ST), and curriculum training (KDE-CT), as described in Section III-C. As shown in Table 5, we applied these three training strategies to MobileNetV3-Small and EfficientNet-B0. We used models in Ensemble-M and Ensemble-E as teacher networks for MobileNetV-3 and EfficientNet-B0, respectively. To verify the effectiveness of ensemble distillation, we compared it to a single model distillation denoted as KD-SM. In the KD-SM method, the students were trained by the best accurate teacher (SEResNext-101). For all KD experiments, a student network was trained from scratch by the teacher networks pre-trained with SD-HZ dataset. Gaussian noise training (GNT$\sigma_{0.5}$) and adversarial noise training (ANT$^{3\times3}$) proposed by [38] were utilized with the same parameters used in the paper for fair comparison with the other method. We added the Gaussian noise to input images with a standard deviation of 0.5 along with the data augmentation pipelines used in Section IV-A and denoted it as GNT$\sigma_{0.5}$. For the ANT, we jointly trained the noise generator with 3 x 3 kernels along with the classifier, which generates adversarial noise against the skin diseases classifier. As a baseline, we used the model trained on clean images following Section IV-B. Note that all models were trained using only clean images of the SD-HZ dataset.

The results are presented in Table 6. The proposed KDE-CT achieved the best performance in mCE among the different methods to improve the corruption robustness. KDE-CT significantly reduced the mCE errors when compared with the baseline by achieving the mCE of 67.6 (−31.0) and 56.7 (−39.5) for both MobileV3-Small and EfficientNet-B0, respectively. When CT is not used for KDE, the decrease in mCE was lower than that in KD-SM. It indicates that naive KDE can result in lower robustness, and an appropriate training strategy should be used. The mCE of KD-SM is lower than GNT$\sigma_{0.5}$ and ANT$^{3\times3}$ for MobileNetV3-Small and similar to GNT$\sigma_{0.5}$ and ANT$^{3\times3}$ for MobileNetV3-Small.

Notably, the single student network distilled using KDE-CT showed better corruption robustness than the teacher network of DNN ensembles, with the mCE improvement of 13.5 for MobileNetV3-Small and 13.0 for EfficientNet-B0. This implies that the curriculum strategy for KDE (KDE-CT) yielded better convergence of the student networks by finding better local minima. Moreover, KDE-CT enhanced the overall accuracy compared with the baseline, as evidence by the ACC values of 93.5 (+1.5) and 92.8 (+1.2) for MobileNetV3-small and EfficientNet-B0, respectively. The single mobile models require much lower computational cost than the ensemble of the neural network, as MobileNetV3-Small trained with KDE-CT has 8x lower latency, 94x number of the parameters, and 549x smaller MACs than Ensemble-E. Thus, considering the possible input

TABLE 5. Configurations of student network and teacher networks for KDE.

Student Network	Teacher Networks	
MobileNetV3-small	$\mathcal{T}_1$	MobileNetV3-small
	$\mathcal{T}_2$	DenseNet-121
	$\mathcal{T}_3$	ResNext-101
	$\mathcal{T}_4$	SEResNext-101
EfficientNet-B0	$\mathcal{T}_1$	EfficientNet-B0
	$\mathcal{T}_2$	DenseNet-121
	$\mathcal{T}_3$	ResNext-101
	$\mathcal{T}_4$	SEResNext-101

TABLE 6. Comparison of different methods to improve the robustness against common visual corruptions (SD-HZ-C), where the models are trained using only clean images (SD-HZ). Here, ACC is the overall accuracy (higher is better). The values in the Noise, Blur, Weather, and Digital columns are corruption errors, and the mCE is the mean of these values (lower is better). The underlined KDE-CT significantly enhanced the corruption robustness of the models when compared with other methods.

		(%)		Noise			Blur				Weather				Digital			
Architecture	**Method**	**ACC**	**mCE**	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
MobileNetV3-Small	Baseline	92.0	98.6	43	102	109	67	62	149	110	107	115	67	110	78	125	125	110
	GNT$\sigma_{0.5}$	92.9	83.7	25	63	63	53	36	104	81	113	111	107	106	57	127	104	103
	ANT$^{3\times3}$	92.1	81.7	31	75	84	28	41	93	80	102	113	115	98	59	108	101	98
	KD-SM	93.2	85.4	25	64	63	68	50	99	102	94	121	72	102	104	87	122	108
	KDE-AT	93.4	78.6	31	92	77	52	37	78	90	96	95	109	94	118	79	78	52
	KDE-ST	93.2	68.0	28	77	73	38	24	58	77	92	91	81	77	114	71	74	44
	KDE-CT	**93.5**	**67.6**	25	68	66	28	28	61	48	101	97	102	86	109	76	75	44
EfficientNet-B0	Baseline	91.6	96.2	33	83	79	75	86	104	127	82	98	79	97	44	149	190	117
	GNT$\sigma_{0.5}$	**94.1**	94.5	41	98	100	80	81	86	115	88	77	125	91	76	124	149	88
	ANT$^{3\times3}$	92.8	90.4	26	77	60	44	61	81	81	85	113	119	98	76	138	171	124
	KD-SM	93.3	70.1	33	90	89	25	25	62	44	97	95	81	110	106	76	76	42
	KDE-AT	92.4	85.6	49	143	126	31	32	66	67	95	104	83	101	125	90	102	69
	KDE-ST	92.8	67.2	24	63	58	26	27	64	49	90	102	106	85	106	79	83	45
	KDE-CT	92.8	**56.7**	19	56	52	17	18	51	27	78	98	88	74	80	68	78	44

TABLE 7. Comparison between state-of-the-art methods and our method for skin disease classification from clinical images in terms of the accuracy for all classes (ACC), accuracy for herpes zoster (ACC-HZ), number of trainable parameters in the model (Params.), number of test classes (No. of test classes), and number of images (No. of images) for training, test, and HZ training. Test on corruption indicates whether the model was tested under various visual corruptions. Params. for the state-of-the-art methods are approximate values, as detailed parameters are not specified.

Method	**Model**	**Test on Corruption**	**ACC** (all)	**ACC -HZ**	**Params.** (million)	**No. of test classes**	**No. of images** (train / test / HZ train)
Burlina et al. [59]	ResNet-50	No	0.84	0.89	∼24	4	1,094 / 500 / 410
Burlina et al. [60]	ResNet-50	No	0.816	0.81	∼24	4	2,706 / 549 / 594
Liu et al. [11]	InceptionV4	No	0.71	-	∼43	27	64,837 / 3,707 / 0
Han et al. [12]	Ensemble of SENet, SE-ResNet-50, VGG-19	No	0.448	0.36	∼287	134	220,680 / 2,201 / 1,047
Ours	MobileNetV3-small (KDE-CT)	Yes	0.935	0.74	1.5	4	5,860 / 1,676 / 289

noise from the user and the computational limitations owing to the mobile setting, the proposed KDE-CT can be useful for the mobile diagnosis of skin diseases.

D. COMPARISON WITH STATE-OF-THE-ART METHODS

We compared our methods with state-of-the-art methods proposed for skin disease classification from clinical images, and the results are shown in Table 7. Note that comparing the clinical skin disease diagnosis methods can be challenging because the dataset, training methods, and evaluation metrics are different in each case. For example, some studies focus on accurate classification of a small number of skin diseases [59], [60] whereas the others aim to train a more general model on a large number of classes and datasets [11], [12]. Therefore, we compared the performance of these methods based on the statistics of the dataset and classes. Specifically, in Table 7, we list the accuracy of the overall test class (ACC), accuracy for herpes zoster (ACC-HZ), and the number of model parameters (Params.) with the number of test classes (No. of test classes) and the number of images for training, testing, and herpes zoster.

When compared with the model proposed by Burlina *et al.* [59], [60], which used ResNet-50 for the 4-ary classification (HZ, erythema migrans, tinea corporis, and normal skin), we achieved an ACC of 0.935 (+0.095) and an ACC-HZ of 0.74 (-0.15). The performance difference originated from the different sample distributions of our SD-HZ dataset and that used by Burlina *et al.*, as our SD-HZ contains more skin disease images (5,860 images) but fewer HZ images (289 images) than that in [59], [60]. However, a comparison between the performance of the ResNet-50 [45] used in [59], [60] and our MobileNetV3-Small with KDE-CT considering the same training methods and datasets shows that we achieved better performance (93.5 ACC, 67.5 mCE) than ResNet-50 (93.7 ACC, 106.6 mCE) while using ∼15x fewer parameters (Tables 2, 4 and 6). Liu *et al.* [11] achieved satisfactory performance, with an overall accuracy of 0.71 over 27 classes using single deep convolutional networks. However, their model could not diagnose herpes zoster, as their dataset did not include this disease.

Han *et al.* [12] trained DNNs on a large number of clinical images (220,680 images) to classify 134 skin disorders. However, although a large ensemble model consisting of three deep convolutional networks was utilized in [12], the overall accuracy was significantly lower than that achieved in this study. The ACC of 0.448 and ACC-HZ of 0.36 are not sufficient for mobile diagnosis of herpes zoster. In contrast, our model achieved significantly better performance than that

in [12] in terms of ACC and ACC-HZ while maintaining a low computational cost by focusing on certain common skin diseases. This implies that the proposed KDE-CT method can be used to train an accurate and mobile model to diagnose herpes zoster with competitive performance and computational cost when compared with state-of-the-art methods. Our study is the first to evaluate the robustness of DNNs against various visual corruptions systemically on a reasonably large dataset.

V. CONCLUSION

This study aimed to distinguish HZ from the other skin diseases for mobile applications. We established an SD-HZ dataset comprising 413 single images and 8,345 clinical images. We also constructed an SD-HZ-C dataset with 15 types of corruption with five severity levels for each type to evaluate robustness against input visual corruption. A total of 13 different DNNs were trained on SD-HZ images and were evaluated on both clean and corrupted images. The results showed that corruption error should be considered along with accuracy when selecting an appropriate model for mobile skin disease diagnosis. In this regard, MobileNet-V3-small represents a reasonable choice considering efficiency as well as accuracy. We proposed KDE-CT to enhance the robustness of DNN by progressively changing the teacher network and verified that it can be an effective solution for improving the corruption robustness while retaining satisfactory performance on clean images. By applying KDE-CT, a small student model can achieve higher accuracy and lower mCE than the model trained with other KDE methods, which is significant for real-world applications. Thus, it is possible to train an accurate, robust, and mobile DNN for diagnosis of skin diseases from clinical images and it has potential to be used for mobile dermatology under user-level mobile conditions. However, KDE-CT has hyperparameters that need to be manually adjusted depending on the model and datasets. As future work, we are planning to examine a more general teacher selection algorithm in KDE.

ACKNOWLEDGMENT

(Seunghyeok Back and Seongju Lee contributed equally to this work.)

REFERENCES

[1] B. P. Yawn and D. Gilden, "The global epidemiology of herpes zoster," *Neurology*, vol. 81, no. 10, pp. 928–930, Sep. 2013.

[2] E. Koshy, L. Mengting, H. Kumar, and W. Jianbo, "Epidemiology, treatment and prevention of herpes zoster: A comprehensive review," *Indian J. Dermatology, Venereology, Leprology*, vol. 84, no. 3, p. 251, 2018.

[3] R. H. Dworkin, R. W. Johnson, J. Breuer, J. W. Gnann, M. J. Levin, M. Backonja, R. F. Betts, A. A. Gershon, M. L. Haanpää, M. W. McKendrick, and T. J. Nurmikko, "Recommendations for the management of herpes zoster," *Clin. Infectious Diseases*, vol. 44, no. 1, pp. S1–S26, 2007.

[4] E. Paek and R. Johnson, "Public awareness and knowledge of herpes zoster: Results of a global survey," *Gerontology*, vol. 56, no. 1, pp. 20–31, 2010.

[5] R. W. Johnson, M.-J. Alvarez-Pasquin, M. Bijl, E. Franco, J. Gaillat, J. G. Clara, M. Labetoulle, J.-P. Michel, L. Naldi, L. S. Sanmarti, and T. Weinke, "Herpes zoster epidemiology, management, and disease and economic burden in europe: A multidisciplinary perspective," *Therapeutic Adv. Vaccines*, vol. 3, no. 4, pp. 109–120, Jul. 2015.

[6] S. Pathan, K. G. Prabhu, and P. C. Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review," *Biomed. Signal Process. Control*, vol. 39, pp. 237–262, Jan. 2018.

[7] P. Schmid, "Segmentation of digitized dermatoscopic images by two-dimensional color clustering," *IEEE Trans. Med. Imag.*, vol. 18, no. 2, pp. 164–171, Feb. 1999.

[8] P. Wighton, T. K. Lee, H. Lui, D. I. McLean, and M. S. Atkins, "Generalizing common tasks in automated skin lesion diagnosis," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 4, pp. 622–629, Jul. 2011.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[10] H. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, and L. Uhlmann, "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018.

[11] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, and V. Gupta, "A deep learning system for differential diagnosis of skin diseases," *Nature Med.*, vol. 26, no. 6, pp. 900–908, 2020.

[12] S. S. Han, I. Park, S. Eun Chang, W. Lim, M. S. Kim, G. H. Park, J. B. Chae, C. H. Huh, and J.-I. Na, "Augmented intelligence dermatology: Deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders," *J. Investigative Dermatology*, vol. 140, no. 9, pp. 1753–1761, Sep. 2020.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[14] A. Kwasigroch, M. Grochowski, and A. Mikolajczyk, "Neural architecture search for skin lesion classification," *IEEE Access*, vol. 8, pp. 9061–9071, 2020.

[15] S. Mishra, S. Chaudhury, H. Imaizumi, and T. Yamasaki, "Assessing robustness of deep learning methods in dermatological workflow," 2020, *arXiv:2001.05878*. [Online]. Available: http://arxiv.org/abs/2001.05878

[16] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.

[17] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artif. Intell. Med.*, vol. 56, no. 2, pp. 69–90, Oct. 2012.

[18] A. Kjoelen, M. J. Thompson, S. E. Umbaugh, R. H. Moss, and W. V. Stoecker, "Performance of AI methods in detecting melanoma," *IEEE Eng. Med. Biol. Mag.*, vol. 14, no. 4, pp. 411–416, Jul. 1995.

[19] Y. Cheng, R. Swamisai, S. E. Umbaugh, R. H. Moss, W. V. Stoecker, S. Teegala, and S. K. Srinivasan, "Skin lesion classification using relative color features," *Skin Res. Technol.*, vol. 14, no. 1, pp. 53–64, 2008.

[20] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscipl. Reviews: Data Mining Knowl. Discovery*, vol. 8, no. 4, 2018, Art. no. e1249.

[21] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

[22] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*. [Online]. Available: http://arxiv.org/abs/1902.03368

[23] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, Jan. 2020, Art. no. 100864.

[24] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[26] F. Perez, S. Avila, and E. Valle, "Solo or ensemble? Choosing a CNN architecture for melanoma classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2775–2783.

[27] F. Xie, Y. Lu, A. C. Bovik, Z. Jiang, and R. Meng, "Application-driven no-reference quality assessment for dermoscopy images with multiple distortions," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 6, pp. 1248–1256, Jun. 2016.
[28] X. Sun, J. Yang, M. Sun, and K. Wang, "A benchmark for automatic visual classification of clinical skin disease images," in *Computer Vision—ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 206–222.
[29] J. Yang, X. Sun, J. Liang, and P. L. Rosin, "Clinical skin lesion diagnosis using representations inspired by dermatologist criteria," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1258–1266.
[30] T.-T. Do, T. Hoang, V. Pomponiu, Y. Zhou, Z. Chen, N.-M. Cheung, D. Koh, A. Tan, and S.-H. Tan, "Accessible melanoma detection using smartphones and mobile image analysis," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2849–2864, Oct. 2018.
[31] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, "Data augmentation for skin lesion analysis," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis* (Lecture Notes in Computer Science). Granada, Spain: Springer, 2018, pp. 303–311.
[32] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A comprehensive survey on model compression and acceleration," *Artif. Intell. Rev.*, vol. 53, no. 7, pp. 5113–5155, 2020.
[33] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learn. Represent. Learn. Workshop*, 2015, pp. 1–9. [Online]. Available: http://arxiv.org/abs/1503.02531
[34] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1285–1294.
[35] Z. Shen, Z. He, and X. Xue, "Meal: Multi-model ensemble via adversarial learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 4886–4893.
[36] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7538–7550.
[37] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.
[38] E. Rusak, L. Schott, R. S. Zimmermann, J. Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel, "A simple way to make neural networks robust against diverse image corruptions," 2020, *arXiv:2001.06057*. [Online]. Available: http://arxiv.org/abs/2001.06057
[39] J. Lee, T. Won, T. Kwan Lee, H. Lee, G. Gu, and K. Hong, "Compounding the performance improvements of assembled techniques in a convolutional neural network," 2020, *arXiv:2001.06268*. [Online]. Available: http://arxiv.org/abs/2001.06268
[40] F.-F. Xue, J. Peng, R. Wang, Q. Zhang, and W.-S. Zheng, "Improving robustness of medical image diagnosis with denoising convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, in Lecture Notes in Computer Science, Shenzhen, China: Springer, 2019, pp. 846–854.
[41] J. Yang, X. Wu, J. Liang, X. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Self-paced balance learning for clinical skin disease recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 2832–2846, Aug. 2020.
[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[44] 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. (CVPR)*, Jun. 2016, pp. 2818–2826.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[47] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.
[48] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2820–2828.
[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[50] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
[51] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[52] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," in *Proc. Interspeech*, Aug. 2017, pp. 3697–3701.
[53] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
[54] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.
[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[56] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125
[57] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 2–15.
[58] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*. Paris, France: Physica-Verlag HD, 2010, pp. 177–186.
[59] P. M. Burlina, N. J. Joshi, E. Ng, S. D. Billings, A. W. Rebman, and J. N. Aucott, "Automated detection of erythema migrans and other confounding skin lesions via deep learning," *Comput. Biol. Med.*, vol. 105, pp. 151–156, Feb. 2019.
[60] P. M. Burlina, N. J. Joshi, P. A. Mathew, W. Paul, A. W. Rebman, and J. N. Aucott, "AI-based detection of erythema migrans and disambiguation against other skin lesions," *Comput. Biol. Med.*, vol. 125, Oct. 2020, Art. no. 103977.

SEUNGHYEOK BACK (Graduate Student Member, IEEE) received the B.S. degree in mechanical engineering from the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2018, where he is currently pursuing the Ph.D. degree with the School of Integrated Technology, under the intelligent robotics program. His current research interests are image processing based on deep learning for robotic manipulation and health care application.

SEONGJU LEE received the B.S. degree in mechanical engineering from the Gwangju Institute of Science and Technology (GIST), in 2018, and the M.S. degree in intelligent robotics from the School of Integrated Technology, GIST, where he is currently pursuing the Ph.D. degree in intelligent robotics. His research interest is deep learning for health care, especially automatic sleep stage classification.

SUNGHO SHIN received the B.S. degree in environmental science and engineering from the Gwangju Institute of Science and Technology (GIST), in 2018, where he is currently pursuing the Ph.D. degree with the School of Integrated Technology, under intelligent robotics program. His current research interests are deep learning application to vision and signal, and medical data analysis using the deep learning architecture.

YEONGUK YU received the B.S. degree in computer engineering from the Hanbat National University, Daejeon, South Korea, in 2020. He is currently pursuing the M.S. degree in intelligent robotics with the School of Integrated Technology, Gwangju Institute of Science and Technology (GIST). His current research interest is out-of-distribution data detection for deep learning applications.

TAEKYEONG YUK is currently pursuing the B.S. degree in electrical engineering and computer science with the Gwangju Institute of Science and Technology (GIST). He is currently doing an internship at the Artificial Intelligence Laboratory, School of Integrated Technology. His current research interest is computer vision using deep learning.

SAEPOMI JONG received the M.D. degree from Jeonbuk National University, in 2016. Her research interests are occupational and preventive medicine for the health of the general public with the application of artificial intelligence.

SEUNGJUN RYU received the M.D. and Ph.D. degrees in medicine (neurosurgery) from Yonsei University, South Korea, in 2012 and 2017, respectively. From 2013 to 2017, he was a Neurosurgery Resident at the Yonsei University Health System. From 2017 to 2020, he completed the additional Ph.D. (biomedical engineering) with the Institute of Integrated Technology, Gwangju Institute of Science and Technology (GIST). Since 2020 September, he has been a Research Professor at Yonsei University. His research interests include neuromodulation using machine learning and deep learning to neurologic disease screening for digital healthcare. He is a member of the Korean Spinal Neurosurgery Society.

KYOOBIN LEE (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mechanical engineering from the Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2008. From 2008 to 2010, he was a Postdoctoral Scholar with the Center for Neuroscience, KIST. From 2012 to 2017, he was a Principal Researcher with the Samsung Advanced Institute of Technology. Since 2017, he has been an Assistant Professor with the School of Integrated Technology, Gwangju Institute of Science and Technology (GIST). His research interests include vision recognition using deep learning, robot control application using computer vision and split learning for neural networks of cloud computing applications. He is the Editor of the Korea Robotics Society Review.