Enhanced Skin Condition Prediction Through Machine Learning Using Dynamic Training and Testing Augmentation

**TRYAN ADITYA PUTRA [iD], SYAHIDAH IZZA RUFAIDA [iD],
AND JENQ-SHIOU LEU [iD], (Senior Member, IEEE)**
Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan

Corresponding author: Syahidah Izza Rufaida (syahidah.izza@gmail.com)

ABSTRACT In recent years, deep learning has taken the spotlight in automated medical bioimaging. However, the performance of current state-of-the-art score stems primarily from well-tuned parameters and architecture. There is still only limited research focused on dynamic data augmentation, even in the fields of machine learning and computer vision. In this study, we propose a dynamic training and testing augmentation capable of increasing performance significantly. The searching augmentation framework used in this study requires fewer GPU hours than a conventional search algorithm, which needs to train a new model every time augmentation is proposed. Speeding up of the search algorithm is achieved by using Bayesian optimization on a trained model, so we do not have to train a new model every time a new augmentation policy is proposed. The performance of our method is compared with that of a single model and the ensemble model that happens to be the winner of the ISIC 2019 challenge. Furthermore, we use the latest compact yet significantly accurate network architecture EfficientNet as the backbone system. Our method delivers a superior result, and this study also shares the searched augmentation policy utilized, which requires extraordinary resources. Thus, other researchers can use the searched augmentation policies for dermoscopic images to improve performance.

INDEX TERMS Machine learning, augmentation, skin cancer, dermoscopic images.

I. INTRODUCTION

Skin cancer diagnosis has increased significantly over the last decade: 54% from 2009 to 2019 [1]. Skin cancer is an abnormal condition in which skin cells multiply rapidly due to DNA mutation or genetic defects. This abnormality is predominantly caused by protracted exposure to ultraviolet radiation from the sun [2].

Traditionally, dermatologists diagnose skin cancer via physical examination and biopsies. The skin samples taken from the suspected spot during the biopsy are sent to a laboratory for examination. However, obtaining the final diagnosis may take about one week. Meanwhile, about 10,000 people in the United States are diagnosed with skin cancer everyday [1]. The traditional process is time consuming, and the survival rate for early-detected skin cancer is about 98%.

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han [iD].

Unfortunately, with late detection, the survival rate decreases dramatically to 23% [3].

More than two people die every hour due to a skin cancer disease, and the number of deaths in 2019 is estimated at around 7,230 people (4,740 men and 2,490 women) [1]. Furthermore, from age 15 to 39, which is the productive age range, men are 55% more likely to die from melanoma than women in the same age group [4].

Deep Learning has achieved significant breakthroughs in medical image segmentation, such as brain tumor segmentation and retinal vessel segmentation [5]–[7]. Skin cancer segmentation is a challenging area due to the presence of artifacts such as hair, reflection, air and oil bubbles [8], [9]. Another crucial challenge with medical problems is the limited training data.

Data augmentation is one of the essential keys to overcoming the challenge of limited training data. Augmentation that expands data diversity can make a machine learning

model aware of invariance in the data domain. Only a few studies have focused on improving augmentation techniques [10], [11]. However, there is a lot of attention directed at improving machine learning architectures to enhance performance [12]–[15]. For example, most winners of the largest competition on image datasets [16] held annually since 2010 achieved state-of-the-art score by improving the model architecture [12], [13], [17]–[19]. All the winners have used the same augmentation technique, with small changes, as the one proposed in 2012 [17].

There are many problems that come with using image datasets; from medical issues to everyday objects, such as disease, character, and animal classification. Several studies have explored the best augmentation for a single dataset. However, there is no research exploring whether the augmentation can be combined or transferred between datasets. To the best of our knowledge, only Cubuk et al. have explored the possibility of transferring an augmentation policy to another dataset [10]. This study has inspired us to continue the exploration of augmentation policies for images in the medical-domain.

Not only that we explore the application of the probabilistic augmentation framework to images in the medical-domain, specifically to dermoscopic images, to diagnose skin cancer types. We also expand the idea of probabilistic augmentation from training to inference stage. From the experimental results, the application of probabilistic augmentation can be expanded to cover dermoscopic images. We are able to even push the result further by using probabilistic augmentation at inference stage. We share the searched augmentation for future research to be used. Thus, other researchers that use skin datasets can focus on tuning the model parameters alone while leaving the search process for the best augmentation policy by using our result. A brief overview of the probabilistic augmentation framework is presented in Fig. 1. We use EfficientNet for the backbone network [20]. EfficientNet is the latest and the most compact network with high performance.
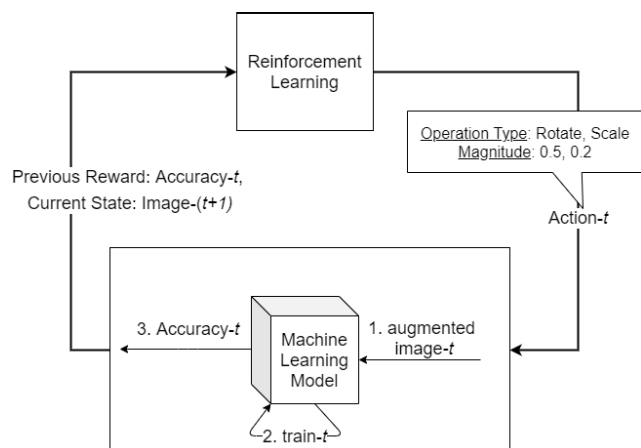
To summarize, the contributions of this paper are as follow:

- We exploit the advance dynamic training augmentation called Fast Autoaugment to train the machine learning model for skin cancer diagnosis.
- We purpose dynamic preprocessing on inference (DPI) in order to adapt the environmental condition when the image was taken.
- We proved the performance of dynamic augmentation and DPI outperform the current state-of-the-art ensemble model for skin diagnosis.
- We present the searched augmentation for future research which need thousand of GPU hours to be found.
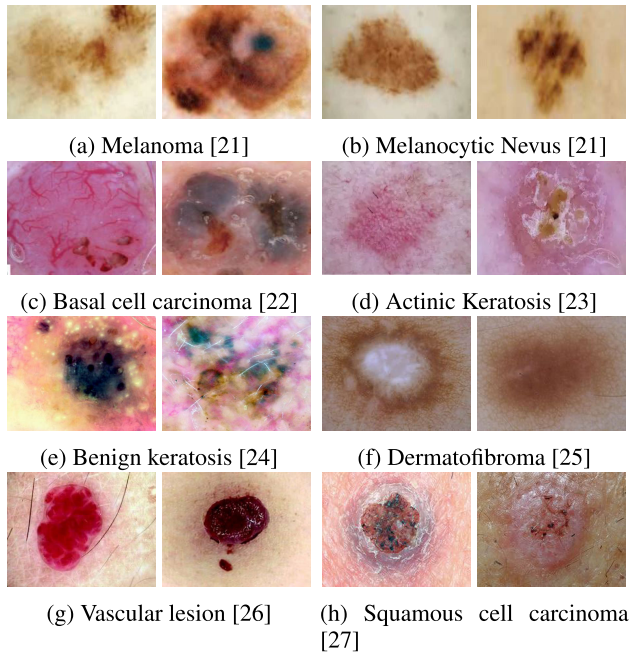
II. SKIN LESIONS

In this section, we explain the general pathological hierarchy of skin lesions and briefly describe the dermoscopic characteristics of eight diagnosis categories from the dataset used in our experiment [28].

Hierarchically, there are two types of skin lesions: melanocytic and non-melanocytic. Based on the origin of the lesion, a melanocytic lesion or melanoma comes from a genetic defect in pigment-producing cells known as melanocytes. A melanoma-type cell can rapidly grow from stage 0 (melanoma in situ) to stage 4 (metastatic melanoma) without proper treatment [29]. Furthermore, based on the degree of malignancy, melanoma lesions can be classified into two categories: benign and malignant. Benign melanoma cells lack the ability to spread or interfere with nearby cells. Oftentimes, the benign melanoma cells will stop growing once they reach a certain size, while malignant melanoma cells never stop growing and can thus press and interfere with surrounding healthy cells [30].

Melanoma is a malignant melanocytic lesion that should be detected as early as possible. However, it is hard to detect melanoma early because its characteristics vary depending on its anatomic site [31]. Generally, the characteristics of melanoma are dark (brown or black) rhomboidal structures and asymmetric pigmented follicular openings (Fig. 2a). Another important characteristic in relation to facial area is the appearance of slate-gray globules and dots. The only effective treatment for melanoma is surgical excision.

A melanocytic nevus (NV) is a benign variant of melanocytic lesions. Unlike the malignant-melanoma, this benign variant, the NV, has a symmetrical structure and color [21]. It is worthy of note that a symmetrical structure in biology refers to a uniform pattern regardless of an irregular shape or the presence of stain. A melanocytic nevus sample can be seen in Fig. 2b.

Basal cell carcinoma (BCC) is a malignant form of non-melanocytic lesions. The BCC cells rarely grow but can spread if left untreated. The dermoscopic descriptions of non-pigmented BCC are arborizing vessels and shiny white-red structureless areas, while the description for pigmented-BCC are: maple leaf-like areas and large blue-gray ovoid

FIGURE 1. Autoaugment Illustration.

(a) Melanoma [21] (b) Melanocytic Nevus [21]

(c) Basal cell carcinoma [22] (d) Actinic Keratosis [23]

(e) Benign keratosis [24] (f) Dermatofibroma [25]

(g) Vascular lesion [26] (h) Squamous cell carcinoma [27]

FIGURE 2. Skin lesion categories.

nests [22]. Dermoscopic samples of BCC are presented in Fig. 2c. There are several treatment options for removing BCC cells: curettage and electrodessication, radiation therapy, photodynamic therapy, and surgery.

Actinic Keratosis (AK) is a variant of mini-squamous cell carcinoma that can be treated without surgery. However, AK should be diagnosed and treated as early as possible before it progresses and evolves into squamous cell carcinoma in situ. AK is commonly found in the facial area, and its characteristics take the form of red pseudo-network (strawberry-network), dotted/glomerular vessels, white structureless areas, a central mass of keratin, and diffuse yellow opaque scales [23], as shown in Fig. 2d.

Benign keratosis (BKL) is a benign type of non-melanocytic lesion. Seborrheic keratoses, solar lentigo, and lichen-planus like keratoses (LPLK) are categorized under the BKL group because they are biologically similar. As shown in Fig. 2e, the dermoscopic criteria for benign keratosis are: milia-like cysts and comedo-like openings [24], [32].

Dermatofibroma is a benign variant of the non-melanocytic type of lesion. The most common characteristics are a central white patch and peripheral pigment network [25]. Dermatofibroma is harmless because it a benign type. Furthermore, optional treatments are available for aesthetic reasons, such as freezing or laser procedures. Sample images are presented in Fig. 2f.

A vascular lesion is a benign variant of the non-melanocytic type of lesion. Angiomas, angiokeratomas, pyogenic granulomas, and hemorrhage are included in this category. As shown in Fig. 2g, the characteristics are a solid purple color with well-circumscribed structures known as red clods or lacunes [31].

TABLE 1. The variant of EfficientNet models.

Model Name	Width Coefficient	Depth Coefficient	Resolution	Dropout Rate
EfficientNet-B0	1.0	1.0	224	0.2
EfficientNet-B1	1.0	1.1	240	0.2
EfficientNet-B2	1.1	1.2	260	0.3
EfficientNet-B3	1.2	1.4	300	0.3
EfficientNet-B4	1.4	1.8	380	0.4
EfficientNet-B5	1.6	2.2	456	0.4
EfficientNet-B6	1.8	2.6	528	0.5
EfficientNet-B7	2.0	3.1	600	0.5

Squamous cell carcinoma is a thickened plaque or nodule. It has varied characteristics depending on the location on which it appears. The characteristics are white circles, keratin, and blood spots [27], shown in Fig. 2h. Primarily, the optional treatment is excision surgery.

III. METHODOLOGY

A. EFFICIENTNET

Recent state-of-the-art architecture is used as backbone in our experiment. Tan and Le proposed a compound-scaling-method to construct an efficient machine learning architecture. In order to get a better performance, many researchers have been trying to scale the width of the network [12], the depth of the network [13], and the resolution of the image [33]. However, none of them define how to balance all the dimensions with a proper ratio. The size of input resolution is a crucial aspect for the model to achieve a superior performance. Furthermore, if the input resolution is increased, then more layers and channels should be added to capture more pattern in order to increase the model's receptiveness.

EfficientNet is a family model that is built from a baseline model which developed through a neural architecture search [34]. The neural architecture search is a framework to search for the best architecture in the search space under certain constrain. The main building block of the network is mobile inverted bottleneck (MBConv) layer which proposed by Sandler *et al.* [35] and Tan *et al.* [36]. The illustration of the baseline model which called as EfficientNet b0 is displayed at Fig. 3. The detail of the scale for other expanded models can be seen in the Table 1.

The performance of EfficientNet is better than all the predecessors with similar computation and memory need by large margin. EfficientNet b4 which has similar number of FLOPs and parameters as ResNet50 [13] performs 6% better (82% vs 76%) at ImageNet dataset. The performance of EfficientNet b4 is the same as NasNet-A [34] with $5.7\times$ lower computational resources and $4.7\times$ lower memory needs.

B. AUTOAUGMENT

AutoAugment uses a reinforcement learning approach to formulate the problem of finding the best augmentation technique for a dataset. In this problem, a state is represented by an image, which can be an original image or an image to which augmentation techniques have been applied. Then,
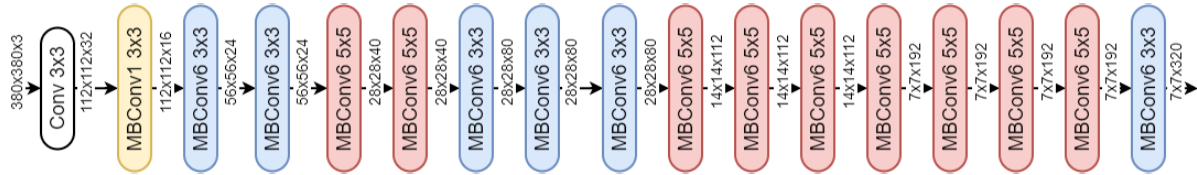
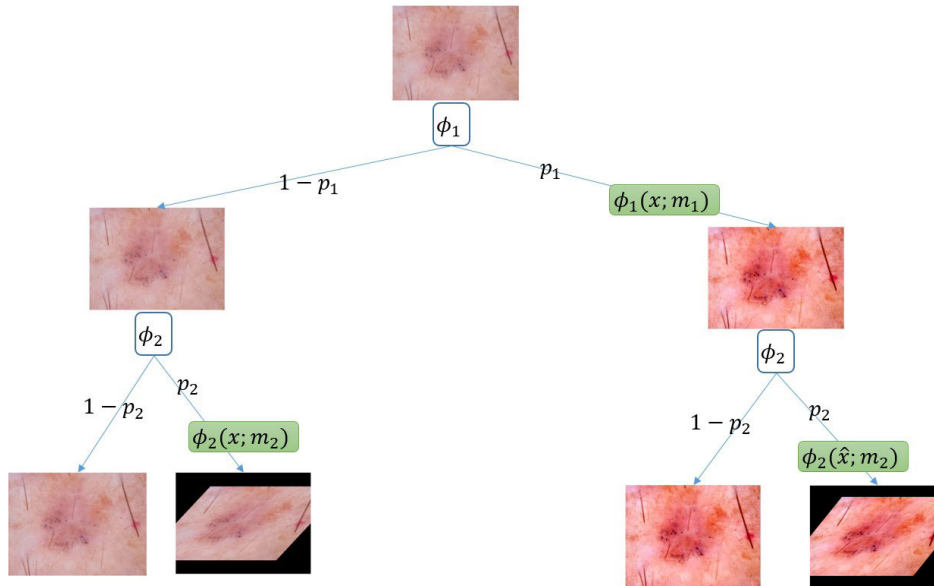**FIGURE 3.** The architecture of EfficientNet b0.

FIGURE 4. Example of random augmentation from the searched policies. Each policy contains two transformations and the probability of its usage.

the action space consists of augmentation techniques for images, such as rotation, translation, and scaling, because the augmentation techniques for images require a magnitude parameter. For example, in rotation transformation, the magnitude of the degree through which an image is to be rotated is required to perform the transformation. Thus, the degree becomes a parameter required for the transformation. The action comes in a pair: an image transformation and a magnitude value as its parameter. A reward in AutoAugment is based on a machine learning models performance on a validation set after applying the augmentation techniques.

For each episode, an RL agent selects the proper augmentation technique to apply to the dataset. Subsequently, a backbone machine learning model is trained with a dataset in which the images have been augmented. Then the final accuracy is used as a reward to the RL agent to improve its next action.

An RL agent will have a sufficient ability to pick the best augmentation policy after several episodes. Because a completely new model is trained in each episode, the time complexity of this framework is expensive. Based on the experiment by Cubuk et al., they needed thousands of GPU hours to improve state-of-the-art score [10].

Lim et al. proposed an improvement, called Fast AutoAugment, that overcomes the time constraint of the AutoAugment framework. Instead of re-training the model each time a new augmentation technique is selected, they only train a model once and pick the best augmentation technique based on inference time [11].

Let us denote ϕ as an image transformation that can be used as an augmentation technique that receives an image x, magnitude m, and probability p. The result of the transformation can either be $\phi(x; m)$ with probability p or the original x itself with probability $1 - p$. The illustration is shown in Fig. 4.

$$\phi(x; m, p) = \begin{cases} \phi(x; m) & \text{with probability } p, \\ x & \text{with probability } 1 - p. \end{cases} \quad (1)$$

Then, several image transformations can be stacked sequentially and annotated as Ψ_n, which contains n augmentation techniques applied sequentially to the input image x. Formally, the definition can be expressed as follows:

$$\Psi_n(x) = \phi_n(\cdots \phi_0(x; , m_0, p_0) \cdots ; , m_n, p_n). \quad (2)$$

At first, a trained dataset δ is divided into k-folds. Each fold consists of a tuple $<\delta_{train}^k, \delta_{val}^k>$. Then, several classifiers f_θ^k are trained concurrently with each dataset δ_{train}^k. Subsequently, using the Bayesian optimization approach [37], the top-n augmentation techniques are selected based on the lowest error rates of classifier f_θ^k when predicting dataset δ_{val}^k.

Algorithm 1 Fast AutoAugment

Data: Classifier f_θ, Dataset δ, Bayesian Opt β, Policies Ψ^{all}, k, t, and n

Result: Policies Ψ^{best}

1 $\Psi^{best} := \emptyset$;

2 $\{< \delta^1_{train}, \delta^1_{val} >, \cdots, < \delta^k_{train}, \delta^k_{val} >\} = $ split(δ, k);

3 **for** $i \leftarrow 0$ **to** k **do**

4   train f_θ with δ^i_{train} ;

5   **for** $j \leftarrow 0$ **to** t **do**

6         $\Psi^j := \beta(\Psi^{all}, f_\theta(\delta^i_{val}))$;

7         $\Psi^j_n := $ extract best(Ψ^j, n) ;

8         $\Psi^{best} := \Psi^{best} \cup \Psi^j_n$;

9   **end**

10 **end**

11 Return Ψ^{best},

All the best augmentation techniques Φ_n are merged from each fold. The final step is to train the entire training set δ with augmentation techniques Φ_{nk}. The complete step by step procedure is outlined in Algorithm 1.

C. BAYESIAN OPTIMIZATION

The primary goal of Bayesian optimization is to find the minimum value of a function $f_\theta(x)$ based on a finite input set X. First, Bayesian optimization uses an approximation function $\hat{f}_\theta(x)$ that represents the probabilistic function $f_\theta(x)$. Then, Bayesian optimization constructs an acquisition function q to exploit the approximation function for picking the next input x of $\hat{f}_\theta(x)$ to be evaluated.

A Gaussian process is selected to be the approximation function $f_\theta(x)$. We define Φ as the cumulative distribution function of the Gaussian distribution. The acquisition function used in our paper is the expected improvement function, which is mathematically defined as:

$$q(x) = \sigma(\hat{f}_\theta(x))[\zeta(x)\Phi(\zeta(x)) + \mathcal{N}(\zeta(x))]. \quad (3)$$

$\zeta(x)$ is the probability of improvement of the current maximum value, which can be calculated using:

$$\zeta(x) = \frac{f(x_{min}) - \mu(x, \{x_n, y_n\})}{\sigma(x, x_n, y_n)}. \quad (4)$$

The $\mu(\cdot)$ and $\sigma(\cdot)$ is the mean and variance of $\cdot$ respectively. In this study, we use the Bayesian optimization to find the best augmentation by using the loss given by the trained model as the target to be minimized.

D. DYNAMIC PREPROCESSING ON INFERENCE (DPI)

In recent years, most researchers have focused on the best network architecture, while only a few have focused on data augmentation. To the best of our knowledge, there is no study focus on dynamic preprocessing framework for inference yet. To achieve dynamic augmentation on inference, we define $P_\vartheta : x_i- > \Psi$ as a convolution neural network function

| (a) MEL | (b) NV | (c) BCC | (d) AK |
| (e) BKL | (f) DF | (g) VASC | (h) SCC |
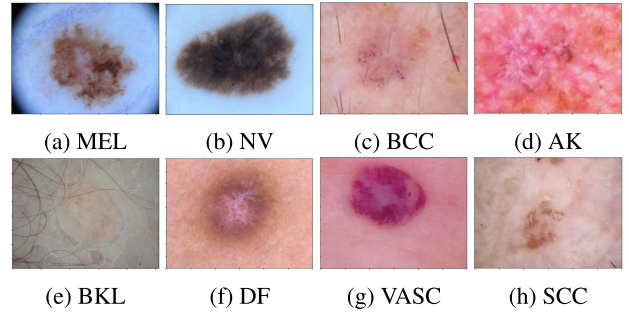
FIGURE 5. A sample from each category.

that maps an image x_i into augmentation space Ψ. The augmentation space Ψ contains all possible augmentations. The training objective can be define as:

$$\min_\vartheta \sum_i \sum_j z_{i,j} \times log(P_{\vartheta,j}(x_i)), \quad (5)$$

where z_i is the top-k augmentation of the i-th data. The index j on $z_{i,j}$ is a boolean indicator whether the j-th augmentation is one of the top-k augmentation. The definition of z_i is mathematically described as:

$$z_i = top_k(-\ell(f_\theta(\Psi(x_i), y_i))), \quad (6)$$

where ℓ is the loss function of the main model. The task of finding the top-k augmentation becomes multilabel classification problem with the top-k augmentation as the label. Because the magnitude of augmentation is continuous, we discretize into several levels of equal chunks. The model for inference augmentation is trained using (x_i, z_i) to understand which augmentation works best for certain images. After successfully training the network with the multilabel data on inference, we use the maximum logit value to select the desired augmentation.

E. DATASET

The International Skin Image Collaboration (ISIC) is the largest dermoscopic dataset in the world with an open license and has a total of 25,331 dermoscopic images with ground-truth labels [28], [31], [38]. The data also comes with clinical information, such as age and gender. The ground-truth labels have 8 different diagnostic categories: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesion (VASC), and squamous cell carcinoma (SCC). A sample from each category in the dataset shown in Fig. 5. The image resolution comes in a 600×450 dimension.

The obvious problem with the dataset is an imbalanced class distribution, shown in Fig. 6. The NV class consists of 12,875 samples, while DF and VASC contain fewer than 300 samples.

IV. EXPERIMENT

In this experiment, we set the augmentation search space as follows: shear, translate, rotate, auto-contrast, invert,
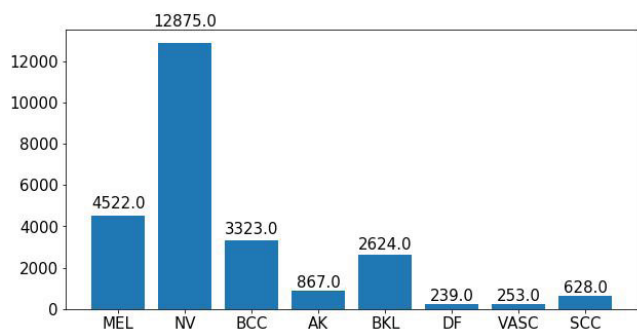
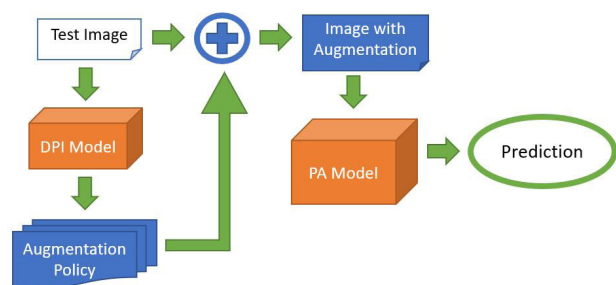**FIGURE 6.** Distribution of class in ISIC dataset.

FIGURE 7. Dynamic Preprocessing Inference (DPI) model is used to select the best augmentation policy for a test image, before being predicted by the main model.
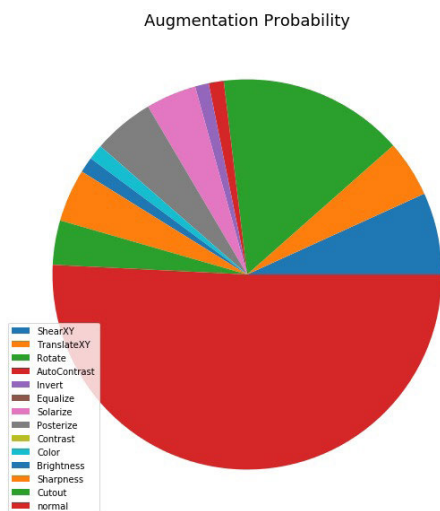

FIGURE 8. Probability of image augmentation based on the searched policies.

FIGURE 9. Average magnitude of each augmentation policy.
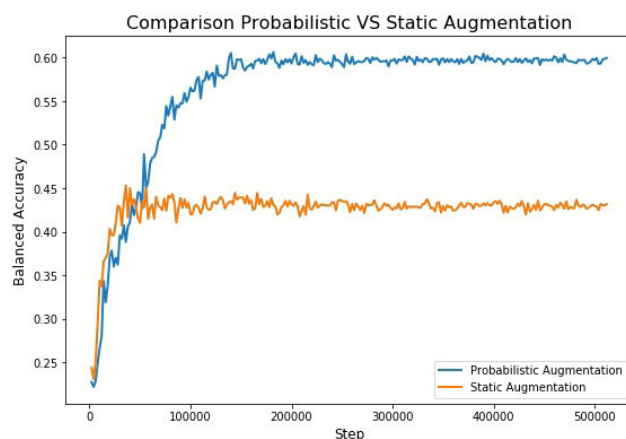
FIGURE 10. Validation accuracy while training progresses.

TABLE 2. Binary classification metrics.

		True Label	
		Positive	Negative
Predictive	Positive	True Positive (TP)	False Positive (FP)
Label	Negative	False Negative (FN)	True Negative (TN)

equalize, solarize, posterize, color, contrast, brightness, sharpness, and cut out. We employ EfficientNet, which is state-of-the-art architecture in deep learning [20] proposed by Tan and Le as a backbone model for classifying dermoscopic images. Precisely, we are using the EfficientNet b4 model with the exact same setting as Tan and Le. We called this model as probabilistic augmentation (PA) model. We used 512,000 steps, with 16 as the batch size. The Adam optimizer [39] is used with a learning rate of 0.0001. We also use an exponential decay of 0.99998. The search for augmentation is done using four independent training models.
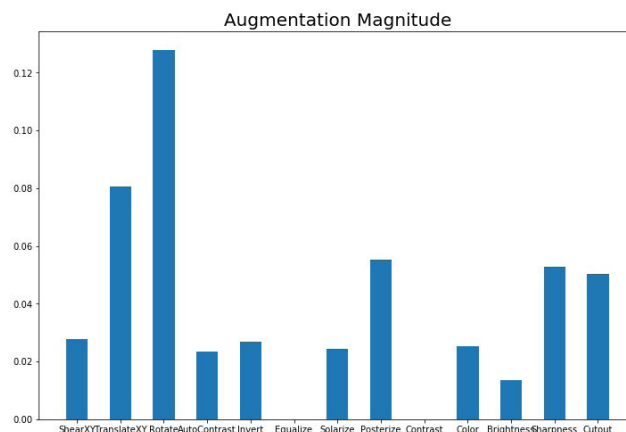
Each independent training model uses two searches. For every search, the top-5 policies with the lowest loss are added to the final augmentation policies. For the final training model, we used the same setup as for the augmentation policy search with a batch size of 64. Using probabilistic augmentation policy, a higher training step is needed to achieve a well-trained model. This is required to ensure stochastic stability in the training phase.

For preprocessing a model on inference, we used the exact same model as the main model, Efficient-net b4. We called the second model as dynamic preprocessing on inference (DPI) model. We train the DPI model as follows: for each possible augmentation, we use 16 levels of magnitude, where the 0 level means no augmentation is applied. Because we have 13 candidates for augmentation with 16 levels, we have $13 \times (16 - 1) + 1 = 196$ possible outputs. The final preprocessing on inference is determined by the

TABLE 3. Comparison result with other studies on the ISIC 2019 challenge vs purpose method based on AUC, accuracy, sensitivity, and F1-Score. Results marked with † are the results measured on the competition test set.

		Average	MEL	NV	BCC	AK	BK	DF	VASC	SCC
AUC	Static Augmentation	0.84	0.77	0.87	0.90	0.83	0.73	0.81	0.96	0.83
	EfficientNet+SerestNext	0.88	0.81	0.89	0.94	0.90	0.78	0.88	0.98	0.88
	EfficientNet+SerestNext †	0.78	0.81	0.88	0.87	0.77	0.76	0.83	0.80	0.74
	PA	0.90	0.83	0.92	**0.95**	0.89	0.86	**0.87**	**0.99**	**0.90**
	PA + DPI	**0.91**	**0.88**	**0.95**	**0.95**	**0.91**	**0.89**	0.85	0.96	0.89
Accuracy	Static Augmentation	0.91	0.84	0.80	0.88	0.94	0.88	0.98	0.98	0.96
	EfficientNet+SerestNext	0.92	0.85	0.82	0.91	0.96	0.89	0.99	0.99	0.97
	EfficientNet+SerestNext †	0.92	0.90	**0.90**	0.89	0.92	**0.93**	0.98	0.98	0.96
	PA	0.93	0.87	0.85	0.93	0.96	0.90	**0.99**	**0.99**	0.97
	PA + DPI	**0.95**	**0.90**	0.87	**0.94**	**0.97**	0.93	**0.99**	**0.99**	**0.98**
Sensitivity	Static Augmentation	0.45	0.36	0.85	0.63	0.36	0.30	0.17	0.75	0.21
	EfficientNet+SerestNext	0.48	0.42	0.91	0.70	0.28	0.38	0.21	0.70	0.28
	EfficientNet+SerestNext †	0.61	**0.68**	0.81	**0.84**	**0.60**	0.56	**0.68**	0.60	**0.52**
	PA	0.61	0.53	0.88	0.75	0.41	0.52	0.60	**0.90**	0.25
	PA + DPI	**0.65**	0.63	**0.93**	0.80	0.47	**0.61**	0.56	0.85	0.31
F1-Score	Static Augmentation	0.40	0.46	0.81	0.58	0.29	0.35	0.16	0.34	0.20
	EfficientNet+SerestNext	0.48	0.51	0.83	0.66	0.34	0.43	0.31	0.51	0.26
	EfficientNet+SerestNext †	0.53	0.69	0.84	0.66	0.42	0.57	0.48	0.51	0.37
	PA	0.57	0.60	0.86	0.73	0.39	0.53	0.48	**0.63**	0.31
	PA + DPI	**0.64**	**0.71**	**0.88**	**0.78**	**0.51**	**0.65**	**0.56**	0.60	**0.42**

maximum logit given by the DPI model. The overall flow of our experiment can be seen in Fig. 7.

To compare our method, we ran two different scenarios. First, we used the searched augmentation policy to train the PA model. Second, we combined the first scenario with dynamic preprocessing on inference. We compared those scenarios with a single model trained using static augmentation and the ensemble model (the ISIC 2019 challenge winner).

V. EVALUATION

In this section, we describe several metrics that can quantify the models performance in our experiment. We divide our metrics into two categories: binary classification metric and multiclass classification metric.

A. BINARY CLASSIFICATION METRIC

Despite the ISIC dataset having eight categories, we can measure the models performance using a binary classification metric. We can calculate the correctness of the predictions for each class separately. Therefore, for each metric, we produce eight different scores for each class.

In a binary classification problem, we can assign a positive-negative class scenario directly. A common metric for binary classification measurement is to calculate how many samples are truly classified as positive (TP), truly classified as negative (TN), falsely classified as positive (FP), and falsely classified as negative (FN) as shown in Table 2.

1) ACCURACY

Accuracy is a metric for determining the quality of the models performance in categorizing both positive and negative classes. The score is calculated by comparing all truly classified data with all data. The formal definition can be written as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

2) SENSITIVITY/RECALL

The sensitivity, also known as the recall score, is used to measure how good the models performance was for discovering positive data from the actual positive data. When the smaller positive sample is categorized as negative, a higher recall score will be achieved. From a medical perspective, it is important to minimize the number of sick persons categorized as healthy because it can jeopardize their life. The recall score can be calculated as follows:

$$Recall = \frac{TP}{(TP + FN)}. \tag{8}$$

3) F1 SCORE

As mentioned previously, the recall score is very important from a medical perspective, and we do not want our model to categorize every sample as positive. Therefore, we also need an F1 score that can balance recall and precision. Precision is used to measure how precise the model's ability is for identifying the actual positive data from positive-predicted
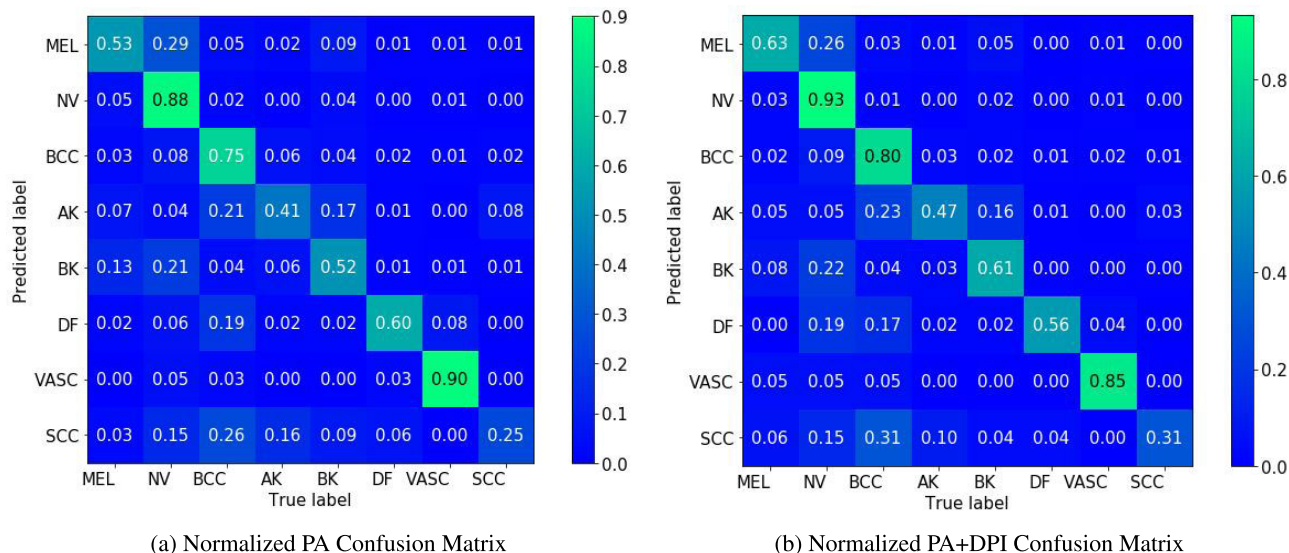
(a) Normalized PA Confusion Matrix

(b) Normalized PA+DPI Confusion Matrix

FIGURE 11. Confidence Matrix Improvement of DPI framework.

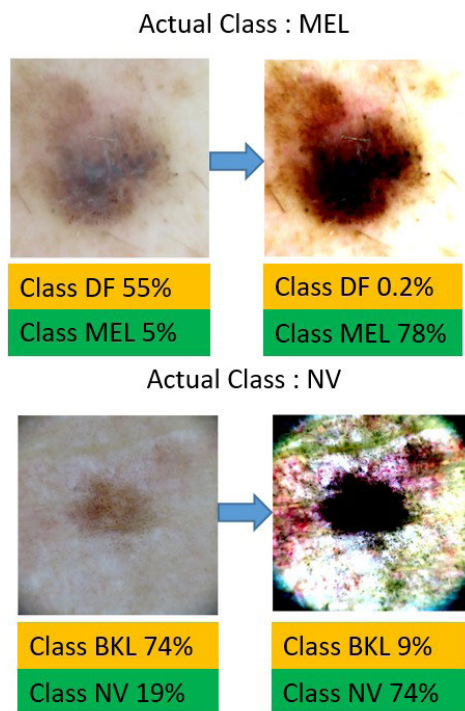**FIGURE 12.** Confidence score with probability Augmentation.

data. The precision score is expressed as:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}. \tag{9}$$

While the formula of F1 score is defined as follows:

$$\text{F1 score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}. \tag{10}$$

4) AREA UNDER CURVE (AUC)

Binary classification comes with a single probability: whether a sample belongs to a positive or negative class.

TABLE 4. Comparison of methods based on balance accuracy score. Result marked with [†] is the result measured on the competition test set.

	Balance Accuracy
Static Augmentation	0.45
EfficientNet+SerestNext	0.48
EfficientNet+SerestNext [†]	0.61
PA	0.61
PA+DPI	**0.65**

Instead of determining a fixed threshold for categorization, an AUC metric is used to measure how well data are ranked based on their prediction probability. The AUC metric is obtained by calculating the area under the receiver operating characteristic (ROC) curve. The ROC is a graph representation of the false-positive rate (FPR) and the true-positive rate (TPR). FPR and TPR are mathematically defined by:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

and

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{12}$$

The final AUC value is a number representing the total area under the ROC curve. The AUC value bounded between 0 and 1 with the higher value is superior.

B. MULTICLASS CLASSIFICATION METRIC
The common problem with medical datasets is a highly imbalanced number of samples for each class. Therefore, to measure the models performance, we cannot use a common accuracy metric that treats each sample equally because it stimulates the model to ignore samples from the minor class. Because we need to treat each class equally, we use balance accuracy to know how well the model performs on the dataset.

TABLE 5.

No.	Augmentation 1	M	P	Augmentation 2	M	P	No.	Augmentation 1	M	P	Augmentation 2	M	P
1	AutoContrast	0.75	0.36	Solarize	0.7	0.52	21	Color	0.58	0.14	Cutout	0.91	0.23
2	TranslateXY	0.66	0.14	TranslateXY	0.78	0.76	22	Brightness	0.62	0.18	Cutout	0.28	0.4
3	Posterize	0.91	0.6	Cutout	0.68	0.65	23	Invert	0.78	0.41	Cutout	0.63	0.32
4	Posterize	0.91	0.59	Cutout	0.49	0	24	ShearXY	0.41	0.03	Cutout	0.52	0.36
5	Posterize	0.59	0.74	Posterize	0.44	0.67	25	Brightness	0.1	0.15	Sharpness	0.79	0.61
6	Color	0.1	0.77	Cutout	0.52	0.98	26	Posterize	0.46	0.76	Sharpness	0.5	0.47
7	Rotate	0.83	0.18	Rotate	0.53	0.75	27	Sharpness	0.6	0.5	Autocontrast	0.59	0.13
8	Color	0.66	0	Rotate	0.7	0.96	28	Brightness	0.24	0.15	Solarize	0.1	0.88
9	Sharpness	0.53	0.4	Rotate	0.47	0.86	29	Brightness	0.13	0.55	Solarize	0.25	0.46
10	Invert	0.61	0.16	Rotate	0.53	0.77	30	Sharpness	0.53	0.34	Sharpness	0.41	0.26
11	TranslateXY	0.17	0.13	ShearXY	0.09	0.91	31	TranslateXY	0.89	0.25	Rotate	0.95	0.92
12	TranslateXY	0.21	0.18	ShearXY	0.04	1	32	TranslateXY	0.81	0.17	Rotate	0.97	0.96
13	Posterize	0.49	0.57	ShearXY	0.5	0.79	33	TranslateXY	0.99	0.27	Rotate	0.61	0.99
14	TranslateXY	0.16	0.28	ShearXY	0.04	0.94	34	TranslateXY	0.12	0.2	TranslateXY	0.67	0.12
15	TranslateXY	0.18	0.09	ShearXY	0.96	0.2	35	TranslateXY	0.89	0.25	Solarize	0.71	0.85
16	Rotate	0.05	0.99	Rotate	0.65	0.06	36	Rotate	0.69	0.38	Rotate	0.89	0.99
17	Solarize	0.16	0.16	Autocontrast	0.23	0.33	37	Solarize	0.21	0.31	ShearXY	0.95	0.79
18	Rotate	0.06	0.86	Rotate	0.67	0.07	38	AutoContrast	0.3	0.17	Rotate	0.05	0.23
19	Posterize	0.61	0.16	Rotate	0.83	0.93	39	Rotate	0.74	0.63	Invert	0.75	0.34
20	Sharpness	0.83	0.93	Sharpness	0.73	0.28	40	Color	0.68	0.11	Rotate	0.02	0.82

1) BALANCE ACCURACY

Balance accuracy is a weighted accuracy score based on the proportion of samples from each class to the total sample. Mathematically, balance accuracy can be written as follows:

$$\text{ACC}_{\text{balance}}(y, \hat{y}, n, c) = \frac{1}{c} \sum \frac{\sum 1(y_i = \hat{y}_i)}{n_i}. \quad (13)$$

We can get an overall score that represents the ability to identify each class equally regardless of the imbalanced samples.

VI. RESULT

We employ the search for the best set of augmentation policies using Fast AutoAugment using Bayesian-optimization. The searched augmentation probability and magnitude can be seen in Fig. 8 and Fig. 9. The procedure to run the search requires around 300 hours of GPU time (Tesla V100) for completion. Thus, we present the details of the searched augmentation in Appendix A to be used for future research.

Based on the results presented, the evidence that augmentation plays a significant role in medical image prediction is undeniable. The comparison of accuracies between static and PA while training can be seen in Fig. 10. We refer to the term "static augmentation" as using a few augmentation policies. The augmentation policies are limited to crop, resize, and horizontal flip. In Fig. 10, static augmentation during training phase provides a faster convergence rate than probabilistic augmentation. However, the weight is not stochastically tuned because we only use a few augmentations, which leads to memorization instead of generalization [10], [40]. Due to not being stochastically tuned, the accuracy cannot grow further, unlike training using probabilistic augmentation. A low convergence rate is a drawback of employing the

considerably massive possibility of augmentations. Greater augmentation possibilities will lead to lower convergence rates.

As part of the multiclass classification, the confussion matrix is one of the most important visualization metrics. We can see in Fig. 11a and 11b that high accuracy is concentrated in classes with a high number of samples. At a glance, combining both PA and DPI gives a better confusion matrix in Fig. 11b than using only PA, as can be seen in Fig. 11a.

The entire experiment using PA and PA plus DPI is compared with a single model trained using static augmentation and the ensemble model that is the first place solution in the ISIC 2019 challenge, which does not use an additional dataset. The ensemble method uses several EfficientNets [20] and SerestNexts [41]. As can be seen in Table 3 and 4, our methods can supersede other methods on all average metrics, including AUC, accuracy, sensitivity, and F1 score. Since the competition does not release label for the test set, we showed the ensemble model performance metrics on both cross validation and the competition leaderboard. We can see in the Fig. 12 the example of images after dynamic augmentation applied. The probability of the actual class is increased significantly and able to alter the network outcome.

VII. CONCLUSION

This study proposes a technique for exploring the possibility of having multiple probabilistic training augmentations and dynamic augmentation on inference. The searching framework uses Bayesian optimization to reduce GPU hours compared to that required for reinforcement learning-based augmentation search [10]. We present the searched augmentation in Appendix A for future research. In addition to the first work to apply augmentation search on medical dataset, our paper also proposed novel technique to perform dynamic

preprocessing on the inference. We use two models to accomplish our goal: the first model is for predicting the skin condition, and the second model is built to predict the best augmentation for the inference stage. We use EfficientNet b4 for both models. The performance of PA and PA plus DPI proved to be better than that of the normal model and the ensemble model that is the ISIC 2019 challenge winner. Although the proposed method has high accuracy, more steps are needed to train the model. The low convergent rate of our method is due to achieving stochastically tuned weight, which is rarely achieved using minimum augmentation.

APPENDIX
THE SEARCHED POLICY
See Table 5.

REFERENCES

[1] (2019). *Cancer Facts Figures 2019*. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf

[2] D. M. Parkin, D. Mesher, and P. Sasieni, ''13. Cancers attributable to solar (ultraviolet) radiation exposure in the UK in 2010,'' *Brit. J. Cancer*, vol. 105, no. S2, pp. S66–S69, Dec. 2011.

[3] R. Z. Conic, C. I. Cabrera, A. A. Khorana, and B. R. Gastman, ''Determination of the impact of melanoma surgical timing on survival using the national cancer database,'' *J. Amer. Acad. Dermatology*, vol. 78, no. 1, pp. 40–46, Jan. 2018.

[4] D. E. Fisher and A. C. Geller, ''Disproportionate burden of melanoma mortality in young US men,'' *JAMA Dermatology*, vol. 149, no. 8, p. 903, Aug. 2013.

[5] N. Ibtehaz and M. Sohel Rahman, ''MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation,'' 2019, *arXiv:1902.04049*. [Online]. Available: http://arxiv.org/abs/1902.04049

[6] J. Jian, F. Xiong, W. Xia, R. Zhang, J. Gu, X. Wu, X. Meng, and X. Gao, ''Fully convolutional networks (FCNs)-based segmentation method for colorectal tumors on T2-weighted magnetic resonance images,'' *Australas. Phys. Eng. Sci. Med.*, vol. 41, no. 2, pp. 393–401, Apr. 2018.

[7] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, ''Dermoscopic image segmentation via multistage fully convolutional networks,'' *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2065–2074, Sep. 2017, doi: 10.1109/TBME.2017.2712771.

[8] A. A. Adeyinka and S. Viriri, ''Skin lesion images segmentation: A survey of the state-of-the-art,'' in *Proc. 6th Int. Conf. Mining Intell. Knowl. Explor. (MIKE)*, Cluj-Napoca, Romania, Dec. 2018, pp. 321–330, doi: 10.1007/978-3-030-05918-7_29.

[9] A. Pennisi, D. D. Bloisi, D. Nardi, A. R. Giampetruzzi, C. Mondino, and A. Facchiano, ''Skin lesion image segmentation using delaunay triangulation for melanoma detection,'' *Comput. Med. Imag. Graph.*, vol. 52, pp. 89–103, Sep. 2016, doi: 10.1016/j.compmedimag.2016.05.002.

[10] 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)*, Long Beach, CA, USA, Jun. 2019, pp. 113–123.

[11] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, ''Fast autoaugment,'' in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 6662–6672.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[13] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[14] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 5987–5995, doi: 10.1109/CVPR.2017.634.

[15] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, ''Regularized evolution for image classifier architecture search,'' *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, vol. 33, pp. 4780–4789, Jul. 2019, doi: 10.1609/aaai.v33i01.33014780.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ''ImageNet large scale visual recognition challenge,'' *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[18] M. D. Zeiler and R. Fergus, ''Visualizing and understanding convolutional networks,'' in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, Sep. 2014, pp. 818–833, doi: 10.1007/978-3-319-10590-1_53.

[19] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1409.1556

[20] M. Tan and Q. V. Le, ''EfficientNet: Rethinking model scaling for convolutional neural networks,'' in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, May 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html

[21] C. Rosendahl, A. Cameron, I. Mccoll, and D. Wilkinson, ''Dermatoscopy in routine practice:'Chaos and clues','' *Austral. Family Physician*, vol. 41, p. 482, Jul. 2012.

[22] A. Lallas, Z. Apalla, G. Argenziano, C. Longo, E. Moscarella, F. Specchio, M. Raucci, and I. Zalaudek, ''The dermoscopic universe of basal cell carcinoma,'' *Dermatol. Practical Conceptual*, vol. 4, pp. 11–24, Jul. 2014.

[23] I. Zalaudek, J. Giacomel, K. Schmid, S. Bondino, C. Rosendahl, S. Cavicchini, A. Tourlaki, S. Gasparini, P. Bourne, J. Keir, H. Kittler, L. Eibenschutz, C. Catricalà, and G. Argenziano, ''Dermatoscopy of facial actinic keratosis, intraepidermal carcinoma, and invasive squamous cell carcinoma: A progression model,'' *J. Amer. Acad. Dermatol.*, vol. 66, no. 4, pp. 589–597, Apr. 2012.

[24] H. P. Soyer, G. Argenziano, R. Hofmann-Wellenhof, and R. H. Johr, Eds., *Seborrheic Keratosis Including Lichen Planus-Like Keratosis*. Berlin, Germany: Springer, 2007, pp. 313–328.

[25] P. Zaballos, S. Puig, A. Llambrich, and J. Malvehy, ''Dermoscopy of dermatofibromas: A prospective morphological study of 412 cases,'' *Arch. Dermatol.*, vol. 144, no. 1, pp. 75–83, Jan. 2008.

[26] P. Zaballos, M. Carulla, F. Ozdemir, I. Zalaudek, J. Bañuls, A. Llambrich, S. Puig, G. Argenziano, and J. Malvehy, ''Dermoscopy of pyogenic granuloma: A morphological study,'' *Brit. J. Dermatol.*, vol. 163, pp. 1229–1237, Dec. 2010.

[27] C. Rosendahl, A. Cameron, G. Argenziano, I. Zalaudek, P. Tschandl, and H. Kittler, ''Dermoscopy of squamous cell carcinoma and keratoacanthoma,'' *Arch. Dermatol.*, vol. 148, no. 12, pp. 1386–1392, Dec. 2012, doi: 10.1001/archdermatol.2012.2974.

[28] 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)*, Washington, DC, USA, Apr. 2018, pp. 168–172, doi: 10.1109/ISBI.2018.8363547.

[29] A. Bhattacharya, A. Young, A. Wong, S. Stalling, M. Wei, and D. Hadley, ''Precision diagnosis of melanoma and other skin lesions from digital images,'' in *Proc. AMIA Summits Transl. Sci.*, Jul. 2017, pp. 220–226.

[30] P. A. Lio and P. Nghiem, ''Interactive atlas of dermoscopy,'' *J. Amer. Acad. Dermatol.*, vol. 50, no. 5, pp. 807–808, 2004.

[31] P. Tschandl, C. Rosendahl, and H. Kittler, ''The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,'' 2018, *arXiv:1803.10417*. [Online]. Available: http://arxiv.org/abs/1803.10417

[32] R. P. Braun, ''Dermoscopy of pigmented seborrheic keratosis: A morphological study,'' *Arch. Dermatol.*, vol. 138, no. 12, p. 1556, Dec. 2002.

[33] Y. Huang, Y. Cheng, A. Bapna, O. Firat, M. Xu Chen, D. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu, and Z. Chen, ''GPipe: Efficient training of giant neural networks using pipeline parallelism,'' 2018, *arXiv:1811.06965*. [Online]. Available: http://arxiv.org/abs/1811.06965

[34] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–16, [Online]. Available: https://openreview.net/forum?id=r1Ue8Hcxg

[35] 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.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4510–4520. http://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html

[36] 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)*, Long Beach, CA, USA, Jun. 2019, pp. 2820–2828. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Tan_MnasNet_Platform-Aware_Neural_Architecture_Search_for_Mobile_CVPR_2019_paper.html

[37] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 2960–2968. [Online]. Available: http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms

[38] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "BCN20000: Dermoscopic lesions in the wild," 2019, *arXiv:1908.02288*. [Online]. Available: http://arxiv.org/abs/1908.02288

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[40] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–15, [Online]. Available: https://openreview.net/forum?id=Sy8gdB9xx

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

TRYAN ADITYA PUTRA received the bachelor's degree in computer engineering and the master's degree in electrical engineering from Universitas Indonesia. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, National Taiwan University of Science and Technology. His research interests include machine learning, artificial intelligence, data mining, and mathematical modeling.

SYAHIDAH IZZA RUFAIDA received the bachelor's and master's degrees in computer science from Universitas Indonesia. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, National Taiwan University of Science and Technology. Her research interests include machine learning, data mining, and image processing.

JENQ-SHIOU LEU (Senior Member, IEEE) received the B.S. degree in mathematics and the M.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1991 and 1993, respectively, and the Ph.D. degree on a part-time basis in computer science from National Tsing Hua University, Hsingchu, Taiwan, in 2006. He was a Research and Development Engineer with Rising Star Technology, Taiwan, from 1995 to 1997. He was an Assistant Manager with the Telecommunication Industry (Mobital Communications and Taiwan Mobile), from 1997 to 2007. In February 2007, he joined the Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology as an Assistant Professor. From February 2011 to January 2014, he was an Associate Professor. From February 2014 to July 2017, he was a Professor and the Vice Chairman. Since August 2017, he has been a Professor and the Chairperson. His research interests include heterogeneous network integration, mobile service and platform design, distributed computing (P2P, cloud computing), and green and orange technology integration. He has published extensively in these areas, with 58 SCI indexed journal articles, 57 conference papers or book chapters, and led 12 MOST project, 12 industry-academia projects, and two cross-university projects in the past ten years.

• • •