Can self-training identify suspicious ugly duckling lesions?

Mohammadreza Mohseni*[1, 2], Jordan Yap[2], William Yolland[2], Arash Koochek[4], and M Stella Atkins[1, 3]

[1]School of Computing Science, Simon Fraser University
[2]MetaOptima Technology Inc
[3]Department of Skin Science and Dermatology, University of British Columbia
[4]Banner Health

{mmohseni, stella}@sfu.ca, {jordan, william}@metaoptima.com, arash.koochek@bannerhealth.com

Abstract

One commonly used clinical approach towards detecting melanomas recognises the existence of Ugly Duckling nevi, or skin lesions which look different from the other lesions on the same patient. An automatic method of detecting and analysing these lesions would help to standardize studies, compared with manual screening methods. However, it is difficult to obtain expertly-labelled images for ugly duckling lesions. We therefore propose to use self-supervised machine learning to automatically detect outlier lesions. We first automatically detect and extract all the lesions from a wide-field skin image, and calculate an embedding for each detected lesion in a patient image, based on automatically identified features. These embeddings are then used to calculate the L2 distances as a way to measure dissimilarity. Using this deep learning method, Ugly Ducklings are identified as outliers which should deserve more attention from the examining physician. We evaluate through comparison with dermatologists, and achieve a sensitivity rate of 72.1% and diagnostic accuracy of 94.2% on the held-out test set.

1. Introduction

In order to diagnose skin cancers such as malignant melanoma, careful visual inspection of all the patient's skin lesions, particularly melanocytic nevi, must be undertaken. To this end, dermatologists have introduced several approaches which can be helpful in detecting melanomas, including the well-known ABCD dermatology criteria (Asymmetry, Border irregularity, Color irregularity, Diameter>6mm)[26, 1] and 7-point checklist[4, 5]. Another useful clinical indicator is the existence of Ugly Duckling (UD) lesions, which look different from the other lesions on the same patient[12]. The existence of a UD lesion is strongly correlated with the existence of melanoma [11, 33], although there is considerable inter-observer variability in visually selecting UD lesions from whole body imaging [16, 32]. After identification of UDs, clinically concerning lesions can then be inspected with a close-up examination, for example under dermatoscopic inspection. However, converting the clinical view of a small 1.5mm diameter nevus into a suitable digital representation for imaging and automatically identifying suspicious lesions is challenging.

Automatic UD detection can be treated as a form of outlier detection problem whose data comes from wide-field (clinical) images of the same patient, typically acquired during total body photography (TBP) also known as full body imaging. In TBP, overview images are typically taken with cameras placed 20-50 cm away from the skin surface, showing enough context to identify the body part, mimicking the visual inspection made by a clinician.

However a limitation of UD analysis from TBP images is lesion size. With current camera resolutions, very small lesions less than 1.5mm diameter cannot be readily detected in a TBP image of one body part. And just like a physical examination, a closer-up view is needed for lesion diagnosis.

An example of a typical TBP image taken with a smartphone camera from about 50cm away is shown in Fig 1. The filesize after this image has been compressed using JPEG is 187 KB.

Converting the clinical view of a small 1.5mm diameter nevus into a suitable digital representation for imaging and diagnosing suspicious nevi is challenging. In our experience, the lesion should occupy an area of at least 20x20 pixels to be reliably detected by algorithms and for auto-

*Corresponding Author

Figure 1: Typical image of part of a back, 1640 x 1116 pixels, used with permission from the SD-260 dataset [39]

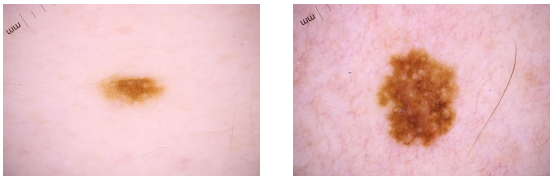
mated analysis. Therefore, a small 1.5mm diameter lesion requires resolution of around 0.075mm/pixel at the imaging surface. A wide-field digital image taken at about 20cm from the subject with a 10 MP smartphone camera, allows digital resolution of around 0.075mm/pixel [9]; smaller lesions cannot be reliably detected or analysed [34]. Artifacts from reflections, lack of polarization, image compression, motion blurring, lighting, and shadows all contribute to the challenge of both identifying UDs from a single digital wide-field image, as well as in detecting changes in lesions found in a series of TBP images taken over time.

Difficulties in obtaining high enough resolution TBP images from personal cameras such as smartphones for automated analysis of small lesions, has also led to difficulties in obtaining expertly-labelled digital TBP images. In our experience, the expert reader cannot zoom the image enough to make an analysis, unlike in a clinical environment where the physician can look closer at suspicious lesions using a magnifying dermoscope. For example, dermoscopic images are shown in Fig 2.The size of these dermoscopic images are 6000 x 4000 pixels, JPEG Size 1 MB, providing much more detail for machine learning classification algorithms.

Even very small lesions such as in Fig 2a, where the nevus is about 3mm wide, occupy about 30,000 pixels.

A few commercial systems can provide automated analysis of digital TBP images, where the images are taken with several fixed cameras for super-high resolution. One is the Canfield 3D Vectra system, where the lesion images are automatically stitched together, and then suspicious lesions can be viewed clinically with a dermoscope [36]. Another system which supports TBP using mobile cameras is provided by Dermengine [8]. An overview of such systems is provided by the International Society for Digital Imaging of the Skin (ISDIS) [ISDIS 2021 [18]].

Automated analysis of TBP "wide-field" digital images taken with smartphones remains challenging, because of the



(a) Dermoscopic nevus (b) Dermoscopic melanoma

Figure 2: Two dermoscopic images from the same patient, each of size 6000x4000 pixels (publicly available skin images from ISIC [37]). Note the mm scale ruler in the top left corner of each image.

difficulty of classifying and labelling such small lesions in the images. We hypothesised that outlier and suspicious UD lesions could be identified using a self-supervised learning method, leading to identification of suspicious lesions which can then be chosen for further detailed analysis (by computer algorithm and/or examination by a board-certified dermatologist).

In this paper we describe our objective, unbiased, and reproducible method to automatically identify UDs in TBP images, achieving accuracy comparable to that of expert dermatologists. By using our approach, expert dermatologists can be helped to identify which lesions deserve further examination in patients with numerous atypical lesions. These tasks are repetitive and time-consuming, with poor intra-observer and inter-observer reliability and consistency. Furthermore, as Schlessinger *et al*. suggested [31], our method can also be used in objective tasks needed for various outcome measures in clinical trials involving TBP images.

2. Related Work

With the recent advances in deep learning and its promising results, deep learning found its way to various areas of science and dermatology was no exception. Esteva *et al*. showed how promising deep learning can be in skin cancer classification [10]. Liu *et al*. extended this success to classification of a large number of general dermatology and skin cancer classes [24]. The contribution of deep learning has not been limited to classification. It has also contributed to segmentation and detection tasks.

One still unsolved challenge in deep learning is outlier detection. There have been many solutions proposed to overcome this challenge but no final solution is known yet. In the literature, this problem is known by different names such as outlier detection, anomaly detection, or one-class classification. A large body of research has demonstrated that finding a good lower dimension representation can lead to promising solutions to outlier detection problems

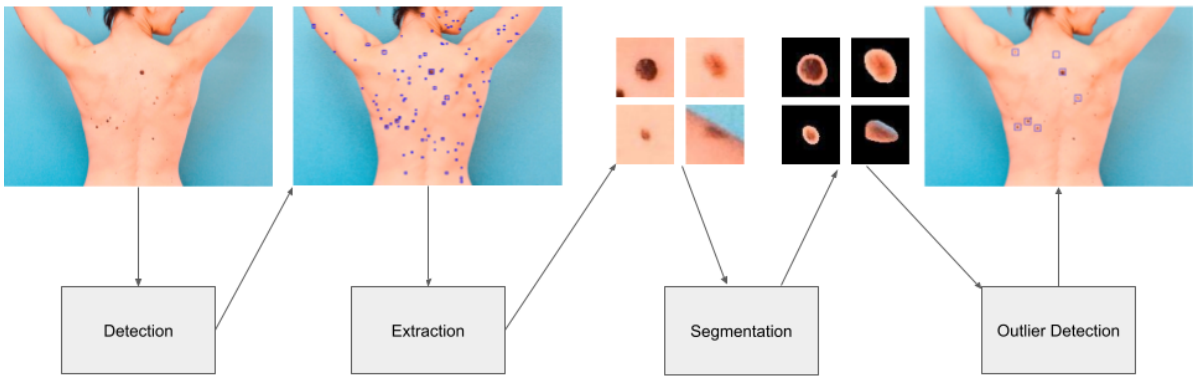


Figure 3: UD Detection Pipeline

[30, 7, 2]. Perera *et al*. showed that representations learned for one-class classification, can be transferred from an unrelated task[28]. Autoencoders are a family of neural networks which learn low-dimensional representation from the input domain. Variational autoencoders (VAE) are a subtype of autoencoders which control the latent space by enforcing a prior [20]. Higgins *et al*. showed that VAEs are able to learn disentangled features[14] and introduced parameter β and β-VAE as a means for disentangled representation learning[15].

Despite the recent advances in machine learning and deep learning, UD lesion detection still remains a challenge in computer aided dermatology. Soenksen *et al*. recently published a supervised machine learning method to identify UD lesions using the features of a pretrained network [34, 6]. Three board-certified dermatologists labelled UDs by rank order in 135 TBP patient images. Their results show between 83% and 88% agreement between the top 3 UDs ranked by the dermatologists and the top 3 outlier lesions ranked by the algorithm.

In this paper we show that it is possible to skip the long process of collecting labelled data and training a classifier. We propose self-training as an approach which enables researchers to reproduce our results easily; and comparably does as good job as using the labelled data.

3. Method

In order to detect and classify each lesion as an inlier or UD, we developed a multi-stage pipeline which consists of three main modules (shown in Fig 3). First, we pass the input TBP image to the detection module. This module detects all the lesions in the TBP image, crops a fixed sized window centered on the lesion and passes them to the segmentation module. The segmentation module segments the lesions from the skin and prepares the segmented images for the outlier detection module which is the last step.

3.1. Detection

The detection module used was based on the Single Shot MultiBox Detector (SSD) [22]. We used Resnet-18 [13] as our backbone along with a Feature Pyramid Network (FPN) [21] for improved multi-scale detection performance. The detection network was trained using 512x512 images as input to the network.

To detect lesions across a large TBP image, first the full-sized image is split up into 512x512 tiles with a 50% overlap. Tiles are then sent through the detection module to obtain the predicted locations of lesions. Detections are then aggregated across all tiles and non-max suppression is performed to obtain the final set of lesion detections for the entire TBP image.

3.2. Segmentation

The segmentation module is a smaller variant of the U-Net architecture[29], with only 6.2% of the trainable parameters as the original model. We use no batch normalization[17] in the model, similar to the original paper, and unlike some of the recent implementations in deep learning libraries[27]. The model runs on 64x64 RGB images, cropped around detected lesions, and produces a 64x64 probability map. An optimal threshold is determined on a validation set for binary lesion/background segmentation.

The model is trained on thousands of manually segmented lesions along with a collection of skin-only patches all of which are sourced from a set of internal full-body images spanning various skin-types, and body locations. We found that without the true negative patches, the model produced false positive segmentation masks when presented with detection errors (skin-only patches). We train with random rotations, flips, crops, and color-jitter, and optimize for a pixel-wise cross-entropy loss using the Adam[19] optimizer with a fixed step-down schedule and an initial learning rate of $1e^{-3}$.

3.3. Outlier Detection

Variational autoencoders tend to have higher reconstruction loss on anomalous samples[3, 25, 23]. This fact suggests that by training a VAE on the extracted lesions from one patient, UDs end up with high reconstruction loss value. The intuition behind this statement is that VAEs try to reconstruct common looking lesions (the majority) as perfectly as possible. UD lesions (which are the minority) are less prioritized, thus leading to higher reconstruction losses.

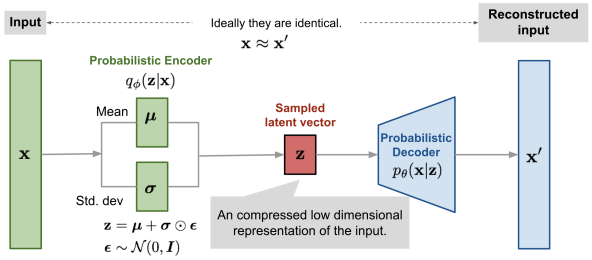


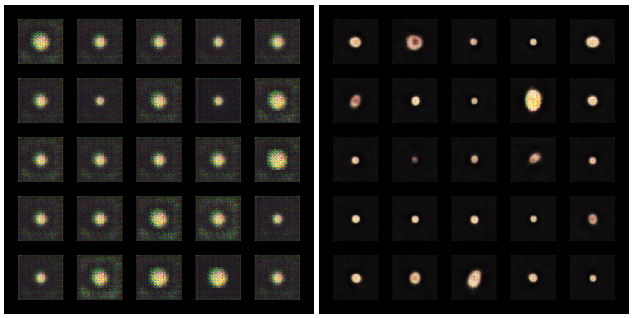
Figure 4: An overview of a Variational Autoencoder (VAE) architecture (image taken from [38])

In our first experiments we self-trained a VAE on extracted lesion images of one patient, and by defining a threshold on the reconstructed losses, we classified each lesion as either common looking or as a UD. In order to train the VAE, we enforce it to reconstruct the input image in the output as shown in Fig 4.

This approach's limitation is that for each TBP image query, the VAE needs to be trained for almost 130 epochs which takes $\approx$ 2 minutes. We tried to overcome this issue by doing pre-training to speed things up.

3.4. Fast Outlier Detection

We observed when training the VAE from scratch, in the initial epochs VAE is learning the basics of describing the lesions as shown in Fig 5a. With the increase in the number of epochs, our model learns more about the input domain and reconstructs better quality images in the output (as shown in Fig 5b). In order to decrease the time spent for the model to learn the basic features, we self-trained a base VAE model on 300 TBP images. In the query time when a TBP image is given, we fine-tune the base model on the lesion images from the TBP image for a few epochs and then calculate reconstruction loss and features from lesion images. Using this approach, we were able to identify UDs in just a few seconds and get results as accurate as before.



(a) Reconstructed lesion images at Epoch 13 (b) Reconstructed lesion images at Epoch 121

Figure 5: Two unordered sets of reconstructed images at different epochs. The model has more information regarding input domain in epoch 121 compared to epoch 13.

Additionally, we observed that by incorporating a β-VAE's disentangled latent space and calculating the distance between each lesion's extracted features and mean of features belonging to the lesions in the same TBP image, we still can identify UDs and optionally can skip the finetuning phase.

4. Data

We evaluated our UD detection algorithm on 75 TBP images in total. 32 images were sourced from the SD-198 [35], and SD-260 datasets [39], and an additional 43 were collected internally from various clinics. UD lesions in these images were labelled by a board certified dermatologist. TBP images used in this experiment contained a varying number of lesions ranging from 10 to 182. Overall, we extracted a total of 4628 lesion images across all 75 TBP images. 53 of the images contained at least one lesion labelled as UD, with an average of 1.44 UDs on each TBP image. The validity of data was manually checked after the segmentation module.

5. Results

The UD detection problem can be considered either as ranking prediction problem or binary classification problem. In order to evaluate our algorithm's performance, we used ranking evaluation metrics and binary classification evaluation metrics. In order to find the rankings, first we calculated an embedding for each lesion and then calculated the L2 distance between each embedding to the mean of embeddings. By sorting these distances, we generated the rankings. In the binary classification setting, we defined a threshold on each lesion in order to determine whether a lesion should be called UD or not. Using the following

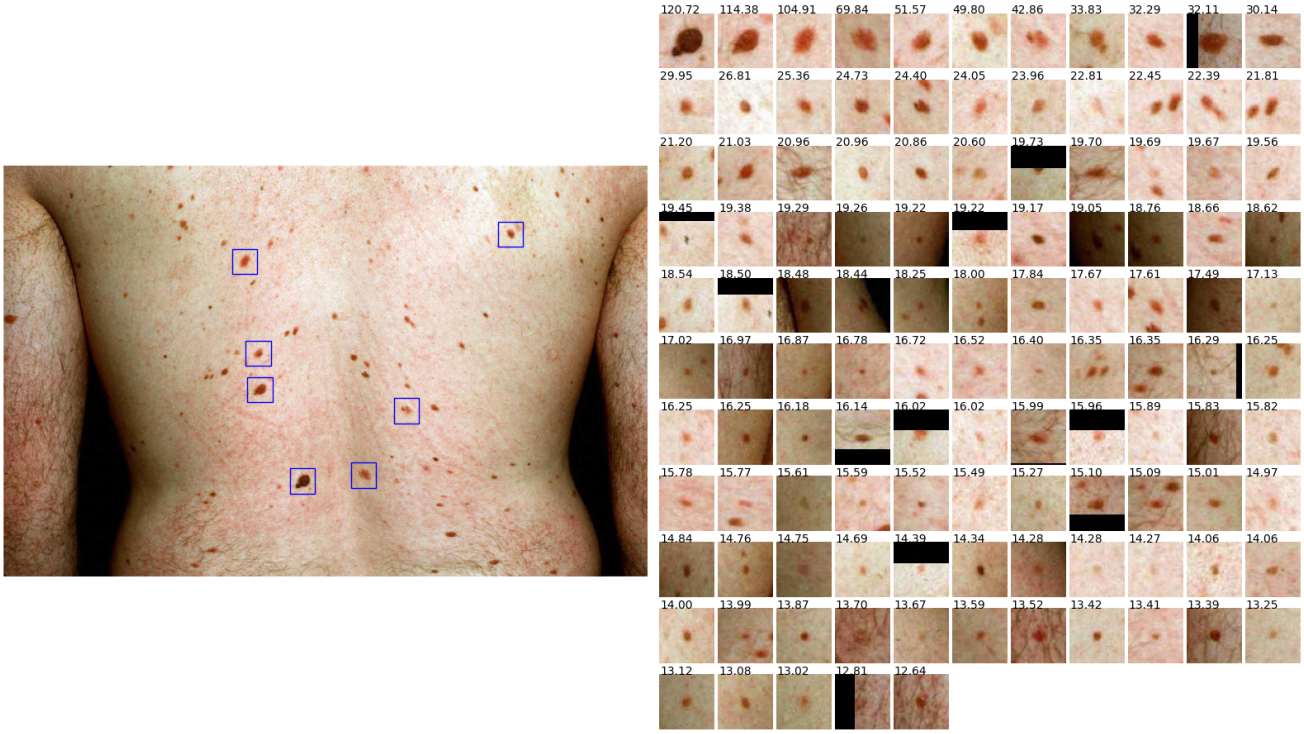


Figure 6: Automatic analysis of all the lesions in the image, with an ordered list of all the distance rankings, highest at top left hand side. The outliers which had a distance higher than our defined threshold of 38.63, are highlighted in blue on the image. The mean and standard deviation of the distances were 21.60 and 17.03 respectively.

formula, we obtained the threshold for each TBP image:

$$threshold = mean(distances) + \min \begin{cases} mean(distances) \\ std(distances) \end{cases} \quad (1)$$

Fig 6 shows the results of automated analysis of all the lesions extracted from a TBP image. The 7 lesions outlined in blue rectangles are the outliers according to our threshold.

5.1. Ranking Evaluation Metrics

We used average precision (AP) and reciprocal rank (RR) as our primary metrics. Additionally we used top-3 and top-7 agreement (Agr.) from [34]. We calculated each metric for all TBP images and report the average over all TBP images. In order to calculate top-3 agreement, for each lesion labelled as UD, we found its rank in our results. If the ranking was less than or equal to 3, we called success (TBP image counted as 1), otherwise we called it failure (TBP image counted as 0). We calculate and report the average of success over TBP images. Top-7 agreement is also calculated in a similar manner. Since not every TBP image contains a UD lesion, we only calculate top-3 and top-7 agreement for the TBP images containing at least one lesion labelled as UD.
Results are shown in Table 1.

MAP	MRR	Top-3 Agr.	Top-7 Agr.
0.659	0.721	86.79%	94.34%

Table 1: UD Ranking Evaluation

5.2. Binary Classification Metrics

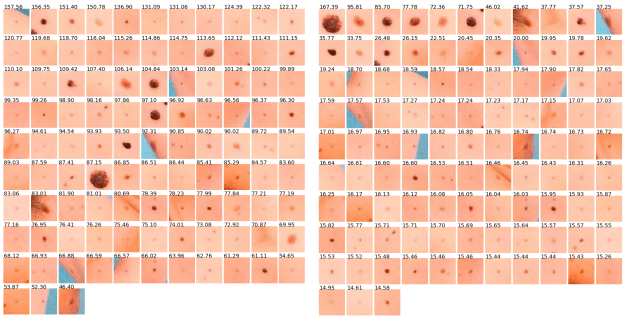
Binary classification predictions can be obtained by applying a threshold on the distances calculated for each lesion.Results are shown in Table 2. On average our model predicts 4.25 UDs per TBP image. In order to evaluate our method, we measured accuracy, sensitivity, and specificity over all extracted lesions and also averaged per TBP image.

6. Discussion

It is sometimes difficult to distinguish pigmented nevi from other skin conditions, such as lentigos and ephelides (freckles) which often occur on sun-exposed areas of the upper back, face, back of hands, and forearms. Also, differentiating small nevi (those less than 2 mm in diameter) from ephelides is not easy even clinically. Depending on

Metric \ Evaluation Type	Over All Lesions (Micro Avg.)	Averaged Over TBPs (Macro Avg.)
Accuracy	94.16%	94.23%
Sensitivity	72.07%	71.91%
Specificity	94.70%	94.95%

Table 2: UD Binary Classification Evaluation. It should be noted that when calculating average sensitivity over TBP images, only images with at least one UD lesion were considered.



(a) Before segmentation (b) After segmentation

Figure 7: Ranking predicted before and after incorporating segmentation module. Following the same rules as Fig 6.

the image resolution, our algorithm automatically detects all the skin lesions larger than around 1.5mm diameter, and ranks them in order of similarity to each other, providing an objective measure of the outlier nature of each lesion. Our algorithm provides an opportunity to perform quick and unbiased screening for ugly duckling lesions in TBP images, especially useful for patients with many lesions. Additionally, our method can work on all types of skin lesions, which means it is not limited to pigmented lesions.

Our algorithm also benefits much from incorporation of segmentation module. Before incorporating segmentation module our results were heavily influenced by noise. Some removed noises were shadow, skin color around a lesion, and proximity to the body edge. With the help of the segmentation module, our model made much better predictions as it became more robust to noise and avoided many false positives. Fig 7 shows how our ranking was improved with the help of segmentation module. It is important to mention that segmentation module also resolves the need for additional normalization methods to fix the lighting. Although normalization methods are mostly helpful, but sometimes they can drastically decrease the quality of images in some edge-cases.

Limitations of our method include difficulties in detecting UD lesions when the skin around the lesion has useful information for making the decision (e.g. scars after excision), as we are not able to incorporate additional domain knowledge in our decision making. We also remove hairs during detection and segmentation, which means a loss of a clue such as a hair growing out of a lesion, which indicates it might have been a congenital nevus. Since our model does not have domain knowledge, it can not leverage these facts to make better decisions.

7. Conclusion and Future Work

We have developed an algorithm to quickly identify outlier lesions in total body photography images, at even the relatively low resolution obtained by imaging backs with mobile smartphone cameras. Higher resolution images of lesions on arms, etc. can be analysed with even more accuracy. This opens up the opportunity to use this as a screening aid for patients to submit their photographs using teledermatology.

Future work will involve further evaluation of the method against a larger image set of expertly labelled images. We may also use different distance measures other than the L2 to determine outliers. Also we hope the result of this UD outlier detection project can improve classification of skin lesions extracted from TBP images. We hope introduction of our pipeline opens the path for future research in the domain of TBP images.

References

[1] Naheed R Abbasi, Helen M Shaw, Darrell S Rigel, Robert J Friedman, William H McCarthy, Iman Osman, Alfred W Kopf, and David Polsky. Early diagnosis of cutaneous melanoma: revisiting the abcd criteria. *JAMA*, 292(22):2771–2776, 2004. 1

[2] Charu C. Aggarwal. *Outlier Analysis*. Springer International Publishing, 2 edition, 2017. 3

[3] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015. 4

[4] Giuseppe Argenziano, Gabriella Fabbrocini, Paolo Carli, Vincenzo De Giorgi, Elena Sammarco, and Mario Delfino. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of der-

matoscopy and a new 7-point checklist based on pattern analysis. *Archives of Dermatology*, 134(12):1563–1570, 1998. 1

[5] G. Betta, Giuseppe DI Leo, Gabriella Fabbrocini, Alfredo Paolillo, and M. Scalvenzi. Automated application of the 7-point checklist diagnosis method for skin lesions: Estimation of chromatic and shape parameters. *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, page 1818–1822, 2005. 1

[6] Judith S. Birkenfeld, Jason M. Tucker-Schwartz, Luis R. Soenksen, José A. Avilés-Izquierdo, and Berta Marti-Fuster. Computer-aided classification of suspicious pigmented lesions using wide-field images. *Computer Methods and Programs in Biomedicine*, 195:105631, Oct. 2020. 3

[7] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Robust, Deep and Inductive Anomaly Detection. In Michelangelo Ceci, Jaakko Hollmén, Ljupčo Todorovski, Celine Vens, and Sašo Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 36–51, Cham, 2017. Springer International Publishing. 3

[8] Dermengine. Dermengine total body photography. = https://www.dermengine.com/en-ca/total-body-photography/. Accessed: 2021-03-11. 2

[9] Bogdan Dugonik, Aleksandra Dugonik, Maruška Marovt, and Marjan Golob. Image Quality Assessment of Digital Image Capturing Devices for Melanoma Detection. *Applied Sciences*, 10(8):2876, Apr. 2020. 2

[10] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, Feb. 2017. 2

[11] Caroline Gaudy-Marqueste, Yanal Wazaefi, Yvane Bruneu, Raoul Triller, Luc Thomas, Giovanni Pellacani, Josep Malvehy, Marie-Françoise Avril, Sandrine Monestier, Marie-Aleth Richard, Bernard Fertil, and Jean-Jacques Grob. Ugly Duckling Sign as a Major Factor of Efficiency in Melanoma Detection. *JAMA Dermatology*, 153(4):279–284, 04 2017. 1

[12] J J Grob and J J Bonerandi. The 'ugly duckling' sign: identification of the common characteristics of nevi in an individual as a basis for melanoma screening. *Archives of dermatology*, 134(1):103–4, Jan 1998. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 3

[14] Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*, 2016. 3

[15] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Journal on Learning Representations*, 2016. 3

[16] Muneeb Ilyas, Collin M. Costello, Nan Zhang, and Amit Sharma. The role of the ugly duckling sign in patient education. *Journal of the American Academy of Dermatology*, 77(6):1088–1095, 2017. 1

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 3

[18] International Society for Digital Imaging of the Skin ISDIS. Isdis. https://isdis.org/total-body-photography/. Accessed: 2021-03-11. 2

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 3

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016. 3

[23] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps. Towards Visually Explaining Variational Autoencoders. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8639–8648, June 2020. ISSN: 2575-7075. 4

[24] Yuan Liu, Ayush Jain, Clara Eng, David H. Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, Vishakha Gupta, Nalini Singh, Vivek Natarajan, Rainer Hofmann-Wellenhof, Greg S. Corrado, Lily H. Peng, Dale R. Webster, Dennis Ai, Susan J. Huang, Yun Liu, R. Carter Dunn, and David Coz. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, June 2020. 2

[25] Yuchen Lu and Peng Xu. Anomaly Detection for Skin Disease Images Using Variational Autoencoder. *arXiv:1807.01349 [cs, stat]*, July 2018. arXiv: 1807.01349. 4

[26] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B. Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *J of American Academy of Derm.*, 30(4):551–559, 1994. 1

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 3

[28] P. Perera and V. M. Patel. Learning Deep Features for One-Class Classification. *IEEE Transactions on Image Processing*, 28(11):5450–5463, Nov. 2019. 3

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 3

[30] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In Marc Niethammer, Martin Styner, Stephen Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen, editors, *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, pages 146–157, Cham, 2017. Springer International Publishing. 3

[31] Daniel I. Schlessinger, Guillaume Chhor, Olivier Gevaert, Susan M. Swetter, Justin Ko, and Roberto A. Novoa. Artificial intelligence and dermatology: opportunities, challenges, and future directions. *Seminars in Cutaneous Medicine and Surgery*, 38(1):E31–37, Mar. 2019. 2

[32] Alon Scope, Stephen W. Dusza, Allan C. Halpern, Harold Rabinovitz, Ralph P. Braun, Iris Zalaudek, Giuseppe Argenziano, and Ashfaq A. Marghoob. The "Ugly Duckling" Sign: Agreement Between Observers. *Archives of Dermatology*, 144(1):58–64, 01 2008. 1

[33] Alon Scope and Ashfaq A. Marghoob. The "Ugly Duckling" Sign: The "ugly duckling" sign: An early melanoma recognition tool for clinicians and the public. *Melanoma Letters*, 25:1–3, 2007. 1

[34] Luis R. Soenksen, Timothy Kassis, Susan T. Conover, Berta Marti-Fuster, Judith S. Birkenfeld, Jason Tucker-Schwartz, Asif Naseem, Robert R. Stavert, Caroline C. Kim, Maryanne M. Senna, José Avilés-Izquierdo, James J. Collins, Regina Barzilay, and Martha L. Gray. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine*, 13(581):eabb3652, Feb. 2021. 2, 3, 5

[35] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A Benchmark for Automatic Visual Classification of Clinical Skin Disease Images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 206–222, Cham, 2016. Springer International Publishing. 4

[36] Canfield Scientific Imaging Systems. Canfield. https://www.canfieldsci.com/imaging-systems/vectra-wb360-imaging-system/. Accessed: 2021-03-11. 2

[37] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161, Aug. 2018. 2

[38] Lilian Weng. From autoencoder to beta-vae. *lilianweng.github.io/lil-log*, 2018. 4

[39] J. Yang, X. Wu, J. Liang, X. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang. Self-Paced Balance Learning for Clinical Skin Disease Recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2832–2846, Aug. 2020. 2, 4