Transfer Learning Using a Multi-Scale and Multi-Network Ensemble for Skin Lesion Classification

Article *in* Computer Methods and Programs in Biomedicine · March 2020

DOI: 10.1016/j.cmpb.2020.105475

6 authors, including:

Amirreza Mahbod
Medical University of Vienna
23 PUBLICATIONS **344** CITATIONS
SEE PROFILE

Chunliang Wang
KTH Royal Institute of Technology
84 PUBLICATIONS **1,960** CITATIONS
SEE PROFILE

Georg Dorffner
Medical University of Vienna
285 PUBLICATIONS **4,006** CITATIONS
SEE PROFILE

Rupert C Ecker
TissueGnostics GmbH
50 PUBLICATIONS **1,463** CITATIONS
SEE PROFILE

Transfer Learning Using a Multi-Scale and Multi-Network Ensemble for Skin Lesion Classification

Amirreza Mahbod, Gerald Schaefer, Chunliang Wang, Georg Dorffner, Rupert Ecker, Isabella Ellinger

---

This is the preprint version. The original paper is published by Computer Methods and Programs in Biomedicine journal on March 2020. When citing this work, please cite the original article available from:

https://doi.org/10.1016/j.cmpb.2020.105475

Transfer Learning Using a Multi-Scale and Multi-Network Ensemble for Skin Lesion Classification

Amirreza Mahbod[a,b,*], Gerald Schaefer[c], Chunliang Wang[d], Georg Dorffner[e], Rupert Ecker[b], Isabella Ellinger[a]

[a] *Institute for Pathophysiology and Allergy Research, Medical University of Vienna, Vienna, Austria*
[b] *Research and Development Department of TissueGnostics GmbH, Vienna, Austria*
[c] *Department of Computer Science, Loughborough University, Loughborough, United Kingdom*
[d] *Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden*
[e] *Section for Artificial Intelligence and Decision Support, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria*

Abstract

Background and Objective: Skin cancer is among the most common cancer types in the white population and consequently computer aided methods for skin lesion classification based on dermoscopic images are of great interest. A promising approach for this uses transfer learning to adapt pre-trained convolutional neural networks (CNNs) for skin lesion diagnosis. Since pre-training commonly occurs with natural images of a fixed image resolution and these training images are usually significantly smaller than dermoscopic images, downsampling or cropping of skin lesion images is required. This however may result in a loss of useful medical information, while the ideal resizing or cropping factor of dermoscopic images for the fine-tuning process remains unknown.

Methods: We investigate the effect of image size for skin lesion classification based on pre-trained CNNs and transfer learning. Dermoscopic images from the International Skin Imaging Collaboration (ISIC) skin lesion classification challenge datasets are either resized to or cropped at six different sizes ranging from 224×224 to 450×450. The resulting classification performance of three well

*Corresponding author
Email address: amirreza.mahbod@tissuegnostics.com (Amirreza Mahbod)

established CNNs, namely EfficientNetB0, EfficientNetB1 and SeReNeXt-50 is explored. We also propose and evaluate a multi-scale multi-CNN (MSM-CNN) fusion approach based on a three-level ensemble strategy that utilises the three network architectures trained on cropped dermoscopic images of various scales.

Results: Our results show that image cropping is a better strategy compared to image resizing delivering superior classification performance at all explored image scales. Moreover, fusing the results of all three fine-tuned networks using cropped images at all six scales in the proposed MSM-CNN approach boosts the classification performance compared to a single network or a single image scale. On the ISIC 2018 skin lesion classification challenge test set, our MSM-CNN algorithm yields a balanced multi-class accuracy of 86.2% making it the currently second ranked algorithm on the live leaderboard.

Conclusions: We confirm that the image size has an effect on skin lesion classification performance when employing transfer learning of CNNs. We also show that image cropping results in better performance compared to image resizing. Finally, a straightforward ensembling approach that fuses the results from images cropped at six scales and three fine-tuned CNNs is shown to lead to the best classification performance.

Keywords: Skin cancer, dermoscopy, medical image analysis, deep learning, image resolution, image cropping, transfer learning.

1. Introduction

Skin cancer is the most common malignancy in the white population and the incidence rates of both malignant melanoma (MM) and non-melanoma skin cancer are increasing on a global scale [1]. MM is one of the most lethal types of skin cancer and worldwide 55,000 people die from MM annually which corresponds to 0.7% of all cancer deaths, although incidence and mortality rates differ widely by country. Early detection is an important factor to increase the overall survival and cure rates for patients with melanoma [2].

Due to their morphological pattern, only 65% – 80% of melanomas are cor-

rectly diagnosed using clinical inspection by an experienced physician [3]. The ABCD rule which assesses asymmetry (A), border (B), colour (C), and diameter (D) of a lesion is often employed for this purpose [4]. Supportive imaging techniques such as dermoscopy, which, through magnification and illumination, makes sub-surface structures more visible, are commonly used and have been shown to reduce screening errors and to lead to better detection of melanomas of clinically atypical appearance [5]. Depending on the dermatologist's experience, dermoscopy can improve the diagnostic accuracy for melanoma detection by up to 50% compared to pure visual inspection [6]. However, diagnostic accuracy can vary greatly among individuals with different experiences [7]. Consequently, there is high interest in the development of semi or fully automatic computer-aided diagnosis (CAD) systems [8, 3] which can be employed in screening programmes or as a second, independent opinion. Such CAD approaches can be based on classical image processing techniques or on advanced machine learning paradigms [9, 10]. Until recently, skin lesion classification was commonly conducted in four steps: (1) image pre-processing, (2) segmentation of the lesion, (3) extraction of hand-crafted features from the segmented lesion and its border and (4) training a classifier for decision making [9]. However, performance of these methods was highly dependent on the segmentation task, which is difficult due to fuzzy borders for some lesions as well as artefacts such as skin hairs and bubbles, and the hand-crafted features, while varying lighting conditions and other factors add further challenges [9].

With the advent of convolutional neural networks (CNNs) and their excellent performance in natural image analysis [11], the current trend is to use these deep models for medical image analysis including dermoscopic image classification [12]. A summary of recent CNN-based methods for skin lesion classification is reported in [3]. Although there are various approaches for performing this task, the methods can be broadly divided into two main categories.

In the first group, either well-known CNN architectures or novel architectures are trained from scratch. The weights of the networks in these approaches are either initialised randomly or with initialisation methods such [13] or [14].

However, since a large number of annotated data are required for training a CNN from scratch, this approach is not often used, especially in medical image analysis where ground truth data is often limited. Examples of training CNNs from scratch for skin lesion classification can be found in [15] and [16]. As reported in these papers, training CNNs from scratch can be a powerful method when the number of training images is high. In [15], more than 129,000 images, mostly derived from private datasets, were used for training, resulting in a classification accuracy that outperformed experienced dermatologists (72% three-way accuracy compared to 65.8% three-way accuracy of two dermatologists). In contrast, in [16] a limited number of training images from the ISIC 2016 challenge [17] were used and the obtained results were relatively poor (66% MM accuracy by training from scratch compared to 81% MM accuracy by fine-tuning a pre-trained CNN).

The second category makes use of transfer learning where the weights of a trained model (which we refer to as backbone model in this paper) for a specific application (e.g., natural image classification) are used for another application (in our case, skin lesion classification). Various well-known pre-trained backbone models, such as AlexNet [18], VGGNet [19], ResNet [20] and its derivatives (e.g., ResNeXt [21]), GoogLeNet [22] and its derivatives (e.g. Inception models [23]), DenseNet [24], NasNet [25], SeNet [26], EfficientNet [27] and mixed models such as Inception-ResNet [28] or SeResNeXt [26], which have different depths, building blocks and architectures can all be used for various transfer learning applications. Transfer learning is widely used in medical image analysis for segmentation, classification and detection across different imaging modalities including radiological images, histological images as well as dermoscopic images [12, 29, 30].

Transfer learning for skin lesion analysis has been used in two ways. Pre-trained CNNs can be employed as optimised feature extractors. Deep features are usually extracted from the last fully connected (FC) layers or last convolutional layers of a pre-trained network and are then used to train a classifier such as a support vector machine (SVM) or a multi-layer perceptron

(MLP) [16, 31, 32]. Alternatively, pre-trained CNNs can be fine-tuned for skin lesion classification [16, 33, 34]. The most common way to fine-tune a pre-trained CNN is to replace the FC layers of the network with a number of new FC layers [10] with the number of neurons in the last layer matching the number of classes in the dataset (i.e., the number of skin lesion types). After this replacement, the whole network is re-trained. However, usually the first to mid layers of the network have a lower learning rate (or even zero learning rate) while the newly added layers can update their weights to match the patterns of the dataset [10, 35].

Besides these two main possibilities of transfer learning for skin lesion classification, other approaches aimed to boost the classification performance have been introduced. These include the combination of two schemes of transfer learning in a single approach [10], using attention layers to guide the network towards lesion areas [36, 37], feeding wavelet transformed images to fine-tune the models [38], using multiple cropped images and concatenating their features for classification [37, 39], and a two-stage network to perform segmentation and classification [40].

Virtually all transfer learning-based methods have one pre-processing step in common which is resizing the images to smaller sizes, or cropping parts of the images. There are two main reasons for this. First, since the backbone networks are trained on natural images of a certain size, it seems rational that resizing the images to the input size of a pre-trained network (e.g., 224×224 for ResNets or 299×299 for Inception models) should lead to good classification performance [37]. The other reason is the computational limitation that makes it impossible to fine-tune a pre-trained CNN with the original size of skin lesion images which can be several thousand pixels along each image dimension. However, useful information could be lost during the downsampling or cropping process and the optimal size for dermoscopic images to fine-tune a network is unknown. In some previous work, resized images of higher resolutions compared to the original input size of the pre-trained network were utilised for skin lesion classification [32, 39, 41]. In all these cases, a certain pre-trained CNN or a

fixed image resolution was used and the effect of using higher resolution images on the classification performance was not reported.

In this paper, we explicitly investigate the effect of different input image sizes, by using either a cropping or a resizing strategy, on skin lesion classification performance employing three different well-established pre-trained CNNs. For this, we examine the classification performance using input images of six different sizes: 224×224, 240×240, 260×260, 300×300, 380×380, and 450×450 pixels. Moreover, we propose and evaluate a unique multi-scale multi-CNN (MSM-CNN) fusion approach by ensembling the results of three different fine-tuned networks which are trained with cropped images at different scales in a three-level fusion scheme for skin lesion classification. Our fusion scheme is straightforward, easy to be implemented and is shown to yield excellent classification performance on the ISIC 2018 challenge test dataset [42] with a multi-class balance accuracy of 86.2%, making it the currently second best approach in the live leaderboard.

2. Materials and Methods

In this part, we first introduce the dataset used for this study. Afterwards, we describe our proposed approach in detail including image pre-processing, CNN fine-tuning and fusion.

2.1. Datasets

We use three datasets from the ISIC archive to fine-tune the models, including the training and test sets of the ISIC 2016 challenge [17], the training, validation and test sets of the ISIC 2017 challenge [43], and the training set of the ISIC 2018 challenge [42]. Since many images of the ISIC 2016 challenge are duplicated in the ISIC 2017 challenge, we disregard duplicates to prevent any biases in the dataset. We also note that there are a few labelling mismatches in the ISIC 2016 and ISIC 2017 challenge datasets (e.g., the same image labelled as nevus in the ISIC 2016 challenge but as melanoma in the ISIC 2017 challenge)

and remove such cases from both data sets. In total, we remove 1117 images from the two datasets and use the remaining 2912 skin lesion images for training. The extracted images from the two datasets have varying sizes ranging from 767×1022 to 4499×6748 pixels. The training set of the ISIC 2018 challenge is derived from the HAM10000 dataset [44] and includes 10015 skin lesion images of a fixed size of 450×600 pixels all of which are included in our training set. In total, we thus use 12927 dermoscopic skin lesion images in the training phase. The utilised images of the ISIC archive were acquired from different centres and with different dermoscopes and therefore, the resolution (mm/pixel) may vary between images. Although most of the images from the ISIC archive are provided with additional metadata such as age or sex information, the actual image resolution (mm/pixel) is not supplied.

The ISIC 2018 challenge dataset contains seven skin lesion types, namely malignant melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AKIEC), benign keratosis (BKL), dermatofibroma (DF) and vascular lesion (VASC). On the other hand, the images from the former ISIC datasets are labelled with fewer classes, the ISIC 2016 dataset with two (melanoma and nevus) and the ISIC 2017 dataset with three (melanoma, seborrheic keratosis and benign nevi). However, based on the comprehensive information in the ISIC archive[1], we are able to extract the labels for all images to form a complete dataset with seven skin lesion classes similar to the ISIC 2018 labels, 1666 MEL, 8677 NV, 514 BCC, 327 AKIEC, 1486 BKL, 115 DF, and 142 VASC.

For evaluation, we use the ISIC 2018 challenge test set which comprises 1512 skin lesion images of a fixed size of 450×600 pixels. The labels of these test images are kept private and it is only possible to evaluate the performance on the test set through the ISIC live evaluation platform[2].

[1] https://www.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery

[2] https://challenge.isic-archive.com

2.2. Pre-processing

We employ standard pre-processing steps on the images before feeding them to the networks. First, we apply a grayworld colour constancy algorithm to normalise the colours of the images as suggested in [45]. This pre-processing step deals with the various lightening conditions in the images and is widely used for skin lesion analysis [10, 34, 46]. Then, we subtract the mean intensity RGB value of the ImageNet dataset [11] from each individual channel of all training and test images which is a common pre-processing step for transfer learning. To have a similar scale of all skin lesions in the training dataset, the ISIC 2016 and ISIC 2017 challenge images are resized, using bicubic interpolation, from $M \times N$ to $450 \times N'$ where $N' = \frac{N}{M} \times 450$.

In order to investigate the effect of image size on skin lesion classification performance, we explore two different strategies, namely resizing and random cropping of the images, with six pre-defined scales. For the resizing strategy, the images are resized to six different resolutions of 224×224, 240×240, 260×260, 300×300, 380×380, and 450×450 using bicubic interpolation. For non-square images, we use zero padding to preserve the aspect ratio while resizing. For cropping, random crops of the same six sizes are extracted from the images. Random cropping is performed during training as an augmentation to increase the cropping variability. An example of resizing strategy and random cropping strategy from the original data is depicted in Fig. 1.

2.3. Pre-trained CNNs

As mentioned previously, a number of pre-trained models can be used for transfer learning. In this paper, we select three models from the SeNet and EfficientNet families which have shown excellent classification performance for both natural and medical images. EfficientNets [27] are state-of-the-art models for natural image classification, while SeResNeXt [26] is a variation of SeNet which integrates the squeeze and excitation blocks into a ResNeXt model. Both models have been previously successfully used for medical image classification by transfer learning (e.g., in the gold medal winners of the APTOS 2019 Blindness

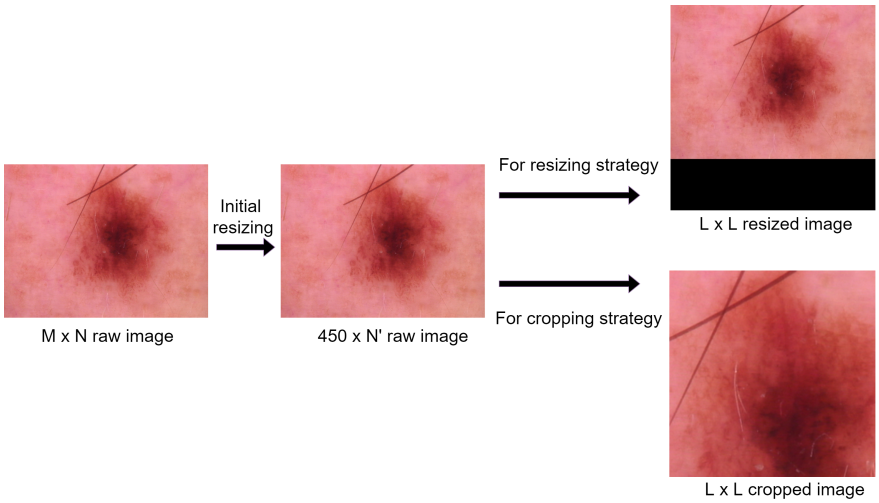

Figure 1: Example of the utilised resizing or cropping strategy. $M \times N$ represents the original image size, $450 \times N'$ represents the image size after initial resizing and $L \times L$ refers to the size of the resized or the cropped image ($L \in \{224, 240, 260, 300, 380, 450\}$).

Detection competition[3]). Although for both architectures networks of varying depths exist, we choose the shallower depths to prevent overfitting to our limited training data, and select EfficientNetB0, EfficientNetB1 and SeResNeXt-50 as our backbone models.

EfficientNet models were originally trained on six image resolutions similar to the ones used for resizing and cropping in this paper, which is the main reason behind the selection of the pre-defined sizes in Section 2.2. All utilised image sizes are identical to the original image sizess of the EffcientNet default input except the largest size (one variation of EfficientNet was trained on 456×456 images but we use 450×450 images since most of the training images have a resolution of 450×600). For the same reason, and also considering the computational burden, we did not investigate the effect of image sizes bigger than 450×450 pixels in this paper.

[3] https://www.kaggle.com/c/aptos2019-blindness-detection

2.4. Fine-tuning

The general workflow that we use for fine-tuning is depicted in Fig 2.

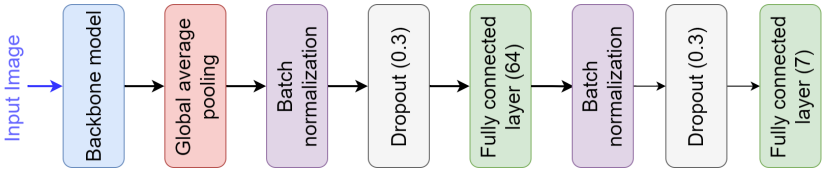

Figure 2: Fine-tuning workflow.

First, we remove the fully connected layers from the backbone model. Then, we add an average pooling layer to the backbone model's output to take care of the various input image resolutions. Two blocks of batch normalisation, dropout and fully connected (FC) layers are then added to the network. In the dropout layers, we use a dropout factor of 0.3, while an FC layer with 64 nodes in the first block and an FC layer with 7 nodes (to match the 7 classes) are employed. The weights of the new FC layers are initialised by the well-known Xiavier initialisation method [13]. We use L2 regularisation in the new dense layers with a fixed value of 0.0001, and adaptive moment estimation (Adam) [47] for the optimisation with an initial learning rate of 0.0001. However, we keep the learning rate of the new FC layers 10 times larger compared to all other weight layers. A weighted focal loss [48] is used to deal with the heavily imbalanced dataset. The class weight for each class is calculated by dividing the number of images for a class by the total number of training images so that under-represented classes have larger weights. We also explored having an equal numbers of images in each training batch to deal with class imbalance but do not use it in the final setting due to insufficient performance. We choose varying batch sizes based on the used network, image resolution and the utilised GPU memory ranging from 8 to 32. Finally, we use 5-fold cross validation and train the networks for 70 epochs while the learning rate is dropped by a factor of 2 at the 40-th, 50-th and 60-th epoch. We monitor the average recall score on the validation set and save the best model based on the validation performance.

The motivation of using the batch normalisation layer, dropout layer and regularisation term in the dense layers is to prevent overfitting to the limited training set [49]. Most of the aforementioned selected hyperparameters are based on our previous work [10] as well as other related work in skin lesion classification [34, 36].

To obtain more robust models, we use various augmentation techniques including random scaling, random rotations, vertical and horizontal flipping, random brightness and contrast shifts, random adaptive histogram equalisation, random cutouts and random manipulations of HSV colour channels. For the cropping strategy, we apply random cropping on the training images, and to calculate the mean recall score on the validation set, we apply central cropping on the validation set images.

As described in Section 2.1, the image resolution (mm/pixel) is unknown. Therefore, we resample all the images to have similar image sizes in terms of pixel by pixel (as described in Section 2.2) and assume there are sufficient training data at different resolutions so that the networks learn sufficiently about them. This issue is further resolved by the applied random scaling augmentation in the training phase.

In the inference phase, as illustrated in Fig. 3, we apply augmentation similar to the training phase augmentation to create 50 augmented images of a single test image. These images are fed to the trained networks to obtain probability vectors. If the maximum value in a probability vector is below 50%, we disregard that vector. Finally, we take the average over the remaining prediction vectors to yield the prediction vector for a sample test image. This step is important for the cropping strategy as random croppings might result in some non-informative part of the image (e.g., from the background) being selected.

2.5. MSM-CNN with three-level fusion

For our proposed MSM-CNN, we develop a three-level fusion scheme illustrated in Fig. 4.

At level one, we train models with cropped images at a fixed size using

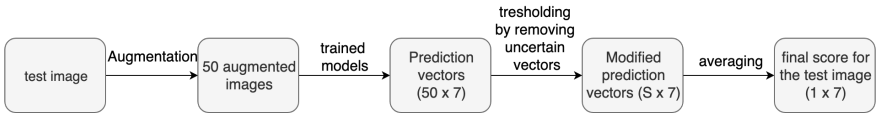

Figure 3: Augmentation in the inference phase. S refers to the number of augmented images where the prediction vectors has the maximum value of at least 0.5.

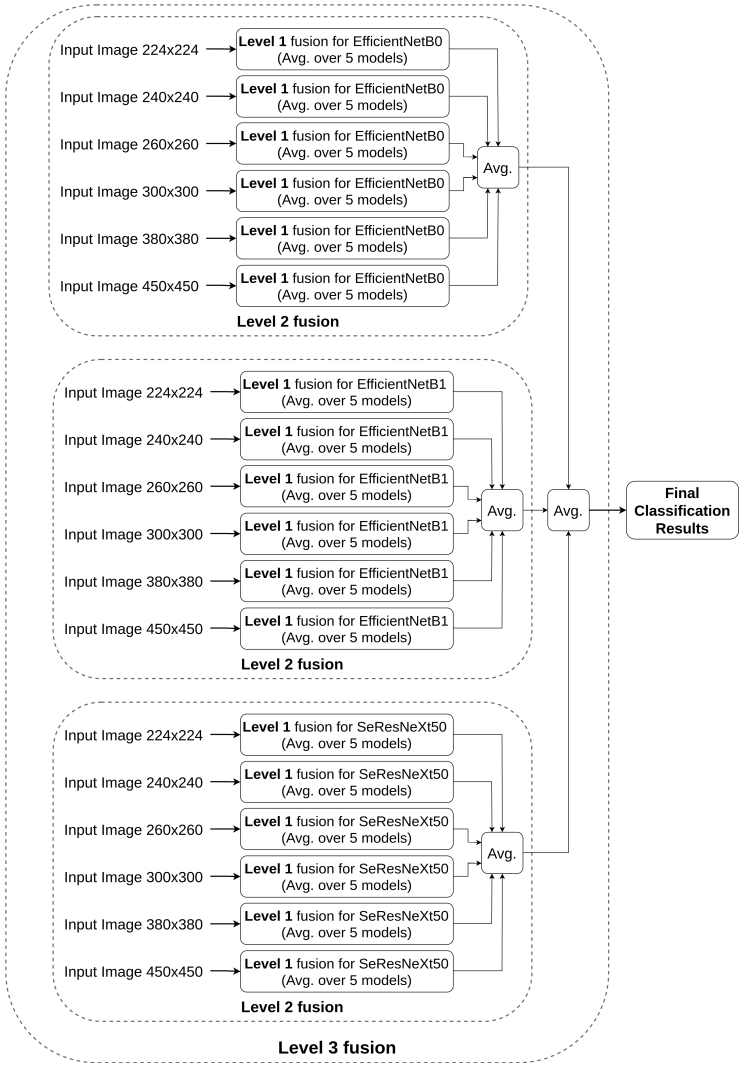

Figure 4: The proposed MSM-CNN three-level fusion scheme. All input images that were fed to the networks were cropped images at various scales.

5-fold cross validation, resulting in 5 models (one for each fold of cross validation). We then apply these sub-models on the test set and calculate the average classification probability vectors. At level two, we also fuse the results from the individual networks trained on the six different image sizes (i.e., 224×224, 240×240, 260×260, 300×300, 380×380, and 450×450). At the third and final fusion level, we fuse the predicted probability vectors of the various architectures to yield the final classification result. For fusion at all three levels, a simple averaging is applied on the network outputs. The final MSM-CNN classification is thus derived from 90 ($5 \times 6 \times 3$) sub-models.

2.6. Evaluation

As suggested for the ISIC 2018 skin lesion classification challenge, we use the average recall score (referred to as balanced multi-class accuracy (BMCA) in the ISIC 2018 challenge) as the main evaluation measure in our experiments. We are therefore able to compare our obtained results with other algorithms that have previously been applied on the same dataset from the live and legacy leaderboards of the challenge. For MSM-CNN third fusion level, we also report average precision, average accuracy and average area under the receiver operating characteristic curve (AUC) and compare them to other state-of-the-art algorithms.

3. Results

As mentioned, the reported results are based on the 1512 test images of the ISIC 2018 skin lesion classification challenge. None of the test set images was used in the training or validation phase. Since the image labels for the test set are kept private by the challenge organisers, we use the ISIC online evaluation platform to obtain our results.

We start the experiments by comparing the effect of image resizing and image cropping on the classification performance for six image scales for one sample network, namely EfficientNetB0. The results of this are shown in Table 1, which shows that cropping consistently outperforms resizing.

Table 1: BMCA results of cropping and resizing strategies for EfficientNetB0 model. The last column report the average over the 6 results.

	224×224	240×240	260×260	300×300	380×380	450×450	average
resizing	75.3	76.1	76.1	77.3	80.3	80.8	77.6 ± 2.3
cropping	83.0	84.3	83.3	83.3	83.6	83.6	83.5 ± 0.4

Consequently, we continue the experiments with the cropping strategy only. The effect of six image sizes on the classification performance of all three networks using the cropping strategy is depicted in Table 2.

Table 2: Classification performance, in terms of BMCA, of cropping strategy for different network models.

	224×224	240×240	260×260	300×300	380×380	450×450	average
EfficientNetB0	83.0	84.3	83.3	83.3	83.6	83.6	83.5 ± 0.4
EffcientNetB1	84.4	83.9	84.0	82.7	81.8	81.4	83.0 ± 1.2
SeResNeXt-50	79.5	79.9	80.9	80.3	82.9	82.3	80.9 ± 1.4

Table 3 shows the results obtained by the higher level fusion (level 2 and 3 fusion in Fig. 4) schemes described in Section 2.5.

Table 3: Results for level 2 and level 3 fusion schemes.

networks	input size	BMCA (%)
level 2 fusion - EfficientNetB0	all sizes	84.6
level 2 fusion - EfficientNetB1	all sizes	84.6
level 2 fusion - SeResNeXt-50	all sizes	83.3
level 3 fusion - MSM-CNN	all sizes	86.2

In order to investigate the effect of each network on the classification performance in our proposed fusion scheme, we report the results obtained from fusing the other two network models in Table 4.

We also evaluate an additional fusion strategy which performs fusion of the

Table 4: Results from removing a model from the fusion scheme.

networks	BMCA (%)
EfficientNetB0+EfficientNetB1	85.1
EfficientNetB0+SeResNeXt-50	85.7
EfficientNetB1+SeResNeXt-50	85.5

three network outputs for a single image resolution and report the results in Table 5.

Table 5: Results of fusing different networks for a single image resolution.

networks	input size	BMCA (%)
fusion of all networks	224×224	84.4
best single network (EffcientNetB1)	224×224	84.4
fusion of all networks	240×240	84.3
best single network (EffcientNetB0)	240×240	84.3
fusion of all networks	260×260	84.9
best single network (EffcientNetB1)	260×260	84.0
fusion of all networks	300×300	84.1
best single network (EffcientNetB0)	300×300	83.3
fusion of all networks	380×380	84.4
best single network (EffcientNetB0)	380×380	83.6
fusion of all networks	450×450	83.9
best single network (EffcientNetB0)	450×450	83.6

In Table 6, we compare the performance of the best performing approach, i.e. our proposed MSM-CNN fusion approach, to both the top three performers of the ISIC 2018 competition legacy leaderboard and the online leaderboard. Although the ranking is based on BMCA scores, we also report results in terms of mean precision, mean accuracy and mean AUC across the seven skin lesion

types. As can be seen from Table 6, as well as from the live leaderboard[4], at the time of writing this paper, our proposed approach is ranked second overall among more than 200 submissions (the leaderboard only lists the current top 200 submissions).

Table 6: Results of ISIC 2018 challenge winners from the legacy leaderboard (rows 1-3), current ISIC 2018 challenge online leaderboard (rows 4-6), and our proposed MSM-CNN (bottom row). ACC=accuracy, PR=precision.

team/ authors	# externally used images	BMCA (%)	avg. ACC (%)	avg. PR (%)	avg. AUC (%)
Nozdryn *et al.*	37807	88.5	95.8	91.7	98.3
Gassert *et al.* [34]	13475	85.6	97.2	93.1	98.7
Zhuang *et al.*	n/a	84.5	96.8	89.1	97.8
IPM-HPC	n/a	86.6	96.3	83.3	97.6
Gassert *et al.* [34]	13475	85.6	97.0	92.7	98.7
Skinai-1	n/a	84.3	96.7	70.9	91.0
MSM-CNN	2912	86.2	96.3	91.3	98.1

Our algorithm is implemented using the Keras deep learning framework (ver. 2.3.1) using Tensorflow (ver. 1.14) as the backend. All experiments were conducted on a single workstation with an Intel Core i7-8700 3.20 GHz CPU, 32 GB of RAM and an nVIDIA Titan V GPU with 12 GB of installed memory. The average inference times for the complete test image dataset for each architecture and each image resolution are reported in Table 7. For the full MSM-CNN approach, based on the final result being obtained through 90 sub-models, it takes 13.9 seconds to classify a single test images.

[4] https://challenge2018.isic-archive.com/live-leaderboards/ (Task 3: lesion diagnosis)

Table 7: Average inference times (in minutes) for the entire test set (1512 images) for each individual model and image resolution.

	EfficientNetB0	EfficientNetB1	SeResNeXt-50
224×224	8.76	12.12	31.22
240×240	8.83	12.30	31.46
260×260	9.04	12.49	32.29
300×300	9.63	13.27	33.03
380×380	11.35	15.21	35.51
450×450	14.31	19.01	40.49

4. Discussion

The main contributions of this paper are, first, investigating the effect of image downsampling and cropping on skin lesion classification performance, and second, proposing a three-level fusion approach that achieves excellent classification performance on the ISIC 2018 challenge test dataset using multiple fine-tuned deep networks and cropped multi-scale skin lesion images.

The results in Table 1 compare the effects of the cropping and resizing strategies on classification performance. It can be seen that at all image resolutions, the cropping strategy outperforms the resizing method. The performance differences become smaller for larger image sizes (7.7% for 224×224 vs. 2.8% for 450×450), which is expected since cropped and resized images become more similar for larger image sizes. On the other hand, for the resizing strategy, classification performance improves for larger image sizes, whereas it is relatively stable across image resolutions for image cropping.

Table 2 allows us to compare the classification performance of the three used networks for all image resolutions. EfficientNetB0 performs slightly better compared to EfficientNetB1 and much better compared to SeResNeXt-50.

The results in Table 3 show the effect of the second- and third-level fusion schemes in our approach. Second-level fusion delivers better performance compared to the average results of the various image sizes (84.6% vs. 83.5%,

84.6% vs. 83.0%, and 83.3% vs. 80.9% for EfficientNetB0, EfficientNetB1 and SeResNeXt-50, respectively). However, compared to the best performance of a network at a singe image resolution, the improvement is relatively minor (84.6% vs. 84.3% (for 240×240), 84.6% vs. 84.4% (for 224×224), and 83.3% vs. 82.9% (for 380×380) for EfficientNetB0, EfficientNetB1 and SeResNeXt-50, respectively. As the single networks yield their best performance at different image sizes, the level two fusion scheme leads to a more robust overall decision by taking into account all network outputs at various image scales. Our proposed MSM-CNN algorithm, which additionally performs level three fusion, combining 90 models, yields the best classification performance of 86.2%, outperforming all single networks and all lower level fusion schemes.

The contribution of each network in the fusion scheme is illustrated in Table 4. As we can see from there, removing each network in turn leads to a degradation in classification performance, suggesting that the chosen networks provide complementary information so that the ensemble of all of them delivers the best results. Another interesting observation from Table 4 is that while SeResNeXt-50 yields the worst classification performance, dropping this model from the fusion scheme degrades the classification performance the most. Since EfficientNetB0 and EfficientNetB1 are from the same network family, their fusion leads to a lesser complementary effect in comparison to the fusion of the EfficientNet and SeResNeXt-50 models.

The results in Table 5 show that fusing different networks at a single image resolution also leads to a slight improvement in classification performance for most image scales. However, the results here are inferior compared to MSM-CNN by at least 1.3%.

The comparison of our approach with other state-of-the-art methods submitted to the ISIC 2018 challenge in Table 6 shows that our MSM-CNN algorithm outperforms all but one existing method each on both the legacy and live leaderboards of the competition, highlighting the excellent performance of our proposed algorithm on this very challenging task. While all reported results are obtained on the ISIC 2018 challenge test dataset, a direct comparison of

the classification performance is not trivial since different training sets are used in the different approaches. However, our method uses fewer external training samples compared to the other approaches (where known). While the challenge leaderboards are ranked based on the BMCA, our approach also delivers results comparable to the other state-of-the-art methods in terms of other performance measures (average accuracy and precision, area under ROC curve) as indicated in Table 6.

Looking at Table 7, it is apparent that the training time required for SeResNeXt-50 is significantly higher compared to the other networks for all image sizes. This is not surprising since SeResNeXt-50 has many more trainable parameters (there are about ~ 4.1, ~ 6.7, and ~ 25.7 million parameters for EfficientNetB0, EfficientNetB1, and SeResNeXt-50, respectively). Also, as expected, inference time is proportional to the image size.

At all three levels of our proposed fusion scheme, we use averaging to fuse the network outputs. In [50], a multi-scale CNN (M-CNN) is presented that uses multiple scale images in one single network. However, as the network width increases drastically, only three convolutional layers are used leading to a very shallow network. Moreover, with a new architecture proposed, it is necessary to train the model from scratch and hence not possible to take advantage of transfer learning, while this approach also does not allow to evaluate the contribution of each image scale to the final classification performance. However, with sufficient computational power, the classification performance of an M-CNN with pre-trained deep models for each image scale can be investigated.

There are some limitations in this work. The biggest limitation of our fusion approach is the large number of utilised sub-models that consequently need significant training time. This may not be suitable for application in a clinical setting even though this process is typically performed offline and the inference time to perform classification is rather short (as shown in Table 7). However, since the different networks are trained independently, this could be done in parallel on suitable hardware (e.g., multiple GPUs), and the outputs fused at the end. The training time and model complexity can be also reduced by removing

one or two networks from the fusion scheme with a slight sacrifice in terms of accuracy (as shown in Table 4).

Another issue that can be further investigated is the resizing or cropping factor. The utilised six image scales were selected based on the EfficientNet original input image sizes. While in this paper, we utilised six pre-defined image scales, the effect of other image resolutions between the minimum and maximum sizes can also be explored (since almost 80% of the training data is of size 450×600 the biggest scale we explore in this paper is 450×450). We also performed some additional experiments on the images resized to smaller sizes of 32×32, 64×64, 128×128 pixels but the obtained results (BMCAs of 56.5%, 59.7% and 68.0%) are much inferior to the ones we employ above indicating that important information is lost at these smaller scales.

Finally, the number of pre-trained networks that we use is limited to three pre-trained CNNs in this paper. We employ some of most advanced pre-trained CNNs which have shown excellent classification performance for medical image classification in the literature but classification performance can likely be further improved by exploiting and adding other networks to the fusion scheme although this would also further increase its computational complexity.

5. Conclusions

In this paper, we have investigated the effect of image size for transfer learning classification performance in the context of skin lesion analysis. The results of our study show that a cropping strategy outperforms image resizing for skin lesion classification, although classification performance improves with image resolution for resizing while it is relatively stable with respect to image size for cropping. We have also presented a three-level fusion approach – MSM-CNN – which combines results from different networks and cropped images at various sizes and have shown it to yield excellent classification performance on the ISIC 2018 skin lesion classification challenge test dataset compared to a number of state-of-the-art algorithms for skin lesion analysis.

Conflict of interest statement

There are no conflicts of interest to disclose for publication of this paper.

Acknowledgements

This work was supported by the EU Horizon 2020 CaSR Biomedicine project, No. 675228. The authors would like to thank the TissueGnostics Research and Development team[5] and especially Dr. Alain Pitiot for valuable suggestions. Moreover, we thank nVIDIA corporation for their generous GPU donation. There is no need for ethical approval for this study.

References

[1] U. Leiter, T. Eigentler, C. Garbe, Epidemiology of skin cancer, in: Sunlight, Vitamin D and Skin Cancer, Springer, 2014, pp. 120–140.

[2] D. Schadendorf, A. C. van Akkooi, C. Berking, K. G. Griewank, R. Gutzmer, A. Hauschild, A. Stang, A. Roesch, S. Ugurel, Melanoma, The Lancet 392 (10151) (2018) 971–984.

[3] T. J. Brinker, A. Hekler, J. S. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. H. Enk, C. von Kalle, Skin cancer classification using convolutional neural networks: Systematic review, Journal of Medical Internet Research 20 (10) (2018) e11936.

[4] W. Stolz, A. Riemann, A. B. Cognetta, L. Pillet, W. Abmayr, D. Holzel, P. Bilek, F. Nachbar, M. Landthaler, O. Braun-Falco, ABCD rule of dermatoscopy: A new practical method for early recognition of malignant melanoma, European Journal of Dermatology 4 (7) (1994) 521–527.

[5] I. Tromme, L. Sacré, F. Hammouch, C. Legrand, L. Marot, P. Vereecken, I. Theate, P. van Eeckhout, P. Richez, J.-F. Baurain, et al., Availability

[5] http://tissuegnostics.com/en/

of digital dermoscopy in daily practice dramatically reduces the number of excised melanocytic lesions: results from an observational study, British Journal of Dermatology 167 (4) (2012) 778–786.

[6] H. Kittler, Dermatoscopy of pigmented skin lesions, Giornale Italiano di Dermatologia e Venereologia 139 (6) (2004) 541–546.

[7] P. H. Youl, B. A. Raasch, M. Janda, J. F. Aitken, The effect of an educational programme to improve the skills of general practitioners in diagnosing melanocytic/pigmented lesions, Clinical and Experimental Dermatology: Clinical dermatology 32 (4) (2007) 365–370.

[8] D. A. Okuboyejo, O. O. Olugbara, A review of prevalent methods for automatic skin lesion diagnosis, The Open Dermatology Journal 12 (1) (2018).

[9] R. B. Oliveira, J. P. Papa, A. S. Pereira, J. M. R. S. Tavares, Computational methods for pigmented skin lesion classification in images: review and future trends, Neural Computing and Applications 29 (3) (2018) 613–636.

[10] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, C. Wang, Fusing fine-tuned deep features for skin lesion classification, Computerized Medical Imaging and Graphics 71 (2019) 19–29.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, ImageNet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[12] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88.

[13] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.

[14] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2015, pp. 1026–1034.

[15] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.

[16] A. R. Lopez, X. Giro-i Nieto, J. Burdick, O. Marques, Skin lesion classification from dermoscopic images using deep learning techniques, in: IASTED International Conference on Biomedical Engineering, 2017, pp. 49–54.

[17] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, Skin lesion analysis toward melanoma detection: A challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC), arXiv preprint arXiv:1605.01397 (2016).

[18] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778.

[21] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5987–5995.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 1–9.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[24] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, IEEE, 2017, pp. 4700–4708.

[25] B. Zoph, V. Vasudevan, J. Shlens, Q. V. Le, Learning transferable architectures for scalable image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 8697–8710.

[26] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018, pp. 7132–7141.

[27] M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, arXiv preprint arXiv:1905.11946 (2019).

[28] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, pp. 4278–4284.

[29] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. van Ginneken, et al., Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 516–524.

[30] A. Mahbod, I. Ellinger, R. Ecker, Ö. Smedby, C. Wang, Breast cancer histological image classification using fine-tuned deep network fusion, in: A. Campilho, F. Karray, B. ter Haar Romeny (Eds.), Image Analysis and Recognition, Springer International Publishing, Cham, 2018, pp. 754–762.

[31] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, I. Ellinger, Skin lesion classification using hybrid deep neural networks, in: International Conference on Acoustics, Speech and Signal Processing, IEEE, 2019, pp. 1229–1233.

[32] J. Kawahara, A. BenTaieb, G. Hamarneh, Deep features to classify skin lesions, in: International Symposium on Biomedical Imaging, IEEE, 2016, pp. 1397–1400.

[33] X. Zhang, S. Wang, J. Liu, C. Tao, Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge, BMC Medical Informatics and Decision Making 18 (2) (2018) 59.

[34] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, A. Schlaefer, Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting, arXiv preprint arXiv:1808.01694 (2018).

[35] J. R. Hagerty, R. J. Stanley, H. A. Almubarak, N. Lama, R. Kasmi, P. Guo, R. J. Drugge, H. S. Rabinovitz, M. Oliviero, W. V. Stoecker, Deep learning and handcrafted method fusion: Higher diagnostic accuracy for melanoma dermoscopy images, IEEE Journal of Biomedical and Health Informatics 23 (4) (2019) 1385–1391.

[36] Y. Yan, J. Kawahara, G. Hamarneh, Melanoma recognition via visual attention, in: A. C. S. Chung, J. C. Gee, P. A. Yushkevich, S. Bao (Eds.), Information Processing in Medical Imaging, Springer International Publishing, Cham, 2019, pp. 793–804.

[37] J. Zhang, Y. Xie, Y. Xia, C. Shen, Attention residual learning for skin lesion classification, IEEE Transactions on Medical Imaging 38 (9) (2019) 2092–2103.

[38] S. Serte, H. Demirel, Gabor wavelet-based deep learning for skin lesion classification, Computers in Biology and Medicine 113 (2019) 103423.

[39] T. DeVries, D. Ramachandram, Skin lesion classification using deep multi-scale convolutional neural networks, arXiv preprint arXiv:1703.01402 (2017).

[40] Y. Li, L. Shen, Skin lesion analysis towards melanoma detection using deep learning network, Sensors 18 (2) (2018) 556.

[41] Z. Yu, X. Jiang, T. Wang, B. Lei, Aggregating deep convolutional features for melanoma recognition in dermoscopy images, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2017, pp. 238–246.

[42] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC), arXiv preprint arXiv:1902.03368 (2019).

[43] 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, 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), arXiv preprint arXiv:1710.05006 (2017).

[44] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Scientific Data 5 (2018) 180161.

[45] C. Barata, M. E. Celebi, J. S. Marques, Improving dermoscopy image classification using color constancy, IEEE Journal of Biomedical and Health Informatics 19 (3) (2015) 1146–1152.

[46] K. Matsunaga, A. Hamada, A. Minagawa, H. Koga, Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble, arXiv preprint arXiv:1703.03108 (2017).

[47] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference for Learning Representations, 2015.

[48] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2017, pp. 2980–2988.

[49] A. Aldwgeri, N. F. Abubacker, Ensemble of deep convolutional neural network for skin lesion classification in dermoscopy images, in: H. Badioze Zaman, A. F. Smeaton, T. K. Shih, S. Velastin, T. Terutoshi, N. Mohamad Ali, M. N. Ahmad (Eds.), Advances in Visual Informatics, Springer International Publishing, Cham, 2019, pp. 214–226.

[50] W. J. Godinez, I. Hossain, S. E. Lazic, J. W. Davies, X. Zhang, A multiscale convolutional neural network for phenotyping high-content cellular images, Bioinformatics 33 (13) (2017) 2010–2019.