# TransSLC: skin lesion classification in dermatoscopic images using transformers.

## SARKER, M.M.K., MORENO-GARCÍA, C.F., REN, J. and ELYAN, E.

### 2022

# TransSLC: Skin Lesion Classification in Dermatoscopic Images Using Transformers

Md Mostafa Kamal Sarker[1][(✉)] , Carlos Francisco Moreno-García[2] ,
Jinchang Ren[1] , and Eyad Elyan[2]

[1] National Subsea Centre, Robert Gordon University, Aberdeen AB21 0BH, UK
{m.sarker,j.ren}@rgu.ac.uk
[2] School of Computing Science and Digital Media, Robert Gordon University,
Aberdeen AB10 7GJ, UK
{c.moreno-garcia,e.elyan}@rgu.ac.uk

**Abstract.** Early diagnosis and treatment of skin cancer can reduce patients' fatality rates significantly. In the area of computer-aided diagnosis (CAD), the Convolutional Neural Network (CNN) has been widely used for image classification, segmentation, and recognition. However, the accurate classification of skin lesions using CNN-based models is still challenging, given the inconsistent shape of lesion areas (leading to intra-class variance) and inter-class similarities. In addition, CNN-based models with massive downsampling operations often result in loss of local feature attributes from the dermatoscopic images. Recently, transformer-based models have been able to tackle this problem by exploiting both local and global characteristics, employing self-attention processes, and learning expressive long-range representations. Motivated by the superior performance of these methods, in this paper we present a transformer-based model for skin lesion classification. We apply a transformers-based model using bidirectional encoder representation from the dermatoscopic image to perform the classification task. Extensive experiments were carried out using the public dataset HAM10000, and promising results of 90.22%, 99.54%, 94.05%, and 96.28% in accuracy, precision, recall, and F1 score respectively, were achieved. This opens new research directions towards further exploration of transformers-based methods to solve some of the key challenging problems in medical image classification, namely generalisation to samples from a different distribution.

**Keywords:** Computer aided diagnosis · Skin lesion classification · Deep learning · Convolutional neural networks · Transformers

## 1 Introduction

Skin cancer is the most common type of cancer worldwide, responsible for $64,000$ fatalities in 2020 [16]. The majority of skin cancers can be treated if diagnosed

early. However, visual inspection of skin malignancies with the human eye during a health screening is prone to diagnostic errors, given the similarity between skin lesions and normal tissues [12]. Dermatoscopy is the most reliable imaging method for screening skin lesions in practice. This is a non-invasive technology that allows the dermatologist to acquire high-resolution images of the skin for better visualisation of the lesions, while also enhancing sensitivity (i.e. accurate identification of the cancer lesions) and specificity (correct classification of non-cancerous suspicious lesions) when compared with the visual inspection. Nonetheless, dermatologists still confront hurdles in improving skin cancer detection, since manual assessment of dermatoscopic images is often complicated, error-prone, time-consuming, and subjective (i.e., may lead to incorrect diagnostic outcomes) [12]. Thus, a computer-aided diagnostic (CAD) system for skin lesion classification that is both automated and trustworthy has become an important evaluation tool to support dermatologists with proper diagnosis outcomes to finalise their decisions.

Over the last decades, several Convolutional Neural Network (CNN) based methods have been presented, delivering better CAD systems that identify the melanoma and non-melanoma skin lesions accurately. Deep neural networks are being used to classify skin cancer at the dermatological level. Examples include [9] using GoogleNet's Inception v3 model, which achieved 72.1% and 55.4% accuracy of the three and nine class respectively, on a Stanford Hospital private dataset. In [22], a fully convolutional residual network (FCRN) is proposed and evaluated on the IEEE International Symposium on Biomedical Imaging (ISBI) 2016 *Skin Lesion Analysis Towards Melanoma Detection Challenge* dataset. This model obtained the $1^{st}$ place on the challenge leaderboard, yielding an accuracy of 85.5%. Moreover, an attention residual learning convolutional neural network (ARL-CNN) was introduced by [23] and evaluated on the ISBI 2017 dataset, achieving an average area-under-curve (AUC) of 91.7%.

Ensemble-based CNN models have also shown superior performance on medical image analysis [5,6] and skin lesion segmentation [15] and classification, as shown in the International Skin Imaging Collaboration (ISIC) datasets 2018 [3], 2019 [10], and the HAM10000 dataset [1]. However, these methods require training several deep learning models to create the ensemble, which requires huge computing power and is not suitable for real-time applications. In summary, it can be said that most methods used for medical image classification, including lesion classification are based on CNN models. However, it was reported that while such model's perform very well on datasets, cross-datasets generalisation is still considered as a key challenge for the computer vision research community [8].

To this end, we aim to address some of the issues above using a single deep learning model to classify skin lesions accurately. We propose the development of vision transformers-based models, as these have proven to be outperforming many image classification tasks [7,14]. In this study, we use a bidirectional encoder representation from the image transformers model to correctly diagnose the skin lesion. The rest of this article is organised as follows. Section 2
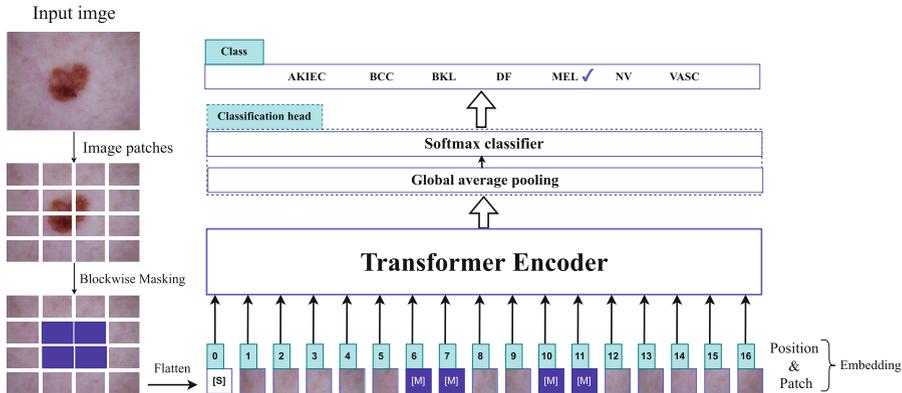
**Fig. 1.** The architecture of the proposed transformers model, *TransSLC*, input image, image patches. A special mask embedding [M] is replaced for some random mask of image patches (blue patches in the figure). Then the patches are fed to a backbone vision transformer and classify. (Color figure online)

describes the materials and the bidirectional encoder representation from the image transformers model in detail. The experimental findings of the CNN and transformer-based models are compared and examined in Sect. 3. Finally, Sect. 4 draws the research conclusions and suggests some future directions.

## 2  Methods and Materials

### 2.1  Image Transformer

In this work, we propose a bidirectional encoder representation from image transformers motivated by BEIT [2]. Figure 1 provides a schematic diagram of the proposed method.Initially, the input skin lesion $224 \times 224$ image is split into an array of 16 image patches, with each patch measuring $14 \times 14$ pixels, as shown in the top-left corner of Fig. 1. In BEIT, a masked image modelling (MIM) task to pretrain vision transformers is proposed for creating the visual representation of the input patches. Therefore, we used a block-wise masking forward by a linearly flatten projection to get the patch embeddings. A special token [S] is added to the input sequence for regularisation purposes. Furthermore, the patch embeddings include standard learnable 1D position embeddings as well. The input vectors of each embeddings are fed into transformers encoder. We then use vision transformers encoder as a backbone network of our model. The encoded representations for the image patches are the output vectors of the final layer of the transformers, which are then fed into the classification head which in turn classifies the input skin lesion image. The classification head consists of two layers: a global average pooling (used to aggregate the representations) and a softmax-based output layer that produces the classification of the distinct the categories.

## 2.2 Model Implementation Setup

As mentioned in the previous section, the proposed *TransSLC* model design is based on the BEIT model presented in [2]. In practice, we utilise a 12-layer transformer encoder, with 768 hidden size and 12 attention heads. A total of 307 feed-forward networks were also implemented for the intermediate size of the network. For our experiment, the input skin lesion image size is set to $224 \times 224$ resolution, with the $14 \times 14$ array of patches having some patches randomly masked. We trained our proposed model for 50 epochs, using the Adam optimiser [13] with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate was set to 0.0001, and a batch size of 8 was used. To ensure a fair comparison with other CNN-based methods, we have used the same experimental settings. Experiments were carried out using Nvidia Tesla T4 16 GB Graphics Processing Unit (GPU) cards, and running the experiment for 50 epochs for all the models below took on average 24 h of training time.

## 2.3 Model Evaluation

Standard evaluation metrics were used to evaluate the performance of the models used in the experiments. These are accuracy, precision, recall, and F1 score. Definitions of these metrics are presented in Table 1.

**Table 1.** Model evaluation metrics to evaluate the models.

| Metric | Formula |
|---|---|
| Accuracy (AC) | (TP+TN)/(TP+TN+FP+FN) |
| Precision (PR) | TP/(TP+FP) |
| Recall (RE) | TN/(TN+FN) |
| F1 Score (F1) | 2.TP/(2.TP+FP+FN) |

TP = True Positives,    TN = True Negatives,
FP = False Positives,    FN = False Negatives.

## 2.4 Dataset

The public and commonly used HAM10000 dataset was used [1] for evaluation purposes. The dataset contains 10,015 images. These images are labelled based on a discrete set of classes representing seven categories: actinic keratoses and intraepithelial carcinoma (AKIEC), basal cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevus (NV), and vascular lesions (VASC). As can be seen in Table 2 the samples distributions is imbalanced. In other words, the number of training images in NV class is 4693 whereas DF and VASC classes have only 80 and 99 images, respectively. This is a common problem in most medical datasets, as well as health-related data [20] where various data sampling methods, as well as algorithmic modifications, are

**Table 2.** The image distribution per class and splits of the HAM10000 dataset.

| Splits | AKIEC | BCC | BKL | DF | MEL | NV | VASC |
|---|---|---|---|---|---|---|---|
| Training | 228 | 359 | 769 | 80 | 779 | 4693 | 99 |
| Validation | 33 | 52 | 110 | 12 | 111 | 671 | 14 |
| Testing | 66 | 103 | 220 | 23 | 223 | 1341 | 29 |
| Total (10,015) | 327 | 514 | 1099 | 115 | 1113 | 6705 | 142 |

employed to handle it [21]. However, for the purpose of this paper, we handled this problem using a simple data augmentation technique. This includes flipping the images horizontally and vertically, random cropping, adaptive histogram equalisation (CLAHE) with varying values for the original RGB images is used to change the contrast. To generate a range of contrast images, we set the CLAHE threshold for the contrast limit between 1.00 and 2.00

## 3 Experimental Results

For comparison purposes with our proposed *TransSLC* model, we have selected several state-of-the-art models, including ResNet-101 [11], Inception-V3 [18], the hybrid Inception-ResNet-V2 [17], Xception [4] and EfficientNet-B7 [19]. These models are considered state-of-the-art, and commonly used in medical image analysis. As can be seen in Table 3, *TransSLC* achieved the top performance reaching an accuracy of 90.22%, precision of 85.33%, recall of 80.62%, and F1 score of 82.53%. It can also be seen that among the selected CNN-based models, EfficientNet-B7 [19] achieved the best results with accuracy of 88.18%, precision of 83.66%, recall of 78.64%, and F1 score of 80.67%, respectively. Thus, our proposed model improves 2.04%, 1.67%, 1.98%, and 1.86% in terms of accuracy, precision, recall, and F1 score, respectively, comparing with CNN-based EfficientNet-B7 [19] model.

**Table 3.** Comparison of the performance (%) of the proposed transformers-based model against different CNN-based models in terms of the accuracy (AC), precision (PR), recall (RE), and F1 score (F1), respectively, on the test dataset.

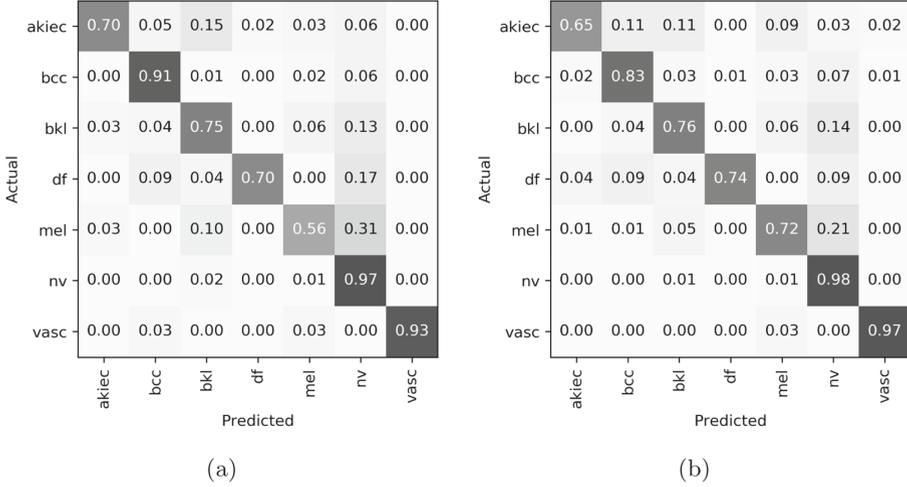| Methods | AC | PR | RE | F1 |
|---|---|---|---|---|
| *CNN-based* | | | | |
| ResNet-101 [11] | 83.04 | 68.86 | 68.06 | 68.06 |
| Inception-V3 [18] | 86.48 | 75.19 | 77.02 | 75.66 |
| Inception-ResNet-V2 [17] | 86.68 | 79.78 | 73.56 | 76.29 |
| Xception [4] | 86.98 | 79.55 | 74.08 | 76.07 |
| EfficientNet-B7 [19] | 88.18 | 83.66 | 78.64 | 80.67 |
| *Transformers-based* | | | | |
| **Proposed *TransSLC*** | **90.22** | **85.33** | **80.62** | **82.53** |

**Fig. 2.** Confusion matrix of (a) CNN-based EfficientNet-B7 Model (b) Transformers-based proposed model (*TransSLC*).

Moreover, Fig. 2 shows a confusion matrix of the 7 classes of the HAM10000 dataset with the test dataset. The confusion matrix in Fig. 2 shows (a) the EfficientNet-B7 [19] with the test dataset has some miss classification, particularly in the MEL types, and (b) that the proposed model, *TransSLC*, with the test dataset, is able to classify the skin lesion types in most of the classes. The CNN-based EfficientNet-B7 [19] model performs well in detecting AKIEC, BCC types of a skin lesion with 5%, and 8% higher than our proposed *TransSLC* model. To classify BKL, DF, MEL, NV, and VASC types of the lesions, the EfficientNet-B7 [19] model performs poorly and significantly fails in MEL types with 15% lower than our proposed model. This is a crucial flaw, as MEL types are deadly for patients. Therefore, CNN-based models have a considerable some limitations when used in real-world clinical settings. In contrast, our proposed model is capable of overcoming this limitation and could potentially be deployed in a real clinical setting. Still, *TransSLC* has some limitations when classifying MEL types, getting this class confused with 1% of AKEIEC, 1% of BCC, and 5% of BKL, 25% of NV types, respectively. Another drawback of the proposed transformers-based model consists of huge number of the parameters which requires large memory (computational capacity) in order to implement.

Figure 3 illustrates the comparison between CNN-based EfficientNet-B7 and proposed model using Receiver Operating Characteristic (ROC) curve. The EfficientNet-B7 yields the area of AKIEC class is 98% which is 2% higher than proposed model. The area of the rest of classes, DF, MEL, NV, and VASC achieved by *TransSLC* improves 1%, 2%, 2%, and 2%, respectively, compared with the EfficientNet-B7 model. The remain area of the BCC and BKL classes are the same for both EfficientNet-B7 and our proposed model. The class-wise
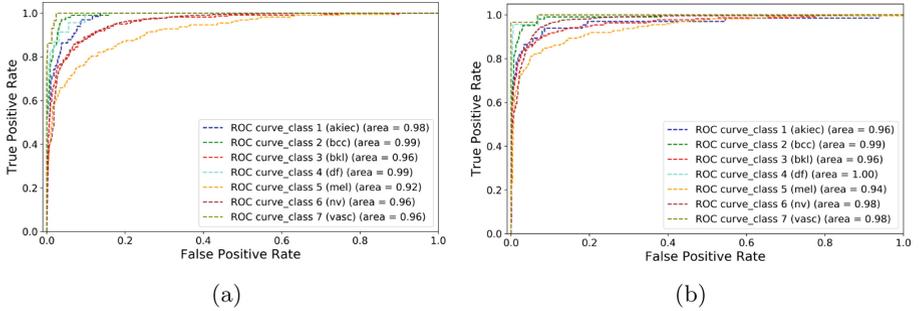
**Fig. 3.** ROC curve (receiver operating characteristic curve) of (a) CNN-based EfficientNet-B7 Model (b) Transformers-based proposed *TransSLC* model.

performance metrics of the proposed transformers-based, *TransSLC*, model is presented in Table 4. The proposed model yields the 86.00%, 78.90%, 84.77%, 89.47%, 79.60%, 93.70%, and 84.8% of accuracy to classify AKIEC, BCC, BKL, DF, MEL, NV, and VASC, respectively.

The performance analysis of several ablation experiments is likely insufficient to assess the benefits and behaviour of the proposed model. Thus, Fig. 4 we depict the activation maps of the CNN-based and transformers-based model. Notice that the EfficientNet-B7 rows show the activation maps, where the model can classify all these images correctly to the corresponding class but activated in over-all regions of the input skin lesion images. More preciously, the skin lesion types can be conformed through some lesion areas only on the dermatoscopic image. The activation maps by the proposed transformers-based *TransSLC* model can remarkably overlay with only the lesion regions, which could signify the presence of lesion type. Finally, we can infer that a transformers-based model would distinguish between important and non-relevant characteristics of skin lesion, as well as learning the appropriate features for each given class.

**Table 4.** The class-wise performance metrics of the proposed transformers-based, *TransSLC*, model for the seven classes of skin lesion classification in terms of the precision (PR), recall (RE) and F1 score (F1), respectively.

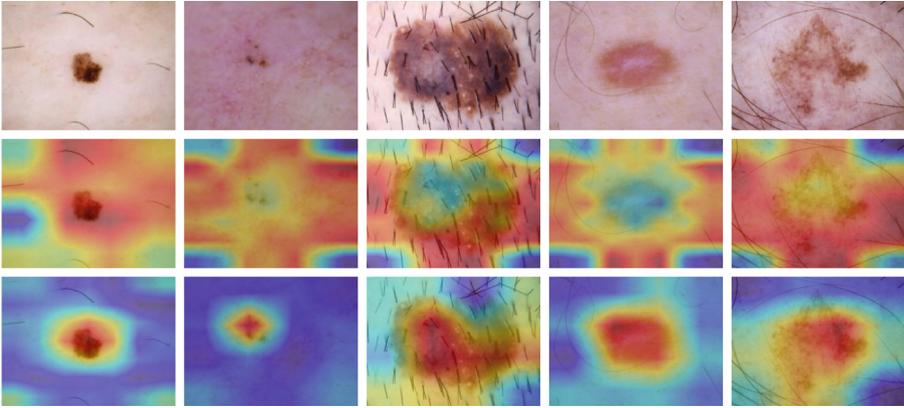| Class type | PR | RE | F1 |
| --- | --- | --- | --- |
| Class 1 (akiec) | 86.00 | 65.15 | 74.14 |
| Class 2 (bcc) | 78.90 | 83.50 | 81.13 |
| Class 3 (bkl) | 84.77 | 75.91 | 80.10 |
| Class 4 (df) | 89.47 | 73.91 | 80.95 |
| Class 5 (mel) | 79.60 | 71.75 | 75.47 |
| Class 6 (nv) | 93.70 | 97.54 | 95.58 |
| Class 6 (vasc) | 84.85 | 96.55 | 90.32 |

**Fig. 4.** Visualisation results of the activation maps. For every column, we show an input image, the corresponding activation maps from the outputs of EfficientNet-B7 and the proposed *TransSLC* model.

## 4 Conclusion

In this paper, we presented *TransSLC*, a transformers-based model able to classify seven types of skin lesions. The proposed method was compared with five popular state-of-the-art CNN-based deep learning models Using the HAM10000 public datasets. Our proposed model achieved the accuracy of 90.22%, precision of 85.33%, recall of 80.62%, and 85.53%, respectively on the test dataset. The proposed model shows the transformers-based model outperforms the traditional CNN-based model to classify different types of skin lesions which can enable new research in this domain. Future work will further explore transformers-based methods performances across other datasets, as well as carrying out cross-datasets evaluation to assess how well the model generalises.

## References

1. Bansal, Nidhi, Sridhar, S..: Skin lesion classification using ensemble transfer learning. In: Chen, Joy Iong-Zong., Tavares, João Manuel R. S.., Iliyasu, Abdullah M.., Du, Ke-Lin. (eds.) ICIPCN 2021. LNNS, vol. 300, pp. 557–566. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-84760-9_47
2. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Bissoto, A., Perez, F., Ribeiro, V., Fornaciali, M., Avila, S., Valle, E.: Deep-learning ensembles for skin-lesion segmentation, analysis, classification: Recod titans at ISIC challenge 2018. arXiv preprint arXiv:1808.08480 (2018)

4. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258 (2017)

5. Dang, T., Nguyen, T.T., McCall, J., Elyan, E., Moreno-García, C.F.: Two layer Ensemble of Deep Learning Models for Medical Image Segmentation. ArXiv (2021). http://arxiv.org/abs/2104.04809

6. Dang, T., Nguyen, T.T., Moreno-García, C.F., Elyan, E., McCall, J.: Weighted ensemble of deep learning models based on comprehensive learning particle swarm optimization for medical image segmentation. In: IEEE Congress on Evolutionary Computing, pp. 744–751. IEEE (2021)

7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

8. Elyan, E., et al.: Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward. Artificial Intelligence Surgery (2022). https://doi.org/10.20517/ais.2021.15

9. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017)

10. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A.: Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. MethodsX **7**, 100864 (2020)

11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

12. Jones, O., et al.: Dermoscopy for melanoma detection and triage in primary care: a systematic review. BMJ Open **9**(8), e027529 (2019)

13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)

15. Sarker, M.M.K., et al.: Slsnet: skin lesion segmentation using a lightweight generative adversarial network. Expert Syst. Appl. **183**, 115433 (2021)

16. Sung, H., et al.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Can. J. Clin. **71**(3), 209–249 (2021)

17. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI Conference on Artificial Intelligence (2017)

18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)

19. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)

20. Vuttipittayamongkol, P., Elyan, E.: Overlap-based undersampling method for classification of imbalanced medical datasets. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds.) Artificial Intelligence Applications and Innovations, pp. 358–369. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_36

21. Vuttipittayamongkol, P., Elyan, E., Petrovski, A.: On the class overlap problem in imbalanced data classification. Knowl.-Based Syst. **212**, 106631 (2021)

22. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans. Med. Imaging **36**, 994–1004 (2017)
23. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Attention residual learning for skin lesion classification. IEEE Trans. Med. Imaging **38**(9), 2092–2103 (2019)