European Journal of Cancer 111 (2019) 148–154

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task

Titus J. Brinker [a,b,*], Achim Hekler [a], Alexander H. Enk [b], Joachim Klode [c], Axel Hauschild [d], Carola Berking [e], Bastian Schilling [f], Sebastian Haferkamp [g], Dirk Schadendorf [c], Stefan Fröhling [a], Jochen S. Utikal [h,i,1], Christof von Kalle [a,1], Collaborators[2]

[a] *National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 460, 69120 Heidelberg, Germany*
[b] *Department of Dermatology, University Hospital Heidelberg, Heidelberg, Germany*
[c] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[d] *Department of Dermatology, University Hospital Kiel, Kiel, Germany*
[e] *Department of Dermatology, University Hospital Munich (LMU), Munich, Germany*
[f] *Department of Dermatology, University Hospital Würzburg, Würzburg, Germany*
[g] *Department of Dermatology, University Hospital Regensburg, Regensburg, Germany*
[h] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[i] *Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany*

Received 22 December 2018; received in revised form 5 February 2019; accepted 6 February 2019
Available online 8 March 2019

KEYWORDS
Melanoma;
Artificial intelligence;
Diagnostics;
Skin cancer

Abstract ***Background:*** Recent studies have demonstrated the use of convolutional neural networks (CNNs) to classify images of melanoma with accuracies comparable to those achieved by board-certified dermatologists. However, the performance of a CNN exclusively trained with dermoscopic images in a clinical image classification task in direct competition with a large number of dermatologists has not been measured to date. This study compares the performance of a convolutional neuronal network trained with dermoscopic images exclusively for identifying melanoma in clinical photographs with the manual grading of the same images by dermatologists.

* *Corresponding author*: National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 460, 69120 Heidelberg, Germany. fax: +496221 566967.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).
[1] These authors contributed equally to this work.
[2] These collaborators are listed in the acknowledgement section.

https://doi.org/10.1016/j.ejca.2019.02.005
0959-8049/© 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Methods: We compared automatic digital melanoma classification with the performance of 145 dermatologists of 12 German university hospitals. We used methods from enhanced deep learning to train a CNN with 12,378 open-source dermoscopic images. We used 100 clinical images to compare the performance of the CNN to that of the dermatologists.

Dermatologists were compared with the deep neural network in terms of sensitivity, specificity and receiver operating characteristics.

Findings: The mean sensitivity and specificity achieved by the dermatologists with clinical images was 89.4% (range: 55.0%–100%) and 64.4% (range: 22.5%–92.5%). At the same sensitivity, the CNN exhibited a mean specificity of 68.2% (range 47.5%–86.25%). Among the dermatologists, the attendings showed the highest mean sensitivity of 92.8% at a mean specificity of 57.7%. With the same high sensitivity of 92.8%, the CNN had a mean specificity of 61.1%.

Interpretation: For the first time, dermatologist-level image classification was achieved on a clinical image classification task without training on clinical images. The CNN had a smaller variance of results indicating a higher robustness of computer vision compared with human assessment for dermatologic image classification tasks.

© 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Skin cancer is the most common malignancy in Western populations, and melanoma specifically accounts for the majority of skin cancer–related deaths worldwide [1]. Despite special training and the use of dermoscopes, dermatologists only rarely achieve clinical test sensitivities greater than 80% [2]. In 2017, Esteva *et al.* were the first to report a deep-learning convolutional neural network (CNN) image classifier that performed as well as 21 board-certified dermatologists when identifying images with malignant lesions [3]. The CNN was trained on both, clinical and dermoscopic images of skin lesions and generated its own diagnostic criteria for melanoma detection during training. Several follow-up publications by other authors have demonstrated dermatologist-level skin cancer classification by using CNN trained on the same kind of images as a training set that would later comprise the test set [4,5,14]. In this work, we trained a CNN with enhanced techniques on dermoscopic images exclusively but tested its performance on classifying suspect lesions as melanoma or atypical nevi captured as clinical close-up images.

The classification results of the CNN were compared with the efforts of German dermatologists in terms of ROC, sensitivity and specificity.

2. Methods

2.1. Data sets

To develop the algorithm, dermoscopic images from melanomas and atypical nevi were obtained from the International Skin Imaging Collaboration (ISIC) image archive and from the HAM10000 dataset by Tschandl et al. [6,15]. This image archive contained a total of 2169 melanomas and 18,566 atypical nevi as of October 17th, 2018. The diagnoses of all melanomas were verified via histopathological evaluation of biopsies. The diagnosis of nevi was made either by histopathological examination (~24%), expert consensus (~54%) or another diagnosis method, such as a series of images that showed no temporal changes (~22%). All images were anonymous and open source.

To compare the performance of the digital automated diagnosis method without training on clinical images with that of dermatologists on a clinical image classification task, we used the MClass-Benchmark for clinical images, [11] which were originally obtained from the MED-NODE database [12]. The melanomas were biopsy-verified; the nevi were declared as benign via expert consensus. For evaluation of the test set by clinicians, an electronic questionnaire was sent to dermatologists at 12 German university hospitals. Each questionnaire comprised 100 clinical images (80 nevi images and 20 melanoma images, each); the melanomas were biopsy verified, and the nevi were determined as benign via expert consensus. The questionnaire recorded factors such as the years of experience in dermatology, performed skin checks, age, sex and the rank within the university hospital or the status as resident physician. For each image, the dermatologists were asked to provide a management decision (treat/biopsy lesion or reassure the patient). Main outcome measures were sensitivity, specificity and the receiver operating characteristics (ROC).

The training and validation images were also selected using a random generator from the set of available images in the ISIC archive, excluding the already selected test images. The ratio of training and validation data was set as 1:10, and the ratio of the two classes was kept at 1:4. This led to a training set consisting of 1888

melanomas and 10,490 atypical nevi and a validation set including 230 melanomas and 1129 atypical nevi. The test, training and validation sets were disjoint.

2.2. Development of the algorithm

From a mathematical perspective, deep neural networks can be interpreted as functions with millions of freely configurable parameters called weights. These weights are adjusted for a given image classification task in such a way that the intensities of the pixels in an input image are mapped to a probability of class label. Because of the huge number of free parameters, training these functions requires a large number of images for which the class is already known. For each image, the output of the function is calculated compared with the given class label, and then the weights are slightly modified to reduce the error. This process is repeated many times for each image in the training set, and the function 'learns' how to precisely predict the class labels given only the pixel intensities of each image. By using training data that adequately represent the possible input space, the result is a function that exhibits large generality when predicting the class labels for unknown images. In this work, we used CNNs which are characterised by a specific architecture. In regular neural networks, every weight, except for that of the first layer, is affected by the dependencies of all pixels. In contrast, CNNs first aggregate local adjacent pixels to recognise local features and then combine them into global features. This constraint on local connections results in faster training and lower model complexity.

In this work, a ResNet50 CNN model was used for the classification of melanomas and atypical nevi. The network parameters were initialised using the weights from the same network architecture trained to classify images in the ImageNet data set [7]. Details on the training procedure are outlined in Appendix 1.

To document the performance of the algorithm and the enhanced training techniques as accurately as possible, we retrained the CNN a total of 10 times, and each training run consisted of 13 epochs.

2.3. Evaluation of the CNN

The trained CNN output a continuous number between 0 and 1 for each input image, which can be interpreted as the probability that a melanoma was present in the input image. For a binary decision task, it is necessary to specify an operating value that if exceeded causes the input image to be classified as melanoma. This parameter selection allows the trade-off between sensitivity and specificity to be adjusted. Two operating values for the algorithm were selected; the first operating value approximated the mean specificity of 69.2% achieved by chief physicians on the test set, while the second operating value corresponded to a sensitivity of 76.7% for detecting melanomas, which is a necessary prerequisite for the application of the algorithm as a screening tool. This high sensitivity was achieved, on average, by resident physicians on the test set of 100 dermoscopic images. To evaluate the algorithm, the receiver operating curve (ROC) was plotted by varying the operating value between 0 and 1 and calculating the corresponding sensitivity and specificity to the one achieved by the dermatologists.

3. Performance measurement of dermatologists

The creation of our benchmark was described in detail in our recent publication [11].

3.1. Ethics approval

The ethics committee of the University of Heidelberg waived the need for ethical approval since all the dermatologists voluntarily participating in the reader study were anonymous, and the training of an artificial intelligence algorithm was conducted with open source images.

4. Results

4.1. Performance of dermatologists

Detailed results of the dermatologists were described elsewhere [11]. An overview of the experience of the 145 dermatologists is given in Fig. 1 with individual results on the benchmark displayed in Fig. 2. Each point represents the results from one dermatologist specified by the sensitivity and specificity achieved. Results are summed up in Table 1.

4.2. Statistical analysis and performance comparison

The mean receiver operating characteristic (ROC) curve over all 50 runs is shown in Fig. 2 (blue line). It was determined by calculating the average predicted class

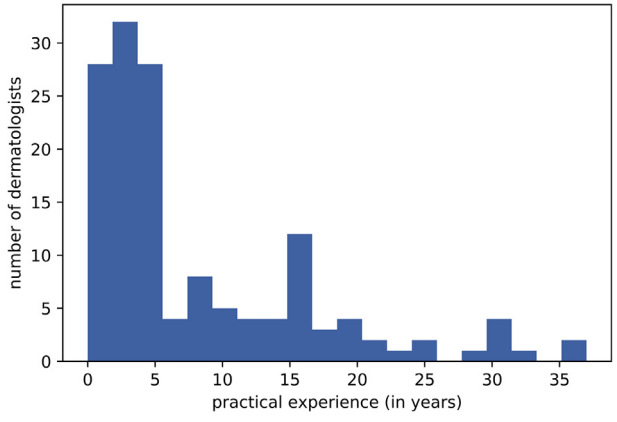

Fig. 1. Distribution of years of experience for participating dermatologists.

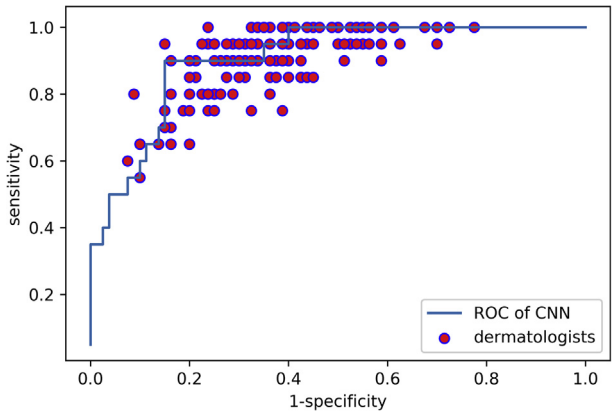

Fig. 2. Receiver operating characteristic (ROC) curve. The blue line represents the mean ROC curve over all 50 test runs. Each point represents the results of one dermatologist for the test set of 100 images. convolutional neural network. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Diagnostic performance of the dermatologists for the test set of 100 clinical images.

	Sensitivity	Specificity	ROC
All participants (n = 145)	89.4%	64.4%	0.769
University hospital (n = 142)	89.4%	62.2%	0.758
Resident physicians (n = 3)	86.7%	73.3%	0.8
Position in hospital hierarchy			
Junior physicians (n = 88)	88.9%	64.7%	0.768
Attendings (n = 16)	92.8%	57.7%	0.753
Senior physicians (n = 35)	89.1%	66.3%	0.777
Chief physicians (n = 3)	91,70%	58.8%	0.753

probability for each test image over all of the 50 runs. The red dots in the figure represent the individual results of the dermatologists, with one dot representing the achieved sensitivity and specificity.

Fig. 3 shows which nevi/melanomas were classified differently by the majority of dermatologists or the CNN. While on the left hand side, the only two melanomas are shown where the two disagreed, and on the right hand side ('nevi'), only a random selection of nevi is shown where disagreement occurred because they were too many.

The mean sensitivity and specificity achieved by the dermatologists with clinical images was 89.4% (range 55.0%–100%) and 64.4% (range 22.5%–92.5%). At the same sensitivity, the CNN exhibited a mean specificity of 68.2% (range 47.5%–86.25%). Among the dermatologists, the attendings showed the highest mean sensitivity of 92.8% at a mean specificity of 57.7%. With the same high sensitivity of 92.8%, the CNN had a mean specificity of 61.1%.

The two operating values of the algorithm, the sensitivity and specificity, were calculated with respect to the class labels documented in the ISIC archive. Using the first operating value at high specificity, approximating the high mean specificity of resident physicians for the test set, the algorithm's mean sensitivity was 86.1%. This value is very close to the resident physicians' corresponding mean sensitivity of 86.7%.

A second operating value for the algorithm was evaluated, based on the high sensitivity of attendings. Using this operating value, the algorithm had a specificity of 61.4%, on average. Compared with the results of the attendings, the corresponding mean specificity of 57.7% was outperformed by the CNN.

5. Discussion

For the first time, a CNN performed on par with the average dermatologist on a clinical image classification task without training on clinical images. The CNN had a smaller variance of results indicating a higher robustness of computer vision compared with human assessment for dermatologic image classification tasks.

Only 19 out of 145 dermatologists achieved a higher sensitivity than the CNN given the individual dermatologist's specificity. In 16 of these 19 cases, the dermatologists achieved an unusually high sensitivity of over 95%.

The rationale of training with dermoscopic images for a clinical classification task was the general idea that state-of-the-art deep-learning algorithms benefit from training with high resolution images even for low-resolution classification tasks. However, the authors were skeptical about the success of the chosen training procedure, because dermatoscopic images not only have just a higher resolution but also show increased visibility of underlying skin layers because of a high-resolution, magnifying camera, which is directly put on the skin of a given patient. Thus, we hypothesised that these patterns might impair training and make the algorithm unprecise for clinical images. With the current state-of-the-art, it is impossible to be able to better understand the decision process of our CNN (i.e. to which extend [a] higher resolution or [b] increased visibility of underlying skin layers for our training set had an impact on the results) [13]. However, this work demonstrates that higher resolution and increased visibility of fine-grained skin patterns in a training set do at least not prevent a CNNs performance on a lower resolution test set.

A digital automated skin diagnosis offers many advantages, including consistent interpretation—because the CNN assigns a distinct class to each specific image every time—and more accurate diagnoses with high sensitivity and specificity. Additionally, by setting the operational value, the trade-off between sensitivity and specificity can be adapted to the requirements of the specific clinical setting. For example, in a screening setting, high sensitivity is desired, so the operating value can be decreased accordingly.

Previous landmark publications comparing the performance of a CNN to dermatologists involved eight,

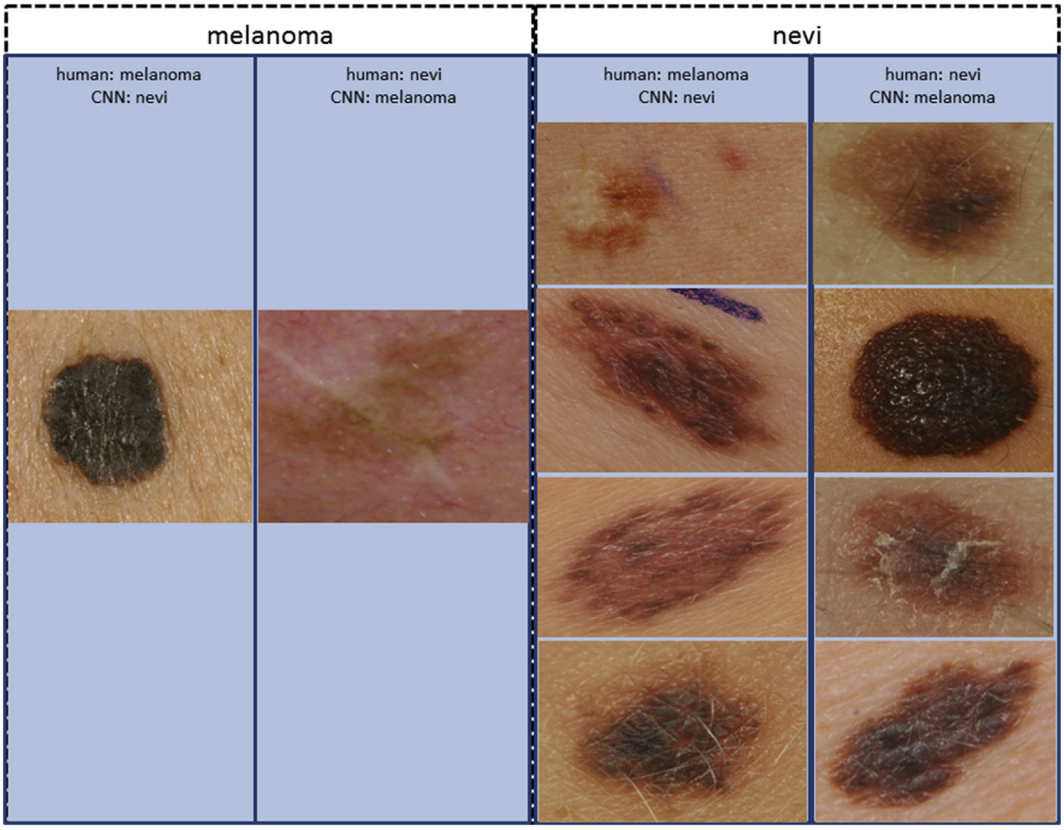

Fig. 3. Different diagnoses of human raters against CNN within our test set. On the left, clinical images of melanoma are shown, and on the right, clinical images of nevi are shown; on the left, the only two melanomas where disagreement occurred are shown; on the left, only a random subselection of nevi is presented (out of 26 nevi images where disagreement occurred in total). CNN, convolutional neural network.

21 or 58 dermatologists [3–5]. This study exceeds these numbers by a magnitude of 145 dermatologists from 12 German university hospitals.

In contrast with previous publications [3–5] that compared the performance of a CNN to that of dermatologists, this was the first stochastic study of this type. We believe that stochastic analysis is absolutely mandatory because the training and evaluation procedure of a CNN includes stochastic components, such as the random splitting of training and validation images, stochastic gradient descent and random initialisation of the parameters. If the test set is known in advance, as in previous studies, it is vital to specify the stochastic characteristics of the results, for example, the first two moments of the distribution and not just a single parameter. Otherwise, there is a risk of adapting the algorithm to the known test set.

When comparing the results of different training runs, it is notable that the quality of the classification differed only slightly. In contrast, the performances of dermatologists showed large variance.

There are some limitations to this system. For instance, clinical encounters with actual patients provide more information than can be provided by images alone. Hänssle *et al.* showed that additional clinical data improve the sensitivity and specificity of dermatologists slightly [5]. Machine learning techniques can also include this information in their decisions. However, even with this slight improvement, the CNN would still perform on par with the dermatologists.

In addition, it needs to be underlined that the training, validation and test set were derived from a mostly fair-skinned population, and the data sets were comprised of images typically seen in clinical practice (randomly obtained). Also, rare melanoma subentities, such as amelanocytic melanomas (=without pigmentation) or lesions on dark-skinned patients, were not present in the sample. Thus, both cases are most likely much harder to detect for our CNN and should be the focus of future research.

6. Conclusion

For the first time, a CNN performed on par with dermatologists on a clinical image classification task without training on clinical images. The CNN had a s-

maller variance of results indicating a higher robustness of computer vision compared with human assessment for dermatologic image classification tasks.

Conflict of interest statement

None declared.

Acknowledgements

The authors would like to thank and acknowledge the dermatologists who actively and voluntarily spend much time to participate in the reader study (claimed to have filled out the anonymous questionnaire with 100 clinical images); some participants did not ask to be mentioned despite their declared participation, and we also thank these colleagues for their commitment. **Berlin (Charité)**: Wiebke Ludwig-Peitsch; **Bonn**: Judith Sirokay; **Erlangen**: Lucie Heinzerling; **Essen**: Magarete Albrecht, Katharina Baratella, Lena Bischof, Eleftheria Chorti, Anna Dith, Christina Drusio, Nina Giese, Emmanouil Gratsias, Klaus Griewank, Sandra Hallasch, Zdenka Hanhart, Saskia Herz, Katja Hohaus, Philipp Jansen, Finja Jockenhöfer, Theodora Kanaki, Sarah Knispel, Katja Leonhard, Anna Martaki, Liliana Matei, Johanna Matull, Alexandra Olischewski, Maximilian Petri, Jan-Malte Placke, Simon Raub, Katrin Salva, Swantje Schlott, Elsa Sody, Nadine Steingrube, Ingo Stoffels, Selma Ugurel, Wiebke Sondermann, Anne Zaremba. **Hamburg**: Christoffer Gebhardt, Nina Booken, Maria Christolouka; **Heidelberg:** Kristina Buder-Bakhaya, Therezia Bokor-Billmann, Alexander Enk, Patrick Gholam, Holger Hänßle, Martin Salzmann, Sarah Schäfer, Knut Schäkel, Timo Schank; **Kiel:** Ann-Sophie Bohne, Sophia Deffaa, Katharina Drerup, Friederike Egberts, Anna-Sophie Erkens, Benjamin Ewald, Sandra Falkvoll, Sascha Gerdes, Viola Harde, Axel Hauschild, Marion Jost, Katja Kosova, Laetitia Messinger, Malte Metzner, Kirsten Morrison, Rogina Motamedi, Anja Pinczker, Anne Rosenthal, Natalie Scheller, Thomas Schwarz, Dora Stölzl, Federieke Thielking, Elena Tomaschewski, Ulrike Wehkamp, Michael Weichenthal, Oliver Wiedow; **Magdeburg:** Claudia Maria Bär, Sophia Bender-Säbelkampf, Marc Horbrügger, Ante Karoglan, Luise Kraas **Mannheim:** Jörg Faulhaber, Cyrill Geraud, Ze Guo, Philipp Koch, Miriam Linke, Nolwenn Maurier, Verena Müller, Benjamin Thomas, Jochen Sven Utikal; **Munich:** Ali Saeed M. Alamri, Andrea Baczako, Carola Berking, Matthias Betke, Carolin Haas, Daniela Hartmann, Markus V. Heppt, Katharina Kilian, Sebastian Krammer, Natalie Lidia Lapczynski, Sebastian Mastnik, Suzan Nasifoglu, Cristel Ruini, Elke Sattler, Max Schlaak, Hans Wolff; **Regensburg:** Birgit Achatz, Astrid Bergbreiter, Konstantin Drexler, Monika Ettinger, Sebastian Haferkamp, Anna Halupczok, Marie Hegemann, Verena Dinauer, Maria Maagk, Marion Mickler, Biance Philipp, Anna Wilm, Constanze Wittmann; **Würzburg:** Anja Gesierich, Valerie Glutsch, Katrin Kahlert, Andreas Kerstan, Bastian Schilling and Philipp Schrüfer.

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Appendix 1

As described in the manuscript, the weights were slightly modified during training to reduce the loss. The loss is mathematically described by a function that models the difference between the class labels predicted by the function for a given parameter setting and actual class labels. The learning rate is a hyperparameter that controls how much these adjustments are made with respect to the gradient of the loss function. In contrast with existing approaches that apply the same learning rate to all layers of the CNN, we used different learning rates for each layer. In particular, slower learning rates were used for layers closer to the input, while faster learning rates were used for layers closer to the output. The intuition behind this enhanced technique, which is called differential learning rates, is that the earlier layers contain more general features, such as edges or gradients. Therefore, their weights do not need to be changed significantly for the new classification task. Thus, the learning rates for the earlier layers are set to low values, resulting in a moderate adjustment of the corresponding weights. In contrast, the later layers contain application-specific features. Consequently, these layers are assigned higher learning rates, which causes the corresponding weights to be modified more in relation to each other compared with the weights of the early layers. To realise this concept, we split the layers into three groups and applied a different learning rate for each group. The first six residual units had a learning rate of 0.009, the subsequent eight residual blocks had a value of 0.003 and the fully connected layers used 0.01. The selection of the specific learning rates was based on practical experience with other image classification tasks.

For each adjustment during training, the parameters normally approach a minimum in the loss function. As the model gets closer to the minimum, it is a common practice to decrease the learning rate stepwise so that the optimisation settles as close as possible to the minimum, instead of overshooting it. In this article, we used a cosine annealing method, which decreases the learning rate based on a cosine function.

The third enhanced training technique addressed the problem that the optimisation process can get stuck in a local, rather than a global, minimum. To overcome this

problem, the learning rate was suddenly increased at some specific time steps, and thus the optimisation process may be able to escape a local minimum and reach the global minimum. This technique is called stochastic gradient descent with restart, an idea shown to be highly effective by Loshchilov *et al.* [8].

References

[1] Schadendorf D, van Akkooi AC, Berking C, Griewank KG, Gutzmer R, Hauschild A, et al. Melanoma. Lancet 2018; 392(10151):971–84.

[2] Carli P, Quercioli E, Sestini S, Stante M, Ricci L, Brunasso G, et al. Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. Br J Dermatol 2003;148(5):981–4.

[3] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;vol. 542(7639):115.

[4] Marchetti MA, Codella NC, Dusza SW, Gutman DA, Helba B, Kalloo A, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018;78(2):270–7. e271.

[5] Haenssle H, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29(8):1836–42.

[6] Gutman D, Codella NC, Celebi E, Helba B, Marchetti M, Mishra N, et al. 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). 2016. *arXiv preprint arXiv:160501397*.

[7] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–52.

[8] Loshchilov I, Hutter F. Stochastic gradient descent with warm restarts. 2016. p. 2–8.

[11] Brinker Titus J, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer 2019;111:30–7.

[12] Giotis I, et al. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst Appl 2015;42(19):6578–85.

[13] Hinton Geoffrey. Deep learning—a technology with the potential to transform health care. Jama 2018;320(11):1101–2.

[14] Brinker Titus Josef, et al. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 2018;20(10).

[15] Tschandl Philipp, Rosendahl Cliff, Kittler Harald. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci data 2018;5: 180161.