IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 July 2023, accepted 24 July 2023, date of publication 28 July 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299819

RESEARCH ARTICLE

Looking Closer to the Transferability Between Natural and Medical Images in Deep Learning

SYAHIDAH IZZA RUFAIDA[1], TRYAN ADITYA PUTRA[1], JENQ-SHIOU LEU[1], (Senior Member, IEEE), TIAN SONG[2], (Member, IEEE), AND TAKAFUMI KATAYAMA[2], (Member, IEEE)

[1]Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan
[2]Department of Electrical and Electrics Engineering, Tokushima University, Tokushima 770-8506, Japan

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

This work was supported by the Taiwan University of Science and Technology (TAIWAN TECH) and Tokushima University (TU) Joint Research Program under Grant TU-NTUST-110-04.

ABSTRACT Transfer-learning has rapidly become one of the most sophisticated and effective techniques in dealing with medical datasets. The most common transfer-learning method uses of a state-of-the-art model and its corresponding parameters as the starting point for new tasks. Recent studies have found that transfer-learning between medical and natural images has minimal advantages, attributed to their different characteristics, even with sufficient data and iterations. This study employs a meta-learning technique, building upon the traditional transfer learning approach, to explore the potential of natural tasks as a starting point for analyzing medical images. In addition, this study investigates the performance of transferring the searched augmentation from natural to medical images. Several studies proposing search algorithms for data augmentation argue that the augmentation techniques can be effectively transferred across different datasets. The results revealed that the transferability between natural and medical images leads to reduced performance owing to the characteristic difference between medical and natural searched augmentation.

INDEX TERMS Data augmentation, medical image dataset, meta-learning, natural images dataset, transfer-learning.

I. INTRODUCTION

Transfer learning is the process of solving new problems using previously acquired knowledge. Deep neural networks commonly employ pre-trained models, which have already learned complex tasks, as a foundation for transfer learning to learn new tasks. The pre-trained model is expected to perform better than models trained from scratch or random initialization. Choosing a proper initialization approach for deep learning is important to prevent exploding or vanishing gradient.

Transfer learning has been applied across various domains, particularly in the medical domain [1], [2], [3], [4]. Most studies have trained state-of-the-art networks [5], [6], [7] that have successfully demonstrated superior performance on the most challenging natural-image datasets: ImageNet [8], containing 1000 categories. These same models have been applied to medical datasets [9], [10], [11] with promising results with medical datasets.

Notwithstanding the popularity of transfer learning [12], researchers do not agree on the significance of its supremacy. Several researchers have demonstrated that the performance difference between random initialization and pre-trained models is not substantial [13], [14]. With more iterations, random initialization can achieve performance that is comparable to that of pre-trained models [13], [15]. In practical applications of transfer learning, fewer training iterations are typically required, making it a cost-effective approach for solving new problems. Given that numerous training iterations can result in network memorization [16], particularly when dealing with small datasets, and since medical datasets are typically small, it is important to exercise

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

caution when training deep neural networks on medical data. Flennerhag et al. suggest that in the transfer learning paradigm, the final weights of the model are often in close proximity to the initial parameters, whereas starting from scratch involves a broader search space for the optimal weights [17].

The benefit of transfer learning from the medical-domain perspective has empirically proved to be minimal [14]. Experimental results of Raghu et al. demonstrate that transfer learning from an ImageNet pre-trained model to medical images offers few benefits. This is because natural images such as ImageNet focus on object-shape recognition [18], [19], whereas medical images predominantly focus on local texture recognition [14]. Medical images are similar to the fine-grained challenge images because medical images comprise the same objects from the same body area. For example, EyePACS retina images comprise human eyes, chest X-rays show the human skeleton, and dermoscopic datasets capture the human skin. The difference is the small anomaly in textures that indicate the characteristics of some diseases. Consequently, recent studies by Kornblith et al. demonstrate that pre-trained ImageNet models do not perform well in fine-grained classification problems [15]. Furthermore, another reason to uphold the limitation of transfer learning is that deep network architectures that are generally used for heavy datasets such as ImageNet are over-parameterized to learn medical imaging datasets, which generally have far fewer categories than ImageNet [9], [11], [20]. Therefore such insights can inform the development of improved transfer learning methods that can be applied across diverse domains.

Meta-learning is the process of learning the training process of several tasks to allow the model to quickly learn a new task [21]. This is similar to how humans never learn from scratch because our previous knowledge helps us to learn new things faster and better. Transfer meta-learning may become a method of transfer learning to achieve high performance with few iterations. In this study, we evaluate the performance of transfer meta-learning in various medical image datasets. To the best of our knowledge only a few used meta-learning for medical image datasets [22], [23].

Another challenge with medical imaging datasets is imbalanced samples. There are generally far more healthy samples than sick samples, especially for rare diseases. One way to tackle this problem is by using data augmentation to enable machine learning to learn a wider variation in the data distribution. Most augmentation techniques involve manual feature engineering [24], [25], [26]. A recent study proposed a method to automatically find optimal augmentation techniques that are transferable between datasets [27]. As a result of using various equipment and instruments to gather medical data, several methods are used differently for analysis, although there may be hidden correlations between them. Similar to the relationship between natural images that enables the transferability of augmentation techniques, we investigate the underlying connections between medical images that facilitate the transfer of these techniques. Despite variations in device acquisition, our findings suggest that augmentation techniques applied to dermoscopic images and X-ray scans can be successfully transferred.

The contributions of this paper are:

- We show that transfer meta-learning is a better method of transfer learning that enables machine-learning models to learn medical images faster.
- However, our findings also reveal that despite its faster convergence rate, transfer learning is outperformed by training from scratch when longer training phases and larger datasets are employed.
- Our study contradicts the assumption of augmentation transferability between datasets, as previously stated in [27] and [28] especially between natural and medical datasets.
- We explored the characteristics of the searched augmentation in medical images and compared it to that of natural images.
- We found the scheme where strong augmentation even hurts the accuracy especially in fine-grained details such as medical image datasets as supported by [29]

The remainder of this paper is organized as follows: Section II presents several adaptations of transfer learning in medical datasets, several studies that highlight transfer-learning drawbacks, some versions of the meta-learning technique, and a few studies that explore augmentation techniques. Section III explains the meta-learning definition and automated augmentation. Section IV presents the dataset used in the experiment. Then, Section V describes the configuration of the experiment. Section VI presents and analyses the experimental results. Finally, the last section concludes the paper.

II. RELATED WORK

Several studies have explored the application of weight transferability on various medical datasets by using a pre-trained model from the ImageNet dataset. The pre-trained models of AlexNet [30] and VGG-net [31] were used to categorize retina images to identify their diabetic retinopathy levels [32]. Using a similar technique, several pre-trained state-of-the-art networks, such as AlexNet, VGG-net, Inception [7], ResNet [5], and DenseNet [33], were applied to classify X-ray images to recognize their pathology types [10], [20], [34]. Ding et al. used a pre-trained Inception v3 architecture to recognize Alzheimer's disease from positron emission tomography images [35]. Recently, a pre-trained Inception model was used to predict embryo quality based on morphological assessments [36]. Furthermore, DenseNet, ResNet, Inception v3, and Inception v4 were trained with dermoscopy images to categorize skin lesions [11].

Several researchers investigated the benefit of transfer learning through pre-trained models [13], [14], [15], [37]. Some believe that pre-trained models can drive the optimal solutions away because of the diversity between learned tasks [17], [38]. The effects of transfer learning in medical

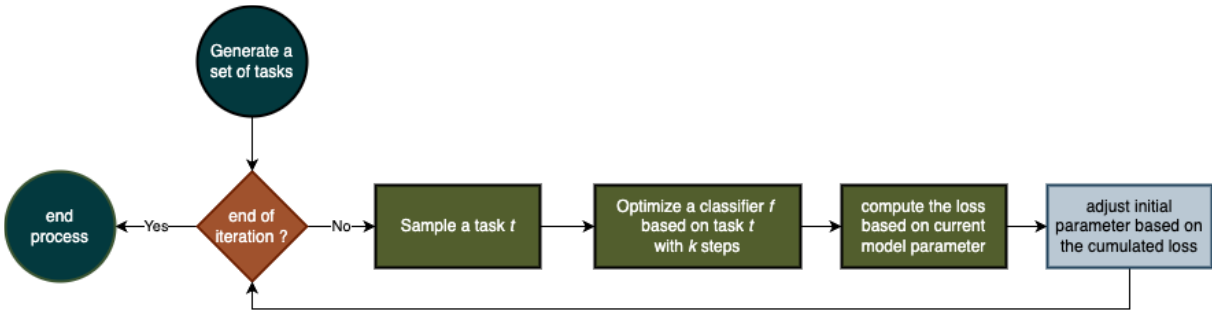


FIGURE 1. Meta-learning flowchart.

imaging are elaborated further in [14]. Meta-learning assists in overcoming such diversity. Meta-learning is the process of learning the training process so that the model can adapt quickly to new tasks [39], [40]. A meta-learner is generally trained to observe how the model updates the parameters when learning different tasks [41], [42]. Three main approaches have been focused on recently. The first is a meta-learner that is trained by comparing new samples to previously learned samples [43], [44]. The second uses the gradient of the model as the RNN [45] input to generate the optimal updates of the parameters during training [40], [46]. The third approach builds a meta-learner to find the optimal starting point and allow the model to learn new tasks faster through limited gradient updates [47], [48]. Meta-learning is frequently used for few-shot learning problems, where the data samples from one class category are limited [44]. Likewise, the medical domain has the same characteristics where the samples for rare diseases are limited [49].

Only a limited number of studies have investigated the use of augmentation techniques in general. Most studies focus on deep-learning architectures to improve machine-learning performance [5], [7], [31]. For the most widely used and common dataset such as ImageNet, the augmentation-technique base has not been changed since 2012 [30]. Autoaugment [27] is the process of automatically searching for the optimal augmentation policy based on the reward score produced by a reinforcement-learning agent [50]. However, the drawback of autoaugment is that the running time is thousands of GPU hours to obtain the most optimal augmentation policy from a single dataset. The scenario is improved by embedding Bayesian optimization [51] as a heuristic approach to obtain the optimal augmentation policy [28].

III. METHODOLOGY

In this section, we describe the general concepts of meta-learning and heuristic augmentation search. We focus on fast autoaugment as an example of a heuristic augmentation search.

A. META-LEARNING

Meta-learning aims to learn the learning process across numerous tasks. Meta-learning increases the flexibility of the learning process and reduces inductive bias. Inductive

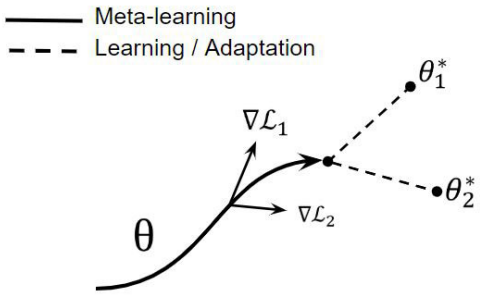


FIGURE 2. The illustration of meta-learning.

Algorithm 1 Meta Learning: FOMAML

Input: Classifier f, Distribution over tasks $p(\pi)$, Parameter θ
Parameter: η steps
Output: Best Initialization Value θ

$\theta_0 =$ initialize θ;
for $i \leftarrow 0$ **to** η **do**
  $\pi_i =$ sample task batch from $p(\pi)$;
  while *not done* **do**
    Evaluate $\triangle_\theta L_{\pi_i}(f(\theta))$
  end
  $\theta = \min_\theta \sum_{\pi_i \sim p(\pi),j} \sum_j L_{\pi_i}(f_{\theta_j})$;
end

bias is the behavior of learning certain tasks if the bias matches the learning problem. The most common way to transfer knowledge in a deep, parameterized model is to use pre-trained parameters as the starting point; therefore, we focus on meta-learning algorithms that focus on the starting point such that the learning process can rapidly adapt to a new task.

To the best of our knowledge, there are two ways of performing meta-learning based on parameter initialization. First, during training, we identify the model parameters that are sensitive to certain tasks. The sensitivity of the parameters can be quantitatively measured by the loss changes corresponding to parameter changes. Based on this, we modify the parameters such that for all possible tasks, the changes are insignificant with respect to the parameters and therefore lead to faster convergence rates. There are

Algorithm 2 Meta Learning: Reptile

Input: Distribution over tasks $p(\pi)$, Parameter θ
Parameter: η steps
Output: Best Initialization Value θ

$\theta_0 =$ initialize θ;
for $i \leftarrow 0$ **to** η **do**
  $\tau =$ sample task batch from $p(\pi)$;
  while *not done* **do**
    θ_k : k steps of optimization algorithm;
  end
  $\theta = \theta_0 + \epsilon(\theta_k - \theta_0)$;
end

two major studies on this approach; FOMAML [48] and Reptile [47]. Second, we employ a technique to navigate across the loss surface to determine the optimal initial point near local minima. The concept of navigating through the loss surface has a different assumption from the first approach. The first approach considers the changes in the parameters as a method of knowing the complexity of the training, while the second approach considers it to be a way of achieving local minima to ensure a more accurate measurement. The evaluation of the complexity of the training process is the main difference between the two approaches. Flennerhag et al., under the second paradigm, proposed the latest and best method so far.

We denote a model f which maps input x to output y. The learning task $\pi = (f_\pi, p_\pi, g_\pi)$ is to learn how to draw a relation between x and y under the distribution $p_\pi(x, y)$. The learning process is done by using a gradient rule update g_π. Consider the initial network parameter is θ_0, as the training progresses, the parameters will change by $\theta_{i+1} = g_\pi(\theta_i)$ until the error converges.

In general, meta-learning has inner and outer parameter updates. As can be seen in Fig. 2, the dotted line is the inner update while the bold line is the outer update. The inner update is an update to learn the given task while the outer update is an update to learn the training process. The main difference between FOMAML and Reptile is in how they perform the outer parameter update. The loss function in FOMAML is mathematically written as

$$\min_{\theta} \sum_{\pi_i \sim p(\pi), j} \sum_{j} L_{\pi_i}(f_{\theta_j}). \tag{1}$$

As can be seen in the equation 1, FOMAML attempts to minimize the overall gradient update using the best initial parameter possible. In contrast, the outer parameter update for Reptile is given as follows:

$$\theta = \theta_0 + \epsilon(\theta_k - \theta_0). \tag{2}$$

where ϵ is a parameter to maintain the distance of the update while the term $\theta_k - \theta_0$ is the substitution of gradient update. Reptile does not need the outer loss function while FOMAML needs it.

Algorithm 3 Meta Learning: Leap

Input: Distribution over tasks $p(\pi)$, Parameter θ
Parameter: η steps
Output: Best Initialization Value θ

$\theta_0 =$ initialize θ;
for $i \leftarrow 0$ **to** η **do**
  $\tau =$ sample task batch from $p(\pi)$;
  ψ_τ^0 : initialize task baseline
  while *not done* **do**
    update baseline ψ_τ^{i+1} based on equation 5;
    following baseline $\theta_\tau^i \leftarrow \psi_\tau^i$
    increment $\triangle \bar{F}(\theta_0, \Psi)$ with gradient pull (eq. 6)
  end
  $\theta = \theta_0 + \triangle \bar{F}(\theta_0, \Psi)$
end

Leap [17] is a method that focuses on loss function navigation. The concept of Leap is to calculate the length of parameter updates during training. The length of parameter updates at iteration t are calculated using the length of gradient updates denoted by γ which is calculated as

$$\text{Length}(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))} dt. \tag{3}$$

Leap calculates the cumulative chordal distance using the following formulation

$$d(\theta_0, M) = \sum_{0}^{K-1} ||\gamma_{i+1} - \gamma_i||_2^p, \text{where } p \in 1, 2, \tag{4}$$

where M is a task manifold. Flennerhag et al. used the following loss function

$$\begin{aligned} &\min_{\theta_0} f(\theta_0) = \mathbb{E}_{\pi \sim p(\pi)}[d(\theta_0; m_\pi] \\ &\text{s.t.} \quad \theta_{i+1} = g_\pi(\theta_i), \\ &\qquad \theta_0 \in \Theta = \cap_\pi(\theta_0 | f_\pi(\theta_k) \leq f_\pi(\Psi_k)), \end{aligned} \tag{5}$$

where Ψ_k is the actual gradient path and θ_k is the final parameter after gradient pull. For the outer update, the update value is incrementally added to the inner training process. The value of the update is defined as follows:

$$\begin{aligned} &\triangle \bar{F}(\theta_0, \Psi) \\ &\quad = -pE_{\pi \sim p(\pi)}[\sum_{i=0}^{k-1} J_i(\theta_0^T) \\ &\qquad \times (\triangle f_i \triangle f(\theta_i) + \triangle \theta_i)(||\gamma_{i+1}^- - \gamma_i||_2^p)^{p-2}], \end{aligned} \tag{6}$$

where j_i is the Jacobian of θ_i with respect to the initialization, $\triangle f_i = f(\psi_{i+1}) - f(\theta_i)$, and $\triangle \theta_i = \Psi_{i+1} - \theta_i$.

B. FAST AUTOAUGMENT

Proposed by Lim et al., fast autoaugment is a technique that improves the previous version of autoaugment, which

Algorithm 4 Fast AutoAugment

Input: Classifier f, Policies Θ_{all}, Dataset $\varkappa$
Parameter: η, t, and k
Output: Augmentation Policies Θ^{best}

$\{\varkappa^1, \cdots, \varkappa^\eta\} = \text{split}(\varkappa)$;
for $i \leftarrow 0$ **to** η **do**
  $\varkappa^i_{train}, \varkappa^i_{val} = \text{split}(\varkappa^i)$;
  train f with $\varkappa^i_{train}$;
  for $j \leftarrow 0$ **to** t **do**
    $\Theta_j :=$ observe $\Theta_{all}(\varkappa^i_{val})$ through α of f ;
    $\Theta^k_j :=$ extract top-$k(\Theta_j)$;
    $\Theta_{best} := \Theta_{best} \cup \Theta^k_j$;
  end
end

required thousands of hours of GPU to find the optimal augmentation policy from one dataset [27]. It uses the lowest inference loss to obtain the optimal augmentation strategy, allowing for a one-time training proces. Denote Θ as an augmentation policy that transforms an input image x with probability ρ, an augmentation policy requires another magnitude input δ reflecting how strong the transformation is done. As an example, a rotation policy requires the degree of rotation for the image to be rotated.

$$\Theta(x, \rho, \delta) = \Theta(x, \delta) \text{ with probability } \rho, \text{ otherwise } x. \quad (7)$$

Multiple augmentation policies can be concatenated, resulting in a sequence of augmentation policies. The final output of fast autoaugment is a set of augmentation policies that maximize the final prediction score.

To find the optimal augmentation strategy for a dataset $\varkappa$, we find augmentation strategies Θ^* which maximize accuracy from an evaluation function α:

$$\alpha(f(\Theta^*(\varkappa_{val}))) < \alpha(f(\varkappa_{val})) \quad (8)$$

First, the dataset $\varkappa$ is divided into η chunks for the sake of generalization. Then, each chunk is partitioned into $\{\varkappa^i_{train}\}^\eta_{i=1}$ and $\{\varkappa^i_{val}\}^\eta_{i=1}$. Afterward, we train the f^i model with $\varkappa^i_{train}$ without any augmentation policy in parallel. The augmentation policies that can increase the accuracy score of dataset $\varkappa^\eta_{val}$ by trained models f^η are selected as the optimal augmentation policies. As can be seen in Algorithm 4, we only train the model η-times. For each trained model, we search the best policy t-times. Meanwhile, autoaugment needs to train the model $(\eta \times t)$-times to obtain the same number of trials thus requiring more GPU hours.

C. EFFICIENT-NET

The Efficient-Net architecture [52] is the baseline network in our experiment. Efficient-Net is a recent state-of-the-art network that achieves high accuracy in ImageNet dataset and outperforms popular state-of-the-art networks such as ResNet [5], Inception [53] and NasNet [54]. The Efficient-Net architecture balances all dimensions of a machine learning architecture such as depth, width, and resolution. To study high-resolution images, the network should have deeper and wider layers. The insight comes from previous studies that discovered a relationship between the width of a network and its depth [55], [56].

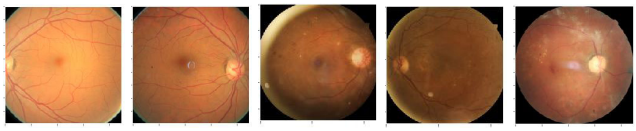


FIGURE 3. Fundus photography images from EyePACS dataset.

Tan and Le proposed a compound scaling mechanism that uses a compound coefficient ϑ to evenly scale the depth, width, and input resolution of the network, which is represented by a constant coefficient ϕ, ω, and Υ respectively. All the constant coefficients can be determined by a small grid search from the simplest model, such that

$$1 \leq \phi \cdot \omega^2 \cdot \Upsilon^2 \approx 2. \quad (9)$$

Note that the floating-point operation per second (FLOPs) of a convolutional network is proportional to ϕ, ω^2 and Υ^2. Therefore, if we increase the network depth twice then the number of FLOPs will also increase twice. However, if the width or resolution is increased twice, the number of FLOPs will increase four times.

$$\text{FLOPs } (f_\theta) = (\phi \cdot \omega^2 \cdot \Upsilon^2)^\vartheta \approx 2^\vartheta. \quad (10)$$

Considering that our computational resources are based on a binary machine, Tan and Le restrict that $\phi \cdot \omega^2 \cdot \Upsilon^2 \approx 2$. Consequently, if Efficient-Net is scaled by ϑ, the number of FLOPs is approximately increased by 2^ϑ.

IV. DATASET

In this section, we introduce two types of datasets: medical and natural image datasets. Each type has its own characteristics depending on its purpose.

A. MEDICAL IMAGE DATASETS

In our experiment, we used three different datasets from medical imaging as the primary benchmark: EyePACS, ChestX-ray8, and Skin ISIC2019 datasets.

1) EyePACS DATASET

Diabetic retinopathy is one of the diseases that can be recognized from retina fundus photography. In this experiment, we used a dataset published by EyePACS which is a clinical institute for diabetic retinopathy (DR) screening in California, USA [9]. All fundus photography images were labeled into five classes: normal, mild DR, moderate DR, severe DR, and proliferative DR according to International Clinical Diabetic Retinopathy [57]. The final label tagged on each image was approved by eight ophthalmologists as the standard reference. Figure 3 displays a sample image from each category. All 9963 images had a 587 × 587 resolution. The distribution of the images per class is shown in Fig. 4.

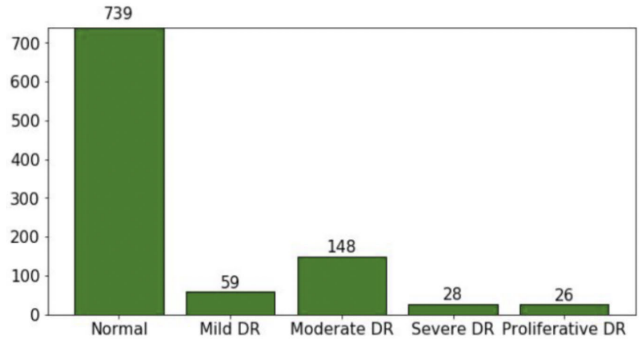


FIGURE 4. Number of images of EyePACS dataset per class.

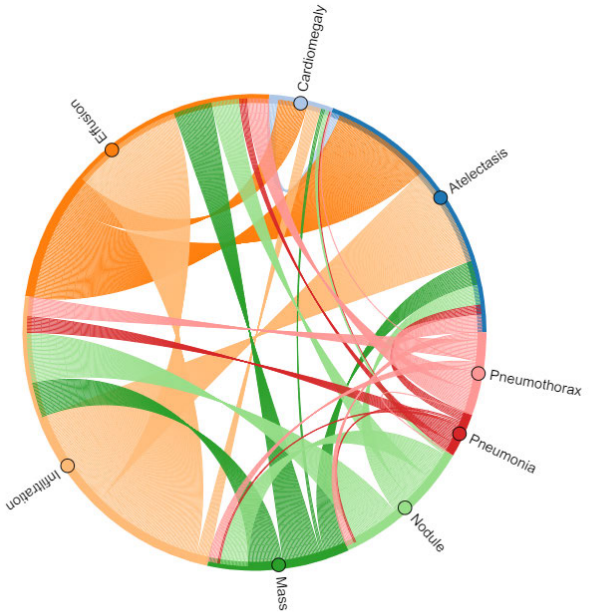


FIGURE 5. Data distribution of Chest X-ray dataset.

2) ChestX-ray8

ChestX-ray8 dataset is a multilabel dataset mined from Picture Archiving and Communication Systems (PACS), which is a warehouse for a tremendous number of X-ray images [10]. The pathology labels listed in ChestX-ray8 include Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax. If no label appears, then the image is categorized as normal or healthy. Original X-ray images with 3000 × 2000 dimensions were resized into 1024 × 1024 dimensions without removing substantial content.

3) SKIN ISIC2019

International Skin Image Collaboration (ISIC) published 25.331 dermoscopic images with an open license in 2019. All images are categorized into eight classes: actinic keratosis (AK), basal cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevus (NV), squamous cell carcinoma (SCC), and vascular lesion (VASC) [58], [59], [60]. Sample images

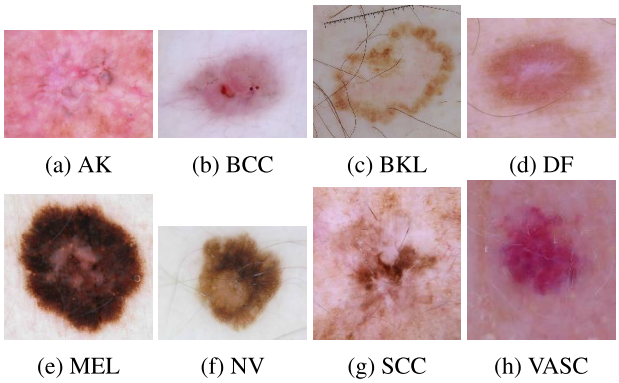


FIGURE 6. ISIC2019 skin lesion sample from each category.

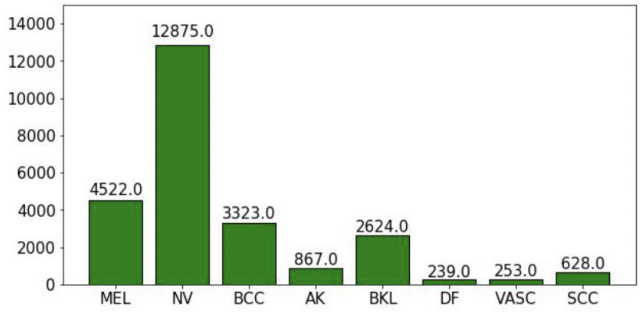


FIGURE 7. Imbalanced data distribution from ISIC2019 skin dataset.

FIGURE 8. Natural object images from CIFAR10 dataset.

are visualized in Fig. 6 while the data distribution is shown in Fig. 7.

B. NATURAL IMAGE DATASET

The natural image dataset consists of natural images from everyday objects such as airplanes, cars, animals, and numbers. We used four different commonly used natural image datasets in our experiment: CIFAR10 and ImageNet datasets.

1) CIFAR10

CIFAR10 is a dataset of 10 categories: airplane, car, bird, cat, deer, dog, frog, horse, ship, and truck [61]. A sample from each category is displayed in Fig. 8. The total number of images is 60.000. The dataset has resolution 32 × 32 pixels.

2) ImageNet

ImageNet is the largest publicly available natural image dataset, used as the primary benchmark in the ImageNet

FIGURE 9. ImageNet dataset sample.

Large Scale Visual Recognition Challenge (ILSVRC), the largest annual computer vision competition since its inception in 2010. The dataset consists of 1.4 million images from 1000 categories [62]. On average, ImageNet dataset has a higher resolution compared with CIFAR10 dataset. In data augmentation, the image is usually resized to 256 × 256 pixels. The sample images are displayed in Fig. 9.

V. EXPERIMENT

To study the feasibility of transfer learning beyond weight reuse, we estimated the quality of meta-learning and augmentation transferability from natural to medical and medical to medical datasets. In addition to studying the quality of transfer learning, we characterized the nature of augmentation search using the heuristic approach. We used varieties of Efficient-Net (E-Net) as backbones for our model.

To validate the transfer performance of meta-learning from natural to medical datasets, the CIFAR10 dataset was utilized for training. For the target task, we used EyePACS, ChestX-ray8, and ISIC2019 datasets. We also added the baseline of transfer learning that uses a pre-trained ImageNet model, called vanilla transfer learning, as the starting point. For a fair comparison of the algorithm performance, we applied resize augmentation in all the considered datasets.

To validate the augmentation transferability between natural and medical images, we applied a searched-augmentation policy from CIFAR10 and ImageNet to EyePACS, ChestX-ray8, and ISIC2019 datasets. We also attempted to accommodate the possibility of transferring the searched augmentation between medical datasets. We used fast autoaugment as the search algorithm to save GPU hours. To further reduce GPU hours, we used a smaller E-Net variant for augmentation search and used a larger variant for end-to-end training.

A. EVALUATION METRICS

Owing to the imbalanced distribution of medical datasets, several types of evaluation metrics are required to fully measure the machine learning performance [63], [64]. In this section, we show and describe the evaluation metrics that we used.

1) AREA UNDER CURVE

To accurately assess the machine learning model's ability to differentiate between positive and negative classes, it is important to compute a metric score that measures both specificity and sensitivity in a balanced way. The area under curve (AUC) score is a commonly used metric for this purpose. We define f as a machine learning model that provides a probability $f(x)$ to categorize whether an input x belongs to a positive class. Then, for each positive sample x^+ and negative sample x^-, AUC score can be calculated as follows:

$$\text{AUC} = \frac{1}{n^+ n^-} \sum_{i}^{n^+} \sum_{j}^{n^-} (f(x_i^+) > f(x_j^-)), \tag{11}$$

where n^+ and n^- represent positive and negative samples, respectively. AUC score measures how good the machine learning is at producing the probability of all positive samples to be always greater than the score of all negative samples.

2) BALANCED ACCURACY

Typically, medical datasets have highly imbalanced samples. We cannot apply the common accuracy score where all samples are treated equally; otherwise, the majority class will dominate the overall accuracy score. Balance accuracy is one of the ways to calculate accuracy fairly. Balance accuracy can be computed through the weighted average accuracy from each class. All classes have the same influence on the overall balance accuracy score. However, the samples from different classes will have different influences depending on the number of class members. Denoting d as the number of classes, where each class has n_i members, balance accuracy can be expressed as follows:

$$\text{Balance\_ACC} = \frac{1}{d} \sum_{i}^{d} \frac{1}{n_i} \sum_{j}^{n_i} \hat{x}_j. \tag{12}$$

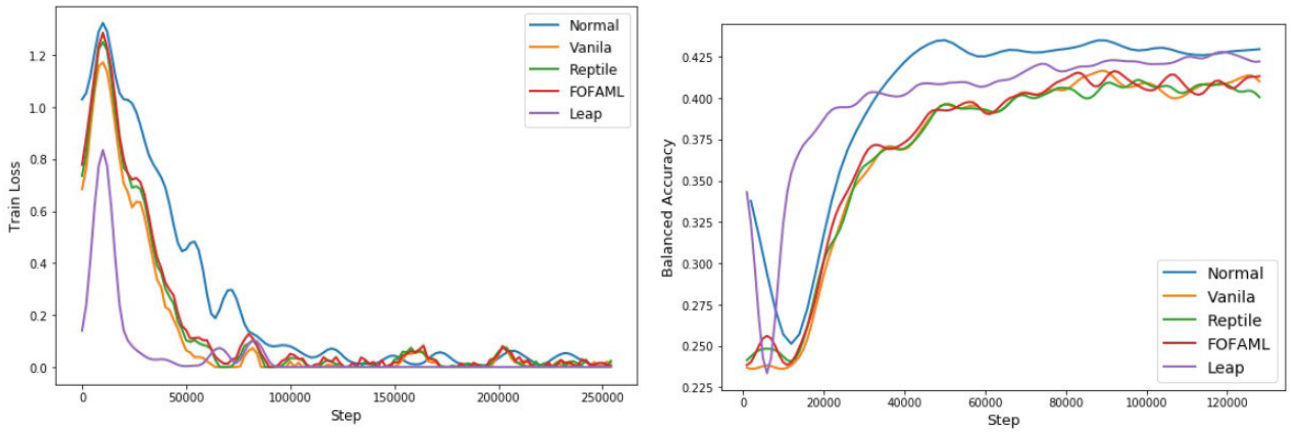
VI. RESULT

We present the results that describe the characteristics and declines encountered in transfer learning in some aspects. We performed two main experiments: transfer meta-learning and augmentation transferability.

A. META-LEARNING

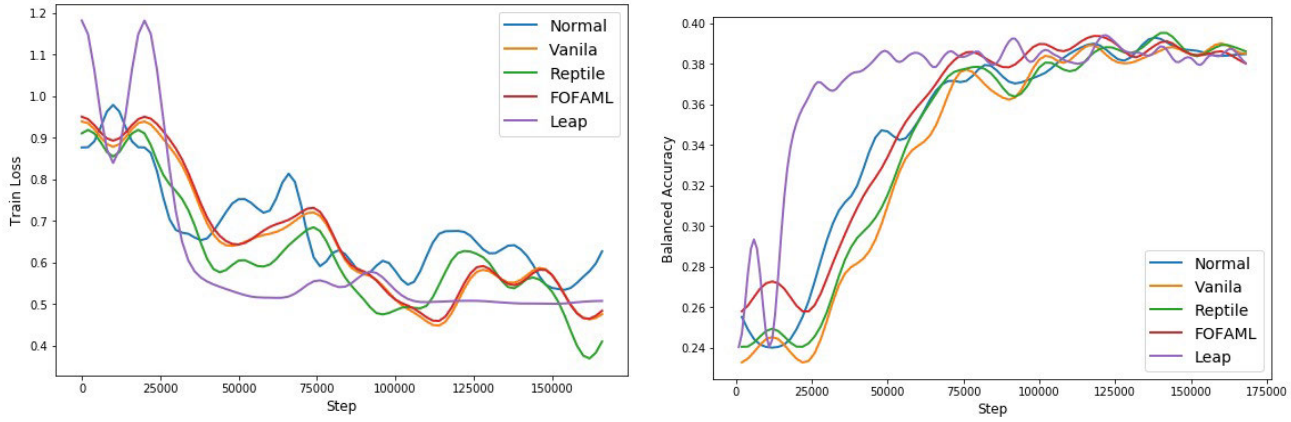
Meta-learning is a deep learning technique that has had limited application in the field of medical image processing. One popular meta-learning approach is finding the best initial point to initiate the training process, closely related to conventional transfer learning that uses pre-trained weights from large models well-tuned on ImageNet. There are three main meta-learning techniques based on weight initialization: Leap, FOMAML, and Reptile.

To investigate the characteristics of the learning behavior, we used the EyePACS and ISIC2019 datasets as baselines and inspected the learning process from the perspectives of training and validation. The results, shown in Fig. 10 and 11, are consistent. Leap had an advantage in terms of convergence rate, while Reptile and FOMAML had fairly poor convergence rates compared to random initialization or normal training. Although Leap had a faster convergence rate, the final performance metrics were quite similar to those of the other methods. This implies that meta-learning helps



(a) Training phase on ISIC2019 (b) Validation phase on ISIC2019

FIGURE 10. Meta-learning performance on ISIC2019 skin dataset. Leap converges faster in the training phase and has superior accuracy in the validation phase compared with the other meta-learning algorithms, including vanilla transfer-learning that uses the ImageNet pre-trained model.



(a) Training phase on EyePACS dataset (b) Validation phase on EyePACS dataset

FIGURE 11. Meta-learning performance on retina EyePACS dataset. Leap has the best convergence rate compared with the other meta-learning algorithms in the training phase and achieves high accuracy faster in the validation phase.

TABLE 1. Augmentation transferability from various dataset to ISIC2019 skin dataset.

Augmentation	AUC Mean	MEL	NV	BCC	AK	BKL	DV	VASC	SCC
Chest-Xray	**0.95**	**0.91**	**0.95**	**0.98**	**0.96**	**0.92**	**0.98**	**1**	**0.95**
CIFAR10	0.94	0.89	0.93	0.97	0.94	0.89	0.97	1	**0.95**
ISIC	0.93	0.88	0.93	0.96	0.92	0.88	0.91	0.99	0.93
ImageNet	0.93	0.87	0.92	0.96	0.93	0.87	0.95	0.98	0.93
No Augmentation	0.86	0.81	0.88	0.9	0.87	0.78	0.82	0.98	0.85

in medical image classification under a limited number of samples and iterations, but with large datasets, the advantage of meta-learning becomes limited.

The results demonstrate the superiority of Leap against the other methods, which aligns with our expectations and is consistent with the findings of Raghu et al. [14]. It can be understood from the results of [14] that transfer learning and normal training are two different solutions with similar local minima. Raghu et al. explored the changes of parameters before and after learning to be insignificant, which is, in fact, the base idea of FOMAML and Reptile techniques. This is also the main reason FOMAML and Reptile have fairly poor performance. On the contrary, in the case of Leap, the concept is to find the starting point with the shortest path to local minima; this is an advantage that helps achieve fast convergence rates in training.

Moreover, the study finds support for these conclusions in the work of [65], which suggests that training from scratch yields better performance compared to using pre-trained models from transfer learning. This additional reference reinforces the superiority of Leap and emphasizes the importance of starting from an optimal initialization point for achieving improved performance in medical image processing. Overall, the investigation of meta-learning techniques in this study sheds light on their potential in medical image analysis, highlighting the advantages of Leap and emphasizing the

TABLE 2. Augmentation transferability from various datasets to ChestX-ray8 dataset.

Augmentation	Mean	0	1	2	3	4	5	6	7
Chest-Xray	0.64	**0.68**	0.77	**0.77**	0.53	0.62	0.56	0.48	0.69
CIFAR10	0.65	0.67	**0.79**	**0.77**	**0.56**	0.64	**0.58**	0.5	0.71
ISIC	0.63	**0.68**	0.73	0.75	0.54	0.62	0.56	0.5	0.65
ImageNet	0.62	0.66	0.77	0.74	0.54	0.61	0.55	0.45	0.61
No Augmentation	**0.67**	**0.68**	0.77	**0.77**	0.55	**0.67**	**0.58**	**0.52**	**0.77**

TABLE 3. Augmentation transferability from various datasets to EyePACS dataset.

Augmentation	Mean	Normal	Mild DR	Moderate DR	Severe DR	Proliferative DR
Chest-Xray	0.5	0.5	0.5	0.5	0.5	0.5
CIFAR10	0.5	0.5	0.5	0.5	0.5	0.5
ISIC 2019	0.73	0.71	0.53	0.7	0.83	0.88
ImageNet	0.5	0.5	0.5	0.5	0.5	0.5
No Augmentation	**0.8**	**0.77**	**0.6**	**0.78**	**0.9**	**0.95**

importance of initialization in achieving fast convergence rates and improved performance.

B. AUGMENTATION TRANSFERABILITY

Searched-augmentation transferability has become one of the most investigated features [27], [28], [29], [66]. Cubuk et al., and Lim et al. performed transfer augmentation between various datasets based on a variety of networks. The transferability results of both studies are surprising [27], [28]. The difference between using the transfer augmentation from any dataset compared to the actual-searched augmentation for specific datasets is negligible. We found out that this may not be the case for medical datasets. As shown in Table 1, 2, and 3, the transferability between natural and medical datasets is not as good as the authors found.

The poor performance of the ImageNet augmentation when applied to medical datasets is consistent across all datasets. This phenomenon occurs because the characteristics of ImageNet are to describe an object, whereas the task of the medical dataset entails closely examining and finding spectral anomalies in the image. Looking further, the heuristic augmentation search in medical datasets was not as efficient as we had anticipated. In some cases, the transfer augmentation from different datasets to a particular dataset outperformed the searched augmentation based on that particular dataset. For example, as presented in Table 1, the performance of augmentation for ChestX-ray8 is better than the performance of augmentation searched for the ISIC dataset. The term "heuristic" in heuristic augmentation search related to the estimation of the augmentation performance on the training phase by using the inference loss. The estimation technique may work on object-shape-oriented problems such as ImageNet to enrich datasets and reduce bias against noise; however, this is not the case on medical datasets, in which spectral anomaly is the primary focus rather than data diversity. Table 2 presented a similar phenomenon, wherein the searched augmentations for the ChestX-ray dataset did not demonstrate superior performance compared to other augmentation techniques. These findings further support the notion that the efficacy of augmentation search methods may vary across different datasets. The

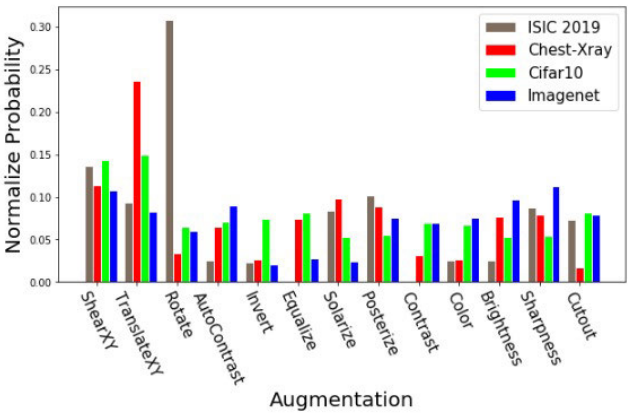


FIGURE 12. Average probability of searched augmentations.

inability of fast autoaugment to accommodate the transfer scenario seems direr on the EyePACS dataset. As can be seen from Table 3, the searched augmentations based on ChestX-ray8, CIFAR10, and ImageNet hinder the learning process from converging, and hence, the value 0.5 in the AUC metric. The shifted dataset caused by augmentation can shift the training paradigm away from the original, thereby affecting the training performance. Recent finding by Wei et al, mention that auto augment have the possibility to eliminate discriminative informations that makes the data indistinguishable to the other classes [29]. The finding is align with our experiment result on Table 3 with no augmentation configuration have better performances than the one with the augmentation approach.

The characteristic of searched augmentation with medical image datasets is different from those of searched augmentation with natural image datasets. As can be seen in Figs. 12 and 13, the searched augmentation policies for natural images are more diverse than those for medical images. ISIC2019 searched augmentation concentrates on the rotation to inform the network that orientation is not a concern for the dermoscopic image in question, while ChestX-ray8 augmentation concentrates on translation to inform the network that shape is not a concern such that the network can focus more on spectral anomalies. To quantify the

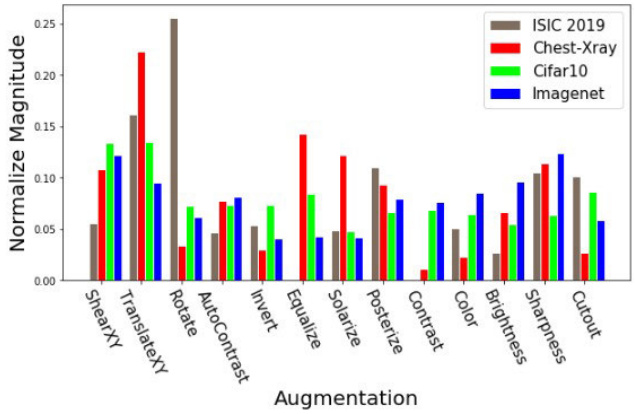


FIGURE 13. Average magnitude of searched augmentations.

TABLE 4. The Gini ratio of all searched augmentations.

	ISIC	Chest-Xray	CIFAR10	ImageNet
Probability	0.51	0.37	0.19	0.23
Magnitude	0.49	0.38	0.16	0.19

TABLE 5. Comparison result of our works.

Dataset	Method	AUC Score
ISIC Skin Dataset	Our work	**0.95**
	Cassidy et al. [67]	0.75
Chest X-ray Dataset	Our work	0.67
	Wang et al. [10]	**0.69**
EyePACS Dataset	Our work	**0.8**
	Gangwar et al. [68]	0.78

concentration of the searched augmentations, we calculated the Gini ratio, as presented in Table. 4. The Gini ratio of the searched-augmentation policy is higher in the case of medical image datasets compared to natural image datasets. This supports the conjecture that with medical image datasets, data diversity is not important.

The comparison results are shown in Table 5. The results obtained on the ISIC Skin dataset demonstrate a significant improvement in performance, with an AUC of 0.95 achieved by this study, whereas Cassidy et al. [67] reported an AUC of 0.75. This substantial difference suggests that the proposed technique outperforms the previous method by a considerable margin, indicating its potential for enhanced analysis of medical images in the dermatological domain. Moving on to the chest X-net dataset, the comparison reveals a relatively small discrepancy between the proposed approach and the work of Wang et al. [10] While the AUC achieved by this study stands at 0.67, Wang et al. reported a slightly higher AUC of 0.69. Although the difference in performance is not as substantial as in the ISIC Skin dataset, it is important to note that the proposed approach still demonstrates competitive performance in comparison to the prior work. Furthermore, when considering the Eyepacs dataset, this study achieved an AUC of 0.8, surpassing the AUC of 0.78 reported by Gangwar and Ravi [68] Although the difference is relatively modest, the proposed approach shows a slightly improved performance, indicating its effectiveness in analyzing medical images in the ophthalmological domain.

VII. CONCLUSION

In this study, we conducted experiments to investigate the potential of meta-learning and data augmentation in medical image classification. Specifically, we explored the impact of choosing the starting point and data augmentation on the performance of deep learning models trained on three large medical image datasets. Our findings suggest that meta-learning can improve the speed of convergence and may be particularly useful in cases with limited amounts of data. However, we also observed that further iterations of meta-learning with normal training can render it unusable. Our results are consistent with previous work on transfer learning [14], which suggests that the weights of a network do not deviate significantly from the initial point. This has an impact on the performance of FOMAML and Reptile, while Leap benefits from it.

Regarding data augmentation, we found that the transferability of searched augmentations from high-resolution natural images did not always meet our expectations. This contradicts previous studies [27], [28] that found searched augmentations to be transferable across datasets. Furthermore, we discovered that the searched augmentations in medical images were more focused on details within the images, rather than removing bias as is the case in natural images.

VIII. DISCUSSION AND FUTURE WORKS

One of the primary challenges in working with medical images is the imbalanced nature of medical datasets. Such datasets often exhibit an uneven distribution of classes, which can lead to a misunderstanding of model performance. Although the accuracy may appear high, the model may not be effectively learning since it tends to predict the majority class. To mitigate this issue, it is essential to tune the loss function, enabling the model to achieve high rewards for detecting minority classes and preventing biased predictions.

Another challenge in handling medical images is their high resolution, which demands significant computational resources. Unlike natural images, random cropping is not a viable option for medical images due to the critical nature of the information contained within them. Instead, proper augmentation techniques are necessary to generate additional training samples. However, as our study revealed, excessively strong augmentation can distort or diminish the semantic information in medical images, thus posing a challenge in finding the right balance between augmentation strength and preserving essential details. Additionally, acquiring medical image datasets can be challenging due to several consent requirements that must be fulfilled before accessing the data. The sensitive nature of medical information necessitates strict adherence to ethical guidelines and patient privacy regulations. These consent protocols and restrictions add an additional layer of complexity to obtaining and utilizing medical datasets, making data acquisition a non-trivial task.

Moving forward, there are promising avenues for future research in medical image classification. One such area is the

exploration of few-shot and zero-shot learning algorithms. These approaches could help detect new unseen class categories, which are essential for discovering novel insights in the medical field. By effectively utilizing limited sample sizes and leveraging meta-learning techniques, few-shot and zero-shot learning algorithms hold potential for addressing the challenges posed by limited positive cases in medical image datasets.

In conclusion, the challenges associated with medical image classification, including class imbalance, resolution considerations, and data acquisition hurdles, require dedicated efforts to develop robust and reliable solutions. Additionally, future research should focus on leveraging few-shot and zero-shot learning algorithms to detect new class categories and broaden the scope of medical image analysis.

REFERENCES

[1] J. Liu, J. Wu, F. Gao, Y. Li, Y. Li, R. Wang, Y. Zhang, T. Li, L. Li, and H. Li, "Development and validation of a deep learning algorithm for the automated detection of small-bowel angioectasias on capsule endoscopy images," *Gastrointestinal Endoscopy*, vol. 94, no. 2, pp. 267–275, 2021.

[2] P. Dutta, T. Roy, and N. Anjum, "COVID-19 detection using transfer learning with convolutional neural network," in *Proc. 2nd Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST)*, Jan. 2021, pp. 429–432.

[3] F. Abdolali, J. Kapur, J. L. Jaremko, M. Noga, A. R. Hareendranathan, and K. Punithakumar, "Automated thyroid nodule detection from ultrasound imaging using deep convolutional neural networks," *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103871.

[4] S. Kachulis, E. Mavrogonatou, C. Kyriakopoulos, L. Mavroeidis, S. Baka, G. Karpathiou, A. Giatromanolaki, C. Tsilikas, G. Galaktidou, and M. E. Froudarakis, "Deep learning-based classification of mesothelioma improves prediction of patient outcome," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021.

[5] 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.

[6] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds. PMLR, Jun. 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[7] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Going deeper with image architectures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10770–10778.

[8] 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.

[9] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[10] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3462–3471, doi: 10.1109/CVPR.2017.369.

[11] 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.

[12] D. Chen, H. Hu, Q. Wang, Y. Li, C. Wang, C. Shen, and Q. Li, "CARTL: Cooperative adversarially-robust transfer learning," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, in Proceedings of Machine Learning Research, vol. 139, M. Meila and T. Zhang, Eds., Jul. 2021, pp. 1640–1650. [Online]. Available: http://proceedings.mlr.press/v139/chen21k.html

[13] K. He, R. Girshick, and P. Dollar, "Rethinking ImageNet pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct./Nov. 2019, pp. 4917–4926, doi: 10.1109/ICCV.2019.00502.

[14] M. Raghu, C. Zhang, J. M. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., Vancouver, BC, Canada, Dec. 2019, pp. 3342–3352.

[15] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2656–2666.

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, pp. 107–115, 2021, doi: 10.1145/3446776.

[17] S. Flennerhag, P. G. Moreno, N. D. Lawrence, and A. C. Damianou, "Transferring knowledge across learning processes," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019. [Online]. Available: https://openreview.net/forum?id=HygBZnRctX

[18] P. Ballester and R. M. Araujo, "On the performance of GoogLeNet and AlexNet applied to sketches," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1124–1128. [Online]. Available: http://dl.acm.org/citation.cfm?id=3015812.3015979

[19] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *Proc. Int. Conf. Learn. Represent.*, 2019. [Online]. Available: https://openreview.net/forum?id=Bygh9j09KX

[20] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. 33rd AAAI Conf. Artif. Intell., 31st Innov. Appl. Artif. Intell. Conf., 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 590–597, doi: 10.1609/aaai.v33i01.3301590.

[21] J. Vanschoren, "Meta-learning," in *Automated Machine Learning—Methods, Systems, Challenges* (The Springer Series on Challenges in Machine Learning), F. Hutter, L. Kotthoff, and J. Vanschoren, Eds. Springer, 2019, pp. 35–61, doi: 10.1007/978-3-030-05318-5_2.

[22] Q. Wei, L. Yu, X. Li, W. Shao, C. Xie, L. Xing, and Y. Zhou, "Meta-learning for bootstrapping medical image segmentation from imperfect supervision," 2023. [Online]. Available: https://openreview.net/forum?id=yd5kGP5_VVE

[23] K. Mahajan, M. Sharma, and L. Vig, "Meta-DermDiagnosis: Few-shot skin disease identification using meta-learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 3142–3151.

[24] 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*, D. Stoyanov, Z. Taylor, D. Sarikaya, J. McLeod, M. A. G. Ballester, N. C. Codella, A. Martel, L. Maier-Hein, A. Malpani, M. A. Zenati, S. De Ribaupierre, L. Xiongbiao, T. Collins, T. Reichl, K. Drechsler, M. Erdt, M. G. Linguraru, C. O. Laura, R. Shekhar, S. Wesarg, M. E. Celebi, K. Dana, and A. Halpern, Eds. Cham, Switzerland: Springer, 2018, pp. 303–311.

[25] E. U. Moya-Sánchez, A. Sánchez, M. Zapata, J. Moreno, D. Garcia-Gasulla, F. Parrés, E. Ayguadé, J. Labarta, and U. Cortés, "Data augmentation for deep learning of non-mydriatic screening retinal fundus images," in *Supercomputing*, M. Torres, J. Klapp, I. Gitler, and A. Tchernykh, Eds. Cham, Switzerland: Springer, 2019, pp. 188–199.

[26] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. Gilad Kusne, and T. Buonassisi, "Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks," *npj Comput. Mater.*, vol. 5, no. 1, p. 60, May 2019.

[27] E. D. Cubuk, B. Zoph, D. Mané, 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.

[28] S. Lim, I. Kim, T. Kim, C. Kim, and S. Kim, "Fast AutoAugment," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds. 2019, pp. 6662–6672. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/6add07cf50424b14fdf649da87843d01-Abstract.html

[29] L. Wei, A. Xiao, L. Xie, X. Zhang, X. Chen, and Q. Tian, "Circumventing outliers of autoaugment with knowledge distillation," in *Computer Vision—ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 608–625.

[30] 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.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*. Computational and Biological Learning Society, 2015, pp. 1–14.

[32] M. Abramoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. Folk, and M. Niemeijer, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative Ophthalmol. Vis. Sci.*, vol. 57, pp. 5200–5206, Oct. 2016.

[33] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[34] T. Peng, Y. Gu, Z. Ye, X. Cheng, and J. Wang, "A-LugSeg: Automatic and explainability-guided multi-site lung detection in chest X-ray images," *Expert Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116873, doi: 10.1016/j.eswa.2022.116873.

[35] Y. Ding, J. Sohn, M. Kawczynski, H. Trivedi, R. Harnish, N. Jenkins, D. Lituiev, T. Copeland, M. Aboian, C. Aparici, S. Behr, R. Flavell, S.-Y. Huang, K. Zalocusky, L. Nardo, Y. Seo, R. Hawkins, M. H. Pampaloni, D. Hadley, and B. Franc, "A deep learning model to predict a diagnosis of Alzheimer disease by using ^{18}F-FDG pet of the brain," *Radiology*, vol. 290, Nov. 2018, Art. no. 180958.

[36] P. Khosravi, E. Kazemi, Q. Zhan, M. Toschi, J. E. Malmsten, C. Hickman, M. Meseguer, Z. Rosenwaks, O. Elemento, N. Zaninovic, and I. Hajirasouliha, "Robust automated assessment of human blastocyst quality using deep learning," *bioRxiv*, 2018, Art. no. 394882. [Online]. Available: https://www.biorxiv.org/content/early/2018/08/20/394882, doi: 10.1101/394882.

[37] M. Huh, P. Agrawal, and A. A. Efros, "What makes ImageNet good for transfer learning?" 2016, *arXiv:1608.08614*.

[38] A. Achille, T. Eccles, L. Matthey, C. Burgess, N. Watters, A. Lerchner, and I. Higgins, "Life-long disentangled representation learning with cross-domain latent homologies," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Montréal, QC, Canada, Dec. 2018, pp. 9895–9905. [Online]. Available: http://papers.nips.cc/paper/8193-life-long-disentangled-representation-learning-with-cross-domain-latent-homologies

[39] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, vol. 48, 2016, pp. 1842–1850. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045390.3045585

[40] M. Andrychowicz, M. Denil, S. G. Colmenarejo, M. W. Hoffman, D. Pfau, T. Schaul, and N. de Freitas, "Learning to learn by gradient descent by gradient descent," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 3981–3989. [Online]. Available: http://papers.nips.cc/paper/6461-learning-to-learn-by-gradient-descent-by-gradient-descent

[41] Y. Bengio, S. Bengio, and J. Cloutier, "Learning a synaptic learning rule," in *Proc. Seattle Int. Joint Conf. Neural Netw. (IJCNN)*, vol. 2, Jul. 1991, p. 969.

[42] S. Bengio, Y. Bengio, J. Cloutier, and J. Gecsei, "On the optimization of a synaptic learning rule," 2007. [Online]. Available: https://api.semanticscholar.org/CorpusID:28783413

[43] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr./May 2018. [Online]. Available: https://openreview.net/forum?id=B1DmUzWAW

[44] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 3630–3638. [Online]. Available: http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning

[45] B. A. Pearlmutter, "Gradient calculations for dynamic recurrent neural networks: A survey," *IEEE Trans. Neural Netw.*, vol. 6, no. 5, pp. 1212–1228, Sep. 1995, doi: 10.1109/72.410363.

[46] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. P. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. 33nd Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, Jun. 2016, pp. 1842–1850. [Online]. Available: http://proceedings.mlr.press/v48/santoro16.html

[47] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018, *arXiv:1803.02999*.

[48] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Sydney, NSW, Australia, Aug. 2017, pp. 1126–1135. [Online]. Available: http://proceedings.mlr.press/v70/finn17a.html

[49] F. Shen, Y. Zhao, L. Wang, M. R. Mojarad, Y. Wang, S. Liu, and H. Liu, "Rare disease knowledge enrichment through a data-driven approach," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, p. 32, Dec. 2019, doi: 10.1186/s12911-019-0752-9.

[50] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Trans. Neural Netw.*, vol. 9, no. 5, p. 1054, Sep. 1998, doi: 10.1109/TNN.1998.712192.

[51] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. 26th Annu. Conf. 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

[52] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, Jun. 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html

[53] 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. [Online]. Available: http://dl.acm.org/citation.cfm?id=3298023.3298188

[54] 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 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

[55] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 6231–6239. [Online]. Available: http://papers.nips.cc/paper/7203-the-expressive-power-of-neural-networks-a-view-from-the-width

[56] M. Raghu, B. Poole, J. M. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "On the expressive power of deep neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Sydney, NSW, Australia, Aug. 2017, pp. 2847–2854. [Online]. Available: http://proceedings.mlr.press/v70/raghu17a.html

[57] American Academy of Ophthalmology. *International Clinical Diabetic Retinopathy Disease Severity Scale Detailed Table*. Accessed: Nov. 21, 2019. [Online]. Available: http://www.icoph.org/dynamic/attachments/resources/diabetic-retinopathy-detail.pdf

[58] A. Mahbod, P. Tschandl, G. Langs, R. Ecker, and I. Ellinger, "The effects of skin lesion segmentation on the performance of dermatoscopic image classification," *Comput. Methods Programs Biomed.*, vol. 197, Dec. 2020, Art. no. 105725, doi: 10.1016/j.cmpb.2020.105725.

[59] 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.

[60] M. Combalia, F. Hueto, S. Puig, J. Malvehy, and V. Vilaplana, "Uncertainty estimation in deep neural networks for dermoscopic image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3211–3220.

[61] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.

[62] 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, Dec. 2015.
[63] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *Int. J. Mach. Learn. Comput.*, vol. 3, p. 224, Apr. 2013.
[64] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *Sci. Rep.*, vol. 12, Apr. 2022, Art. no. 5979. [Online]. Available: https://www.nature.com/articles/s41598-022-09954-8#citeas
[65] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," *Sci. Rep.*, vol. 9, no. 1, p. 6381, Apr. 2019.
[66] D. Ho, E. Liang, X. Chen, I. Stoica, and P. Abbeel, "Population based augmentation: Efficient learning of augmentation policy schedules," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, in Proceedings of Machine Learning Research, Long Beach, CA, USA, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds., Jun. 2019, pp. 2731–2741. [Online]. Available: http://proceedings.mlr.press/v97/ho19b.html
[67] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102305, doi: 10.1016/j.media.2021.102305.
[68] A. K. Gangwar and V. Ravi, "Diabetic retinopathy detection using transfer learning and deep learning," in *Evolution in Computational Intelligence—Frontiers in Intelligent Computing: Theory and Applications*, in Advances in Intelligent Systems and Computing, vol. 1176, Karnataka, India, V. Bhateja, S. Peng, S. C. Satapathy, and Y. Zhang, Eds. Singapore: Springer, Jan. 2020, pp. 679–689, doi: 10.1007/978-981-15-5788-0_64.

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.

TRYAN ADITYA PUTRA received the bachelor's degree in computer engineering and the master's degree in electrical engineering from Universitas Indonesia, in 2015 and 2017, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, National Taiwan University of Science and Technology, in 2021. His research interests include machine learning, artificial intelligence, data mining, and mathematical modeling.

JENQ-SHIOU LEU (Senior Member, IEEE) received the bachelor's degree in mathematics and the master's degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1991 and 1993, respectively, and the Ph.D. degree in computer science from National Tsing Hua University, Hsinchu, Taiwan, in 2006. From 1995 to 1997, he was a Research and Development Engineer with Rising Star Technology, Taiwan. From 1997 to 2007, he was an Assistant Manager in the telecommunication industry, such as Mobital Communications and Taiwan Mobile. 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 Chairperson. Since August 2017, he has been a Professor and the Chairperson. His research interests include heterogeneous network integration, mobile service, platform design, distributed computing (P2P and cloud computing), and green and orange technology integration. He has published extensively in these areas, with 58 SCI-indexed journal articles, 57 conference papers, and book chapters, and led 12 MOST projects, 12 industry-academia projects, and two cross-university projects in the past ten years.

TIAN SONG (Member, IEEE) received the Bachelor of Engineering degree from the Dalian University of Technology, China, in 1995, and the Master of Engineering and Doctor of Engineering degrees from Osaka University, in 2001 and 2004, respectively. He joined Tokushima University, as an Assistant Professor, in 2004. He is currently an Associate Professor with the Department of Electrical and Electronic Engineering, Graduate School of Advanced Technology and Science, Tokushima University. He has over 20 years of research experience in video coding algorithms, VLSI architecture, and system design methodology. His current research interests include deep learning-based object detection, underwater video processing, and medical image processing. He is a member of IEICE and RISP.

TAKAFUMI KATAYAMA (Member, IEEE) received the Bachelor of Engineering, Master of Engineering, and Ph.D. degrees in electrical engineering from Tokushima University, in 2011 and 2019. He was with Renesas Electronics Corporation, from 2012 to 2014. He joined Tokushima University, as an Assistant Professor, in 2019. His current research interests include machine learning, video coding algorithms, and hardware design.