applied sciences

Article

Sample-Efficient Deep Learning Techniques for Burn Severity Assessment with Limited Data Conditions

Hyunkyung Shin [1], Hyeonung Shin [2], Wonje Choi [3], Jaesung Park [3], Minjae Park [4], Euiyul Koh [5] and Honguk Woo [3,*]

[1] Fine Healthcare, Seoul 06069, Korea; hapburn98@naver.com
[2] Bestian Hospital, Cheongju 28161, Korea; skreh12@gmail.com
[3] Department of Computer Science and Engineering, Sungkyunkwan University, Suwon 16419, Korea; wjchoi1995@g.skku.edu (W.C.); v37794@g.skku.edu (J.P.)
[4] Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea; mjpark@kaist.ac.kr
[5] Acryl Inc., Seoul 06069, Korea; kay@acryl.ai
* Correspondence: hwoo@skku.edu

Abstract: The automatic analysis of medical data and images to help diagnosis has recently become a major area in the application of deep learning. In general, deep learning techniques can be effective when a large high-quality dataset is available for model training. Thus, there is a need for sample-efficient learning techniques, particularly in the field of medical image analysis, as significant cost and effort are required to obtain a sufficient number of well-annotated high-quality training samples. In this paper, we address the problem of deep neural network training under sample deficiency by investigating several sample-efficient deep learning techniques. We concentrate on applying these techniques to skin burn image analysis and classification. We first build a large-scale, professionally annotated dataset of skin burn images, which enables the establishment of convolutional neural network (CNN) models for burn severity assessment with high accuracy. We then deliberately set data limitation conditions and adapt several sample-efficient techniques, such as transferable learning (TL), self-supervised learning (SSL), federated learning (FL), and generative adversarial network (GAN)-based data augmentation, to those conditions. Through comprehensive experimentation, we evaluate the sample-efficient deep learning techniques for burn severity assessment, and show, in particular, that SSL models learned on a small task-specific dataset can achieve comparable accuracy to a baseline model learned on a six-times larger dataset. We also demonstrate the applicability of FL and GANs to model training under different data limitation conditions that commonly occur in the area of healthcare and medicine where deep learning models are adopted .

Keywords: self-supervised learning; federated learning; data augmentation; medical image analysis; burn severity

check for updates

Citation: Shin, H.; Shin, H.; Choi, W.; Park, J.; Park, M.; Koh, E.; Woo, H. Sample-Efficient Deep Learning Techniques for Burn Severity Assessment with Limited Data Conditions. *Appl. Sci.* **2022**, *12*, 7317. https://doi.org/10.3390/app12147317

Academic Editor: Valentin Sebastian Schäfer

Received: 23 June 2022
Accepted: 19 July 2022
Published: 21 July 2022

1. Introduction

Deep learning technology has been applied in various situations in the field of image analysis and classification. Medical image analysis is recognized as an important application of deep learning. However, due to the associated high costs (e.g., labeling by domain experts) and complex security aspects of clinical data (e.g., privacy implications and regulation), deep learning-based medical image analysis has not yet been fully adopted in practice.

Recently, there have been increasing attempts to develop sample-efficient deep learning techniques and to apply them in the field of medical image analysis where it is difficult to obtain a sufficiently large training dataset. For example, the the authors of [1] employed self-supervised learning (SSL) to achieve high accuracy in skin lesion classification. In general, SSL-based approaches establish pretrained models using unlabeled datasets, which can be transferred to a specific task using a small amount of labeled data. Unlike SSL, federated learning (FL) focuses on leveraging labeled data from multiple clients to overcome the data limitation conditions of a single client, while supporting privacy-preserving distributed operations on model training. In [2], the Auto-FedAvg approach is described which addresses strict regulation and privacy issues of clinical data by adopting model federation among multiple medical institutions. In contrast to conventional deep learning, for which locally owned data needs to be integrated into a single dataset for model training, in FL, each institution individually conducts local model training on its own data and shares only its local model with a central server or other institutions. That is, the union of locally trained data corresponds to an entire training dataset; however, union operations are not conducted on training data but, instead, on model parameters among multiple FL clients. There have also been a number of studies utilizing synthetic data for deep-learning-based medical image analysis [3–7].

In this paper, we apply these sample-efficient deep learning techniques to develop an inference model for burn severity assessment, which achieves high diagnostic accuracy for distinct degrees of severity under limited training dataset conditions. Burns affect many people every year; for instance, it has been reported that about 1.1 million patients suffered from burn injuries in the United States, and that thousands of deaths resulted from burn injuries and burn-related infections [8]. According to Korean national health insurance service records, each year in South Korea, about half a million people receive burn treatment. A burn injury destroys human skin and adjacent tissues and requires suitable clinical treatment depending on its intensity and severity. A timely and correct diagnosis of burn injuries is recognized as an important factor for patient treatment and recovery [9].

For burn severity assessment, we sought to develop an accurate classification system that would enable assessment of burn-related skin images by learning to extract important features from the images, thereby enabling highly accurate and timely diagnosis based on a single image. To achieve such a high-accuracy model, we first built a dataset of skin burn images, which was sufficiently large to produce convolutional neural network (CNN)-based high-accuracy models through conventional supervised learning. From a deep learning process perspective, it is crucial to establish a high-quality, large-scale dataset in which diagnosis information provided by experienced medical practitioners is used as ground-truth labels. We obtained such a dataset of burn images with diagnostic information through research cooperation with one of the largest medical institutions dealing with burns in South Korea.

We then trained our target models by adopting sample-efficient deep learning techniques and evaluated the model performance. In the evaluation, we deliberately assumed limited dataset conditions and utilized only a small number of training samples (i.e., no more than tens or hundreds of burn images). Our experimental setting was intended to assess the feasibility of sample-efficient deep learning techniques in the context of medical image analysis with data limitations. For this purpose, we used the task of burn severity assessment as the subject of our empirical study.

Specifically, we compared the performance of sample-efficient deep learning models to that of a CNN model learned on the entire dataset (i.e., our sample-sufficient baseline model), to determine whether or not sample-efficient models trained with a small part of the dataset could achieve comparable performance to the sample-sufficient baseline. For example, our SSL-based method utilized a pretrained SSL model (i.e., simCLR [10]), obtaining high accuracy comparable to the sample-sufficient baseline, even when only 16% of the entire dataset was used; the method achieved high sample-efficiency, showing comparable accuracy to a baseline model learned on a dataset six-times larger.

To the best of our knowledge, our work has been the first to implement and evaluate sample-efficient deep learning techniques with real-world burn injury data. The main contributions of this paper are as follows.

- We present sample-efficient deep learning models for burn severity assessment using various techniques. Through extensive experiments, we demonstrate the feasibility of these sample-efficient models with limited training data.
- We build a large dataset containing 13,715 burn images with professionally annotated information, on which not only the high-accuracy burn severity assessment model is trained but also various limited data conditions can be emulated to evaluate the sample-efficient models.
- We discuss the dataset conditions under which sample-efficient deep learning techniques can be productively applied.

The remainder of this paper is organized as follows: The details of the burn severity assessment task, including its datasets and baseline models, are described in Section 2. Several deep learning techniques for limited training data conditions are explained in Section 3. The experimental results for these sample-efficient techniques using burn injury datasets are provided in Section 4. Finally, the study conclusions are presented in Section 5.

2. Burn Severity Assessment

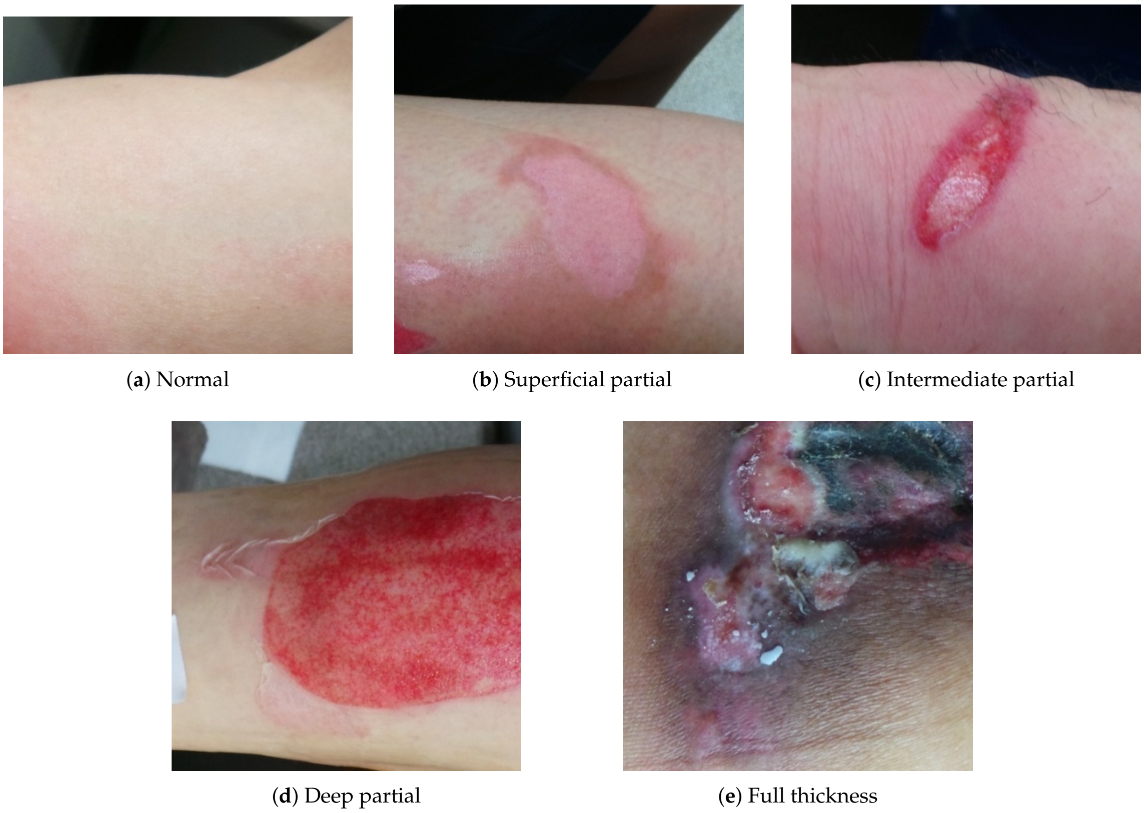
In this section, we describe the problem of burn severity assessment by which a burn image is analyzed and classified into different types.

2.1. Burn Image Dataset

Burn severity is categorized into different types according to the intensity and depth of burn injury elements: *superficial partial* thickness, *intermediate partial* thickness, *deep partial* thickness, *full* thickness, and normal. Each type of burn is defined as follows: (1) superficial partial thickness involves the epidermis and superficial dermis; (2) deep partial thickness involves the epidermis and deeper dermis; (3) full thickness burns destroy the epidermis and dermis; and (4) intermediate partial thickness refers to when it is difficult to determine whether a burn injury is superficial partial or deep partial. It is often non-trivial, even for professionals, to distinguish between superficial partial and deep partial thickness, and, in that case, some burn injuries need to be judged again after a certain period of time. Thus, it is important to judge the burn severity as intermediate partial thickness first and to make a more accurate diagnosis later. The samples of different burn types are presented in Figure 1, where different characteristics for the burn types appear in their images.

As there is no publicly available large burn image dataset that can be used for training burn-related prediction models, we processed numerous burn images and normal skin images collocated from multi-year burn-related medical records. For these, we collaborated with the Bestian Hospital (http://eng.bestianseoul.com/specialized-center/about-the-burn-center/, accessed on 1 June 2022), one of the largest professional burn hospitals in South Korea. Figure 2 illustrates how burn images are professionally annotated with bounding boxes around injuries and their labels in burn types, for which a GUI-based data annotation tool, the *Jonathan Marker* (https://jonathan.acryl.ai/marker, accessed on 1 June 2022) has been used. The tool supports data labeling of different media formats, such as text, images, and video, and enables quality control over data creation pipelines. For each image, a bounding box on the burned area was drawn and labeled by professionals using the tool. For each annotated burn image, a team of three professionals verified its correctness.

In this way, we established a high-quality, large dataset of 13,715 annotated burn image samples that were recorded from outpatients and inpatients during the period from March 2013 to April 2019. We also collected normal skin images that had no annotation. Table 1 illustrates the statistics of the dataset; we used 1700 images for each burn type (i.e., normal, superficial partial, intermediate partial, deep partial, and full thickness) in the training dataset and the other images in the test dataset. We intentionally used a subset of the training dataset (i.e., only using 20, 40, 60, . . . , 170 samples for each burn type) to emulate limited data conditions, as we focus on those problems in practice.


(**a**) Normal (**b**) Superficial partial (**c**) Intermediate partial

(**d**) Deep partial (**e**) Full thickness

Figure 1. Burn samples with different levels of severity: compared to (**a**) *Normal* skin that has no specific features, (**b**) *Superficial partial* thickness and (**c**) *Intermediate partial* thickness appear pink and red, respectively, and furthermore (**c**) appears to have blisters; (**d**) *Deep partial* thickness appears to involve the epidermis and deeper dermis, including some yellow areas related to the deeper dermis damage; (**e**) *Full thickness* appears black and dark brown, involving subcutaneous structure-relevant features.

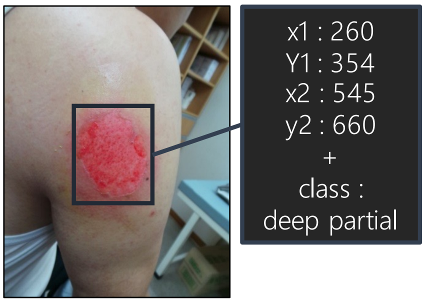


Figure 2. Burn data annotation: the burn injury area is bounding-boxed and annotated by professionals; it is classified as deep partial thickness, considering some white regions related to the deeper dermis damage.

Table 1. Statistics of burn severity dataset: We set the training dataset to have 1700 images for each burn type and deliberately used only a subset of the training dataset for emulating sample-limited training conditions. We included other images than the training samples for the testing dataset. The entire dataset consists of the training and test datasets, and its size (except for normal skin images in the table) is 13,715.

Dataset	Normal	Super.	Inter.	Deep	Full
Training dataset	1700	1700	1700	1700	1700
Test dataset	2779	2033	1770	2386	726

2.2. CNN Models

Deep learning techniques have made significant achievements in image classification for diagnosing skin diseases, e.g., the CNN-based classifier for skin diseases [11], the transfer learning-based skin cancer detector [12], and deep learning-based skin lesion classification [13]. In this research, given the dataset of skin burn images, we employed CNNs for burn severity assessment, as they are known to be superior in image recognition tasks and their pretrained knowledge on large image datasets can be transferred to other tasks, including the analysis of various types of medical images [14,15]. We used ResNet-152 [16], a 152-layer large network variant of ResNet models which have proven high performance in several medical image analysis studies, e.g., burn classification and segmentation [17–19], skin cancer classification [20], CT image improvement [21], and chest X-ray diagnosis [22]. We also note that a pretrained model of ResNet-152 is publicly available and easy to use. Furthermore, as ResNet-152 is commonly used without significant modification for several sample-efficient deep learning techniques which are described in the next section, we evaluated the techniques in a consistent manner for the same deep learning structure.

We set the performance achieved by ResNet-152 as our comparison baseline. For the baseline reference, a ResNet-152 model was trained from scratch on our entire dataset, as shown in Table 1. We also utilized the pretrained ResNet-152 model for which the knowledge was transferred to our burn severity assessment tasks through model fine-tuning with a small number of burn images in the dataset. Unlike end-to-end training from scratch, in transfer learning (TL), the parameters of a pretrained model, except for those in the last layers, are usually retained and not updated. Fine-tuning was used with a small number of samples to update the last layers only to achieve some degree of adaptation, while exploiting the pretrained knowledge. This TL procedure enables rapid domain adaptation between image recognition tasks [23,24].

A pretrained model of ResNet-152 can be established using the ImageNet dataset [25] which contains over 14 million images and is applied as a well-initialized network for task-specific learning, burn severity assessment in our case. In TL, we modified the head (the upper fully connected layer) of a vanilla ResNet-152 structure according to the burn classifications, initializing it using the Kaiming uniform initialization [26].

3. Methods

In this section, we describe several sample-efficient learning techniques, including SSL, FL, and generative adversarial network (GAN)-based data augmentation, which can be used to achieve a more accurate prediction model for burn severity assessment under limited data conditions.

3.1. Self-Supervised Learning

Similar to TL as previously described, SSL is used to create a pretrained model which can be used as a well-initialized model for further target task-specific learning. However, unlike TL, where the pretraining requires a large number of labeled samples, which are not sufficiently available most of the time in many fields, such as medicine, the pretraining in SSL does not require labeled samples.

In principle, the SSL-based approach involves a two-stage learning process. In the pretraining stage, a set of unlabeled images is used to build a pretrained model, and in the training stage (i.e., downstream task learning), a set of labeled burn images, with professionally annotated diagnosis information, is used to adapt the pretrained model. Specifically, we tested two different methods that explore SSL: (1) using a publicly available SSL-based pretrained CNN (SSL-imn) and (2) performing SSL from scratch with unlabeled skin images (SSL-burn). Both methods rely on a contrastive visual learning framework simCLR [10] in which a model is first trained through unsupervised contrastive learning with data augmentation and transformation to learn general representations, and then it is fine-tuned on a task-specific labeled dataset.

In SSL-imn, we used the publicly available pretrained simCLR model based on the ImageNet dataset, which was treated as the starting point for our downstream task learning, burn severity assessment. In SSL-burn, we also created a pretrained simCLR model through contrastive learning on our own unlabeled (unannotated) skin images. This method exploits contrastive loss to learn general representations of skin images through the agreement among differently augmented images from the same input image as well as the distinction among those from different images. In both methods, we used only a small amount of labeled (annotated) burn images to fine-tune the pretrained simCLR model for our task.

The two methods consider limited data conditions differently in terms of the availability of labeled and unlabeled data. More specifically, SSL-burn is intended to incorporate unlabeled images into the pretraining stage, focusing on commonly observed situations in the field of medicine where many medical images are not annotated but stored as part of medical records. SSL-burn can be seen as semi-supervised learning [27], as it makes use of both a large amount of unlabeled images in representational pretraining and a small amount of labeled burn images in task-specific learning to ensure accurate burn diagnosis capability. As such, SSL-burn is considered suitable when a large quantity of unlabeled images can be used.

Figure 3 shows the building blocks in simCLR for the case of SSL-burn where unlabeled skin images are used and augmented for contrastive learning. The encoder learns a function that encodes each image into an embedding vector. For an image x, its positive pairs x_i and x_j have similar embeddings, obtaining $h_i \approx h_j$. We generate these positive pairs by applying transformation functions listed in Section 3.3 on x. (1) The encoder and project head are learned to minimize their contrastive loss, i.e., the cosine similarity of latent vectors z_i and z_j in our implementation. In this pretraining, negative pairs for x are also used, which are generated by the same augmentation functions but on different images other than x. The contrastive loss among those samples augmented from different images is also used for learning. After the contrastive learning, (2) the encoder and linear head are fine-tuned on a dataset of labeled burn images through supervised learning to conduct the burn severity assessment task.

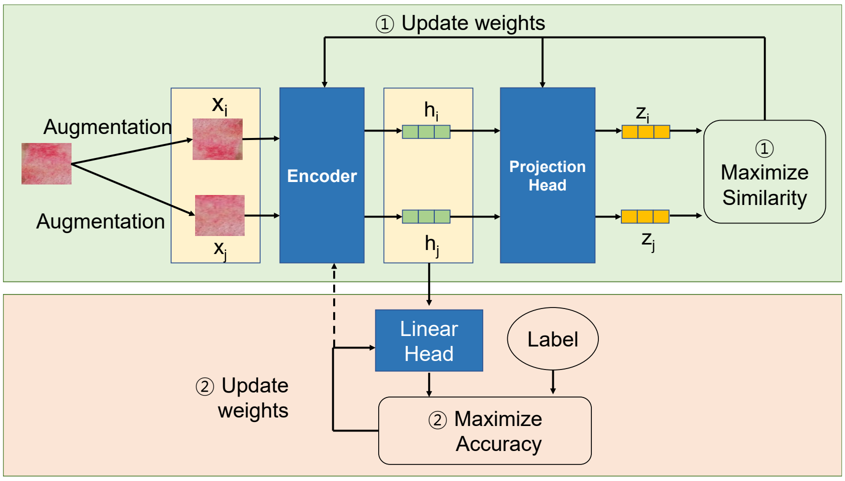


Figure 3. Burn severity assessment based on simCLR [10]: the upper procedure marked with ① denotes self-supervised pretraining on unlabeled images, and the lower procedure marked with ② denotes task-specific supervised fine-tuning on labeled images.

3.2. Federated Learning

While deep learning is promising for image recognition in many fields, existing deep learning models have, to date, only had limited application to medical images. Each medical institution (e.g., hospital) manages its own medical records and rarely shares the datasets with other institutions due to strict regulatory and privacy requirements as well as the expensive process of expert-involving annotation. This data-silo structure makes it difficult to establish a sufficiently large dataset of training samples that are needed to leverage traditional deep learning model architectures [28,29].

FL is intended to leverage the benefits of deep learning based on large datasets, whilst meeting privacy-preserving requirements, by locally training a model without sharing data among multiple learning clients. For example, in mobile device computing, numerous mobile clients participate in the FL process to establish a high-performance global model without sharing the locally collected privacy-sensitive data at the client side with others [30,31]. Recently, n the field of medical image analysis, several studies have demonstrated the feasibility of use of FL in various contexts, e.g., medical image segmentation [2], COVID-19 screening [32], skin lesion classification [33], and paediatric chest X-ray classification [34].

We applied FL to burn severity assessment, assuming a clinical system environment in which each burn medical institution has only a small number of fully annotated burn images and is able to train a model locally with its burn images. Specifically, we used a well-known FL algorithm, FedAvg [31] in which a locally trained model by each FL client (each medical institution) is aggregated via a central server that conducts weighted averaging on the parameters of the models of FL clients. In FL, model aggregation from multiple clients is iteratively conducted for continual model updates, where each aggregation period is called a *round*.

Consider an FL process with K clients. At round t, a server sends its global model (parameters) w^t to the FL clients, which is updated on local data by each client k where $k = 1, \dots, K$. Then, locally updated models w_k^{t+1} are sent back to the server so that an averaging operation is performed as

$$w^{t+1} = \frac{1}{K} \sum_{k=1,\dots,K} w_k^t \tag{1}$$

where we assume the same number of epochs and the same number of training samples across FL clients for simplicity.

With this FL process, the data is not shared among multiple medical institutions and remains distributed, but the locally trained models are shared. When exploiting their own data for model training, this allows medical institutions not only to migrate the risk of security breaches from a technical perspective but also to ensure data asset protection from a management perspective. Moreover, it enables the generalization of trained models alongside more data samples from multiple medical institutions and less restrictive diversity from multiple regions and demographic groups.

In our experiments, we deliberately split our burn image dataset into multiple partial datasets to emulate a clinical environment in which the FedAvg algorithm can be adopted for medical institutions with local data. A model was individually learned on local data at each institution; all the model parameters were locally updated at each round and were then globally aggregated and averaged through a cloud-based federated service to obtain a global model with high accuracy.

3.3. Learning with Data Augmentation

To address the limited availability of training datasets for medical images, several image augmentation heuristics have been tested in deep learning frameworks [35]. For our burn image dataset, we tested the following image transform and augmentation operations that are widely used and implemented in [36]. **RandomBrightnessContrast (RBC)** changes brightness and contrast randomly; **RandomGamma (RG)** changes the gamma value randomly; **ColorJitter (CJ)** changes the brightness, contrast, and saturation randomly; **ISONoise (ISO)** applies camera sensor noise; **GridDropout (DROP)** drops out rectangular regions of an image and the corresponding mask in a grid fashion; **RandomFog (RF)** simulates fog on an image; **ElasticTransform (ET)** applies elastic distortion; and **GridDistortion (GT)** applies grid distortion.

Recently, GAN-based data augmentation approaches have been explored to create synthetic images in various domains. They aim to increase the size and diversity of training datasets and to mitigate the problems of class imbalance and sample deficiency in training datasets. A vanilla GAN architecture includes two neural networks that are trained through performing a minimax game in an adversarial way. A generator network (G) produces synthetic images (i.e., $\hat{x} = G(z)$ for random noise z) and a discriminator network (D) takes those images as input to determine whether they are real or synthetic (i.e., $D(x) = 1$ for real samples and $D(\hat{x}) = 0$ for synthetic samples). The G and D networks are adversaries in that G tries to minimize the following minimax (adversarial) loss function while D tries to maximize it with a maximum log likelihood objective.

$$\min_{G} \max_{D} E_x[\log(D(x))] + E_z[1 - \log(D(G(z)))] \tag{2}$$

By exploiting GAN techniques, very diverse automatic image synthesis is systematically enabled. These techniques are considered promising to reduce the expensive workloads that are required to annotate images professionally, especially for medical image analysis tasks [37–39].

For burn image augmentation, we tested two well-known advanced GAN techniques, CycleGAN [40] and StyleGAN [41], whose ability to synthesize high resolution images is of great interest in a broad spectrum of medical image augmentation and imaging modalities, e.g., synthesizing cardiac MR images [42], high tissue contrast MR images [43], skin lesion images [6,44], and brain CT images [45].

In the CycleGAN architecture, two GANs, where each GAN consists of a pair of generator and discriminator G_X, D_X and G_Y, D_Y, are chained to support automatic image-to-image translation (i.e., X to Y and Y to X) without paired image samples between two different domains X and Y. Figure 4 shows the building blocks of CycleGAN used for our burn image synthesis. Unlike conventional image-to-image translation models (e.g., pix2pix [46]) that require a dataset of paired samples, such as pairs of source images in X and their corresponding target images in Y, the CycleGAN architecture additionally employs a cycle consistency loss function [40] to support unpaired image-to-image translation. Using the loss function defined below, the CycleGAN enforces the generators G_X and G_Y learned to minimize the discrepancy between the source image $x \in X$ (and $y \in Y$) and its reconstructed image by the two generators $x' = G_Y(G_X(x))$ (and $y' = G_X\big(G_y(y)\big)$), providing the automatic translation of input images in the domain X to target images in the domain Y, and vice versa.

$$\min_{G_X, G_Y} E_x[\|G_Y(G_X(x)) - x\|_1] + E_y[\|G_X(G_Y(y)) - y\|_1] \tag{3}$$

In our burn image augmentation scenario, we used normal skin images for the source domain X to generate synthetic burn images of four different burn types corresponding to the target domains Y, assuming a quite small dataset for those burn types. That is, the augmentation was applied to transform a normal skin image into burn injury images with a certain severity level. Thus, we built four CycleGAN models for four different burn types. With these CycleGAN models, we evaluated their augmentation capability and utility for limited datasets. A small set of annotated burn images in the target domain with a large set of unannotated, normal skin images in the source domain can be used as a bootstrap to expand the limited datasets and improve the performance of burn severity classifiers learned on the expanded datasets. This approach could be promising when it is difficult to build a sufficiently large dataset of paired images (in our case, paired images of burn injuries and respective normal skins for different burn types) for data augmentation and subsequent model training.

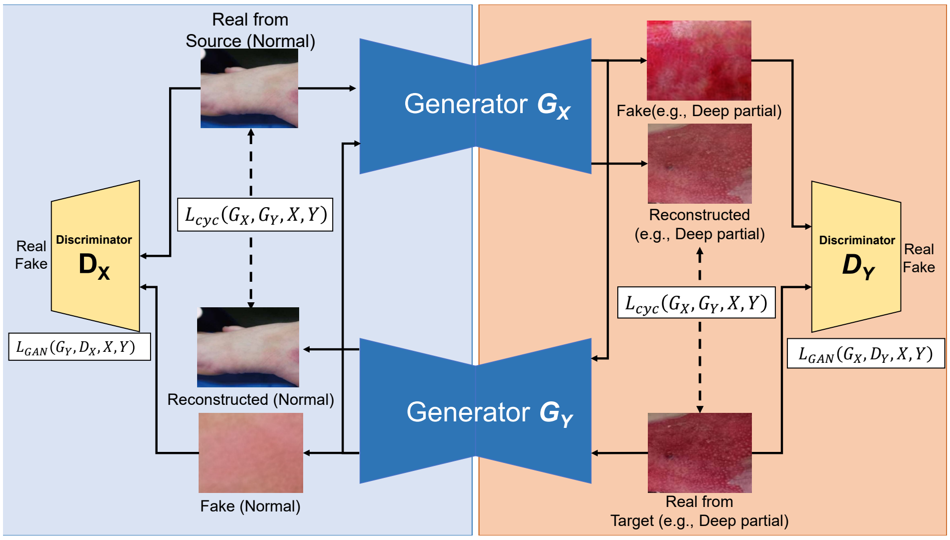


Figure 4. Burn image synthesis based on CycleGAN [40] with two GANs (G_x, D_x and G_y, D_y): L_{cyc} denotes the cycle consistency loss in Equation (3) and L_{GAN} denotes the adversarial loss in Equation (2).

Among several style-based GANs, StyleGAN exploits the benefits of progressive GANs [47] that involve initial learning on low resolution images via a simple neural network and incrementally increase the number of layers in the network for better quality and higher resolution. StyleGAN is now considered to represent a state-of-the-art GAN.

Figure 5 depicts the building blocks of StyleGAN used for our burn image synthesis with different burn types. In the StyleGAN architecture, the generator network is specifically extended to include a mapping function that maps a latent vector ($\mathbf{z}$) to the intermediate latent space ($\mathbf{w}$), while the discriminator architecture is not modified. This structural extension enables control of the style at every level in the generator (synthesis network), combined with injected noise (N) at every level.

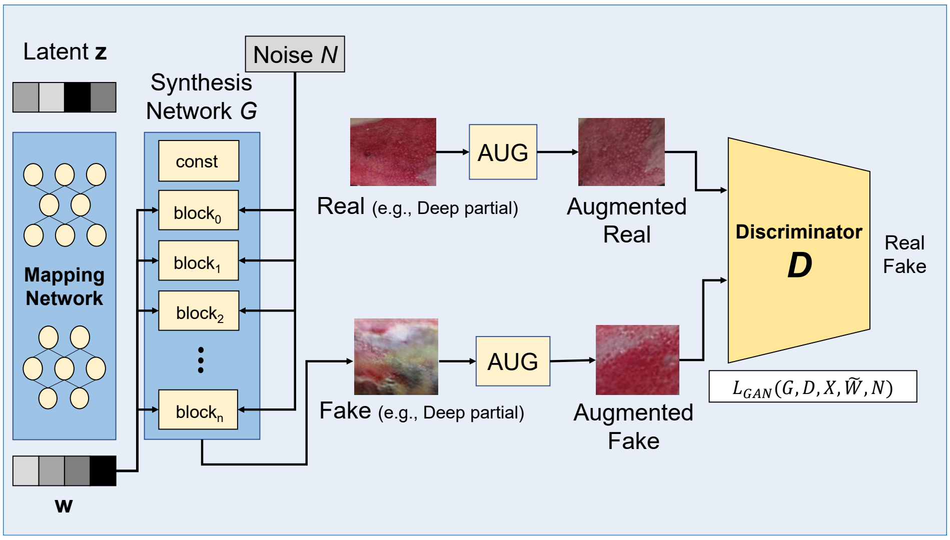


Figure 5. Burn image synthesis based on StyleGAN [41]: the generator structure is extended to include a mapping network and synthesis network to control the image synthesis process, while the discriminator is the same as a conventional GAN.

4. Results

In this section, we evaluate the performance of sample-efficient deep models for burn severity assessment under various data limitation conditions. Our model implementation is based on Python v3.7, PyTorch v1.6.0 [48] and PyTorch-geometric [49], and several open-source projects for ResNet (**ResNet-152**: https://github.com/pytorch/vision, accessed on 1 June 2022, simCLR (**simCLR**: https://github.com/sthalles/SimCLR, accessed on 1 June 2022, https://github.com/Separius/SimCLRv2-Pytorch, accessed on 1 June 2022), CycleGAN (**CycleGAN**: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix, accessed on 1 June 2022), and StyleGAN (**StyleGAN**: https://github.com/NVlabs/stylegan2-ada-pytorch, accessed on 1 June 2022).

4.1. Baseline and TL

To set the baseline performance by a model learned on our training dataset, we tested a vanilla ResNet-152 network trained from scratch. For performance evaluation, we used model accuracy (i.e., $\frac{\text{correct predictions}}{\text{all predictions}}$) for the test dataset.

Figure 6 shows the accuracy of the vanilla ResNet-152 and TL models with respect to various sizes of labeled datasets used in training. We used the performance of vanilla ResNet-152 as a *baseline* for comparison with other sample efficient models. As expected, the larger the dataset, the better the model accuracy. For example, the baseline model learned on a small dataset of 20 samples per class (for short, dataset20) yielded 44.5%, and the model learned on a large dataset of 1700 samples per class (dataset1700) yielded 73.3%. In the following, we use term dataset$K$ where K denotes the number of samples per class.

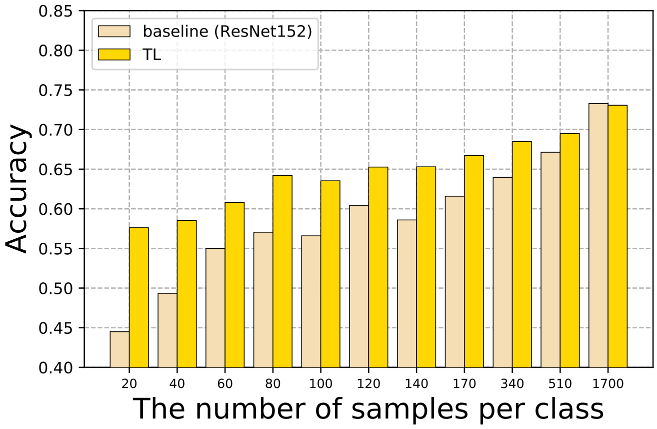


Figure 6. Baseline and TL performance: the graph shows the model accuracy (on the Y-axis) of the vanilla ResNet-152 (baseline) and TL models learned on labeled datasets in various sizes, where the number of training samples per class (burn type) is denoted on the X-axis. For example, 20 and 40 on the X-axis represent models learned on the datasets with 20 and 40 samples per burn type (e.g., superficial partial, intermediate partial, deep partial, full thickness, and normal), respectively.

As there is no publicly available burn image dataset, it is rarely possible to directly compare model accuracy with that of other research studies on burn severity assessment. While the intermediate partial thickness type is difficult for clinicians to correctly evaluate [50] and affects the performance of burn severity assessment models, several studies [9,18,51] have presented such models learned on datasets without much consideration of the intermediate partial thickness type. When we built and tested a model after removing the samples of the intermediate partial thickness type from our dataset, we observed that our baseline accuracy (i.e., 4-class classification) was no lower than 88%, which was much higher than the baseline with the intermediate partial thickness type (i.e., 5-class classification). There have been only a small number of studies processing burn image datasets with the intermediate partial thickness type and training models on those datasets. In [17], a dataset with different burn types, including the intermediate partial thickness type, was used to train deep CNN models for time-independent burn type inference. The models achieved quite high accuracy (e.g., average 73.8∼81.7% in Table 3 in [17]), which can be considered slightly higher than our baseline performance. It should be noted that, as we used our own burn image dataset and sought to adopt sample-efficient learning techniques under data-limited conditions with the dataset, in our experiments, we did not consider it important to compare model performance across different datasets.

We also evaluated TL models that exploited a pretrained ResNet-152 network built on an ImageNet dataset. The TL models showed better accuracy than the respective baseline models in most cases, e.g., 13.1% higher in dataset20 and 9.2% higher in dataset40. However, this gap decreased for larger datasets. It is hypothesized that features relevant to burn types were not fully extracted and represented by the pretrained model due to domain differences between the burn images and diverse images in the ImageNet dataset. The last layer parameter updates in TL might be sufficient to improve the accuracy of low-performance models learned on small datasets. However, the margin of improvement can be less significant for relatively large datasets. In dataset510, the gain of TL from the baseline was small, increasing from 67.1% to 69.5%, and in dataset1700, it barely increased and remained around 73%.

Overall, these results indicate the benefits of TL, particularly for small datasets; they also clarify the limited capability of TL across different domains. The limited performance of several TL models was attributed to the fact that fine-grained, skin-related features required for burn classifications are rarely contained in publicly available pretrained models.

In the following, we concentrate on the data-limited environments in which models are learned on small datasets, such as dataset20∼dataset170, and evaluate SSL, FL, and GAN-based data augmentation techniques on these datasets.

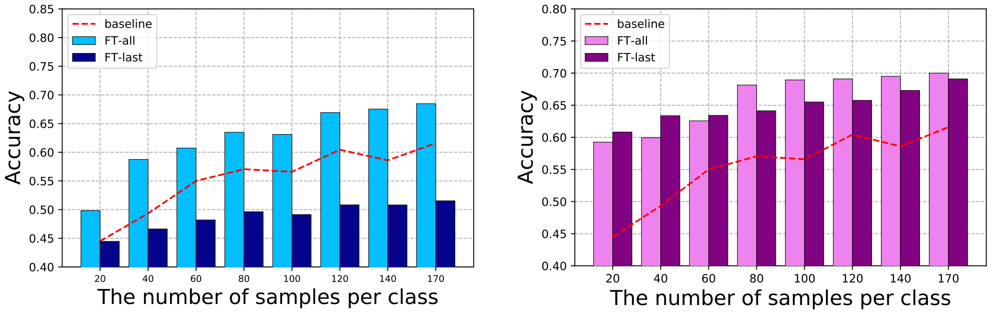
4.2. SSL for Limited Labeled Data

Figure 7 shows the performance of SSL models, where (a) SSL-imn denotes the models utilizing the pretrained simCLR on the ImageNet dataset, and (b) SSL-burn denotes the models pretrained on unlabeled skin images. While these two simCLR models, SSL-imn and SSL-burn, were pretrained on different unlabeled datasets, they were further trained to be fine-tuned on the same datasets of labeled datasets of burn images via supervised learning for various dataset sizes (i.e., dataset20, dataset40, and more, as shown on the X-axis in Figure 7). We also tested two different model updating approaches, including FT-all that updates the model parameters of all layers, and FT-last that updates only the last layer model parameters.

As shown in Figure 7a, FT-all achieved better accuracy than FT-last for all the cases in SSL-imn. The pretrained simCLR in SSL-imn was based on a large-scale, general image dataset, ImageNet, and thus it incurred domain differences when training burn severity assessment models. We also represent the baseline in Figure 6 in the dotted red line for comparison. FT-all showed better accuracy than the baseline, while FT-last showed worse accuracy than the baseline.

In Figure 7b, the SSL-burn models can be seen to show relatively robust accuracy even for small datasets. Both FT-all and FT-last achieved close to 62.6∼68.1% accuracy on dataset60 and dataset80. This was different from SSL-imn (in Figure 7a) where FT-last showed relatively lower performance of about 48.2∼49.6% for the same datasets. This highlights the benefits of SSL-burn which exploits burn image features in pretraining, leading to fast adaptation with a small amount of labeled samples through fine-tuning on entire model parameters.

Interestingly, as shown in Figure 7b, the accuracy of FT-all of SSL-burn started low but increased rapidly alongside more samples, showing that the accuracy achieved by FT-all and FT-last was reversed on dataset80. More importantly, FT-all on dataset80 yielded comparable performance (i.e., 68.1% in accuracy) to the baseline on dataset510, indicating the sample-efficiency of SSL-burn, such that the same level of accuracy to the baseline was able to be achieved by only 16% of the samples, compared to what the baseline uses. In other words, the SSL-burn model learned on a small dataset in data limitation conditions achieved a comparable level of accuracy to the baseline learned on a six-times larger dataset in our experimental settings. SSL-burn was pretrained on unlabeled skin images, and thus its transfer to burn severity assessment models remained in similar domains. Direct comparison between SSL-burn and the baseline on the same dataset also highlighted the sample-efficiency of SSL-burn, showing up to 16.3% and 14% accuracy improvement by FT-last of SSL-burn on dataset20 and dataset40, respectively.



(**a**) Pretrained SSL (SSL-imn)

(**b**) SSL with burn samples (SSL-burn)

Figure 7. SSL performance: the graphs show the model accuracy (on the Y-axis) of the task-agnostic pretrained simCLR (SSL-imn) in (**a**) and the task-specific simCLR pretrained on unlabeled skin images (SSL-burn) in (**b**), respectively, with respect to the number of labeled burn samples per class (on the X-axis) used for task-specific model training on the pretrained simCLR. FT-all denotes models with updating of all model parameters of the pretrained simCLR, and FT-last denotes models with updating of only the last layer model parameters of the pretrained simCLR. The dotted red line corresponds to the baseline in Figure 6.

4.3. FL for Multiple Institutions

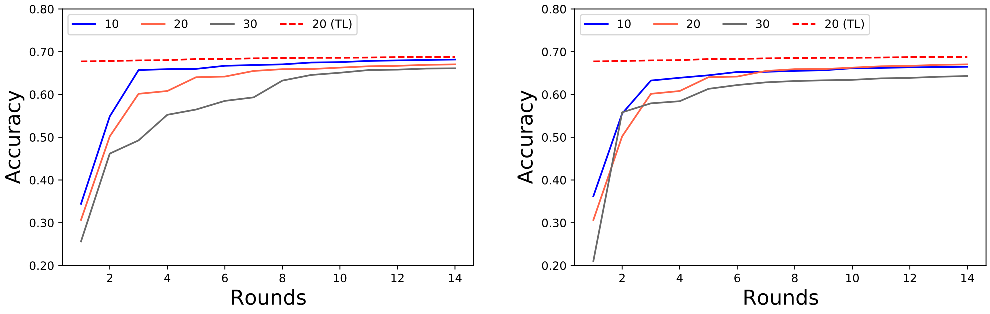
Here, we consider a common situation in the field of medicine where medical institutions individually manage their own medical records and do not share these records with each other. To evaluate FL techniques in this situation, we built an FL simulation environment in which a dataset was spitted to a group of FL clients. Specifically, we used dataset1700 so that each FL client took a partial dataset of the same size as its training dataset. For example, the FL group size was set to 20 by default, and in this default setting, each FL client continuously processed a partial dataset of 85 samples per burn type, i.e., dataset85, for local model training.

Figure 8 shows the accuracy of the FL simulation over federation rounds with respect to various group sizes in (a), and with respect to various numbers of epochs per round in (b), where the performance of FL with the baseline model (without pretrained models) is represented in solid lines, and that of FL with the pretrained model is represented in dotted lines.

As shown in Figure 8a, when converged after certain rounds, FL with the baseline model achieved about 67.7∼68.8% accuracy, slightly better than the baseline on dataset60∼dataset170, showing 55∼61.6% in Figure 6. Note that we compared those on dataset60∼dataset170 with FL models, since different group sizes in FL make each client learn on a local dataset of about 57, 85, or 170 samples per burn type.

Similarly, FL with the pretrained model (denoted as (TL) in Figure 8a,b) with group size 20 and number of rounds 14, achieved 68.8% accuracy, slightly better than the TL model on dataset60∼dataset170 showing 60.8∼66.7% in Figure 6. It was observed that FL with the pretrained model achieved relatively high performance even after the first round. This was also consistent with the superiority of TL over the baseline, particularly for small datasets, which exploits prior knowledge. Overall, these results verify the benefits of FL—the performance of aggregate global models obtained through locally trained models can achieve higher performance than individual local models on datasets of similar sizes.

In FL, the group size and the number of epochs per round are considered important hyperparameters affecting the performance achieved by global models [31]. As shown in Figure 8a, the FL group of 10 clients, where each client has 170 samples per burn type, converged faster and culminated in better performance, compared to the others. We speculate that local training on overly small datasets might negatively affect FL performance. In Figure 8b, it can be seen that the number of epochs per round rarely affected the convergence of the FL models, but an epoch number of 30 was associated with lower performance than the others.



(**a**) Group size

(**b**) Epochs per round

Figure 8. FL performance: the graphs show the model accuracy (on the Y-axis) achieved according to different group sizes in (**a**) and different epochs per round in (**b**), as federation proceeds in rounds (on the X-axis). In (**a**), we tested different FL group sizes from 10 to 30, and in (**b**), we tested different numbers of epochs per round from 10 to 30, where both default values were set to 20. We used the same model as the baseline (the vanilla ResNet-152) for most cases, except for the case of 20(TL) in dotted red lines that used the pretrained ResNet-152 with a group of 20 clients with 20 epochs per round.

4.4. GAN-Based Data Augmentation

We tested several data augmentation techniques including GAN-based ones. Figure 9 shows the performance of baseline models on augmented datasets. For example, 80 on the X-axis corresponds to the models for which the dataset (originally, dataset80) was doubled in size by the conventional transformation methods described in Section 3.3. The models learned on augmented datasets rarely showed better accuracy, compared to their respective baseline models, except for RBC (RandomBrightnessContrast). While it might be difficult for these conventional transformation methods to generate new images which contain adequate relevant features about different burn types, RBC tends to enhance the color characteristics of burn injury parts.

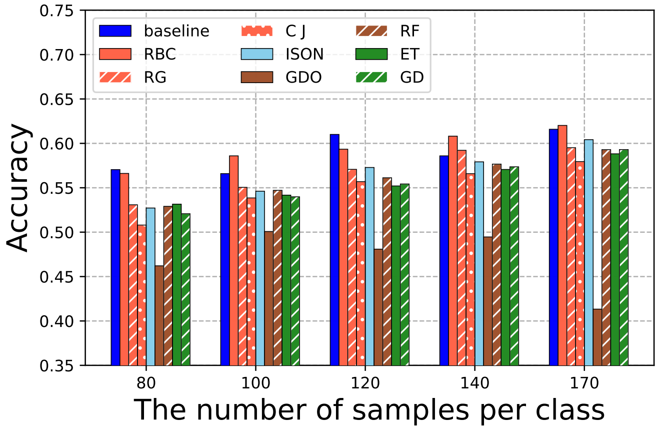


Figure 9. Performance of data augmentation: the graphs show the model accuracy (on the Y-axis) achieved through data augmentation by conventional techniques listed in Section 3.3 with respect to the sizes of training datasets (on the X-axis). Each dataset is augmented to double the size.

In Figure 10, we represent the model accuracy acquired by GAN-based data augmentation methods, where two GAN models are tested to augment datasets such as CycleGAN and StyleGAN. In this experiment, when training the GAN models, labeled burn images were used restrictively. For example, with 80 on the X-axis (dataset80), the respective CycleGAN and StyleGAN were learned on dataset80, as we focused on data-limited conditions. In this way, we ensured our experimental conditions of limited training data such that each GAN model was not exposed to more labeled burn images (the images of different burn types except for normal skin images) than the corresponding burn severity models. We observed performance gains by both GAN-based methods, compared to the baseline, with a larger performance gain observed for CycleGAN than for StyleGAN. We also observed that the gain increased for dataset120 and dataset170 more than for dataset80. This was because the quality of synthetic images generated by the GAN models was dependent on the quantity of samples.

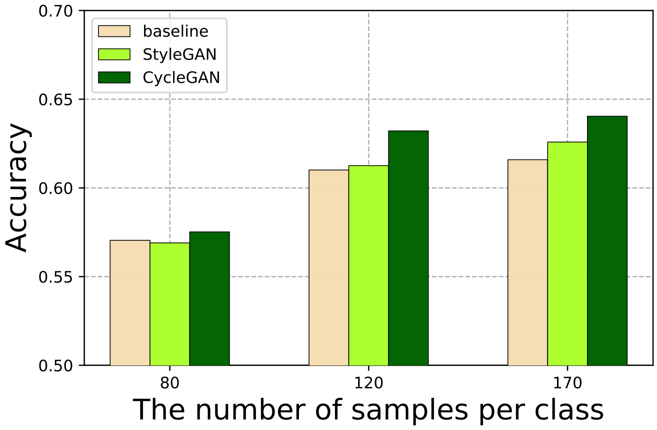


Figure 10. Performance of GAN-based data augmentation models: the graphs show the model accuracy (on the Y-axis) achieved through data augmentation by GAN models with respect to the sizes of training datasets (on the X-axis). Each dataset is augmented to double the size.

In Figure 11, we represent the overall performance comparison among several sample-efficient deep learning techniques. While they can be adopted and optimized differently according to various conditions of limited datasets, we compared their accuracy on dataset170 as a representative case. Overall, the SSL models with pretraining on unlabeled images and full layer fine-tuning (SSL-burn-all) showed better accuracy than the other models. For example, SSL-burn-all on dataset170 achieved 70% while the baseline and TL models achieved 61.6% and 66.7% on the same dataset, respectively. We also applied the CycleGAN-based data augmentation on SSL-burn models (SSL-burn-all+CycleGAN), combining two sample-efficient methods; each demonstrated advantages for improve model accuracy in Figures 7 and 10, respectively. However, despite our expectation, we observed that this rarely improved the model accuracy compared to the respective SSL-burn model. Synthetic images by GAN can improve the model accuracy of low-performance models (e.g., the baseline model), but they rarely do much for non-low-performance models. To generate high-quality synthetic images by GAN techniques, which can improve the model performance for burn image datasets, more research on GAN model optimization is required. More investigation is needed especially for cases when the number of labeled images available for GAN training is quite small.

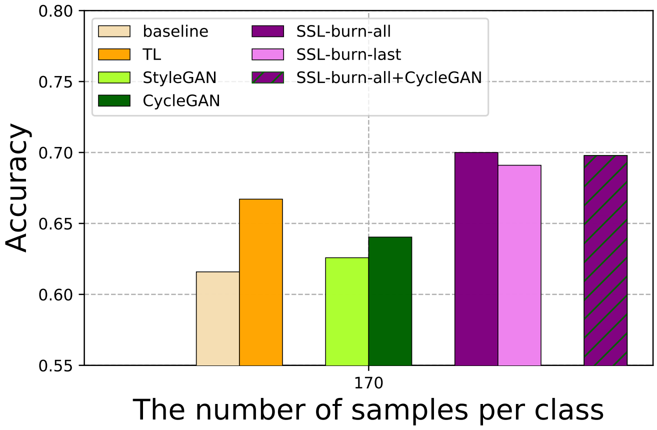


Figure 11. Performance comparison of TL, SSL, and GAN-based data augmentation models: the model accuracy on dataset170 by the baseline and the TL models in Figure 6, the SSL-burn models in Figure 7, and the GAN-based data augmentation models in Figure 10 is compared. SSL-burn-all and SSL-burn-last denote FT-all and FT-last of SSL-burn, respectively. In addition, SSL-burn-all+CycleGAN denotes a combined method of SSL-burn and CycleGAN-based data augmentation.

5. Conclusions

In this study, we developed machine learning models for burn severity assessment. Considering common situations where well-annotated medical images are often not sufficient for model training, we employed and evaluated several sample-efficient deep learning techniques, including TL, SSL, FL, and GAN-based data augmentation for burn severity assessment models. Through extensive experiments with burn images under different dataset conditions, we showed the benefits and limitations of these sample-efficient deep learning techniques to establish the design principles related to specific data conditions in which some techniques can be more effective than others. Specifically, our SSL-based models, which were pretrained on unlabeled images with self-supervised learning schemes and learned on a small labeled dataset task-specifically, achieved comparable performance in burn severity assessment accuracy to a baseline model learned on a six-times larger dataset. This achieved sample-efficiency in model training is an important factor in handling medical image analysis tasks successfully, where only a limited quantity of labeled data is available. To the best of our knowledge, our work is the first to evaluate various deep learning techniques using real-world burn injury images from the perspective of sample-efficiency in model training.

We intend to develop a generalized framework containing sample-efficient deep learning techniques and reference model structures, which can be used to automate the building of deep learning models in the field of medicine. We are also implementing deep learning models for some medical image analysis tasks other than burn severity assessment, in the same vein as this study, focusing on sample-efficiency in model training.

Author Contributions: Conceptualization, H.S. (Hyeonung Shin); Data curation, H.S. (Hyunkyung Shin), H.S. (Hyeonung Shin) and E.K.; Investigation, H.S. (Hyunkyung Shin), H.S. (Hyeonung Shin), W.C., J.P., M.P. and H.W.; Methodology, E.K. and H.W.; Project administration, E.K. and H.W.; Software, W.C., J.P. and M.P.; Writing—original draft, H.S. (Hyunkyung Shin) and E.K.; Writing—review & editing, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Institute for Information & communications Technology Planning & Evaluation (IITP) under grant No. 2021-0-00875, and in part by the ICT Creative Consilience program supervised by the IITP under grant No. IITP-2020-0-01821.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kwasigroch, A.; Grochowski, M.; Mikołajczyk, A. Self-Supervised Learning to Increase the Performance of Skin Lesion Classification. *Electronics* **2020**, *9*, 1930. [CrossRef]
2. Xia, Y.; Yang, D.; Li, W.; Myronenko, A.; Xu, D.; Obinata, H.; Mori, H.; An, P.; Harmon, S.A.; Turkbey, E.B.; et al. Auto-FedAvg: Learnable Federated Averaging for Multi-Institutional Medical Image Segmentation. *arXiv* **2021**, arXiv:2104.10195.
3. Skandarani, Y.; Jodoin, P.M.; Lalande, A. GANs for Medical Image Synthesis: An Empirical Study. *arXiv* **2021**, arXiv:2105.05318.
4. Armanious, K.; Jiang, C.; Fischer, M.; Küstner, T.; Hepp, T.; Nikolaou, K.; Gatidis, S.; Yang, B. MedGAN: Medical image translation using GANs. *Comput. Med. Imaging Graph.* **2020**, *79*, 101684. [CrossRef]
5. Emami, H.; Dong, M.; Nejad-Davarani, S.; Glide-Hurst, C. Generating Synthetic CTs from Magnetic Resonance Images using Generative Adversarial Networks. *Med. Phys.* **2018**, *45*, 3627–3636. [CrossRef]
6. Qin, Z.; Liu, Z.; Zhu, P.; Xue, Y. A GAN-based Image Synthesis Method for Skin Lesion Classification. *Comput. Methods Programs Biomed.* **2020**, *195*, 105568. [CrossRef] [PubMed]
7. Barile, B.; Marzullo, A.; Stamile, C.; Durand-Dubief, F.; Sappey-Marinier, D. Data Augmentation using Generative Adversarial Neural Networks on Brain Structural Connectivity in Multiple Sclerosis. *Comput. Methods Programs Biomed.* **2021**, *206*, 106113. [CrossRef]
8. Abazari, M.; Ghaffari, A.; Rashidzadeh, H.; Badeleh, S.M.; Maleki, Y. A Systematic Review on Classification, Identification, and Healing Process of Burn Wound Healing. *Int. J. Low. Extrem. Wounds* **2022**, *21*, 18–30. [CrossRef]
9. Chauhan, J.; Goyal, P. Deep Learning based Fully Automatic Efficient Burn Severity Estimators for Better Burn Diagnosis. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
10. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020; pp. 1597–1607.
11. Shanthi, T.; Sabeenian, R.; Anand, R. Automatic Diagnosis of Skin Diseases using Convolution Neural Network. *Microprocess. Microsyst.* **2020**, *76*, 103074. [CrossRef]
12. Rashid, J.; Ishfaq, M.; Ali, G.; Saeed, M.R.; Hussain, M.; Alkhalifah, T.; Alturise, F.; Samand, N. Skin Cancer Disease Detection Using Transfer Learning Technique. *Appl. Sci.* **2022**, *12*, 5714. [CrossRef]
13. Kassem, M.A.; Hosny, K.M.; Damasevicius, R.; Eltoukhy, M.M. Machine Learning and Deep Learning Methods for Skin Lesion Classification and Diagnosis: A Systematic Review. *Diagnostics* **2021**, *11*, 1390. [CrossRef] [PubMed]
14. Khan, A.I.; Shah, J.L.; Bhat, M.M. CoroNet: A Deep Neural Network for Detection and Diagnosis of COVID-19 from Chest X-ray Images. *Comput. Methods Programs Biomed.* **2020**, *196*, 105581. [CrossRef] [PubMed]
15. Xie, F.; Yang, J.; Liu, J.; Jiang, Z.; Zheng, Y.; Wang, Y. Skin Lesion Segmentation using High-resolution Convolutional Neural Network. *Comput. Methods Programs Biomed.* **2020**, *186*, 105241. [CrossRef]
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
17. Cirillo, M.D.; Mirdell, R.; Sjöberg, F.; Pham, T. Time-Independent Prediction of Burn Depth using Deep Convolutional Neural Networks. *J. Burn. Care Res. Off. Publ. Am. Burn. Assoc.* **2019**, *40*, 857–863. [CrossRef] [PubMed]
18. Abubakar, A.; Ugail, H.; Bukar, A. Assessment of Human Skin Burns: A Deep Transfer Learning Approach. *J. Med. Biol. Eng.* **2020**, *40*, 321–333. [CrossRef]
19. Chauhan, J.; Goyal, P. Convolution Neural Network for Effective Burn Region Segmentation of Color Images. *Burns* **2021**, *47*, 854–862. [CrossRef]
20. Gouda, N.; Amudha, J. Skin Cancer Classification using ResNet. In Proceedings of the IEEE International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 536–541.
21. Yang, W.; Zhang, H.; Yang, J.; Wu, J.; Yin, X.; Chen, Y.; Shu, H.; Luo, L.; Coatrieux, G.; Gui, Z.; et al. Improving Low-Dose CT Image Using Residual Convolutional Network. *IEEE Access* **2017**, *5*, 24698–24705. [CrossRef]
22. Wang, H.; Xia, Y. ChestNet: A Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography. *arXiv* **2018**, arXiv:1807.03058.
23. Cheplygina, V. Cats or CAT scans: Transfer learning from natural or medical image source data sets? *Curr. Opin. Biomed. Eng.* **2019**, *9*, 21–27. [CrossRef]
24. Reddy, A.S.B.; Juliet, D.S. Transfer Learning with ResNet-50 for Malaria Cell-Image Classification. In Proceedings of the International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 4–6 April 2019; pp. 945–949.
25. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
27. van Engelen, J.E.; Hoos, H. A Survey on Semi-supervised Learning. *Mach. Learn.* **2019**, *109*, 373–440. [CrossRef]

28. Xu, J.; Wang, F. Federated Learning for Healthcare Informatics. *J. Healthc. Inform. Res.* **2021**, *5*, 1–19. [CrossRef] [PubMed]
29. Kaissis, G.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, Privacy-preserving and Federated Machine Learning in Medical Imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311. [CrossRef]
30. Konecný, J.; McMahan, H.B.; Yu, F.; Richtárik, P.; Suresh, A.T.; Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv* **2016**, arXiv:1610.05492.
31. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A.Y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 20–22 April 2017.
32. Feki, I.; Ammar, S.; Kessentini, Y.; Muhammad, K. Federated learning for COVID-19 screening from Chest X-ray images. *Appl. Soft Comput.* **2021**, *106*, 107330. [CrossRef]
33. Bdair, T.; Navab, N.; Albarqouni, S. FedPerl: Semi-supervised Peer Learning for Skin Lesion Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C., Eds.; Springer: Cham, Switzerland, 2021; pp. 336–346.
34. Kaissis, G.; Ziller, A.; Passerat-Palmbach, J.; Ryffel, T.; Usynin, D.; Trask, A.; Lima, I.; Mancuso, J.; Jungmann, F.; Steinborn, M.M.; et al. End-to-end Privacy Preserving Deep Learning on Multi-institutional Medical Imaging. *Nat. Mach. Intell.* **2021**, *3*, 473–484 . [CrossRef]
35. Shorten, C.; Khoshgoftaar, T. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
36. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. [CrossRef]
37. Sandfort, V.; Yan, K.; Pickhardt, P.; Summers, R. Data Augmentation using Generative Adversarial Networks (CycleGAN) to Improve Generalizability in CT Segmentation Tasks. *Sci. Rep.* **2019**, *9*, 100779. [CrossRef]
38. Loey, M.; Smarandache, F.; Khalifa, N.E.M. Within the Lack of Chest COVID-1 X-ray Dataset: A Novel Detection Model Based on GAN and Deep Transfer Learning. *Symmetry* **2020**, *12*, 651. [CrossRef]
39. Kazeminia, S.; Baur, C.; Kuijper, A.; Ginneken, B.V.; Navab, N.; Albarqouni, S.; Mukhopadhyay, A. GANs for Medical Image Analysis. *Artif. Intell. Med.* **2020**, *109*, 101938. [CrossRef]
40. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
41. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4396–4405.
42. Chartsias, A.; Joyce, T.; Dharmakumar, R.; Tsaftaris, S.A. Adversarial Image Synthesis for Unpaired Multi-modal Cardiac Data. In Proceedings of the Simulation and Synthesis in Medical Imaging, Québec City, QC, Canada, 10 September 2017; pp. 3–13.
43. Hamghalam, M.; Wang, T.; Lei, B. High Tissue Contrast Image Synthesis via Multistage Attention-GAN: Application to Segmenting Brain MR Scans. *Neural Netw.* **2020**, *132*, 43–52. [CrossRef] [PubMed]
44. Zhao, C.; Shuai, R.; Ma, L.; Liu, W.; Hu, D.; Wu, M. Dermoscopy Image Classification Based on StyleGAN and DenseNet201. *IEEE Access* **2021**, *9*, 8659–8679. [CrossRef]
45. Yang, H.; Sun, J.; Carass, A.; Zhao, C.; Lee, J.; Xu, Z.; Prince, J. Unpaired Brain MR-to-CT Synthesis Using a Structure-Constrained CycleGAN. In Proceedings of the DLMIA/ML-CDS@MICCAI, Granada, Spain, 20 September 2018; pp. 174–182.
46. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
47. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
48. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
49. Fey, M.; Lenssen, J.E. Fast Graph Representation Learning with PyTorch Geometric. *arXiv* **2019**, arXiv:1903.02428.
50. Johnson, R.M.; Richard, R. Partial-thickness Burns: Identification and Management. *Adv. Ski. Wound Care* **2003**, *16*, 178–189. [CrossRef]
51. Karthik, J.; Nath, G.S.; Veena, A. Deep Learning-Based Approach for Skin Burn Detection with Multi-level Classification. In *Advances in Computing and Network Communications*; Springer: Singapore, 2021; pp. 31–40.