



# Robust Selective Classification of Skin Lesions with Asymmetric Costs

Jacob Carse<sup>1</sup> , Tamás Süveges<sup>1</sup> , Stephen Hogg<sup>1</sup>, Emanuele Trucco<sup>1</sup> ,  
Charlotte Proby<sup>2,3</sup> , Colin Fleming<sup>3</sup>, and Stephen McKenna<sup>1</sup>  

<sup>1</sup> CVIP, School of Science and Engineering, University of Dundee, Scotland, UK  
{j.carse,t.suveges,s.c.z.hogg,e.trucco,s.j.z.mckenna}@dundee.ac.uk

<sup>2</sup> School of Medicine, Ninewells Hospital and Medical School, Dundee, UK  
c.proby@dundee.ac.uk

<sup>3</sup> Department of Dermatology, Ninewells Hospital and Medical School, Dundee, UK  
colin.fleming@nhs.scot

**Abstract.** Automated image analysis of skin lesions has potential to improve diagnostic decision making. A clinically useful system should be selective, rejecting images it is ill-equipped to classify, for example because they are of lesion types not represented well in training data. Furthermore, lesion classifiers should support cost-sensitive decision making. We investigate methods for selective, cost-sensitive classification of lesions as benign or malignant using test images of lesion types represented and not represented in training data. We propose EC-SelectiveNet, a modification to SelectiveNet that discards the selection head at test time, making decisions based on expected costs instead. Experiments show that training for full coverage is beneficial even when operating at lower coverage, and that EC-SelectiveNet outperforms standard cross-entropy training, whether or not temperature scaling or Monte Carlo dropout averaging are used, in both symmetric and asymmetric cost settings.

## 1 Introduction

Automated image analysis of skin lesions has great potential to improve diagnostic decision making and efficiency of clinical workflows in dermatology and primary care. Lesion classifiers that produce class probability distributions could be used to estimate the expected costs of clinical decisions such as whether or not to refer a patient, and thus inform effective decision making. Costs associated with mis-classification are usually asymmetric: deciding that a skin lesion is benign when it is really malignant is more costly than deciding it is malignant when it is benign. Optimal decision making requires predicted class probabilities to be well-calibrated. In addition, a clinically useful system should ascertain

---

**Electronic supplementary material** The online version of this chapter ([https://doi.org/10.1007/978-3-030-87735-4\\_11](https://doi.org/10.1007/978-3-030-87735-4_11)) contains supplementary material, which is available to authorized users.

whether it has been sufficiently well trained to deal with the image under inspection. This is important for robustness and clinical usability. Classifiers should be selective, rejecting images they are ill-equipped to deal with; in particular, not all lesion types will be represented well in training data. Here we investigate methods for selective, cost-sensitive skin lesion classification. We focus on binary classification of malignant versus benign lesions using an experimental setup with test data from disease types represented in the training data as well as types not represented in the training data. Images were sourced from the ISIC 2019 data set [3, 4, 23].

We use empirical coverage and selective cost to evaluate performance and stress that selection and classification decisions must take into account asymmetry of mis-classification costs in the diagnostic setting (Sect. 3). We propose a modification to SelectiveNet [9] which we call EC-SelectiveNet (Sect. 4). SelectiveNet learns representations targeting expected image rejection rates by using two additional heads, a selection head and an auxiliary head, in addition to the usual predictive head; EC-SelectiveNet discards these additional heads at test time and makes selection decisions based on expected costs instead.

We provide empirical evidence that training selective networks for full coverage works well on skin lesion images, even when the desired coverage is lowered, somewhat counter to expectation (Sect. 5). We show that EC-SelectiveNet outperforms corresponding cross-entropy trained networks in both asymmetric and symmetric cost settings, whether or not temperature scaling [10] or Monte Carlo dropout averaging [8] are used (Sect. 5).

## 2 Background

AI systems for teliagnosis with performance comparable to human dermatologists have been demonstrated in some settings [1, 7, 11–13, 16], providing evidence that deep learning can, if appropriately designed and integrated, assist diagnostic decision making effectively. However, deep learning models often overfit, resulting in over-confident predictions, and can struggle to decide which lesion images they are equipped to classify reliably [17]. Nevertheless, it has been noted that simply thresholding the maximum softmax response can be effective for rejecting images and reducing mis-classifications [14].

MC-Dropout [8] can be used to quantify uncertainty. It has been used in medical image analysis [18] including estimation of lesion segmentation quality [6] and provision of selection scores for active learning [2]. It uses dropout [15] at inference time, performing  $M$  forward passes of the model  $f$  on an image,  $x$ . Each pass is treated as a sample in a Bayesian approximation of a Gaussian process. Predictions are averaged to give an expected prediction  $\hat{y} = \frac{1}{M} \sum_{m=1}^M f^m(x)$ . Measures of uncertainty such as sample variance can also be calculated.

Temperature scaling can be used to improve calibration of class probabilities predicted by a network [10]. This can be important when making cost-sensitive decisions. Mozafari et al. [19] used temperature scaling with skin lesions and indicated potential hazards when working with noisy validation data. Nixon et

al. [20] investigated calibration metrics and used temperature scaling. Temperature scaling [10] applies a scaling factor  $T$  to the output logits:  $\hat{y}_i = \frac{\exp(\frac{z_i}{T})}{\sum_j \exp(\frac{z_j}{T})}$ . The value of  $T$  is calculated by minimizing calibration error on a validation set.

SelectiveNet [9] jointly learns a classifier and selection function so that the deep representation can be learned with the expectation that some proportion of images should be rejected. We describe this in Sect. 4.

### 3 Robust Selective Classification

Selective classification is performed using a *selection function* and a *prediction function*. The selection function,  $\sigma(x)$ , indicates whether or not an image  $x$  should be rejected, in which case  $\sigma(x) = 0$ , or selected, in which case  $\sigma(x) = 1$ . Given a data set,  $S$ , of  $N$  images, the empirical *coverage*,  $\phi(\sigma|S)$ , is the proportion of images selected for classification, i.e.,  $\phi(\sigma|S) = \frac{1}{N} \sum_{i=1}^N \sigma(x_i)$ . A classification decision is made for each selected image based on the classifier’s prediction function; each such decision incurs a cost. The empirical *selective cost*, is this cost averaged over the selected images.

In general, mis-classification costs can be specified as a matrix  $C$ , where  $C_{jk}$  is the cost of assigning class  $k$  when the true class is  $j$ . These costs depend on the deployment setting and specifically on factors such as health economics, quality of life considerations, and available treatments. Many reported experiments on classification of dermatology images implicitly use  $C = \mathbf{1} - I$  where  $\mathbf{1}$  is a matrix of ones and  $I$  is the identity matrix. This is unrealistic, as costs are in fact far from symmetric. Indeed, many medical classification tasks have highly asymmetric costs.

In this paper, we consider binary classification with class labels *malignant* (class 1) and *benign* (class 0). In a setting where mis-classifying a malignant lesion as benign is an order of magnitude more costly than mis-classifying a benign lesion as malignant, we have  $C_{1,0} = 10.0$ ,  $C_{0,1} = 1.0$ ,  $C_{1,1} = 0.0$ ,  $C_{0,0} = 0.0$ . These values for the asymmetric costs were deemed reasonable through discussion with dermatologists. The values used should however vary depending on the clinical setting, and further work should be done to investigate this in consultation with general practitioners, patient representative groups and health economists. Here we run experiments under different settings with the cost of mis-classifying a benign lesion as malignant,  $C_{1,0}$ , set to 1 (symmetric costs), 10, and 50 (highly asymmetric costs).

We use cost-coverage curves, showing the trade-off between cost incurred and coverage achieved, to characterize the performance of selective classifiers. A strongly performing selective classifier will have low cost and high coverage. As well as benign and malignant lesions from disease types present during training, we also test using images of benign and malignant lesions of disease types not represented in the training data. A robust system should either reject such data or maintain low selective cost on it.

## 4 SelectiveNet and EC-SelectiveNet

Deep representations can be learned specifically for a situation in which some proportion of data are expected to be rejected. SelectiveNet [9] trains a network end-to-end for a specific target coverage. This is enabled by adding two extra heads to the encoder, in addition to the standard predictive head  $p$ : a selective head  $g$  that outputs a selection score, and an auxiliary head  $a$  that outputs predictions used within the loss function. At test time, select/reject decisions are based on the output of the selective head. Here we propose the Expected-Costs SelectiveNet (EC-SelectiveNet) which modifies the network at test time. Specifically, the additional heads are discarded after training and select/reject decisions are based instead on expected costs computed using predicted class probabilities.

### 4.1 SelectiveNet

The SelectiveNet loss function (Eq. 1), is a combination of two functions ( $L_{p,g}$  and  $L_a$ ) weighted with a hyper-parameter  $\alpha$  to control the relative importance of coverage optimization [9]:

$$L = \alpha L_{p,g} + (1 - \alpha)L_a \quad (1)$$

The first term uses predictive and selective heads (Eq. 2) and combines cross-entropy loss,  $l$ , with coverage. It uses hyper-parameter  $t$  as the target coverage for the model and  $\lambda$  to control the importance of this target coverage. The auxiliary head uses a standard cross-entropy loss for  $L_a$ , and is used to encourage the model to learn robust features from the training data.

$$L_{p,g} = \frac{1}{N\phi(g)} \sum_{i=1}^N l(p(x_i), y_i)g(x_i) + \lambda \max(t - \phi(g|S), 0)^2 \quad (2)$$

### 4.2 Selective Classification Based on Expected Costs

Given any trained classifier that outputs a (calibrated) posterior distribution  $P(c|x)$  over classes given an image  $x$ , the expected costs of classification and rejection decisions can be used to decide *whether to select* and *how to classify* the image. Specifically, in the case of two classes,  $c = 0$  (benign) and  $c = 1$  (malignant), the expected cost of deciding benign is  $R_0 = C_{10}P(c = 1|x)$  and the expected cost of deciding malignant is  $R_1 = C_{01}P(c = 0|x)$ . We should decide that  $x$  is in class 1 if  $R_1 < R_0$ , otherwise  $x$  is in class 0. Suppose that by rejecting an image we incur a cost  $\theta$ . An optimal decision rule is then to reject  $x$  if  $\min(R_0, R_1) > \theta$  and otherwise to decide the class with the lower expected cost. Note that a cost coverage plot can be generated by varying  $\theta$ .

### 4.3 EC-SelectiveNet

Although SelectiveNet directly outputs a selection score, we propose to base selection instead on expected costs computed from the predictive head. We refer to this method as EC-SelectiveNet. The selective head is used during training to guide representation learning but, unlike [9], we discard the selective head along with the auxiliary head at test time.

Optionally, we apply temperature scaling to improve calibration to assist reliable estimation of expected costs. Temperature scaling was applied to the logit outputs of the predictive head  $p$ .

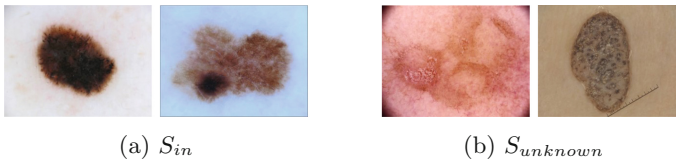
## 5 Experiments

**Dataset and Implementation Setup.** We used data from the ISIC Challenge 2019 [3,4,23] which consists in total of 25,331 images covering 8 classes: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma. We compiled two datasets which we refer to as  $S_{in}$  and  $S_{unknown}$ .

$S_{in}$ : These data were the melanoma, melanocytic nevus and basal cell carcinoma (BCC) images from the ISIC 2019 data. They were assigned to two classes for the purposes of our experiments: *malignant* (melanoma, BCC) and *benign* (melanocytic nevus).  $S_{in}$  was split into training, validation, and test sets consisting of 12432, 3316, and 4972 images respectively.

$S_{unknown}$ : These data consisted of 4,360 ISIC 2019 images from classes that were not present in  $S_{in}$ , namely benign keratosis, dermatofibroma, actinic keratosis, and squamous cell carcinoma. They were assigned to *malignant* or *benign*.  $S_{unknown}$  was not used for training but for testing selective classification performance on images from disease types not represented in the training data.

We refer to the union of the  $S_{in}$  and  $S_{unknown}$  test sets as  $S_{combined}$ . Figure 1 shows example images.



**Fig. 1.** Example images from the test data sets  $S_{in}$  and  $S_{unknown}$ .

All code used for experiments can be downloaded from the project Github repository<sup>1</sup> along with reproduction instructions, trained models and expanded

<sup>1</sup> GitHub Repository: [https://github.com/UoD-CVIP/Selective\\_Dermatology](https://github.com/UoD-CVIP/Selective_Dermatology).

testing metrics. For all experiments we use an EfficientNet [22] encoder with compound coefficient 7, pre-trained on ImageNet [5]. Models were trained using stochastic gradient descent. Cross-entropy loss was used with a two-output softmax. SelectiveNet hyperparameters were  $\alpha = 0.5$  and  $\lambda = 32$  as recommended in [9]. MC-Dropout used a dropout rate of 50% and  $M = 100$  samples. Learning rates were adjusted using a cyclical scheduler [21] that cycled between  $10^{-4}$  and 0.1. Batch size was 8 to enable each batch to fit on our Nvidia RTX 2080TI GPU. Each model was trained for a total of 25 epochs with the weights from the model with the lowest validation loss used for testing.

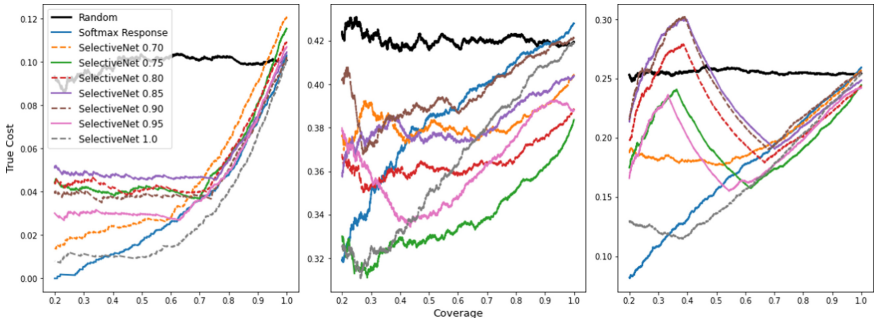
**SelectiveNet: Effect of Target Coverage.** We examined the effect of the SelectiveNet target-coverage parameter,  $t$ , when SelectiveNet’s selection head is used to make selection decisions. Figure 2 shows cost-coverage curves for values of  $t$  ranging from 0.7 to 1.0. These are plotted for  $S_{in}$ ,  $S_{unknown}$ , and  $S_{combined}$ .

We expected to find, in accordance with the intended purpose of this parameter, that lower values of  $t$  would be effective at lower coverage. On the contrary, training with  $t = 1.0$  incurred the lowest test cost on  $S_{in}$  for coverage values as low as 0.2. Costs incurred on  $S_{unknown}$  are higher as expected, and curves show no clear ordering; the  $t = 1.0$  curve, however, does show a clear reduction in cost as coverage is reduced.

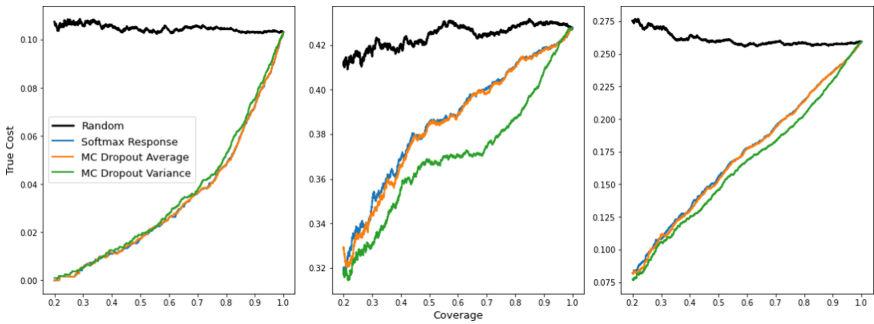
**Does SelectiveNet Training Help?** The extent to which the target coverage  $t$  is enforced is controlled by the weighting parameter  $\lambda$ . Even when set to target full coverage ( $t = 1.0$ ), the model can trade off coverage for cost in extreme cases during training. For this reason, results obtained by SelectiveNet with  $t = 1.0$  will differ from those obtained by training a network without selective and auxiliary heads. We trained such a network using cross-entropy loss, retaining only the softmax predictive head. It made selection decisions at test time based on the maximum softmax output. The resulting cost-coverage curve is plotted in Fig. 2 (labelled ‘softmax’). SelectiveNet trained with a target coverage of 1.0 performed better than a standard CNN with softmax for any coverage above 0.4.

**MC-Dropout, Temperature Scaling, and EC-SelectiveNet.** We investigated the effect of MC-Dropout on selective classification, using the mean and variance of the Monte Carlo iterations as selection scores, respectively. Figure 3 compares the resulting cost-coverage curves with those obtained using a network with no dropout at test time (‘softmax response’). On  $S_{in}$ , using the MC-Dropout average had negligible effect whereas MC variance performed a little worse than simply using the maximum softmax response. In contrast, gains in cost were obtained by MC variance on  $S_{unknown}$  for which model uncertainty should be high.

Figure 4 plots curves for a softmax network using temperature scaling (trained with cross-entropy loss). Although temperature scaling improved calibration it had negligible effect on cost-coverage curves. Figure 4 also shows curves

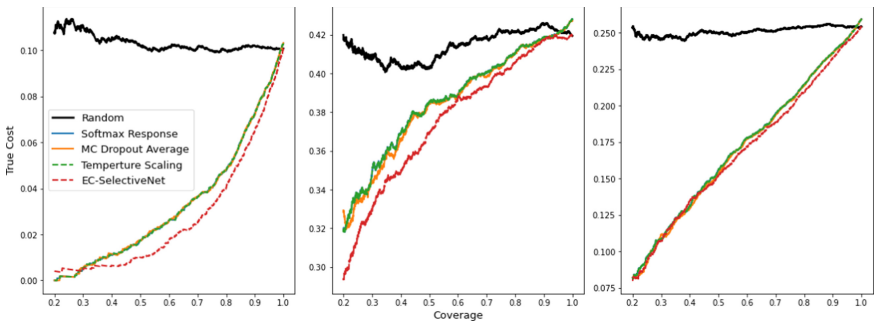


**Fig. 2.** Cost-coverage curves for SelectiveNets trained with different target coverages. From left to right:  $S_{in}$ ,  $S_{unknown}$  and  $S_{combined}$ .



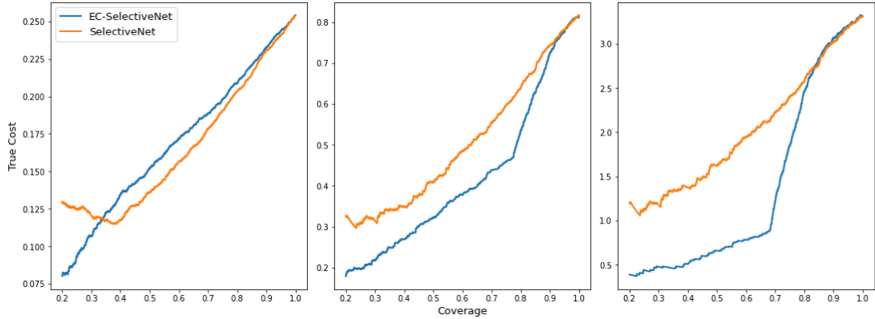
**Fig. 3.** Cost-coverage curves using MC-Dropout on  $S_{in}$ ,  $S_{unknown}$ , and  $S_{combined}$

obtained using EC-SelectiveNet in which the selection head is dropped at test time. EC-SelectiveNet showed a clear benefit on both  $S_{in}$  and  $S_{unknown}$  compared to training a softmax network without the additional heads.



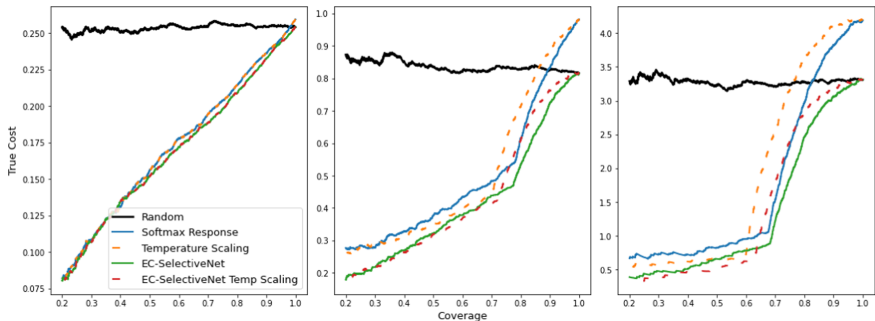
**Fig. 4.** Cost-coverage curves. From left to right:  $S_{in}$ ,  $S_{unknown}$  and  $S_{combined}$ .

**Asymmetric Costs.** We investigated the effect of asymmetric mis-classification costs. Figure 5 compares SelectiveNet with EC-SelectiveNet ( $t = 1.0$ ). They performed similarly when costs were symmetric with SelectiveNet achieving a small cost reduction (approximately 0.015) at middling coverage. However, in the more realistic asymmetric settings, EC-SelectiveNet achieved cost reductions of approximately 0.1 at all coverages below about 0.8.



**Fig. 5.** Cost-coverage curves for SelectiveNet and EC-SelectiveNet. From left to right:  $C_{1,0} = 1$  (symmetric costs), 10, and 50 (highly asymmetric costs)

Figure 6 plots the effect of temperature scaling. Both the softmax response and temperature scaling selection methods are based on the expected costs. The effect of temperature scaling was negligible with symmetric costs. In the asymmetric settings it had a small effect on selective classification. This effect was similar whether using EC-SelectiveNet ( $t = 1.0$ ) or standard network training with cross-entropy loss. In both cases, temperature scaling increased costs at high coverage and reduced costs at low coverage. Figure 6 also makes clear the relative advantage of EC-SelectiveNet.



**Fig. 6.** Cost-coverage curves for cross-entropy training and EC-SelectiveNet combined with temperature scaling. From left to right:  $C_{1,0} = 1$  (symmetric costs), 10, and 50 (highly asymmetric costs)



## 6 Conclusion

This study set out to better understand selective classification of skin lesions using asymmetric costs. In a primary care setting, for example, the cost of misclassifying a life-threatening melanoma is clearly greater than that of misclassifying a benign lesion. We also investigated selective classification with lesion types not adequately represented during training. Generally, EC-SelectiveNet was effective for robust selective classification when trained with a target coverage at (or close to) 1.0. EC-SelectiveNet produced similar or better cost-coverage curves than SelectiveNet.

MC-Dropout averaging made little difference but we note that variance gave encouraging results on  $S_{unknown}$ . Temperature scaling to calibrate output probabilities worsened costs at higher coverage. Future work should investigate use of asymmetric cost matrices in multi-class settings, as well as how so-called out-of-distribution detection methods can help in the context of selective skin lesion classification as investigated here.

**Acknowledgments.** This paper reports independent research funded by the National Institute for Health Research (Artificial Intelligence, Deep learning for effective triaging of skin disease in the NHS, ALAWARD01901) and NHSX. The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health Research, NHSX or the Department of Health and Social Care. This research was also funded by the Detect Cancer Early programme, and the Discovery Institute of Dermatology.

## References

1. Brinker, T.J., et al.: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur. J. Cancer* **113**, 47–54 (2019)
2. Carse, J., McKenna, S.: Active learning for patch-based digital pathology using convolutional neural networks to reduce annotation costs. In: Reyes-Aldasoro, C.C., Janowczyk, A., Veta, M., Bankhead, P., Sirinukunwattana, K. (eds.) *ECDP 2019*. LNCS, vol. 11435, pp. 20–27. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-23937-4\\_3](https://doi.org/10.1007/978-3-030-23937-4_3)
3. Codella, N., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: *IEEE ISBI*, pp. 168–172 (2018)
4. Combalia, M., et al: BCN20000: dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288* (2019)
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255 (2009)
6. DeVries, T., Taylor, G.W.: Leveraging uncertainty estimates for predicting segmentation quality. In: *Conference on Medical Imaging with Deep Learning (MIDL)* (2018)
7. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–8 (2017)

8. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML), vol. PMLR 48, pp. 1050–1059 (2016)
9. Geifman, Y., El-Yaniv, R.: SelectiveNet: a deep neural network with an integrated reject option. In: Proceedings of the 36th International Conference on Machine Learning (ICML), vol. PMLR 97, pp. 2151–2159 (2019)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), vol. PMLR 70, pp. 1321–1330 (2017)
11. Haenssle, H.A., Fink, C., Schneiderbauer, R., 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.* **29**(8), 1836–1842 (2018)
12. Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I., Chang, S.E.: Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J. Inv. Dermatol.* **138**(7), 1529–1538 (2018)
13. Han, S.S., et al.: Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J. Inv. Dermatol.* **140**(9), 1753–1761 (2020)
14. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
15. Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving neural networks by preventing co-adaptation of feature detectors. [arXiv:1207.0580](https://arxiv.org/abs/1207.0580) (2012)
16. Kawahara, J., Hamarneh, G.: Visual diagnosis of dermatological disorders: human and machine performance. [arxiv:1906.01256](https://arxiv.org/abs/1906.01256), 6 (2019)
17. Mårtensson, G., et al.: The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Med. Image Anal.* **66**, 101714 (2020)
18. Mobiny, A., Singh, A., Van Nguyen, H.: Risk-aware machine learning classifier for skin lesion diagnosis. *J. Clin. Med.* **8**(8), 1241 (2019)
19. Mozafari, A.S., Gomes, H.S., Leão, W., Janny, S., Gagné, C.: Attended temperature scaling: a practical approach for calibrating deep neural networks. [arXiv preprint arXiv:1810.11586](https://arxiv.org/abs/1810.11586) (2018)
20. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: CVPR Workshops, vol. 2 (2019)
21. Smith, L.: Cyclical learning rates for training neural networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 464–472. IEEE (2017)
22. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML), vol. PMLR 97, pp. 6105–6114 (2019)
23. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 (2018)