A Modified Inception-v4 For Imbalanced Skin Cancer Classification Dataset

Taha Emara[1], Heba M. Afify[1,2], Fatma Helmy Ismail[1,3], Aboul Ella Hassanien[1,4]

[1]Scientific Research Group in Egypt.

[2]Electrical Engineering Department, Egyptian Academy for Engineering and Advanced Technology, Egypt. hebaaffify@yahoo.com

[3]Faculty of Computer Science, Misr International University, Egypt. fatma.helmy@miuegypt.edu.eg

[4]Faculty of Computers and Artificial intelligence ,Cairo University, Egypt. aboitcairo@gmail.com

***Abstract*—Deep learning architectures, especially deep convolutional neural networks (CNN) achieve high accuracy on object classification and localization tasks. Achieving such high accuracy requires powerful devices. In this paper, rather than an ensemble of multiple complex models, a single Inception-v4 model is adapted to classify extracted from the HAM10000 dataset. The proposed model is enhanced by employing feature reuse using long residual connection in which the features extracted from earlier layers are concatenated with the high-level layers to increase the model classification performance. The dataset used in this study is imbalanced; therefore, a data sampling approach is used to mitigate the data imbalance effect. The proposed architecture achieves an accuracy of 94.7% using the provided test set at the official benchmark for the International Skin Imaging Collaboration (ISIC) 2018.**

***Keywords*—deep learning, convolution neural network , inception v4, skin cancer classification.**

I. Introduction

Skin cancer is considered one of the most deadly dermatological diseases and is caused by unregulated cell growth on the skin surface [1]. There are two major classes of skin cancer, melanoma skin cancer and non-melanoma skin cancer. According to skin cancer statistics, melanoma skin cancer is a skin malignant growth that can result in death. Generally, several factors raise the risk of skin cancer including alcoholic drinks, overweight, sunburns, radiation exposure, family history, weakened immune systems, and genetic disorders [2]. Therefore, skin cancer diagnosis are based on skin lesion imaging techniques such as dermoscopy examination which dermatologists use to evaluate skin pigmentation in terms of colour and texture because it avoids surface reflections on the skin and allows access to deeper layers [3]. The accuracy of manual dermoscopic images diagnosis by dermatologists is approximately 60 % [4]. However, manual diagnosis suffers from issues such as low skin images contrast, the high similarity of visual images and wide variations in skin conditions [5]. Thus, computer-based diagnosis of dermoscopic images can be useful for improving the accurate recognition rate of skin lesions [6]. Early detection of skin cancer is critical to aid dermatologists in making valid decisions for skin treatments and leads to decreased costs for the treatment of advanced stages of diseases while reducing mortality rates from melanomas. Dermoscopic algorithms [7] [8] [9] like chaos & clues, 3-point checklist, ABCD (asymmetry, border, colour, diameter) rule, and CASH (colour, architecture, symmetry, homogeneity) have been proposed to simplify the detection of different classes of skin cancer.

II. Related work

Many approaches have been proposed that use traditional machine learning algorithms and deep learning models to address the problem of skin cancer discrimination [10]. Traditional machine learning consists of sequential steps including preprocessing, segmentation, feature extraction, and classification [11]. Traditional machine learning is applied to skin images by using algorithms such as the K-nearest neighbours (KNN) [12], support vector machines (SVM) [13] and fuzzy systems [14] which can achieve good classification. However, there are some limitations to using traditional machine learning including information loss, errors, excessive time consumption and poor segmentation results. Therefore, there is a demand to improve the classification accuracy by applying deep learning architecture to address the different resolutions of images. In [15], a convolutional neural network (CNN) architecture was proposed to detect melanoma and benign images that were divided into 136 training and 34 testing images. This model achieved an accuracy of 81% . Matsunaga et. al. [16] implemented a CNN that operated on skin images to classify three types of skin cancer. This model used data augmentation and the Keras library and won first place in the 2017 International Symposium on Biomedical Imaging (ISBI) competition. Brinker et.al. [17] reviewed previous research approaches on distinguishing skin lesions using a CNN. Seog Han et. al. [18]presented a CNN technique to classify 12 lesion types extracted from two Korean hospitals. Kawahara et. al. [19] described a CNN model able to accept different image resolutions as an input that achieved with an accuracy of 79.5% on the public Dermofit Image Library. Menegola et. al.[20] proposed skin cancer classes using the 2017 International Skin Imaging Collaboration (ISIC) which was used as a skin image archive. Recently, another study demonstrated a CNN model created to classify skin lesions into benign or malignant lesions using a regularizer algorithm. This model achieves an accuracy of 97.49% [21]. Refianti et. al. [22] explored CNN and LeNet-5 architectures to distinguish skin cancer diseases and reported that the size of training data and the number of epochs used during the training have an impact on the taxonomy accuracy. In addition, it was reported that performing a comparative analysis of various classification algorithms for skin lesion images is complex because of the use of non-public datasets

978-1-7281-5260-8/19/$31.00 ©2019 IEEE

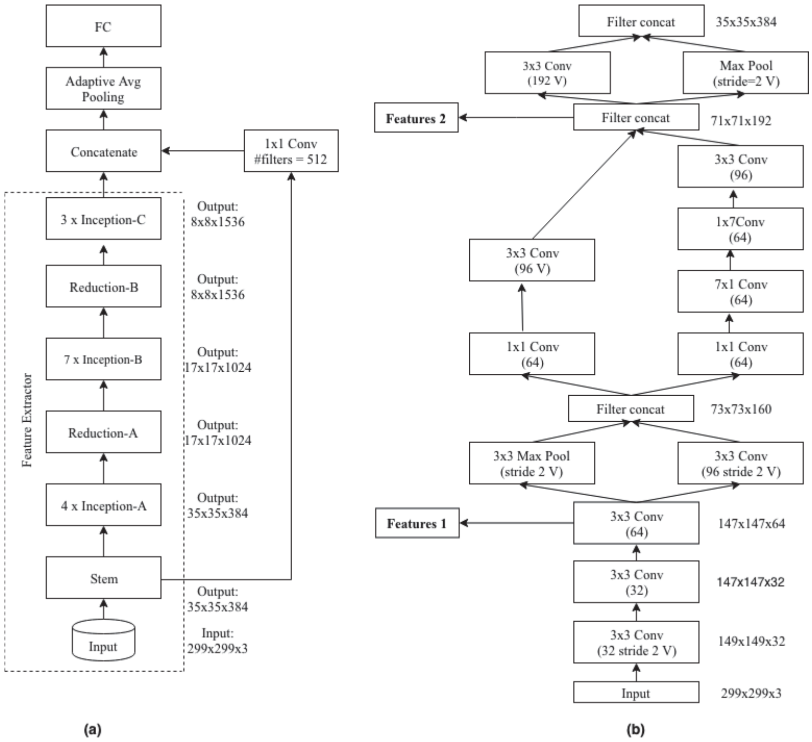


Fig. 1: (a) Proposed modification to Inception-v4 architecture; (b) The schema for the stem module of Inception-v4 architecture, Features 1 are the features to be concatenated with the output features from Inception-v4 after applying 1×1 convolution. The concatenation of Features 2 is also tested.

in research papers. Additionally, the Inception V3 model was presented to classify skin cancer images achieving an accuracy of 85% using a standard ISIC dataset[23]. Gessert et al. [24] and Milton et al. [25] applied an ensemble of many CNN architectures rather than using a single architecture to achieve high accuracy on the ISIC 2018 challenge dataset. From their work, it is found that employing ensemble of multiple CNN models improves the accuracy but with a large sacrifice in computational efficiency as shown in Section IV-F.

III. System Architecture

Deep convolutional networks have been the key to most of the recent advances in image recognition, object detection, and semantic segmentation tasks. The inception architecture is known as one of the deep convolutional models which has high abilities with low computational costs. The original Inception network was one of the most famous learning models for image recognition [26].

A. *Transfer learning*

Learning algorithms can exploit the concept of transfer learning to exchange knowledge between tasks. Deep learning methods are time consuming and need large training datasets. In literature, the existing CNN models working on large datasets can be be replace by other recognition tasks [24, 25]. In our work we use this technique with a modified Inception-V4 architecture trained on ImageNet dataset. And also, the last classifier layer is replaced with a classifier to discriminate the 7 skin cancer diseases instead of the 1000 classes of the ImageNet dataset.

B. *Modified Inception-v4*

We extend the design of inception-v4 but add a residual connection to fuse low-level feature to high level feature which will improve the accuracy of the proposed architecture. The modified inception-v4 architecture is shown in figure 1 and consists of the following stages:

1) **Input**: the input dataset is described in the next section.
2) **Stem**: the stem is used to preprocess the data before it enters the Inception module.
3) **Feature reuse**: This stage is the main contribution of the proposed model. Because the earlier layers of the CNN generate general features (low level features) as colour blobs or edges, reusing these features in combination with high level features can improve model classification accuracy. As the classification result may depend on low level feature such as affected skin colour or the lesion shape, these features may be ignored by later layers. Therefore, including them may improve the classification accuracy [27]. In the proposed model, the output features from the third convolution in the stem module are concatenated with the extracted features from the Inception-v4 model after applying a 1x1 convolution with 512 filters.
4) **Inception layers**: These layers allow the internal

layers to pick and choose which filter size is most relevant to learning the required information.

5) **Reduction blocks**: The Reduction modules between the three Inception modules act as pooling layers.

IV. Experimental results

A. Dataset description

In this paper, 10,015 dermatoscopic microscopy images were extracted from the Human Against Machine with 10000 Training Images (HAM10000) dataset [28]. These images include seven disease classes as shown in figure 2. Each microscopy image is an RGB colour image with a size of 450×600. The dataset was divided into two parts, 90% and 10% used for training and validation, respectively. The dataset images were resized to 299×299 to be compatible with the Inception-v4 architecture. Being this dataset an unbalanced dataset as shown in table I makes the model biased to the classes with high number of samples. To mitigate the unbalance in a given dataset, a data sampling approach can be used such that, some instances of the majority classes are removed (under-sampling) or some traditional instances of minority classes are supplied (over-sampling) to balance the class sizes in a given dataset [29]. Although these methods solve the problem of unbalanced data in a dataset, they have some drawbacks: for example, over-sampling causes the learning algorithm to overfit because of sample duplication in and under-sampling causes information loss, while, in addition, both techniques require extra data analysis and preprocessing. Instead, weighted random sampler method is applied which is officially implemented in Pytorch [30], in which a weight is defined for every instance by the multiplicative inverse of the number of instances in its class. This weight represents the probability of that this instance will be randomly drawn. Thus, it acts to offset the action of over-sampling for classes with smaller number of samples and under-sampling for classes with greater number of samples.

TABLE I: Distribution of seven classes for the HAM10000 images dataset

Pigmented Lesions	number of images
Melanoma	1113
Melanocytic Nevus	6705
Basal cell carcinoma	514
Actinic keratosis	327
Benign keratosis	1099
Dermatofibroma	115
Vascular lesion	142

B. Computing Environment

The experiments were done on a machine equipped with an Intel Core i7-8700 @ 3.2GHZ, 16GB of memory, and an NVIDIA GTX1080Ti GPU card. Our machine runs Ubuntu 18.04 and PyTorch [30] version 0.4.1 with CUDA 9.0 and cudnn 7.0.5.

TABLE II: Performance of Inception-v4 model with different configurations for the convolution layer which used for concatenating low level feature with the extracted features, on the validation set

Model Options	Model Accuracy
Baseline	82.89%
Features 1 $[1 \times 1, 1024]$	80%
Features 1 $[1 \times 1, 512]$	86.17%
Features 1 $[3 \times 3, 512]$	80.1%
Features 1 along with Features 2	72%

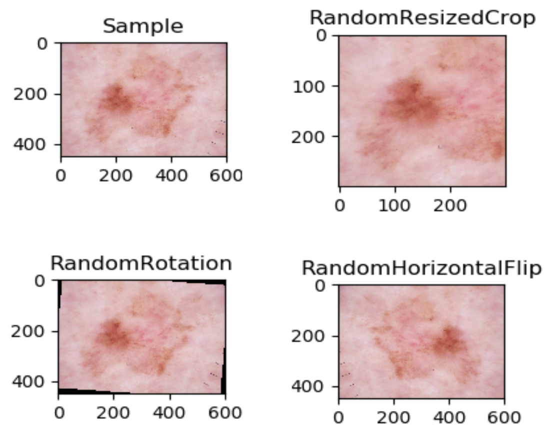


Fig. 3: Visualization of different data augmentation techniques

C. Training Protocol

Stochastic gradient Descent was used with an initial learning rate of 10^{-3} and with momentum value of 0.9. Multiple learning rate policies were applied in which the learning rate changes after every seven epochs. The learning rate of the current epoch was calculated as shown in equation 1 with a power of 0.9.

$$current\_learning\_rate = initial\_learning\_rate * (\frac{1 - current\_epoch}{max\_epochs})^{power} \quad (1)$$

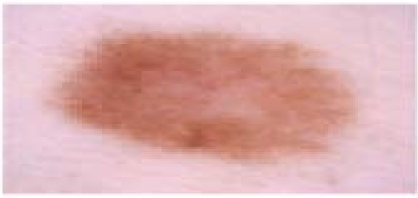
D. Data augmentation

Data augmentation is a crucial part in training deep networks, because they need large amounts of data to achieve high accuracy. Therefore, multiple augmentation techniques are applied, including random crop, random rotation between 0and 15, and random horizontal flipping as show in figure 3.

E. Model Options

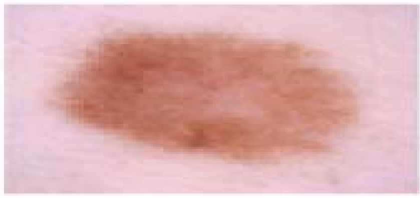
Baseline Model: The initial experiment was conducted using a baseline architecture that employs Inception V4 after replacing the classification layer which was trained to classify 1000 object categories of the Imagenet dataset [31] with a different layer to discriminate dataset classes. This model achieves an accuracy on the validation set of 82.89% as listed in the first row of table II.

Employing Feature Reuse: In this experiment, the effect of employing low level features from the bottom layer with



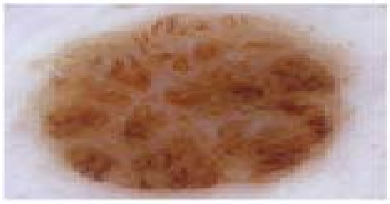
(a) sample of melanocytic nevus lesion

(b) sample of dermatofibroma lesion



(c) Sample of melanoma lesion

(d) sample of actinic keratosis lesion



(e) sample of benign keratosis lesion

(f) sample of basal cell carcinoma lesion

(g) sample of vascular lesion

Fig. 2: Seven classes for the HAM10000 images dataset

different configurations was investigated. First, the output features from the third convolution in the stem module (Features 1) were concatenated with the extracted features from the Inception-v4 model after applying a 1x1 convolution with 1024 filters. This configuration yielded an accuracy of 80%. Then, the number of filters was reduced from 1024 to 512 which improves the accuracy over that of the baseline model by 3.28%. It was apparent beforehand that the more the low level features are included, the more the model overfits, resulting in a reduction in model performance. A 3x3 convolution was applied instead of a 1x1 convolution to refine the low level features before concatenation however, that model yields an accuracy of 80.1%. The low level features from the third convolution (Features 1) of the stem module were concatenated with the output of the second concatenation filter (Features 2) in the stem module to enrich the features. This approach did not improve the model performance instead, it reduced the performance by nearly 10% as shown in the last row of table II.

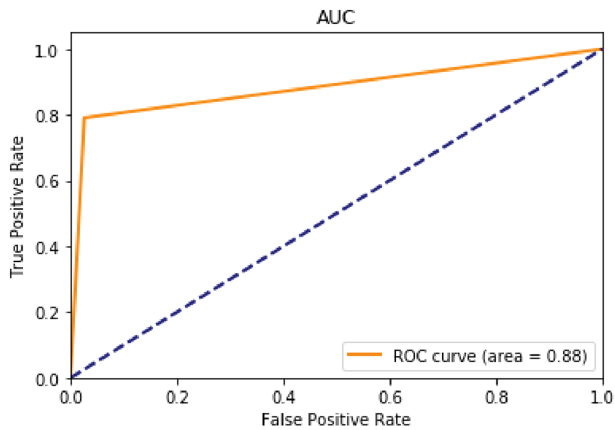


Fig. 4: Area Under The Curve on validation set

F. Benchmark results

The proposed CNN architecture achieves an Area Under the Curve (AUC) of 0.88 on the validation set as shown in Figure 4. The model with the best accuracy in the validation set was adopted and tested on the official benchmark of the ISIC 2018 challenge. The results are listed in table III. Although many approaches on the official benchmark achieve higher accuracy than the proposed model, the proposed model achieves comparable results while using only one model rather than an ensemble consisting of multiple models. In Table IV, we compare our proposed model with similar work with the aspect of accuracy and computational efficiency by measuring the number of floating point operations (FLOPS). For example, our model rises the accuracy by 13% compared to the Milton et al.[25] model while it reduces model complexity by 85.92% as shown in Table IV.

TABLE III: The evaluation metrics of the proposed architecture on the official benchmark of ISIC 2018 challenge Test dataset

Disease	Accuracy	AUC	Specificity	Sensitivity
MEL	0.904	0.765	0.945	0.585
NV	0.882	0.876	0.847	0.904
BCC	0.963	0.880	0.975	0.785
AKIEC	0.977	0.807	0.987	0.628
BKL	0.928	0.833	0.966	0.700
DF	0.986	0.850	0.995	0.705
VASC	0.987	0.854	0.994	0.714
Average	0.947	0.838	0.958	0.717

TABLE IV: Evaluation of the proposed model and similar works on the official test set and validation set. Note, Giga floating point operations (GFLOPs) is measured in ensemble models by adding the GFLOPs of each model in the ensemble.

Previous work	accuracy of test set	accuracy of val set	model	GFLOPs
Gessert et al.[24]	0.970	-	Ensemble of 54 models	377.8
Milton et al.[25]	-	0.73	Ensemble of 4 models	333.7
The proposed model	**0.947**	**0.8617**	**Only Inception-v4**	**46**

V. Conclusion

Skin cancer images are still mostly diagnosed visually through dermoscopic analysis. However, the early detection of different skin diseases requires advanced computational models as well as high classification accuracy. The strategy used in this paper presents a deep CNN based on the Inception-v4 model and train it to classify the seven types of skin lesions in unbalanced proportions extracted from HAM10000 dataset. The proposed model achieves an accuracy of 94.7% on test sets and 86% on validation sets. The designing of a deep learning model depend not only on providing high accuracy on a given dataset, but also providing a computationally efficient model to be run on low-power devices as smart phone. In the future, deep learning algorithms with high accuracy and performance will be applied as histopathological examination tool to improve human skin health.

References

[1] Esteva, A., Kuprel, B., Novoa, R. A. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542 (7639), 115118.

[2] Fabbrocini G, Triassi M, Mauriello MC, Torre G, Annunziata MC, De Vita V, et al. Epidemiology of skin cancer: Role of some environmental factors. Cancers (Basel) 2010 Nov 24;2(4):1980-1989.

[3] M. R. Hamblin, P. Avci, and G. K. Gupta, Imaging in Dermatology. Academic Press, 2016.

[4] Kittler, H., Pehamberger, H., Wolff, K., Binder, M.: "Diagnostic accuracy of dermoscopy". In: The Lancet Oncology. vol. 3, no. 3, pp. 159.165. 2002.

[5] Yang, J., Sun, X., Liang, J. and Rosin, P.L., 2018. Clinical Skin Lesion Diagnosis using Representations Inspired by Dermatologist Criteria. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1258-1266).

[6] Giotis, I., Molders, N., Land, S., et al. (2015). MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Systems with Applications, 42, 65786585.

[7] Argenziano G, Soyer HP. Dermoscopy of pigmented skin lesions: A valuable tool for early diagnosis of melanoma. Lancet Oncol 2001 Jul;2(7):443-449.

[8] Braun, R.P., Rabinovitz, H.S., Oliviero, M., Kopf, A.W., Saurat, J.H.: "Dermoscopy of pigmented skin lesions." In: J Am Acad Dermatol. vol. 52, no. 1, pp. 109-21. 2005.

[9] Rezze, G.G., Soares de S, B.C., Neves, R.I.: "Dermoscopy: the pattern analysis". In: An Bras Dermatol., vol. 3, pp. 261-8. 2006.

[10] Danilo Barros Mendes, Nilton Correia da Silva. Skin Lesions Classification Using Convolutional Neural Networks in Clinical Images, 2018.

[11] Oliveira RB, Papa JP, Pereira AS, Tavares JMRS. Computational methods for pigmented skin lesion classification in images: Review and future trends. Neural Comput Appl 2016 Jul 15;29(3):613-636.

[12] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions," in Color Medical Image Analysis. Springer, 2013, pp. 63–86.

[13] S. Gilmore, R. Hofmann-Wellenhof, and H. P. Soyer, "A support vector machine for decision support in melanoma recognition," Experimental dermatology, vol. 19, no. 9, pp. 830–835, 2010.

[14] B. Salah, M. Alshraideh, R. Beidas and F. Hayajneh, Skin Cancer Recognition by Using a Neuro-Fuzzy System, Cancer Informatics, 2011.

[15] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. Soroushmehr, M. Jafari, K. Ward, and K. Najarian, "Melanoma detection by analysis of clinical images using convolutional neural network," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016.

[16] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, "Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble," 2017.

[17] Titus Josef Brinker, Achim Hekler, Jochen Sven Utikal, Niels Grabe, Dirk Schadendorf, Joachim Klode, Carola Berking, Theresa Steeb, Alexander H Enk, Christof von Kalle . Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review, J Med Internet Res 2018;20(10):e11936

[18] S. Seog Han, M. Shin Kim, W. Lim, G. Hun Park, I. Park, and S. Eun Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," The Journal of Investigative Dermatology, 2018.

[19] Kawahara J, Hamarneh G. Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers. In: Proceedings of the 7th International Conference on Machine Learning in Medical Imaging (MLMI 2016). 2016 Presented at: 7th International Conference on Machine Learning in Medical Imaging (MLMI 2016); October 17, 2016; Athens, Greece.

[20] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle, "Knowledge transfer for melanoma screening with deep learning," in Proceedings- International Symposium on Biomedical Imaging, 2017.

[21] Marwan Ali Albahar. Skin Lesion Classification Using Convolutional Neural Network With Novel Regularizer, IEEE Access , 7: 38306 – 38313, 2019.

[22] Rina Refianti, Achmad Benny Mutiara, Rachmadinna Poetri Priyandini. Classification of Melanoma Skin Cancer using Convolutional Neural Network. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 3, 2019.

[23] Bhavya Sai V, Narasimha Rao G, Ramya M, Sujana Sree Y, Anuradha T. Classification of skin cancer images using TensorFlow andinception v3., International Journal of Engineering and Technology, 7 (2.7), (2018) 717-721.

[24] Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., and Schlaefer, A. (2018). Skin Lesion Diagnosis using Ensembles, Unscaled Multi-Crop Evaluation and Loss Weighting. arXiv preprint arXiv:1808.01694.

[25] Milton, M. A. A. (2019). Automated Skin Lesion Classification Using Ensemble of Deep Neural Networks in ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection Challenge. arXiv preprint arXiv:1901.10802.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision.arXiv preprint arXiv:1512.00567, 2015.

[27] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In Proceedings of the 27th International Conference on Neural Information Processing Systems

- Volume 2 (NIPS'14), Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), Vol. 2. MIT Press, Cambridge, MA, USA, 3320-3328.

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

[29] Mirza, B., Lin, Z. and Toh, K.A., 2013. Weighted online sequential extreme learning machine for class imbalance learning. Neural processing letters, 38(3), pp.465-486.

[30] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... & Lerer, A. (2017). Automatic differentiation in pytorch.

[31] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.