Skin Lesion Classification Using Machine Learning

Sai Preetham Sata
Hamburg University of Technology
sai.sata@tuhh.de

Harsha Bajje Thippeswamy
Hamburg University of Technology
harsha.bajje.thippeswamy@tuhh.de

Sebastian Engelhardt
Hamburg University of Technology
sebastian.engelhardt@tuhh.de

Tim Neuwerk
Hamburg University of Technology
tim.neuwerk@tuhh.de

Abstract—**[TN] In this work we present two different methods for the ISIC 2019 Skin Lesion Classification Challenge. The goal is to predict the class of a skin lesion based on dermoscopic images. Herefor a dataset of around 25000 images of eight different classes are given for training. Furthermore for most images additional meta data is provided which consists of information of the age and gender of the patient and position of the skin lesion on the human body. The final test set contains an additional ninth class, which is unknown.**

In our first approach we are using Support Vector Machine for the classification. Therefor we also consider image preprocessing and feature extraction. These features describe the shape and texture of the lesion.

In our second approach we are using deep learning to face the problem. We used the pretrained Convolutional Neural Network EfficientNet and applied transfer learning.

The second approach promises significantly better results than the first. We achieved a sensitivity score based on the training data of 85.31%.

I. Introduction [TN, SE]

Skin cancer is one of the most common types of cancer. It makes up more then half of the worldwide cancer diagnoses. Melanoma is the most dangerous type of skin cancer that can arise through different causes. A mole can change over time with an increase in size, irregular edges, change in color, itchiness or skin breakdown [1]. The primary cause of melanoma is the influence of ultraviolet light exposure from the sun or other sources like tanning devices. Its incidence and mortality rates have been increasing in the last decades [2] and therefore it represents an important public health problem. In 2012 out of 232,000 people that were diagnosed with Melanoma 55,000 people died.

However, if skin cancer is diagnosed in an early stage, it has very high curing rates. To detect cancerous skin dermatologists usually evaluate a skin lesion with the so called ABCD rule [3]. In the first step the "**a**symmetry, **b**order, **c**olors, and **d**iameter" criteria are approximately estimated. In the second step each criteria is multiplied by a given weight factor to get the total dermoscopy score. This shows the importance of the shape and texture of the lesion in skin cancer diagnoses.

In this work we will use machine learning techniques to classify the type of skin lesion based on skin lesion images. For the required data we use the "ISIC 2019: Training" dataset that was published by the International Skin Imaging Collaboration (ISIC). This dataset contains 25,331 images of 8 different types of skin lesions. The test dataset also contains additional skin images that don't fall in any of there categories. Those images are referred to as unknown. The dataset is composed of three separate sources. Part of the data originates from the HAM1000 dataset [4]. Those images are of size 600×450 and were manually cropped around the lesion by the dataset creators. Some of the images are also preprocessed by histogram equalization. The second dataset is the BCN20000 dataset [5] containing high resolution images (1024×1024). Not all images of that dataset are cropped, meaning there are images present with large black borders and the skin only visible in a circle in the middle. The last dataset is the MSK dataset [6] with images of different sizes originating from several sources. Table I shows the name of the lessions together with the number of images of these classes in the training dataset.

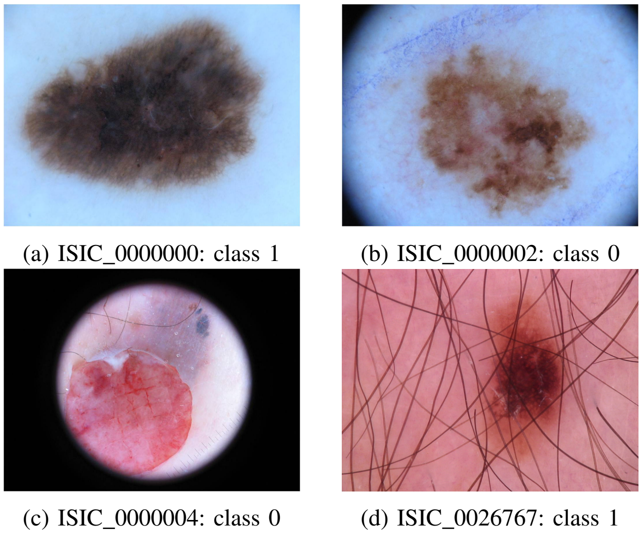
TABLE I: Number of samples per class in the training dataset

Lesion	Number of samples
0-Melanoma	4522
1-Melanocytic nevus	12875
2-Basal cell carcinoma	3323
3-Actinic keratosis	867
4-Benign keratosis	2624
5-Dermatofibroma	239
6-Vascular lesion	253
7-Squamous cell carcinoma	628
8-Unknows	0

As seen in the table, the dataset is heavily imbalanced. To get meaningful evaluation results, we rate our approaches using the balanced multi-class accuracy, wich is equivalent to the macro-average sensitivity, as our main metric [7].

Four images and the belonging classes are illustrated in figure 1. There is also additional data for most images called meta data. This contains information of the age and gender of the belonging patient and the position of the skin lesion on the patients body. For the skin lesion classification we will present two different approaches:

In section II we are using Support Vector Machine (SVM) for the classification. Therefor data preprocessing and feature extraction is necessary. These features describe the shape and texture of the image and are used as input data. In



(a) ISIC_0000000: class 1 (b) ISIC_0000002: class 0

(c) ISIC_0000004: class 0 (d) ISIC_0026767: class 1

Fig. 1: Four images of the “ISIC 2019: Training” dataset.

the following section III we are using a Convolutional Neural Network (CNN) for the classification. CNNs have become increasingly important in the domain of medical image analysis. Here the image itself is used as the input data. Furthermore we will investigate the prediction results by taking the additional meta data into account. One of the main challenges is the classification of the unknown class in the test dataset. Our approach for that is shown in section IV. In the last section V we will discuss the two different approaches and summarize our results.

II. Skin Lesion Classification Using Support Vector Machine [TN]

In this section we present a skin lesion classification approach by using Support Vector Machine (SVM). Therefor we first extract features out of the images that are then used as input data. Afterwards based on the extracted features we train and evaluate a SVM by making use of the Sklearn library [8].

A. Feature Extraction [SPS]

Feature extraction is the process of extracting relevant features that give information about the color,shape,texture etc. of the skin lesion.For extracting the features from the skin lesion it is mandatory to segment the skin lesion from the image and then extract the features of the segmented lesion.In order to segment the lesion from the image , three segmentation techniques has been implemented.Initially an elliptical mask is applied to the image, and the skin lesion is extracted using the elliptical mask.Then features are extracted for the masked image and prepared as a data set.In the second segmentation technique which works on algorithm termed as watershed algorithm, the skin lesion is segmented based on peaks and valleys that are created based on the intensity gradient across the image.This kind of segmentation technique works really well on the images with no hair.But the images with hair , this type of segmentation did not work properly with hair as shown in the figure below.So after this segmentation technique the features are extracted and second data set is created.The final type of segmentation is done using k means clustering.It forms segmented regions using k clusters.The features for this kind of segmentation are extracted and then third data set is created.Based on the classification results from SVM, K means clustering technique found to give a better evaluation parameters like the high accuracy and high sensitivity.The features that were extracted after segmentation are listed as Moments, Huemoments,PCA features,Mean and Standard deviation of the color channels in the images,haralick features.

1) Moments and Hu moments: Image moments are a weighted average of image pixel intensities. Image moments capture information about the shape of a lesion in a binary image because they contain information about the intensity I(x,y) as well as position x and y of the pixels.The central moments are translation invariant. In other words, no matter where the lesion is in the image, if the shape is the same, the moments will be the same. Hu Moments (or rather Hu moment invariants) are a set of 7 numbers calculated using central moments that are invariant to image transformations. The first 6 moments have been proved to be invariant to translation, scale, and rotation, and reflection. While the 7th moment’s sign changes for image reflection.

2) Haralick features: The basis for these features is the gray-level co-occurrence matrix . This matrix is square with dimension , where Ng is the number of gray levels in the image. Element [i,j] of the matrix is generated by counting the number of times a pixel with value i is adjacent to a pixel with value j and then dividing the entire matrix by the total number of such comparisons made. Each entry is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j.

3) PCA features: PCA is essentially a method that reduces the dimension of the feature space in such a way that new variables are orthogonal to each other (i.e. they are independent or not correlated).By reducing the dimension of the feature space, we have fewer relationships between variables to consider less likely to overfit the final model.

B. Learning and Testing [TN]

Based on the generated table with all extracted features and the corresponding skin lesion classes from the section above we use Sklearn to train the SVM. The classifier is set as follows: *estimator =SVC(C=10, kernel=’rbf’, gamma=’auto’, probability=True, class_weight=„balanced“, break_ties=True, decision_function_shape=’ovr’).* Here the regularization parameter C, the kernel, the degree of the polynomial kernel function and the gamma were tuned beforhand by trying out different combinations. To handle the unbalanced dataset we use class weights which are inversely proportional to its frequency in the datasat.
The ROC curves which are created by plotting the true positive rate against the false positive rate at various threshold settings

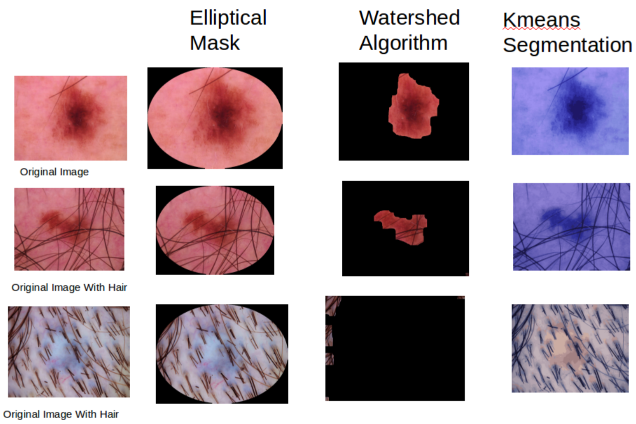


Fig. 2: Different Types of Segmentation

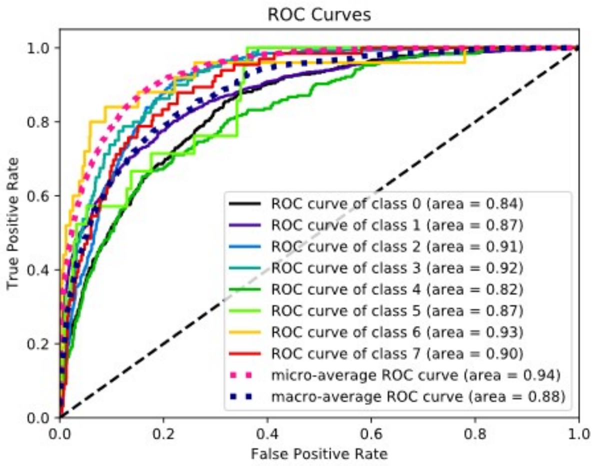


Fig. 3: ROC Curves of the testing results

are shown in figure 3. In total we achieved a sensitivity score (true positive rate) of 51.558%.

C. *Consideration of Meta Data [TN]*

As already mentioned in the introduction there exsists additional data for most images called meta data. This contains information about the age, gender and the position of the skin lesion on the patients body. To make use of this additional data we convert each type into a one-hot vector and then attach all three vectors as additional input data. This improves the sensitivity score to 56.558%.

III. Skin Lesion Classification Using Convolutional Neural Networks [SE]

For the second part of this project, we used convolutional neural networks (CNNs) to classify the images containing skin lesions. We apply transfer learning to pretrained networks on the ImageNet dataset. All training and testing was done on a Nvidia Tesla P100 with 16GB of video memory via Google Colab. The model and its input pipeline was developed using Tensorflow and evaluated using the same Sklearn functions as in section II.

A. *Architecture [SE]*

We use the EfficientNet models for our prediction pipeline. The base model B0 uses the standart ImageNet input size of 224 × 224, the models B1 to B7 scale that base architecture and allow much larger input sizes. The models are scaled uniformly in resolution, width and depth. This helps them to achieve state of the art performance on standart datasets while being smaller and faster then conventional architectures [9].

In our work, we manly use the models B3 and B4 which have an input size of 300 × 300 and 380 × 380 respectively. Those models outperformed resnet and densenet architectures in initial testing by a large margin. While the step from model B2 to B3 brought a significant performance improvement, the results of B3 and B4 were relatively similar. Larger models could not be trained enough epochs because of resource constraints.

For our final model, we only add global average pooling and one dense classification layer with softmax activation to the pretrained network. We tried adding additional fully connected layers in between so that the number of neurons don't decrease so fast (1536 to 8 for B3) but this did not bring an improvement.

B. *Data Augmentation and Input Strategies [SE]*

We used the tensorflow.data API to handle the large dataset efficiently. The batches of training data are automatically loaded in the background only when they are needed. To increase the variety of the dataset, we applied online data augmentation. These included randomly flipping and rotating images, as well as random changes in contrast brightness and hue. The larger images are then resized so that they fit the original size of the images from the HAM10000 dataset. The aspect ratio is kept at this operation. The input for the network consists of random crops of same size of these augmented images.

A different strategy without augmentation is applied to the validation and test data. We use equally spaced crops as seen in the last years challenge [10]. Each image is divided into six by six overlapping cops. All crops share the size of of the training data.

C. *Learning and Testing [SE]*

The training is done using the Adam optimizer. We apply a learning rate schedule to reduce the learning rate every 20 epochs. This helped the loss to decrease continuously during training procedure.

The imbalanced dataset is handled as before by class weights. Every sample gets a weight assigned inversely proportional to its frequency in the dataset. The strategie proofed to be much more efficient in our tests than oversampling the less frequent classes. A batch size of 32 is used when ever possible and reduced for larger models to fit our GPU memory. In the first stage, we only train our final classification layer for 50 epochs. After that, the lower convolutional layer of the base model are unfrozen and also fitted to our dataset. This fine tuning of the more specialized convolution kernels gave us a big performance improvement. We train these layers using a low learning rate in several stages for an additional 75 epochs. Our final model was an ensemble of the B3 and B4 model. The predictions of both networks were averaged a the end. That technic improved the accuracy by tree percent points. The final model was able to archive a validation accuracy of 85.31%.

D. Consideration of Meta Data [HBT]

The Neural Networks with multiple inputs when the network require data from multiple sources or in different formats. For example, networks that require image data captured from multiple sensors at different resolutions. In many researches, the use of different type of data along with the image data for classification in deep learning has shown the considerable improvement in image classification [11], [12]. The metadata is merged within the same pixel matrix of the image in each RGB layer which enriches the extraction of features, has shown significant improvement in accuracy in this research [12]. The image metadata is used nonparametrically to generate neighbourhoods of related images using Jaccard similarities, then used a deep neural network to blend visual information from the image and its neighbours in this research [11]. McAuley and Leskovec pioneered the study of multilabel image annotation using metadata, and demonstrated impressive results using only metadata. In hope of improvement of accuracy mixed data approach is used.

Including the metadata into the neural netwok is challenging. In contrast to the SVM based approach, the metadata is here in a different format then the image data. For that reason we created two separate networks and merged them. One branch is a Multi-Layer Perceptron (MLP) model that is designed to handle metadata and a second branch is a cnn for image data. Finally, these two branches are concatenated to have a final model for training.

The sex and position information of the metadat is converted to one-hot numpy arrays, the age is used as an numerical value. The MLP model consists of 4 layers with relu activation function with (128,64,32,8) neurons respectively. It has two inputs, age and sex. The conventional customized CNN model with 3 layers, with 3x3 filter size and (16,32,64) depth respectively and stride of 2 is chosen. Padding is done to the images to maintain the same length. Another CNN model with EfficientNet B4 for transfer learning followed by GlobalMaxPooling and 2 Dense layers with (512,8) neurons with SoftMax activation to the pretrained network. The transfer learning is performed using both the EfficientNet model and conventional CNN for better approximation of expected outputs. The cnn is trained the same way as before.

Figure 4 shows the Multi-Input model which combines both networks.

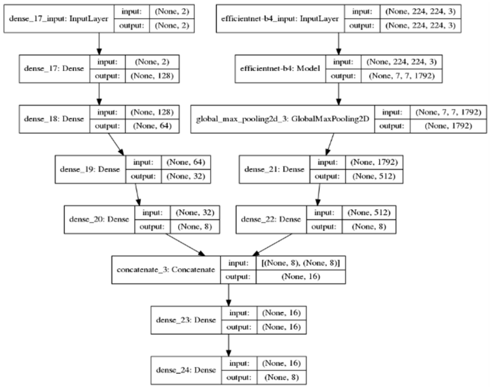


Fig. 4: Multi Input model that includes both CNN and MLP branches to handle mixed data.

Due to technical problems concerning the connection of tensorflow and keras, we could not use the image input pipeline described in section III-B. A validation accuracy of 53.63% was achieved using the Multi-Input model. There is huge requirement of improvement in image data pre-processing techniques, hyper-parameter tuning, extracting features from image related to metadata for multi-input model as compared CNN model by considering only image data. Since the images are loaded using Open CV at onetime without using any Data Loading Pipeline, the training was taking very long time for each epoch. Normalizing is done at once for a whole image array is not efficient way and proved by experimenting it. The accuracy is better for CNN model with a pre-trained EfficientNet model. Since the computational problems caused by OpenCV library in high resolution image data loading as input to multi-input model requires implementation of precise image processing techniques and a precise regularization with the loss obtained from model. The learning has to be regulated by considering metadata with respect to image data. Image augmentation and few other techniques may enhance the accuracy.

IV. Handling of Unknown Data [SE]

As mentioned in section I, the test dataset contains images that don't belong to any of the categories of the training dataset. There is no information about the number of unknown images, the only given information is, that these images still show human skin and skin lesions. The performance of our models on the test dataset was not revealed before the submission deadline. That conditions made it hard to tune the testset classification accuracy.

The automated evaluation system for this challenge required for each image a vector where each element is equal to the estimated probability for the respective class. In contrast to the real ISIC 2019 challenge, this vector is supposed to sum up to 1. For that reason it was obvious to use a softmax activation function for the neural network approach. We assumed that the maximum value of the predicted class probabilities (referred to as confidence in the following) will be low for images of the unknown category. Literature sows that this assumption is not necessary true [13]. To test if this approach is still usable in this situation, we set all images showing basal cell carcinoma (class 2) to unknown and trained the model on the remaining seven categories.

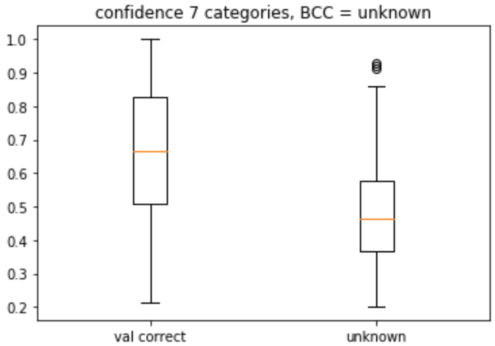


Fig. 5: comparison of distribution of confidence for known and unknown images

Figure 5 shows the result of this test. The bar for ”val correct” referrers to the confidence values for the correctly classified images in the validation dataset, ”unknown” to the confidence for the images of the class that was set to unknown. It can be seen, that our assumption holds for this test. The confidence for unknown images is in average significantly lower. Both classes are however not perfectly separated. With no additional information it is not easy to assign a threshold for the confidence under which the lesion is assumed to be unknown. With no feedback for the performance on the test dataset we didn’t see any evaluation method that would allow us to optimize this hyperparameter. Given the time constraints however, we decided that this approach is the best method available for us to handle the unknown class.

We wanted to use this approach also in phase one using a SVM but the SVC class in sklearn does only predict classes and does not output probabilities for them. The `predict_proba` method wich can for other classifiers output probability estimates was not applicable for SVMs. It uses Platt scaling and internal cross validation to create probabilities [8]. However they are not consistent with the normal class estimations and produced much worse results. For that reason we followed the advice in the documentation and used the values of the decision function for our confidence threshold.

V. Discussion and Conclusion [SE]

One of the most important factors for the final score of the 2019 ISIC challenge was the handling of unknown data. At the time of writing, we do not have a performance score for the complete test dataset. An email asking for scoring was not was not answered. Our best evaluation score of 85.31% for the deep learning approach is in the top range of evaluation scores that were claimed by participants of the ISIC challenge in there papers. The performance on the test dataset was in general much lower. With that limitations, we can not rate if our models would achieve good final results and are able to precisely detect the unknown images.

We were able to achieve much better results using deep-learning and CNNs than with the simple classifiers. However it could not be determined if the limiting factor was the feature extraction or the support vector machine. A downside of the CNN based approach is the resource usage. We trained the SVM on a single core of a desktop CPU for under ten minutes while the CNN required several hours of training on an enterprise grade GPU. The required time for image prediction using the trained model scaled the same way. This aspect can become a problem when deploying the model for practical use as it enforces a powerful cloud computing infrastructure.

We belief that an automated machine learning based system for skin lesion classification could be applied in a medical context. The main use for a product like this would be an assistance to human doctors. To little guarantees can be given in order to convert this models in stand allone diagnostics systems. In last years ISIC challenge which had no unknown category, the majority of algorithms was able to beat human experts [14]. This experiment has not yet been done for the current challenge. Together with legal challenges to certify a medical product, we conclude that a hybrid approach of algorithm and human would be the best usage.

Out of distribution detection is an active filed of research. Some theoretical problems of a softmax confidence based approach like adversarial attacks [15] are not relevant for this problem. However state of the are methods are usually more complex and can deliver more reliable results. These include diffrent loss functions and calibration or apply a two stage approach. In that case a unsupervised model is used to detect unknown images before the main model classifies the remaining ones [16].

In our input pipeline for the deep-learning approach, we scale down larger images to match the size of the HAM1000 dataset. This is a waste of information. In future work one should try to scale up the lower resolution images insteat. Given sufficient hardware, it would be possible to use larger EfficientNet Models and take larger crops.

Appart from the segmentation in the first phase, we did not apply and preprocessing other then the described data augmentation. It has been shown, that techniques like Color Constancy can improve the robustness and accuracy, specially when dealing with data from diffrent sources [17].

References

[1] I. Zaqout, “Diagnosis of skin lesions based on dermoscopic images using image processing techniques,” in *Pattern Recognition* (A. Zak, ed.), ch. 6, Rijeka: IntechOpen, 2019.

[2] M. Lens and M. Dawes, “Global perspectives of contemporary epidemiological trends of cutaneous malignant melanoma,” *The British journal of dermatology*, vol. 150, pp. 179–85, 03 2004.

[3] W. Stolz, A. Riemann, A. Cognetta, and O. Braun-Falco, “Abcd rule of dermatoscopy : a new practical method for early recognition of malignant melanoma,” *Eur. J. Dermatol.*, vol. 4, pp. 521–527, 1994.

[4] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific Data*, vol. 5, 03 2018.

[5] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, “Bcn20000: Dermoscopic lesions in the wild,” 2019.

[6] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, and A. Halpern, "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)," *CoRR*, vol. abs/1710.05006, 2017.

[7] L. Mosley, "A balanced approach to the multi-class imbalance problem," 2013.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019.

[10] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer, "Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting," 2018.

[11] J. Johnson, L. Ballan, and L. Fei-Fei, "Love thy neighbors: Image annotation by exploiting image metadata," in *Proceedings of the IEEE international conference on computer vision*, pp. 4624–4632, 2015.

[12] J.-S. Ruiz-Castilla, J.-J. Rangel-Cortes, F. García-Lamont, and A. Trueba-Espinosa, "Cnn and metadata for classification of benign and malignant melanomas," in *International Conference on Intelligent Computing*, pp. 569–579, Springer, 2019.

[13] U. Ozbulak, W. D. Neve, and A. V. Messem, "How the softmax output is misleading for evaluating the strength of adversarial examples," *CoRR*, vol. abs/1811.08577, 2018.

[14] P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, A. Lallas, J. Lapins, C. Longo, J. Malvehy, M. A. Marchetti, A. Marghoob, S. Menzies, A. Oakley, J. Paoli, S. Puig, C. Rinner, C. Rosendahl, A. Scope, C. Sinz, H. P. Soyer, L. Thomas, I. Zalaudek, and H. Kittler, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *The Lancet Oncology*, vol. 20, no. 7, pp. 938 – 947, 2019.

[15] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," 2018.

[16] E. Daxberger and J. M. Hernández-Lobato, "Bayesian variational autoencoders for unsupervised out-of-distribution detection," 2019.

[17] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, pp. 1146–1152, May 2015.