CANCER

Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images

Luis R. Soenksen[1,2,3,4,5]*, Timothy Kassis[6], Susan T. Conover[2], Berta Marti-Fuster[2,5], Judith S. Birkenfeld[2,5], Jason Tucker-Schwartz[2,5], Asif Naseem[2,5], Robert R. Stavert[7,8,9], Caroline C. Kim[10,11], Maryanne M. Senna[9,12], José Avilés-Izquierdo[13], James J. Collins[2,3,4,6,14,15], Regina Barzilay[16,17], Martha L. Gray[2,4,5,17]

A reported 96,480 people were diagnosed with melanoma in the United States in 2019, leading to 7230 reported deaths. Early-stage identification of suspicious pigmented lesions (SPLs) in primary care settings can lead to improved melanoma prognosis and a possible 20-fold reduction in treatment cost. Despite this clinical and economic value, efficient tools for SPL detection are mostly absent. To bridge this gap, we developed an SPL analysis system for wide-field images using deep convolutional neural networks (DCNNs) and applied it to a 38,283 dermatological dataset collected from 133 patients and publicly available images. These images were obtained from a variety of consumer-grade cameras (15,244 nondermoscopy) and classified by three board-certified dermatologists. Our system achieved more than 90.3% sensitivity (95% confidence interval, 90 to 90.6) and 89.9% specificity (89.6 to 90.2%) in distinguishing SPLs from nonsuspicious lesions, skin, and complex backgrounds, avoiding the need for cumbersome individual lesion imaging. We also present a new method to extract intrapatient lesion saliency (ugly duckling criteria) on the basis of DCNN features from detected lesions. This saliency ranking was validated against three board-certified dermatologists using a set of 135 individual wide-field images from 68 dermatological patients not included in the DCNN training set, exhibiting 82.96% (67.88 to 88.26%) agreement with at least one of the top three lesions in the dermatological consensus ranking. This method could allow for rapid and accurate assessments of pigmented lesion suspiciousness within a primary care visit and could enable improved patient triaging, utilization of resources, and earlier treatment of melanoma.

INTRODUCTION

Melanoma is a type of malignant tumor responsible for more than 70% of all skin cancer–related deaths worldwide. In 2019, there were an estimated 96,480 patients newly diagnosed with melanoma, with a reported 7230 deaths in the United States alone (*1*, *2*). Typically, patients presenting only with localized primary cutaneous melanomas of ≤1 mm thickness have an excellent prognosis (>90% 5-year survival rate) (*1*, *3*). For patients with thicker tumors, however, melanoma survival rates decrease to 62 and 18% for stages III and IV, respectively (*1*, *3*). Furthermore, studies evaluating the economic burden of melanoma estimate a 20-fold increase in treatment cost from early- to late-stage melanoma (*4*), accounting for additional healthcare expenses that could potentially be reduced through early detection and treatment. Although recent immunotherapies such as programmed cell death protein 1 (PD-1) inhibitors have improved clinical outcomes, they still constitute substantial treatment costs (*5*). Visual inspection of patients to identify lesions that exhibit features clinically concerning for skin cancer or suspicious pigmented lesions (SPLs), is a long-standing dermatological practice that belongs to a group of visual tasks known as outlier lesion macroscreening (*6*). For years, the assessment of SPL features such as asymmetry, border unevenness, color distribution, diameter, and evolution (collectively known as the ABCDE criteria) have constituted the cornerstone of early-stage melanoma screening. These visual descriptors, in combination with risk assessments from the patient's medical history, full-body inspection, nevi density, and lesion saliency (ugly duckling), help dermatologists identify lesions for skin biopsy and histopathologic evaluation, the gold standard in melanoma diagnosis (*7*). In particular, highly accurate and skilled clinical detection of melanoma appears to rely heavily on unconscious visual pattern and comparative "ugly duckling" recognition rather than simplified algorithms of ABCDE morphologic criteria (*8*–*10*).

In recent years, several studies have suggested that access to dermatological screenings can correlate with earlier detection of melanomas and subsequently improved prognosis (*11*–*13*). On the basis

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA. [2]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA. [3]Wyss Institute for Biologically Inspired Engineering, Harvard University, 3 Blackfan Cir, Boston, MA 02115, USA. [4]Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02139, USA. [5]MIT linQ, Massachusetts Institute of Technology Cambridge, MA 02148, USA. [6]Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, 02139, MA, USA. [7]Division of Dermatology, Cambridge Health Alliance, 1493 Cambridge Street, Cambridge, MA 02139, USA. [8]Department of Dermatology, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, MA 02215, USA. [9]Department of Dermatology, Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA. [10]Pigmented Lesion Program, Newton Wellesley Dermatology Associates, 65 Walnut Street Suite 520 Wellesley Hills, MA 02481, USA. [11]Department of Dermatology, Tufts Medical Center, 260 Tremont Street Biewend Building, Boston, MA 02116, USA. [12]Department of Dermatology, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114 [13]Department of Dermatology, Hospital General Universitario Gregorio Marañón, Calle del Dr. Esquerdo 46, 28007 Madrid, Spain. [14]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [15]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. [16]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA. [17]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology Cambridge, MA 02148, USA.
*Corresponding author. Email: soenksen@mit.edu

of this research, various pilot programs in Europe have been deployed to measure the effects of large-scale skin cancer screening initiatives at the primary care level (*14*, *15*). Such programs have shown the potential to provide reductions in patient treatment costs and mortality when performing full-body examinations in large numbers of high-risk individuals, leading to a majority of experts and patient advocacy groups in support of melanoma screening policies (*14*, *15*). Despite these encouraging results, such programs have not been adopted in most countries (both developed and developing), primarily because of the high initial cost of implementation and the lack of scalable tools to help primary care physicians in the identification of cancerous skin lesions at a population level (*16*). The challenges that these initiatives face are complex but primarily stem from the fact that malignant melanoma is a relatively rare disease (13 per 10,000 persons) that is also difficult to confirm solely through visual inspection (*17*). Even among experts, only 0.8% of detected SPLs are confirmed to be malignant through biopsy (*14*). This situation has profound implications for screening policies in high-risk regions, where millions of pigmented lesions would need to be evaluated to prioritize regional expert evaluations and biopsies for SPLs so that melanoma treatments could be efficiently directed to reduce the burden of this disease. Considering these conditions, it becomes clear why providing indiscriminate dermatological referrals to the general population for pigmented lesion screenings is cost-prohibitive, impractical, and mostly controversial as an effective public health measure (*18*).

In the past decade, advances in smartphone technologies have increased access to high-quality personal cameras and robust mobile computing systems for a wide range of applications, including dermatology. However, the images produced by commonly used personal devices have long been considered suboptimal for use in skin cancer computer-aided diagnosis (CAD). This is, in part, due to difficulties in segmentation, ABCD feature extraction, and accurate lesion classification in the presence of image artifacts (*19*, *20*). Furthermore, these systems rely on the user to drive the appropriate identification of pigmented lesions for image acquisition and analysis (*21*). Thus, traditional pigmented lesion CAD systems have been of little use in large-scale melanoma screening initiatives, as most have been developed to work only with dermoscopy and single-lesion near-field photography, both of which require specialized illumination and training (*22*). Such strict requirements are impractical for most real-use scenarios at the primary care level, which are limited in both time and capacity to image a large number of lesions per patient carefully. The presence of multiple potentially suspicious lesions at different scales, spurious nonskin regions such as clothing, uneven illumination, angled surfaces, and obstructing hair have all been reported to lead to poor accuracy in traditional skin cancer CAD systems (*23*, *24*).

More recently, deep convolutional neural networks (DCNNs) have been used in next-generation CAD systems to overcome many of the challenges associated with automated dermatology evaluations. For example, seminal study (*25*) used a deep neural network model based on Google's Inception v3 architecture and ImageNet transfer learning for the classification of 2032 different skin diseases. In this implementation, a pretrained network was fine-tuned using 129,450 dermatological images and then tested along with 21 board-certified dermatologists on biopsy-proven clinical images. Upon evaluation, this DCNN was capable of delivering an average accuracy of 72.1 ± 0.9% for three aggregated skin disease classes (benign, malignant, and nonneoplastic), whereas when a subset of the validation dataset was assessed by two dermatologists they scored 65.78 ± 0.22% (*25*). Such results, as well as other investigations on deep learning for dermatology (*26*–*30*), have informed the speculation that DCNN-based models can reach comparable or even superior diagnostic accuracy compared with board-certified dermatologists in specific visual tasks.

Despite the potential of implementing deep learning in clinical dermatology, previous demonstrations using DCNNs have not been trained to specifically address some of the more practical real-world challenges present in rapid, multilesion analysis for large-scale melanoma screenings. For example, most DCNN-based CAD systems rely on the assumption that the user has the time, training, and incentives to appropriately preselect all relevant lesions in a patient that are worthy of analysis. Furthermore, the classification in these systems is inferred only at the single-lesion level in comparison with the training dataset (*25*–*27*), without any consideration for other relevant interlesion dependencies, such as feature saliencies (also known as the ugly duckling criteria), used by expert dermatologists when conducting efficient SPL evaluations (*6*, *7*, *31*, *32*). Thus, here, we present a DCNN system optimized for the identification and classification of SPLs in wide-field images (photographs depicting multiple lesions from large body parts). Our DCNN system has been designed to generate marked overlays of suspiciousness classifications at the single-lesion level, as well as ugly duckling heatmaps showing intrapatient lesion saliencies (Fig. 1). With this system, we hope to provide a scalable solution to improve dermatological referrals at the primary care level, which attends to the naïve wide-field nature of these observations and the often overlooked ugly duckling criteria.

RESULTS

Dataset for wide-field suspicious skin lesion detection

We generated an image dataset to train DCNN models for SPL detection and classification in wide-field dermatological images by combining open-access dermatology repositories, web scraping outputs, and deidentified clinical images from 133 patients at the Hospital Gregorio Marañón (Madrid, Spain) (fig. S1). This dataset contained a total of $n_{baseline}$ = 33,980 individually labeled and nonoverlapping image crops divided into six classes (Fig. 2A), including backgrounds (n_b = 8888), skin edges (n_{se} = 2528), bare skin sections (n_{sk} = 10,935), nonsuspicious pigmented lesions type A (NSPL-A) ($n_{nspl\text{-}a}$ = 10,759) of low priority, NSPL-B ($n_{nspl\text{-}b}$ = 1110) of medium priority, and SPLs (n_{spl} = 4,063). The included background images span a variety of fabrics, furniture, walls, and other objects commonly found in primary care and home care settings. The NSPL-A class was aggregated from nine distinct pigmented lesion subtypes where low-priority management is typically indicated. Similarly, the NSPL-B class includes images from the other five skin lesion subtypes where dermatological referral or follow-up are usually indicated to better assess patient risk of skin cancer. Last, the SPL class consisted of melanomas stages 0 to IV (n_m = 2906), squamous cell carcinomas (n_{scc} = 589), and 568 basal cell carcinomas (n_{bcc} = 568) for which biopsy or excision is usually recommended (Fig. 2B). Nondermoscopy images were used for 99.15% of the NSPL-A set, 82.79% of the NSPL-B set, and 90.02% of the SPL set (Fig. 2C). Additional information on pigmented lesion class taxonomy is provided in table S1. All pigmented lesion class labels (NSPL-A, NSPL-B, and SPL) were confirmed visually by consensus of three board-certified

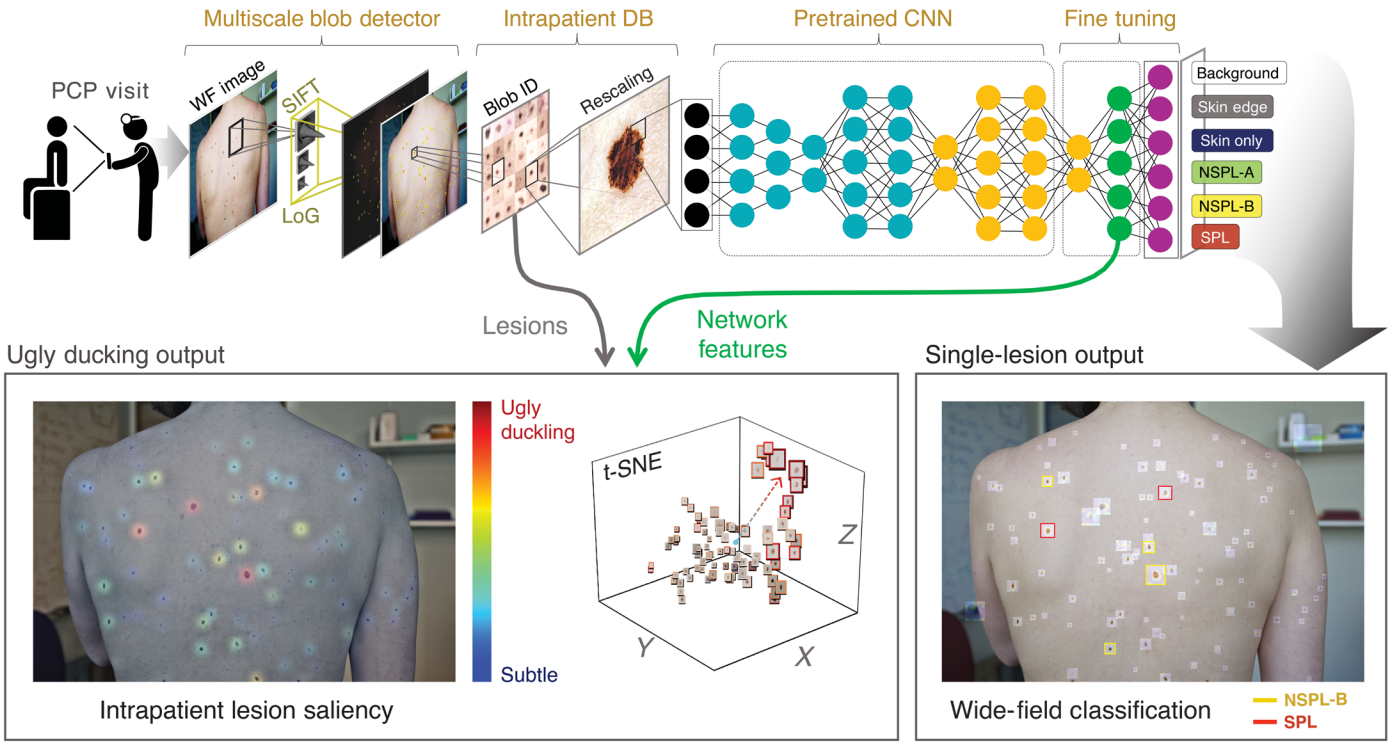


Fig. 1. Wide-field DCNN and deep ugly duckling saliency layout. Our system data flow is shown from left to right. First, a wide-field patient image was acquired by the user (or primary physician) and fed to the algorithm. Then, a blob detection algorithm on the basis of Laplacian of Gaussians (LoG) and scale-invariant feature transformation (SIFT) was used to detect all blob-like regions to accelerate analysis. Detected blobs were cropped and stored in an intrapatient repository. Stored images were fed into a deep classifier developed using an ImageNet pretrained convolutional neural network (CNN) architecture (for example, VGG16 and Xception) and fine-tuned on our own dataset of 33,980 images comprised of six different classes (SPLs, nonsuspicious pigmented lesions type A (NSLP-A), NSLP-B, skin, skin edges, and backgrounds). Detected pigmented lesions were classified as suspicious or nonsuspicious considering the single-blob class probabilities generated by the dense layer of the network (single-lesion output) and also using a multilesion saliency ranking or score (ugly duckling output) calculated using the deep features from the CNN. Our multilesion saliency score is a patient-dependent metric of pigmented lesion oddness calculated using the geometric distance of deep features from all moles in a single patient. Results are presented in the form of an output image with suspicious regions of interest, a saliency heatmap, and complete lesion montages to assist clinical users when conducting referral decisions. PCP, primary care physician; WF, wide field; *t*-SNE, *t*-distributed stochastic neighbor embedding.

dermatologists (R.R.S., C.C.K., and M.M.S.). The skin edge, bare skin, NSPL-A, NSPL-B, and SPL classes encompassed a range of Fitzpatrick skin tones (types I to VI) (table S2). Nonlesion-related classes (backgrounds, skin edge, and bare skin) were included in this specific dataset to allow for the training of DCNN models capable of discriminating pigmented lesions from other features commonly observed in wide-field dermatological images.

DCNN training, validation, and testing

We trained various DCNN models for SPL classification to assess differences in performance resulting from architecture and data preparation strategies (Fig. 3). First, we trained a baseline DCNN model with three convolutional layers (fig. S2, A and B) using our curated six-class dataset ($n_{baseline}$ = 33,980) with a randomized percentage split for training (60%, n_{train} = 20,388), validation (20%, n_{val} = 6796), and testing (20%, n_{test} = 6796) (fig. S3). In the testing set n_{test}, 71.5% of lesions corresponded to melanoma, whereas 14.0 and 14.5% corresponded to basal and squamous cell carcinomas, respectively. When evaluated on the testing set using receiver operator curves (ROCs) with a one-versus-all binarization strategy for each class, this model reached a micro-averaged area under the curve (AUC_{micro}) across all six classes of 0.975 (95% confidence interval, 0.896 to 0.988) [sensitivity = 0.890 (0.885 to 0.895), specificity = 0.899 (0.894 to 0.904), and accuracy = 84.62% (83.8 to 85.4%)], with an SPL AUC_{spl} = 0.945 (0.940 to 0.949) (Fig. 3, A and B). Similarly, this baseline DCNN architecture was also trained on a ~10× nonoverlapping augmented dataset with class balancing (n_{aug} = 300,000) (fig. S4). This baseline model with data augmentation showed a slight reduction in measured performance with AUC_{micro} = 0.957 (0.956 to 0.958) [sensitivity = 0.878 (0.771 to 0.881), specificity = 0.892 (0.780 to 0.895), and accuracy = 78.42% (78.1 to 78.7%)] and AUC_{spl} = 0.911 (0.909 to 0.913) (Fig. 3, C and D), likely due to reduced model overfitting.

Given that the augmentation and class balancing constitute desirable approaches to improve DCNN generalization capacity, we decided to use this dataset to train all further models. A third DCNN was then trained using transfer learning from the VGG16 ImageNet pretrained network (*33*) in conjunction with our six-class augmented dataset (fig. S5). This approach leverages ImageNet's 14-million image dataset grouped into 21,841 classes (*34*, *35*) to extract visual features and facilitate classification in situations where reduced amounts of images are available. For this architecture, using a bottleneck training strategy for feature extraction and weight adjusting of the last fully connected layers (fig. S5, A and C), our model reached

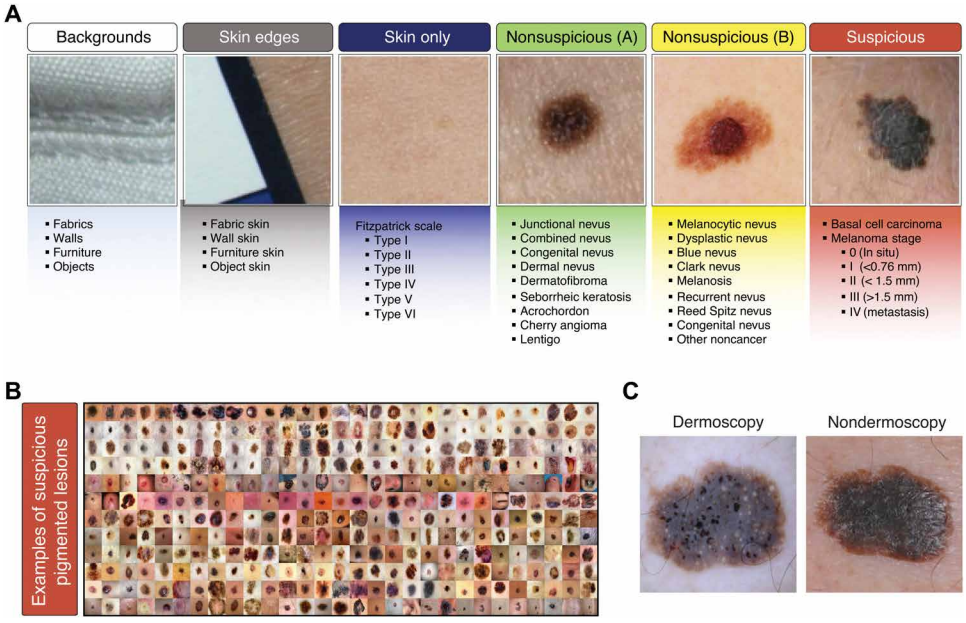


Fig. 2. Database taxonomy and an example set of the dataset. (**A**) Our database included a total of 33,980 images divided into six classes. These classes included backgrounds ($n_b = 8888$), skin edges ($n_{se} = 2528$), skin ($n_{sk} = 10{,}935$), NSPL-A ($n_{nspl\text{-}a} = 10{,}759$), NSPL-B ($n_{nspl\text{-}b} = 1110$), and SPLs ($n_{spl} = 4063$). Our background dataset included a variety of fabrics, furniture, and other common objects present in evaluation rooms. The skin dataset included crops with Fitzpatrick skin tones types I to VI. NSPL-A included images from six distinct skin lesion subtypes, where no particular management is generally indicated. NSPL-B included images from the other seven skin lesion subtypes where follow-up is often indicated to assess evolution. (**B**) The SPL class was only composed of melanoma (stages 0 to IV) and basal cell carcinoma images for which biopsy or excision is recommended. (**C**) Example of pigmented lesion imaged with dermoscopy and nondermoscopy techniques.

its highest performance with $AUC_{micro} = 0.97$ (0.969 to 0.971) [sensitivity = 0.903 (0.9 to 0.906), specificity = 0.899 (0.896 to 0.902), and accuracy = 86.56% (86.3 to 86.8%)] and $AUC_{spl} = 0.935$ (0.933 to 0.937) when evaluated on the augmented testing set (Fig. 3, E and F). Fine-tuning of additional VGG16 network layers, such as the weights of convolutional block 5, was also attempted (fig. S5, B and C) but did not lead to improved classification performance as compared to the VGG16 transfer learning bottleneck (VGG16-BTF) model (fig. S6).

A fourth DCNN transfer learning model based on the ImageNet's pretrained Xception network (*35*) was also created to compare the performance of this deeper network to the VGG16 architecture in our dataset. When using a bottleneck training strategy (fig. S7, A and C), our Xception transfer learning model only reached an $AUC_{micro} = 0.858$ (0.857 to 0.858) [sensitivity = 0.837 (0.606 to 0.841), specificity = 0.770 (0.766 to 0.774], and accuracy = 61.05% (60.7 to 61.4%)] and $AUC_{spl} = 0.827$ (0.825 to 0.829) (Fig. 3, G and H) for the augmented testing set. Fine-tuning of Xception weights in earlier network layers, such as the last two groups of convolutional blocks (fig. S7, B and C), did not improve performance and led to a substantial deterioration in classification accuracy across all classes (fig. S8). The fact that, here, our baseline and VGG16-BTF networks outperformed the deeper Xception transfer learning models indicates that shallower neural networks (~16 layers) may be better suited for this specific problem and data constraints as compared with deeper networks. As compared to a previous DCNN system with an average accuracy of 72.1% for suspiciousness-like classification endpoints (malignant versus nonneoplastic and benign) (*25*), our VGG16-BTF model reached 79.9% average accuracy, with the advantage of including mostly nondermoscopy training images, which are more representative of dermatological screening events at the primary care level than dermoscopy.

SPL and ugly duckling computer-aided identification

In this section, we present a proof-of-concept demonstration of our integrated DCNN SPL and ugly duckling computer-aided identification system (Fig. 1), which aims to allow for rapid detection and ranking of pigmented lesions according to their levels of suspiciousness in wide-field images (Fig. 4A). In this system, our VGG16-BTF model was used to extract features and calculate patient-independent probabilities of suspiciousness for each pigmented lesion similarly to previous DCNN dermatological tools. However, in our implementation, the extracted features were used in a secondary stage to calculate a quantitative ugly duckling metric based on the geometric distance (cosine) of each lesion's feature vector compared to the averaged feature center of all visible lesions in a particular patient-specific wide-field image. Such DCNN-based "oddness" lesion ranking can be interpreted as a field-of-view patient-dependent metric that can be normalized and presented in lesion montages and ugly duckling heatmaps, as shown in Fig. 4B. This constitutes the first-reported quantifiable definition of the ugly duckling criteria and serves as a way to leverage deep learning networks to overcome the challenging and time-consuming task of characterizing the fine-grained disparities among all the pigmented lesions in a single patient.

We evaluated our ugly duckling scoring method using 135 wide-field dermatological images from 68 individuals depicting large body parts (arm, full back, and full stomach) in comparison with the assessment of three board-certified dermatologists (R.R.S., C.C.K., and M.M.S.) tasked with ranking lesion oddness (Fig. 4, C to E). In these wide-field images, the number of detected lesions considered for analysis ranged from 5 to 239, spanning a wide range of lesion counts that includes the typical number of lesions seen in most dermatological patients (*36*). We measured the percent agreement between our ugly duckling algorithm and the dermatological consensus (average ranking of three expert dermatologists) by considering at least one common lesion between the predicted (top-u) and the expert-identified (top-k) list of lesions as a successful assessment. Under this definition of agreement, when selecting a top three ugly duckling algorithm ranking ($u = 3$), as compared to a top 10 dermatological consensus ranking ($k = 10$), we found a 96.3% (83.67 to 97.57%) agreement for all evaluated wide-field images (Fig. 4C). A more conservative and clinically relevant agreement of 82.96% (67.88 to 88.26%) was found when selecting for $u = 3$ and $k = 3$ as ranking parameters in Fig. 4D. Furthermore, we also explored the effect of selecting more restrictive agreement definitions by substituting the original ≥ 1 common agreement rule with an ≥ 2 or ≥ 3 common lesion requirement

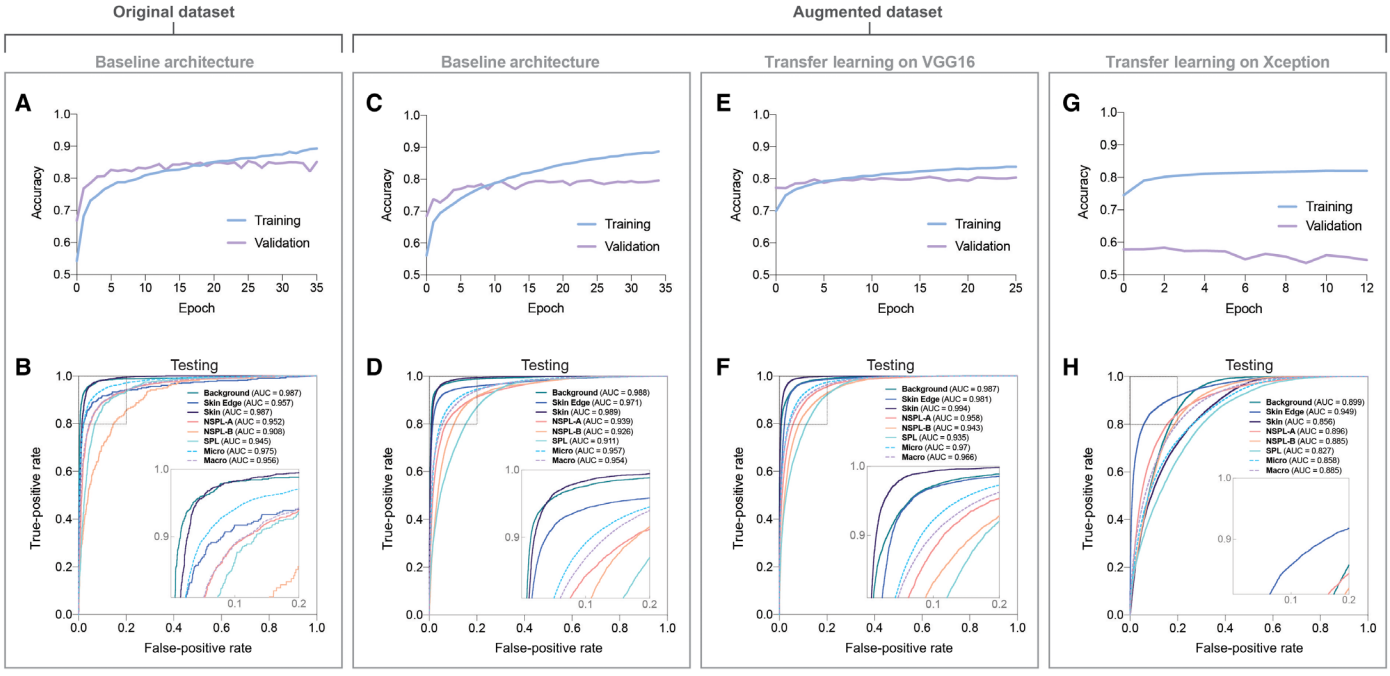


Fig. 3. Training, validation, and testing of core explored DCNN models. The top row show recorded accuracy on training and validation sets per epoch for the original dataset with baseline architecture (**A** and **B**) and a 10× augmented dataset with baseline architecture (**C** and **D**), transfer learning on VGG16 (**E** and **F**), and on Xception (**G** and **H**). The bottom row shows multiclass receiver operation curves (ROCs) for the testing set showing true-positive rate against false-positive rate per class on each of the corresponding dataset architecture. A "macro" average is presented computing the simple aggregated average of all ROCs, as well as a "micro" average, which aggregates such contributions considering any class imbalances. AUC (area under the curve) values are presented for each curve, with higher AUCs denoting better performance. After training and validation of all models, the highest performing trained architecture (VGG16 with transfer learning) was selected as the basis for full system integration and subsequent use.

for a $u = 3$ and $k = (1$ and $10)$ parameter sweep (Fig. 4D). Although the overall agreement was high (>82.96%) for all values of $k > 3$ when considering our one common lesion criteria, it dropped considerably as we increased this requirement. This indicates that although our algorithm is reliable in assessing the oddest lesions in every image, the predicted lower-ranking lesions do not correlate as strongly with the consensus. Last, we compared our ugly duckling algorithm with the dermatological consensus and the individual dermatologists across multiple ranking parameter options by calculating the normalized volume under the surface (VuS) agreement values for all top-k versus top-u parameter sweeps (Fig. 4E). In this analysis, our ugly duckling algorithm appeared to emulate the majority fraction of the expert dermatological consensus VuS (0.88), as well as a majority fraction of the assessment performed by any given individual dermatologists (0.860 ± 0.026). The VuS fractions of each individual dermatologist as compared with every other dermatologist were 0.936 ± 0.026. Because the dermatological consensus is itself derived from the averaged rankings of the three evaluating experts, comparing each individual dermatologist against the consensus would not be appropriate and, consequently, is not included in the performance matrix of Fig. 4E. A summary of all SPLs and ugly duckling identification outputs with varying degrees of performance are provided as part of fig. S10. A complete montage of analysis outputs for the 135 wide-field clinical images is also provided in data files S1 and S2.

DISCUSSION

In this work, we demonstrated a computer-aided system for evaluating the suspiciousness of pigmented skin lesions from wide-field images containing dozens to hundreds of pigmented skin lesions. Although there have been many recent examples on detection and classification of dermatological images using computer vision, these are typically implemented to analyze lesions individually with limited interlesion context. In standard clinical assessments, however, dermatological inspections usually consider various visible lesions to generate a suspiciousness assessment that informs closer inspection or biopsy. Our DCNN-based system has a design focus on primary care use and formalizes the ugly duckling saliency metric as part of its implementation with high agreement with expert dermatologists. If widely distributed, such a system could reduce the need for nondermatologists to manually select lesions that need to be inspected more closely by an expert. We accomplished this by generating wide-field lesion suspiciousness maps designed to efficiently inform primary care providers on referral decisions. Our system was also optimized to operate with nondermoscopy wide-field images, acquired with consumer-grade cameras, to effectively distinguish suspicious from nonsuspicious lesions with complex backgrounds, unevenly illuminated skin, and over large body areas. Our results suggest that such a dermatology support system could be used to rapidly assess patients with hundreds of lesions in a single visit reducing human intervention. This potential efficiency in time and human resources constitutes an advantage over previously reported DCNN systems in dermatology that require single-lesion imaging. Furthermore, although CAD systems have been reported to use consumer-grade cameras and be capable of distinguishing melanoma in both dermoscopy (*37*–*41*) and nondermoscopy (*42*–*45*), with sensitivities and specificities ranging from 77 to 98% (*46*), such previous results are all based on highly unbalanced training datasets

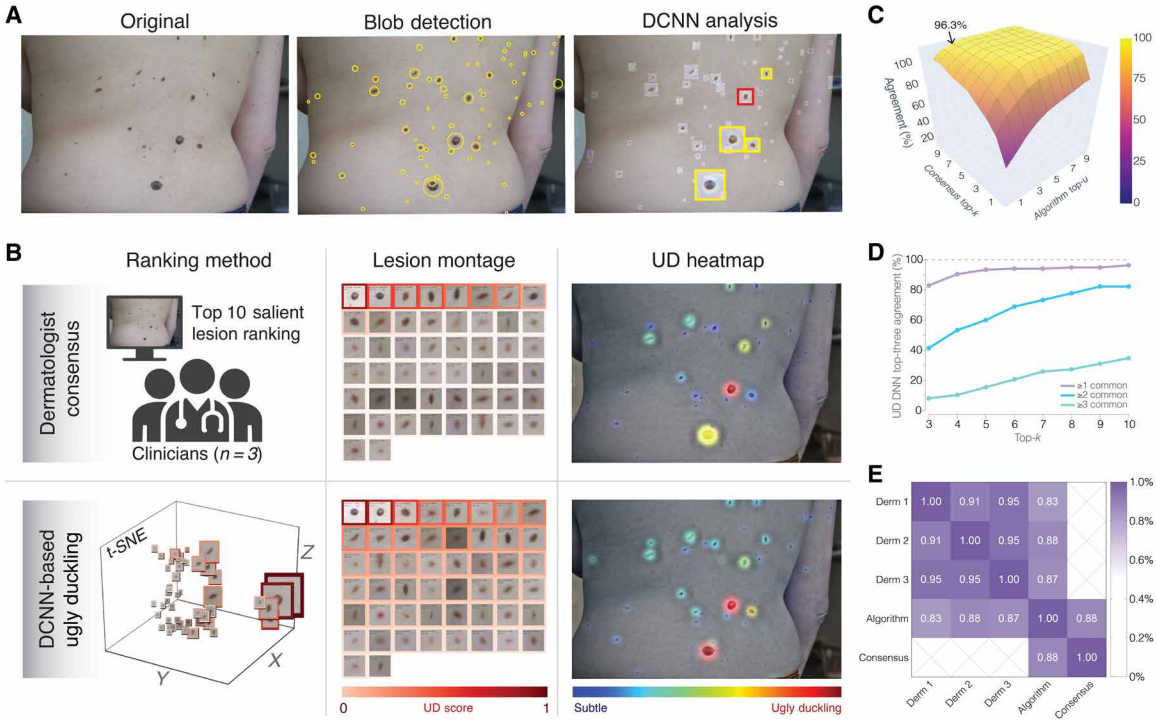


Fig. 4. DCNN system for SPL and ugly duckling identification using wide-field images. (**A**) Example wide-field image with multiple pigmented lesions on the back of a female subject. A SIFT-based blob detection algorithm provides key points at multiple scales for localization and cropping of single-lesion images. DCNN is used on each rescaled single-lesion image for class inference, and the activation map is overlaid over the original wide-field image. Pigmented lesions classified as NSPL-B are marked with yellow, whereas those classified as SPL are marked in red. (**B**) Comparison of intrapatient lesion ranking and ugly duckling (UD) heatmaps generated from dermatological consensus and from DCNN extracted features. The *t*-SNE graph shown visually represents the clustering of all lesions in the field of view for the user. Color-coded pigmented lesion montages and UD heatmaps are shown for the consensus and DCNN-based scorings. (**C**) Sweep surface of percentage agreement for allowed *n* (algorithm-dependent) and *k* (dermatologist-dependent) rank values in the 1 to 10 range, with the indicated numerical accuracy for a top-$u = 3$ ranking and top-$k = 10$ with at least one common lesions between the two ranked lists (96.3%). (**D**) Percentage agreement for top-$u = 3$ when considering at least one, two, or all three lesions to be common. (**E**) Pairwise compilation of normalized VuS values for each *k*, *n* sweep surface for individual dermatologists, the dermatological consensus, and the DCNN UD algorithm. The dermatological consensus is derived from the averaged rankings of the three evaluating experts. Matching with at least one common lesion per set is considered to constitute a true-positive sample for this calculation. The normalized agreement of our DCNN UD algorithm compared with individual dermatologists was 0.86 ± 0.03, and 0.88 as compared with the dermatological consensus.

with few malignant pigmented lesions (usually $n < 100$) and under the assumption that physicians are able or willing to perform single-lesion image acquisitions, limitations that we have improved upon or addressed in this research. Our best performing VGG-BTF model demonstrated a performance $AUC_{micro} = 0.97$ and all-class accuracy = 79.94% after being trained mostly on nondermoscopy images.

Although our algorithmic approach and study are both promising, there are several limitations that can be improved in future work. First, because of the nature of our classification task is based on suspiciousness as opposed to malignancy, we selected the consensus of three board-certified dermatologists as our ground truth. However, our dataset could potentially benefit from the addition of biopsy information to evaluate whether our algorithm missed any malignancies and to best compare our system to other CAD systems in dermatology focused on melanoma detection. Another aspect that could be improved is the scope of the dataset to cover a larger variety of acquisition strategies including different cameras, settings, and photographers. In this sense, additional training data from more sites could allow for improved generalizability potential in our models across a wider variety of environments and conditions. This could also allow us to more explicitly formulate an out-of-distribution test set for our investigated classification task, which is now derived from a held-out image subset coming from a similar distribution as our training and validation sets. In addition, the performance of the system might be affected and, therefore, could be better characterized by using it with extreme imaging conditions such as low light, out-of-focus images, and possibly larger imaging distance. Although we realize that the use of different imaging hardware and our integrated CAD system needs more validation in clinical settings, several of our design decisions align well with the expected clinical workflow of these tools. In particular, the capacity of our CAD system to run multilesion analysis for a wide-field image could be helpful in the task of passively running SPL screening during primary care visits and other type of consultations. Furthermore, by leveraging multiple wide-field images our system could be adapted to implemented redundancy as well as to incorporate, pose estimation and three-dimensional mapping to improve traceability of lesions in clinical practice. Although future work needs to be conducted to address these limitations and the mentioned clinically relevant improvements, the presented results suggest that deep learning systems adapted for wide-field analysis are a feasible and potentially attractive approach to provide full-body dermatological triaging of suspicions pigmented lesions for primary care settings.

Complete skin screenings typically consist of large body surface examinations by certified dermatologists. As part of these evaluations, clinicians recognize and compare a variety of lesion features, including asymmetry, border unevenness, color distribution, diameter, evolution according to ABCDE criteria, and multilesion saliency often referred to as the ugly duckling feature. Comparing fine-grained similarities in these lesions is a challenging task that requires the extraction of between-class and within-class patterns. Along with these examinations, a handheld dermoscope is often used to allow the physician to observe in detail salient cutaneous lesions to improve clinical diagnostic capacity. The lesions that are deemed suspicious for skin cancer are typically biopsied and sent for histopathologic evaluation, which is still considered the gold standard in melanoma diagnosis. Existing shortages in the dermatology workforce in recent years have often led to substantial wait times for patients seeking dermatologic care in the United States (*47*) and substantial bottlenecks in care for all patients, including patients with SPLs (*48*). Increasing the referral volume of low-risk patients via indiscriminate screening events would only exacerbate these difficulties in access. Therefore, in the context of dermatological evaluations, tools that facilitate identification of patients with SPLs in primary care settings such as the one presented in this work could allow for optimization of referrals and improved triaging. In particular, the use of wide-field photography for automatic SPL identification becomes an attractive option for quick assessment of high-risk patients, assisting nonexperts in the correct identification of SPLs to refer patients to an expert dermatologist or conduct a biopsy.

From a policy perspective, implementing large-scale melanoma screening programs is not only likely to be a complicated task but rather an infeasible one in most resource-constrained healthcare systems around the world. In the United States, for example, there are less than 12,000 practicing dermatologists (*49*), and with fewer than 15 visits per 100 individuals annually (*50*), it is expected that most dermatology practices across the nation are already too saturated and time-constrained to provide additional screening services. Unlike dermatologists, primary care physicians such as family practitioners and internists already attend to about 330 million patients per year in the United States (*50*). These substantial coverage rates place primary care providers in a prime position to execute meaningful melanoma screening programs in large cohorts of patients (*51*). Unfortunately, most of the providers are now not trained to perform pigmented lesion assessments (*52*) and tend to have short turnaround times in which to examine patients for many high-priority diseases (*53*, *54*). This situation has led to reports concluding poor diagnostic and referral accuracy for providers conducting direct visual assessments in melanoma screenings (*55*–*57*). Considering the reach of primary care providers, convenient and scalable tools for SPL detection at the primary care level could increase appropriate dermatological referrals and earlier treatment for patients with melanoma. To address this need, here, we have presented an automated deep-learning classification system capable of identifying and ranking suspicious pigmented skin lesions from wide-field images, which could allow for rapid melanoma screenings during primary care visits. We have selected this data flow and analysis modality to enable fast evaluations of multiple lesions within large skin regions at the primary care, with minimal equipment or training. This wide-field DCNN classification strategy allows our system to overcome a variety of challenges regarding the differentiation of pigmented lesions from base skin and complex backgrounds, as well as to provide a holistic analysis of patient's risk for melanoma to guide dermatological referral and biopsies. DCNN systems such as that developed in this work can be used to extract information in pigmented lesions images, which outperform handcrafted visual features such as ABCDE criteria for the support of a variety of classification tasks as seen in other systems (*58*). Furthermore, the use of CNNs promises superior classification robustness compared with traditional image pigmented lesion classification methods, even in the presence of obstructions, shadows, and geometric and chromatic aberrations.

Our approach is intended to improve the likelihood that a large number and variety of SPLs and NSPLs are evaluated in a single visit. The intention of these classification systems has, in general, been to help dermatologists differentiate among borderline lesion diagnosis and to assist nondermatologists with faster access to specialists through teleconsultation. However, to maximize sensitivity and avoid missing melanomas, the vast majority of lesions on a high-risk patient should be assessed. These thorough evaluations now require a substatantial time investment if not using specialized full-body imaging tools. Because of substantial time constraints in primary care practice, the use of computer-based melanoma screening in a nonexpert setting has been limited exclusively to research settings. Despite the importance of these results, a recently proposed model (*25*) was implemented and trained to assume a use case where suspicious lesions were preselected by the observer and then imaged individually for analysis. Moreover, in this implementation, the probability of malignancy and therefore suspiciousness is only determined at the single-lesion level without considering interlesion dependencies. To address this challenge of SPL screening in uncontrolled settings, we constructed a deep learning model and computer-aided system capable of processing wide-field skin examinations without the need to excessively burden primary care physicians or technicians with time-consuming tasks such as lesion localization, image segmentation, and preliminary classification.

Despite the promising performance presented by our proof-of-concept system for wide-field SPL identification, the generated implementation holds various limitations. For instance, although we observed acceptable behavior from the blob detector used in our test samples, particularly during ugly duckling evaluations, the accuracy of said blob detection framework was not directly evaluated here. However, considering that the blob detector constitutes a widely used implementation of the scale-invariant feature transformation (SIFT)–Laplacian of Gaussians (LoG) algorithm in the field of computer vision, available from OpenCV libraries used by hundreds of thousands of researchers worldwide, it is expected to be robust. To mitigate this potential uncertainty, all the detected and analyzed blobs in wide-field images for ugly duckling analysis are provided in the Supplementary Materials to allow for direct analysis of these outputs by the reader. Furthermore, in our system, the established filtering parameters extract most blob-like points from the field of view, all of which were ultimately analyzed and confirmed to be pigmented lesions (NSLP-A, NSPL-B, and SPL) by the selected DCNN inference model, minimizing the number on unevaluated lesions from this stage.

MATERIALS AND METHODS

Study design

The overall goal of our study was to demonstrate the feasibility of detecting and classifying suspicious pigmented skin lesions from

wide-field input images. We focused on the use of deep neural networks to perform both single-lesion classification and outlier (ugly duckling) detection based on the extracted features from all visible lesions in a patient wide-field image. We first examined the effect of using data augmentation on the accuracy of a simple convolutional neural network architecture. Next, we explored potential improvements in accuracy by leveraging transfer learning in various architectures with higher capacity than the baseline model. Then, we selected the best-performing network and evaluated its use as an integrated multilesion detector and feature extractor in wide-field images for ugly duckling detection, as to compared with the consensus of three certified dermatologists. Throughout the study, we exploited a need-driven approach for the design of our computer-aided identification system for primary care use, along with an independent testing set randomly selected before any data augmentation or analysis. Sample size of our single-lesion dataset and wide-field images for ugly duckling evaluation was determined on the basis of availability, allowing for accuracy and agreement calculations. Conclusions were drawn on the basis of multiclass ROCs and agreement plots with the measured dermatological consensus.

Dataset compilation and image acquisition

Our dataset consisted of n_{baseline} = 33,980 manually curated images from various sources including publicly available atlases of pigmented lesions (*59*–*61*), single-lesion and wide-field images collected via web scraping in conventional search engines (Google, Yahoo, and Bing) using QImageScraper version 1.4 (https://github.com/stereomatchingkiss/QImageScraper), and nonoverlapping image crops from wide-field dermatological images from 133 individual patients recruited at Hospital Gregorio Marañón (Madrid, Spain). A STARD (Standards for Reporting Diagnostic Accuracy) diagram depicting the image collection process and sample distribution for this study is shown in fig. S1. Lesion images collected from wide-field images were curated using crops obtained using a multilevel nonoverlapping sliding-window process (maximum window size = 299 × 299 and maximum horizontal padding = 150 and maximum vertical padding = 299 pixels) applied over the collected wide-field images. This process generated unique image sections or crops that were then divided into six categories as follows: 8888 backgrounds, 2528 skin-edge images, 10,935 bare-skin patches, 10,759 low-priority NSPL-A, 1110 medium-priority NSPL-B, and 4063 high-priority SPLs. SPLs included 568 basal cell carcinomas, 589 squamous cell carcinomas, and 2906 melanomas (90% nondermoscopy). For all pigmented-lesion classes, dermoscopy and nondermoscopy images were included to generalize the analysis of both imaging modalities. Nondermoscopy images were primarily obtained using a wide-field technique, defined here as the acquisition of an image including at least one pigmented lesion using a personal camera or smartphone, and were taken at least 10 cm away from a patient. Wide-field images, including multiple lesions, bare-skin, and nonskin regions, were cropped at multiple scales, with each crop placed in its respective class. Single-pigmented lesions in wide-field images were cropped, considering a 1:3 ratio between the lesion's average radius and its surrounding skin region. This cropping process was assisted by the standard blob detection algorithm on the basis of a SIFT, using LoG according to the specifications presented in fig. S9. All images within the pigmented-lesion classes in the database were then independently evaluated by three board-certified dermatologists (R.R.S., C.C.K., and M.M.S.). Lesion classifications differing among raters were resolved by consensus.

From the 15,932 images corresponding to pigmented lesions, a subset of 4800 single-lesion images was extracted from 600 clinical wide-field images from 133 consenting patients evaluated at the Gregorio Marañón Hospital (Madrid, Spain), in collaboration with the Massachusetts Institute of Technology (MIT). The clinical images obtained at the Gregorio Marañón Hospital were captured using an Olympus E-420 camera (10 M pixels, 14- to 42-mm lens) at a distance of 0.2 m from the patient and anonymized before processing and analysis. The camera was operated by an expert dermatologist specialized in melanoma (J.A.-I.) while performing full-body skin examinations. Illumination was not controlled during image acquisition, and any artifacts present in these images were not corrected before analysis. At a 20-cm distance using this camera, the image pixel size was confirmed to be around 67 μm using a positive United States Air Force (USAF) 1951 resolution test target (Thorlabs). This resolution is comparable to high-end smartphone cameras and eye reading/viewing resolution at the same distance (~58 to 72 μm at 0.2 to 0.25 m from visual target). A distance range between 0.2 and 0.5 m also corresponds to the approximate distance at which lesions are visually evaluated by expert dermatologists during full-body examinations. All other lesions and nonlesion images were obtained from online dermatological image repositories, published pigmented lesion atlases, and online scraping. Once compiled, all these images were verified as an integrated image corpus by three board-certified dermatologists.

Human subjects

The images acquired from patients at the Department of Dermatology Gregorio Marañón Hospital were obtained under a clinical protocol (promoter: ILP_AP_HGUGM v1/code: 126/16) and reviewed and approved by the Ethical Committee from Hospital General Universitario Gregorio Marañón (Madrid, Spain) and Committee on the Use of Humans as Experimental Subjects from the Massachusetts Institute of Technology (Cambridge, MA, USA) under reference no. 1501006861. Inclusion criteria for participants were signed informed consent, aged greater than 18 years old, and assured mental integrity. Exclusion criteria were marks on the subject's body that would prevent full anonymization.

Data taxonomy

Background images (class 0) included various types of fabric patches, furniture, and walls. Skin-edge images (class 1) consisted of skin-background intersection crops manually selected from dermatological wide-field photographs. Skin images (class 2) included crops from dermoscopy, near-field, and wide-field photographs of patients with Fitzpatrick tones I to VI. Skin images in this class included hair obstructions, folds, wrinkles, freckles, nail sections, nonpigmented formations, as well as nonhomogeneous illumination and other artifacts. NSPL-A images (class 3) included low-priority pigmented lesions with a high likelihood of being benign. Lesions in this category include benign melanocytic nevi, dermal nevi, junctional nevi, combined nevi, congenital nevi, seborrheic keratoses, acrochordons, cherry angiomas, dermatofibroma, and lentigo. Because of their low probability of being skin cancer, NSPL-A lesions can often be evaluated by nondermatologist primary care providers to visually confirm this low-priority classification, particularly in places where access to expert dermatologic care is limited. NSPL-B images (class 4) include medium-priority pigmented lesions that should be tracked over time or considered for biopsy. Lesions in

this category are unlikely to be skin cancer; however, given constraints in input image quality and classification accuracy, there is reasonable doubt that prevents this pigmented lesion from being wholly excluded from malignancy diagnosis. An in-person evaluation with dermoscopy is recommended for patients with these types of lesions, which should be prioritized secondarily to those with high-priority lesions. Lesions included in this category include melanocytic, dysplastic, blue, Clark, recurrent, Reed Spitz, and congenital nevi. SPL images (class 5) include high-priority SPLs with features indicative of skin cancer. Urgent referral to a dermatologist is recommended for further evaluation, because it is likely that a biopsy will be necessary for definitive diagnosis and management. Patients presenting with SPLs should be given the highest priority within the constraints of an existing system of care. Lesions commonly diagnosed in this category are melanoma, melanoma in situ, basal cell carcinoma, and squamous cell carcinoma.

Data preprocessing

Images used for model training were preprocessed to ensure they were all in Portable Network Graphics (PNG) format, cropped, and rescaled to standard input size of 299 × 299 pixels. Preprocessed images were then transformed from RGB (red, green, and blue) to hue, saturation, and value representation to adjust the V channel via contrast-limited adaptive histogram equalization (CLAHE) (*62*) and then transformed back to RGB space. This ensures consistent individual image contrast and normalizes illumination across images in the base dataset (fig. S3). Rescaling of images to different input sizes depending on the requirements of the selected DCNN architecture (150 × 150 pixels) was carried out at run time during training or inference without any further modification from this base dataset. Randomized transformations for data augmentation included rotations (0° to 30°), horizontal flips, vertical flips, horizontal shifts (0 to 10%), vertical shifts (0 to 10%), color-channel shifts (0 to 10%), and zooming (0 to 20%) (fig. S4). If transformations required pixel filling to match the original image size, then the nearest pixel reflection was used. Randomized dataset splitting into training (60%), validation (20%), and testing (20%) sets were performed before data augmentation to prevent overlap of augmented images across training and testing sets.

Baseline DCNN model

The baseline DCNN model (fig. S2, A and B) was trained using the collected dataset with CLAHE preprocessing and random split into training (60%), validation (20%), and testing (20%) sets (fig. S3). The architecture of this network consisted of three sets of convolutional layers with rectified linear unit (ReLU) activation, followed by corresponding max-pooling layers. A flattening layer and a dense layer with ReLU activation were then applied before a final dropout layer (p_{drop} = 0.5). The output of this network was then connected to a final six-neuron dense layer (one neuron per class) with Softmax activation to output the final class probability vector of the DCNN (fig. S2, A and B). Loss was calculated using categorical cross-entropy, with a standard Adam optimizer (learning rate = 0.001, β_1 = 0.9, β_2 = 0.999, and $\varepsilon = 1 \times 10^7$, no AMSGrad) and accuracy across all categories as the metric to evaluate performance. No other model hyperparameter was manually tuned. The training process began by training for a total of 100 epochs. The validation subset was used during every training epoch to assess and prevent overfitting through an early stopping DCNN training routine. Thus, by epoch 35 of this training process, the algorithm detected that accuracy was not changing for more than 10 consecutive iterations, activating an early-stopping callback to avoid overfitting (Fig. 3A). ROCs are presented for all six classes (Fig. 3B), as well as the equally aggregated ROC average (macro) and the aggregated ROC average weighting for the individual contributions of each class (micro). Therefore, the micro-averaged ROCs result from adjusting the class-aggregated macro-averaged curves by the relative number of images from each class presented during testing.

Baseline DCNN model with data augmentation and class balancing

The same baseline DCNN model architecture was also trained using an augmented dataset with 300,000 images separated into six balanced classes that were created through random transformations (for example, rotation, scaling, and translations). By applying this randomized transformation scheme (fig. S4A), we balanced the classes and artificially enhanced our training set to be trained with a larger number of unseen images (fig. S4B). Such database augmentation was performed after randomly separating into training (60%), validation (20%), and testing (20%) sets, which was intended to reduce overfitting and allow better generalization capability for the networks trained with this augmented dataset. By epoch 35, the algorithm detected that accuracy was not changing for more than 10 consecutive epochs, activating an early-stopping callback used to avoid overfitting (Fig. 3C). ROCs are presented for all six classes (Fig. 3D), as well as the equally aggregated ROC average (macro) and the aggregated ROC average weighting for the individual count contributions of each class (micro).

VGG16 transfer learning DCNN model

Our VGG16 transfer learning model (fig. S5) was trained using the balanced and augmented dataset (fig. S4A). For the bottleneck transfer learning approach (fig. S5A), we fixed the first 15 layers of the pretrained VGG16 network to retrain the final flattened, dense, dropout, and SoftMax activation layers for the desired six classes. Loss was calculated using categorical cross-entropy, a standard Adam optimizer (learning rate = 0.001, β_1 = 0.9, β_2 = 0.999, and $\varepsilon = 1 \times 10^7$, no AMSGrad) and accuracy across all categories as the metric to evaluate performance. No other model hyperparameter was manually tuned. The training began by instructing a total of 100 epochs. The validation subset was used during every training epoch to assess and prevent overfitting through an early stopping DCNN training routine. Thus, by epoch 25 of this training event, the algorithm detected that accuracy was not changing for more than 10 consecutive iterations, activating an early-stopping callback to avoid overfitting, and converging at an accuracy value (Fig. 3E). ROCs are also presented for all six classes (Fig. 3F), as well as the equally aggregated ROC average (macro) and the aggregated ROC average weighting for the individual count contributions of each class (micro).

Xception DCNN transfer learning model

The Xception transfer learning model (fig. S7) was trained using the balanced and augmented dataset (fig. S4A). For the bottleneck transfer learning approach (fig. S7A), we fixed the first 126 layers of the pretrained Xception network (a Keras improvement on Google's Inception v3 from ImageNet) to retrain the final convolutional and dense layers for the desired classes. Loss was calculated using categorical cross-entropy, with a standard Adam optimizer (learning

rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1 \times 10^7$, no AMSGrad) and accuracy across all categories as the metric to evaluate performance. No other model hyperparameter was manually tuned. The training began by instructing a total of 100 epochs. The validation subset was used during every training epoch to assess and prevent overfitting through an early stopping DCNN training routine. Thus, by epoch 12 of this last training event, the algorithm detected that accuracy was not changing for more than 10 consecutive iterations, activating an early-stopping callback to avoid overfitting, converging at 61.05% accuracy (Fig. 3G). ROCs are also presented for all six classes (Fig. 3H), as well as the equally aggregated ROC average (macro) and the aggregated ROC average weighting for the individual count contributions of each class (micro).

DCNN system for SPL identification

In our DCNN SPL computer-aided identification system, a wide-field image from the patient's body acquired at the time of the visit is fed to the algorithm (Fig. 4A). This image is then processed by a blob detection algorithm on the basis of a SIFT, using LoG according to the specifications presented in fig. S9. This routine detects all blob-like regions (Fig. 4A) in the image that resemble a pigmented lesion for classification or discrimination. Nonoverlapping image patches are then cropped around all detected regions and centered at each key point with at least a 50% margin from the blob diameter (d), with crop dimensions: (height = $d \times 2$) × (width = $d \times 2$). These square crops derived from the original wide-field image are then stored in an intra-patient database for further analysis. Once stored, each single-blob cropped image is rescaled to a suitable size (150 × 150 × 3 pixels) to be classified using our VGG16-TF model. All pigmented lesions confirmed to contain skin (NSPL-A, NSPL-B, and SPL) were labeled and color-coded according to their single-lesion class probabilities generated by the DCNN. All pigmented lesions confirmed by the DCNN-based algorithm are placed into a secondary database to calculate our ugly duckling criteria through saliency assessment and ranking.

DCNN ugly duckling score definition

Given that our system was designed to use wide-field images to guide patient referral, we propose to improve the evaluation of pigmented lesions by considering both the patient-independent probability of each lesion being malignant, as well as the ugly duckling criteria mostly overlooked by other DCNN-based systems. Here, we define the ugly duckling criteria as the patient-dependent probability of each lesion being suspicious given its disparities to all other observable lesions in the wide-field image. Such disparities can be measured and scored using naïve features (*63*) extracted through rule-based saliency algorithms (fig. S9) or by leveraging the features extracted by the DCNN for all evaluated lesions. Given that DCNN features can span a high-dimensional vector space useful in image comparison tasks (*58*), we decided to use the features extracted by our trained VGG16-TF DCNN to generate scores of pigmented lesion similarity. These scores can then be combined with the DCNN classification outputs into a single suspiciousness representation or map that integrates this information for primary care physicians and other clinical personnel performing full-body SPL screenings. The saliency or ugly duckling score that enables this was calculated using a geometric distance (cosine distance) from its DCNN output feature vector with respect to the averaged geometric feature center of all observable lesions in the wide-field image. This distance was then normalized and used to overlay a ranking heatmap over all the pigmented lesions in the image ("UD heatmap" in Fig. 4B).

DCNN ugly duckling score validation

Lesions from 135 wide-field dermatological images acquired from 68 individuals were detected and assigned a numeric label using the previously described blob detection algorithm (Fig. 4A). All board-certified dermatologists were confirmed to have at least 10 years of experience assessing pigmented lesions and asked to rank up to 10 lesions by "visual oddness," starting from the oddest. The consensus was then derived by taking the average scoring of all rankings for each lesion. To quantify agreement, we compare the top-k–ranked lesions (from most to least "odd") as evaluated by the dermatologists with the top-u–ranked lesions as predicted by our DCNN-based ugly duckling algorithm. From this agreement metric, surface plots can be generated by performing parametric sweeps across different n (allowable algorithm ranking) and k (allowable consensus ranking) lists. A sample surface plot of percentage agreement for all n and k values ranging from 1 to 10 can be seen in Fig. 4C. Upon consultation with expert dermatologists, we determined that a top-three predicted ranking system ($u = 3$) with at least one common value from the top-three ranked consensus list ($k = 3$) constituted a reasonable and clinically meaningful measure of accuracy for this specific system. Extending this evaluation, we also calculated agreement scores for our system when using more conservative ranking approaches, such as a two- or three-lesion match requirement for agreement labeling (Fig. 4D). A pairwise comparison of the normalized VuS for individual dermatologists, the dermatological consensus, and our DCNN ugly duckling algorithm is provided in Fig. 4E.

Training and evaluation code

All models were trained using TensorFlow 1.13.1 and Keras 2.1.3 on a Google Cloud virtual instance with Ubuntu 16.04 operating system, 250-gigabyte SSD, 8 CPUs, 30-gigabyte RAM, 2 NVIDIA Tesla K80 GPUs, CUDA 8.0 (Nvidia-384), cuDNN 6.0, Python 3.5, and OpenCV 3.1. The code needed to reproduce the results presented here is provided in the Supplementary Materials. All code is dated, documented, and referenced using Jupyter notebooks and marked down to facilitate reproduction of these results and can be found at DOI: 10.5281/zenodo.4292573.

SUPPLEMENTARY MATERIALS

stm.sciencemag.org/cgi/content/full/13/581/eabb3652/DC1

Materials and Methods

Fig. S1. STARD diagram of data aggregation.

Fig. S2. Baseline DCNN model architecture.

Fig. S3. Preprocessing and splitting of the base database into training, validation, and testing sets.

Fig. S4. Data augmentation strategy.

Fig. S5. Transfer learning DCNN model architecture based on VGG16.

Fig. S6. Training, validation, and testing of fine-tuned VGG16 DCNN model.

Fig. S7. Transfer learning DCNN model architecture based on Xception.

Fig. S8. Training, validation, and testing of fine-tuned Xception DCNN model.

Fig. S9. Blob detection and naïve saliency calculation.

Fig. S10. Selected samples of DCNN ugly duckling outputs as compared with naïve saliency and dermatological consensus.

Table S1. Taxonomy of pigmented lesions included in our study's baseline dataset.

Table S2. Distribution of Fitzpatrick skin tones along all skin-relevant classes in the base dataset.

Data file S1. Montage of analysis outputs for wide-field images, numbers 1 to 35.

Data file S2. Montage of analysis outputs for wide-field images, numbers 36 to 70.

Data file S3. Montage of analysis outputs for wide-field images, numbers 71 to 105.

Data file S4. Montage of analysis outputs for wide-field images, numbers 106 to 135.

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. American Cancer Society, *Cancer Facts & Figures 2019* (American Cancer Society, 2019).
2. R. L. Siegel, K. D. Miller, A. Jemal, Cancer statistics, 2019. *CA Cancer J. Clin.* **69**, 7–34 (2019).
3. American Cancer Society, *Cancer Prevention & Early Detection Facts and Figures 2019–2020* (American Cancer Society, 2019).
4. G. P. Guy Jr., S. R. Machlin, D. U. Ekwueme, K. R. Yabroff, Prevalence and costs of skin cancer treatment in the U.S., 2002–2006 and 2007–2011. *Am. J. Prev. Med.* **48**, 183–187 (2015).
5. A. J. Klink, B. Chmielowski, B. Feinberg, S. Ahsan, D. Nero, F. X. Liu, Health care resource utilization and costs in first-line treatments for patients with metastatic melanoma in the United States. *J. Manag. Care Spec. Pharm.* **25**, 869–877 (2019).
6. A. Scope, A. Marghoob, The "ugly duckling" sign: An early melanoma recognition tool for clinicians and the public. *Melanoma Lett.* **25**, 1–3 (2007).
7. J. Grob, J. Bonerandi, The "ugly duckling" sign: Identification of the common characteristics of nevi in an individual as a basis for melanoma screening. *Arch. Dermatol.* **134**, 103–104 (1998).
8. J. Gachon, P. Beaulieu, J. F. Sei, J. Gouvernet, J. P. Claudel, M. Lemaitre, M. A. Richard, J. J. Grob, First prospective study of the recognition process of melanoma in dermatological practice. *Arch. Dermatol.* **141**, 434–438 (2005).
9. M. Ilyas, C. M. Costello, N. Zhang, A. Sharma, The role of the ugly duckling sign in patient education. *J. Am. Acad. Dermatol.* **77**, 1088–1095 (2017).
10. C. Gaudy-Marqueste, Y. Wazaefi, Y. Bruneu, R. Triller, L. Thomas, G. Pellacani, J. Malvehy, M.-F. Avril, S. Monestier, M.-A. Richard, B. Fertil, J.-J. Grob, Ugly duckling sign as a major factor of efficiency in melanoma detection. *JAMA Dermatol.* **153**, 279–284 (2017).
11. F. Durbec, F. Vitry, F. Granel-Brocard, D. Lipsker, F. Aubin, G. Hédelin, S. Dalac, F. Truchetet, C. Michel, M.-L. Batard, B. Domissy-Baury, J.-M. Halna, J. L. Schmutz, C. Delvincourt, G. Reuter, S. Dalle, P. Bernard, A. Danzon, F. Grange, The role of circumstances of diagnosis and access to dermatological care in early diagnosis of cutaneous melanoma: A population-based study in France. *Arch. Dermatol.* **146**, 240–246 (2010).
12. N. M. Fisher, J. V. Schaffer, M. Berwick, J. L. Bolognia, Breslow depth of cutaneous melanoma: Impact of factors related to surveillance of the skin, including prior skin biopsies and family history of melanoma. *J. Am. Acad. Dermatol.* **53**, 393–406 (2005).
13. M. L. Pennie, S. L. Soon, J. B. Risser, E. Veledar, S. D. Culler, S. C. Chen, Melanoma outcomes for Medicare patients: Association of stage and survival with detection by a dermatologist vs a nondermatologist. *Arch. Dermatol.* **143**, 488–494 (2007).
14. A. Waldmann, S. Nolte, M. Weinstock, E. Breitbart, N. Eisemann, A. C. Geller, R. Greinert, B. Volkmer, A. Katalinic, Skin cancer screening participation and impact on melanoma incidence in Germany–An observational study on incidence trends in regions with and without population-based screening. *Br. J. Cancer* **106**, 970–974 (2012).
15. C. Curiel-Lewandrowski, S. C. Chen, S. M. Swetter; Melanoma Prevention Working Group-Pigmented Skin Lesion Sub-Committee, Screening and prevention measures for melanoma: Is there a survival advantage? *Curr. Oncol. Rep.* **14**, 458–467 (2012).
16. G. Merlino, M. Herlyn, D. E. Fisher, B. C. Bastian, K. T. Flaherty, M. A. Davies, J. A. Wargo, C. Curiel-Lewandrowski, M. J. Weber, S. A. Leachman, M. S. Soengas, M. M. Mahon, J. W. Harbour, S. M. Swetter, A. E. Aplin, M. B. Atkins, M. W. Bosenberg, R. Dummer, J. E. Gershenwald, A. C. Halpern, D. Herlyn, G. C. Karakousis, J. M. Kirkwood, M. Krauthammer, R. S. Lo, G. V. Long, G. M. Arthur, A. Ribas, L. Schuchter, J. A. Sosman, K. S. Smalley, P. Steeg, N. E. Thomas, H. Tsao, T. Tueting, A. Weeraratna, G. Xu, R. Lomax, A. Martin, S. Silverstein, T. Turnham, Z. A. Ronai, The state of melanoma: Challenges and opportunities. *Pigment Cell Melanoma Res.* **29**, 404–416 (2016).
17. K. A. Freedberg, A. C. Geller, D. R. Miller, R. A. Lew, H. K. Koh, Screening for malignant melanoma: a cost-effectiveness analysis. *J. Am. Acad. Dermatol.* **41**, 738–745 (1999).
18. K. J. Wernli, N. B. Henrikson, C. C. Morrison, M. Nguyen, G. Pocobelli, P. R. Blasi, Screening for skin cancer in adults: Updated evidence report and systematic review for the US Preventive Services Task Force. *JAMA* **316**, 436–447 (2016).
19. K. Ramlakhan, Y. Shang, in *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence* (IEEE, 2011), pp. 138–141.
20. L. Ballerini, R. B. Fisher, B. Aldridge, J. Rees, in *Color Medical Image Analysis* (Springer, 2013), pp. 63–86.
21. J. S. Birkenfeld, J. M. Tucker-Schwartz, L. R. Soenksen, J. A. Avilés-Izquierdo, B. Marti-Fuster, Computer-aided classification of suspicious pigmented lesions using wide-field images. *Comput. Methods Programs Biomed.* **195**, 105631 (2020).
22. A. Masood, A. Ali Al-Jumaily, Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms. *Int. J. Biom. Imaging* **2013**, 1–22 (2013).
23. B. Rosado, S. Menzies, A. Harbauer, H. Pehamberger, K. Wolff, M. Binder, H. Kittler, Accuracy of computer diagnosis of melanoma: A quantitative meta-analysis. *Arch. Dermatol.* **139**, 361–367 (2003).
24. M. Burroni, R. Corona, G. Dell'Eva, F. Sera, R. Bono, P. Puddu, R. Perotti, F. Nobile, L. Andreassi, P. Rubegni, Melanoma computer-aided diagnosis: Reliability and feasibility study. *Clin. Cancer Res.* **10**, 1881–1886 (2004).
25. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
26. T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, S. Fröhling, J. S. Utikal, C. Kalle, A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *Eur. J. Cancer* **111**, 148–154 (2019).
27. J. Zhang, Y. Xie, Y. Xia, C. Shen, Attention residual learning for skin lesion classification. *IEEE Trans. Med. Imaging* **38**, 2092–2103 (2019).
28. Y. Fujisawa, Y. Otomo, Y. Ogata, Y. Nakamura, R. Fujita, Y. Ishitsuka, R. Watanabe, N. Okiyama, K. Ohara, M. Fujimoto, Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *Br. J. Dermatol.* **180**, 373–381 (2019).
29. X.-Y. Zhao, X. Wu, F.-F. Li, Y. Li, W.-H. Huang, K. Huang, X.-Y. He, W. Fan, Z. Wu, M.-L. Chen, J. Li, Z.-L. Luo, J. Su, B. Xie, S. Zhao, The application of deep learning in the risk grading of skin tumors for patients using clinical images. *J. Med. Syst.* **43**, 283 (2019).
30. Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, A deep learning system for differential diagnosis of skin diseases. *Nat. Med.*, 1–9 (2020).
31. J. D. Jensen, B. E. Elewski, The ABCDEF Rule: Combining the "ABCDE rule" and the "ugly duckling sign" in an effort to improve patient self-screening examinations. *J. Clin. Aesthet. Dermatol.* **8**, 15 (2015).
32. A. Scope, S. W. Dusza, A. C. Halpern, H. Rabinovitz, R. P. Braun, I. Zalaudek, G. Argenziano, A. A. Marghoob, The "ugly duckling" sign: Agreement between observers. *Arch. Dermatol.* **144**, 58–64 (2008).
33. K. Simonyan, A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556 (2014).
34. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 248–255.
35. F. Chollet, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 1251–1258.
36. K. Cooke, G. Spears, D. Skegg, Frequency of moles in a defined population. *J. Epidemiol. Community Health* **39**, 48–52 (1985).
37. H. Iyatomi, H. Oka, M. E. Celebi, M. Hashimoto, M. Hagiwara, M. Tanaka, K. Ogawa, An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm. *Comput. Med. Imaging Graph.* **32**, 566–579 (2008).
38. M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, R. H. Moss, A methodological approach to the classification of dermoscopy images. *Comput. Med. Imaging Graph.* **31**, 362–373 (2007).
39. D. Ruiz, V. Berenguer, A. Soriano, B. SáNchez, A decision support system for the diagnosis of melanoma: A comparative approach. *Exp. Syst. Appl.* **38**, 15217–15223 (2011).
40. M. Zortea, T. R. Schopf, K. Thon, M. Geilhufe, K. Hindberg, H. Kirchesch, K. Møllersen, J. Schulz, S. O. Skrøvseth, F. Godtliebsen, Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. *Artif. Intell. Med.* **60**, 13–26 (2014).
41. L. K. Ferris, J. A. Harkes, B. Gilbert, D. G. Winger, K. Golubets, O. Akilov, M. Satyanarayanan, Computer-aided classification of melanocytic lesions using dermoscopic images. *J. Am. Acad. Dermatol.* **73**, 769–776 (2015).
42. P. G. Cavalcanti, J. Scharcanski, Automated prescreening of pigmented skin lesions using standard cameras. *Comput. Med. Imaging Graph.* **35**, 481–491 (2011).
43. P. G. Cavalcanti, J. Scharcanski, G. V. Baranoski, A two-stage approach for discriminating melanocytic skin lesions using standard cameras. *Exp. Syst. Appl.* **40**, 4054–4064 (2013).
44. J. F. Alcón, C. Ciuhu, W. Ten Kate, A. Heinrich, N. Uzunbajakava, G. Krekels, D. Siem, G. De Haan, Automatic imaging system with decision support for inspection of pigmented skin lesions and melanoma diagnosis. *IEEE J. Select. Topics Signal Process.* **3**, 14–25 (2009).
45. W.-Y. Chang, A. Huang, C.-Y. Yang, C.-H. Lee, Y.-C. Chen, T.-Y. Wu, G.-S. Chen, Computer-aided diagnosis of skin lesions using conventional digital photography: A reliability and feasibility study. *PLOS ONE* **8**, e76212 (2013).
46. M. Ramezani, A. Karimian, P. Moallem, Automatic detection of malignant melanoma using macroscopic images. *J. Med. Signals Sens.* **4**, 281–290 (2014).
47. A. B. Kimball, J. S. Resneck Jr., The US dermatology workforce: A specialty remains in shortage. *J. Am. Acad. Dermatol.* **59**, 741–745 (2008).
48. M. W. Tsang, J. S. Resneck Jr., Even patients with changing moles face long dermatology appointment wait-times: A study of simulated patient calls to dermatologists. *J. Am. Acad. Dermatol.* **55**, 54–58 (2006).

49. A. M. Glazer, A. S. Farberg, R. R. Winkelmann, D. S. Rigel, Analysis of trends in geographic distribution and density of US dermatologists. *JAMA Dermatol.* **153**, 322–325 (2017).
50. P. C. Beatty, D. K. Cherry, C.-J. Hsiao, E. A. Rechtsteiner, National ambulatory medical care survey: 2007 summary. *Natl. Health Stat. Report* **2010**, 1–32 (2010).
51. S. M. Strayer, P. Reynolds, Diagnosing skin malignancy: Assessment of predictive clinical criteria and risk factors.(Research findings that are changing clinical practice). *J. Fam. Practice* **52**, 210–218 (2003).
52. E. Wise, D. Singh, M. Moore, B. Hayes, K. B. Biello, M. C. Dickerson, R. Ness, A. Geller, Rates of skin cancer screening and prevention counseling by US medical residents. *Arch. Dermatol.* **145**, 1131–1136 (2009).
53. M. Tai-Seale, T. G. McGuire, W. Zhang, Time allocation in primary care office visits. *Health Serv. Res.* **42**, 1871–1894 (2007).
54. R. Young, S. Burge, K. Kumar, J. Wilson, D. Ortiz, A time-motion study of primary care physicians' work in the electronic health record era. *Fam. Med.* **50**, 91–99 (2018).
55. P. H. Youl, P. D. Baade, M. Janda, C. B. Del Mar, D. C. Whiteman, J. F. Aitken, Diagnosing skin cancer in primary care: How do mainstream general practitioners compare with primary care skin cancer clinic doctors? *Med. J. Australia* **187**, 215–220 (2007).
56. G. Argenziano, S. Puig, Z. Iris, F. Sera, R. Corona, M. Alsina, F. Barbato, C. Carrera, G. Ferrara, A. Guilabert, D. Massi, J. A. Moreno-Romero, C. Muñoz-Santos, G. Petrillo, S. Segura, H. P. Soyer, R. Zanchini, J. Malvehy, Dermoscopy improves accuracy of primary care physicians to triage lesions suggestive of skin cancer. *J. Clin. Oncol.* **24**, 1877–1882 (2006).
57. S. Menzies, J. Emery, M. Staples, S. Davies, B. McAvoy, J. Fletcher, K. Shahid, G. Reid, M. Avramidis, A. Ward, R. C. Burton, J. M. Elwood, Impact of dermoscopy and short-term sequential digital dermoscopy imaging for the management of pigmented lesions in primary care: A sequential intervention trial. *Br. J. Dermatol.* **161**, 1270–1277 (2009).
58. J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, Learning fine-grained image similarity with deep ranking. *Proc. IEEE Conf. Comp. Vision Pattern Recog.* **2014**, 1386–1393 (2014).
59. G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara, R. Hofmann-Wellenhof, M. Landthaler, S. W. Menzies, H. Pehamberger, D. Piccolo, H. S. Rabinovitz, R. Schiffner, S. Staibano, W. Stolz, I. Bartenjev, A. Blum, R. Braun, H. Cabo, P. Carli, V. De Giorgi, M. G. Fleming, J. M. Grichnik, C. M. Grin, A. C. Halpern, R. Johr, B. Katz, R. O. Kenet, H. Kittler, J. Kreusch, J. Malvehy, G. Mazzocchetti, M. Oliviero, F. Ozdemir, K. Peris, R. Perotti, A. Perusquia, M. A. Pizzichetta, S. Puig, B. Rao, P. Rubegni, T. Saida, M. Scalvenzi, S. Seidenari, I. Stanganelli, M. Tanaka, K. Westerhoff, I. H. Wolf, O. Braun-Falco, H. Kerl, T. Nishikawa, K. Wolff, A. W. Kopf, Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet. *J. Am. Acad. Dermatol.* **48**, 679–693 (2003).
60. N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (IEEE, 2018), pp. 168–172.
61. P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 (2018).
62. S. M. Pizer, J. D. Austin, J. R. Perry, H. D. Safrit, J. B. Zimmerman, in *Application of Optical Instrumentation in Medicine XIV and Picture Archiving and Communication Systems* (International Society for Optics and Photonics, 1986), vol. 626, pp. 242–250.
63. L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259 (1998).

Acknowledgments: We thank the M+Visión IDEA^3 faculty panel for guidance and advice in developing this project, namely, E. Adalsteinsson, D. Burstein, R. San Jose, A. Muñoz-Barrutia, and B. Vakoc. In addition, we would like to thank L. Giancardo for assistance during project development and our collaborators at the Universidad Rey Juan Carlos, N. Malpica, and E. Viaña for helpful contributions in data analysis. We thank the Hospital General Universitario Gregorio Marañón (Madrid, Spain) for providing dermatological images, as specified in the STARD diagram of fig. S1, under the data usage agreement no. 160523. We especially thank all participating patients. **Funding:** This work was supported by the Abdul Latif Jameel Clinic for Machine Learning in Health (to J.J.C., R.B., and L.R.S.). This work was also supported by the Consejería de Educación, Juventud y Deportes de la Comunidad de Madrid through the Madrid-MIT M+Visión Consortium and the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no. 291820 (to L.R.S., B.M.-F., J.S.B., J.T.-S., A.N., and M.L.G.), and the Mexico CONACyT grant 342369/40897 (to L.R.S.). This work was also supported by EU FP7-PEOPLE-2011-COFUND Program within the M+Visión Project of Fundación Madri+d from Comunidad de Madrid (to B.M.-F. and J.S.B.) the Ramón Areces Foundation (L.S., B.M., J.B., J.T., A.N., and M.G.), and the DOE training grant DE-SC0008430 (to B.M.-F. and J.S.B.). **Author contributions:** L.R.S. and T.K. designed, constructed, and tested the DCNN models; performed analysis; and wrote the manuscript. S.T.C. designed models and wrote the manuscript. L.R.S., B.M.-F., J.S.B., J.T.-S., A.N., and J.A.-I. collected the image database and revised the manuscript. J.A.-I., R.R.S., C.C.K., and M.M.S. verified the image database, clinically assessed its outputs and analyses, and revised the manuscript. L.R.S., T.K, S.T.C., B.M.-F., J.S.B., J.T.-S., and A.N. planned and performed experiments, wrote code, and analyzed the data. M.L.G. directed the research plan and edited the manuscript. R.B. and J.J.C. oversaw the research and edited the manuscript. **Competing interests:** S.T.C. is chief executive officer and stockholder in LuminDx Inc., a for-profit startup developing artificial intelligence (AI) for applied dermatology. The other authors declare that they have no competing interests. **Data and materials availability:** All data associated with this work can be found in the manuscript or the Supplementary Materials. A deidentified version of the dataset as specified in the STARD diagram of fig. S1 may be provided for noncommercial research purposes upon reasonable request. Correspondence and requests for materials should be addressed to L.R.S.

Submitted 2 March 2020
Resubmitted 17 August 2020
Accepted 8 January 2021
Published 17 February 2021
10.1126/scitranslmed.abb3652

Citation: L. R. Soenksen, T. Kassis, S. T. Conover, B. Marti-Fuster, J. S. Birkenfeld, J. Tucker-Schwartz, A. Naseem, R. R. Stavert, C. C. Kim, M. M. Senna, J. Avilés-Izquierdo, J. J. Collins, R. Barzilay, M. L. Gray, Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Sci. Transl. Med.* **13**, eabb3652 (2021).

Science Translational Medicine

Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images

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

Sci Transl Med **13**, eabb3652.
DOI: 10.1126/scitranslmed.abb3652

Finding the odd one out

Early identification of skin cancer is key to improving patient outcome. Soenksen *et al.* built a deep convolutional neural network that examines lesions from a given patient present in wide-field images, including those taken with cell phone cameras. Rather than evaluate a single lesion at a time looking for predetermined signs of neoplasia, the algorithm identifies lesions that differ from most of the other marks on that patient's skin, flagging them for further examination and ranking them in order of suspiciousness. The algorithm performed similarly to board-certified dermatologists and could potentially be used at primary care visits to help clinicians triage suspicious lesions for follow-up.

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Translational Medicine* is a registered trademark of AAAS.

Science Translational Medicine
AAAS

stm.sciencemag.org/cgi/content/full/13/581/eabb3652/DC1

Supplementary Materials for

Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images

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, Martha L. Gray

*Corresponding author. Email: soenksen@mit.edu

Published 17 February 2021, *Sci. Transl. Med.* **13**, eabb3652 (2021)
DOI: 10.1126/scitranslmed.abb3652

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at stm.sciencemag.org/cgi/content/full/13/581/eabb3652/DC1)

Data file S1 (.jpg format). Montage of analysis outputs for wide-field images, numbers 1 to 35.

Data file S2 (.jpg format). Montage of analysis outputs for wide-field images, numbers 36 to 70.
Data file S3 (.jpg format). Montage of analysis outputs for wide-field images, numbers 71 to 105.
Data file S4 (.jpg format). Montage of analysis outputs for wide-field images, numbers 106 to 135.

MATERIALS AND METHODS

Dermatological image database and expert class labeling verification

Image dataset aggregation was conducted and controlled as specified in the STAndards for the Reporting of Diagnostic (STARD) diagram of fig. S1. After initial data aggregation, an image duplicate search was conducted using Gemini 2.5.8 (MacPaw Inc.) to ensure no image duplicates were included throughout our analyses. All expert dermatologists (R.S., C.K., & M.S.) were asked to label lesions according to their degree of confidence in malignancy. All pigmented lesions being assessed as "benign" with high confidence were labeled as NSPL-A. This low priority class entails that a patient having only this lesion does not need to be seen/followed by an expert dermatologist in the following 1-3 months and/or it is unlikely the lesion would need to be inspected with dermoscopy. Other lesions triggering considerable uncertainty of a benign assessment with the recommendation to be followed by an expert dermatologist within 1-3 months were classified as NSPL-B. Borderline lesions that are most likely benign, but with at least one expert dermatologist recommending inspection via dermoscopy or monthly follow-up to assess evolution were also assigned an NSPL-B label. All other lesions with high confidence of being malignant, with a recommended biopsy intervention for pathological confirmation, were labeled as SPL. This class included melanomas, basal cell carcinomas (BCC), squamous cell carcinomas (SCC), and other malignant lesions. Dermatologists were asked to make their "best guess" even in lesion images of low resolution. In traditional teledermatology, the general protocol in the presence of a low-resolution lesion image is to ask the user or patient for a better resolution image or to consider it for a referral to a dermatologist automatically; nonetheless, due to the low probability of reassessment in the case of primary care visits, we prepared our dataset to work even in the presence of low-resolution images to the limit of human dermatological assessments. All classification labels (NSLP-A, NSPL-B and SPL) were generated by expert evaluation and majority consensus from R.S. M.M. and C.K. Pigmented lesions in our baseline dataset that differed in classification among all expert reviewers (n_d=15) during primary evaluation were resolved by follow-up revision among all reviewers. No images from the baseline dataset were removed from analysis due lack of consensus. After revision, the number of individually labeled image crops was n_{baseline}=33,980 divided into six classes comprising backgrounds (n_b = 8,888), skin edges (n_{se} = 2,528), bare skin sections (n_{sk} = 10,935), non-suspicious pigmented lesions type A ($n_{nspl\text{-}a}$ = 10,759) of low priority, non-suspicious pigmented lesions type B ($n_{nspl\text{-}b}$ = 1,110) of medium priority, and suspicious pigmented lesions (n_{spl} = 4,063). Although clinically relevant, the distribution of the suspicious pigmented lesion (SPL) class consisting of melanomas stages 0-IV (n_m = 2,906), squamous cell carcinomas (n_{scc} = 589), and 568 basal cell carcinomas (n_{bcc} = 568) in our database does not actually emulate the proportion of such type of lesions seen in usual clinical practice. In turn, this distribution was chosen to ensure our trained DNN algorithms could reach high sensitivity and specificity in melanoma, which is the most clinically relevant SPL subclass. A notable concern among previous work in CAD DNNs for dermatology point to the low number of melanomas generally used to train and test these systems, which we aim to address by aggregating a larger number of those samples in our database than usually targeted. Although melanomas are rare relative to BCC and SCC, they still carry the greatest risk for potential harm if under or misidentified, therefore purposely increasing the number of this subclass of suspicious pigmented lesions is a design feature of our work towards exploring a clinically useful tool.

Fitzpatrick skin type assessment

The Fitzpatrick skin types were obtained for both single-lesion and wide-field images using an automated grading pipeline with subsequent expert validation. First, all images in RGB format were transformed to 8-bit HSV (hue, saturation, value) color space using OpenCV (OpenCV.org). Then binary masks for rough segmentation of visible skin regions were produced using a thresholding filter to allow for channel pixel values H = [0-20], S = [48-255] and V = [0-255]. Once such skin regions were segmented, pixel-wise vector quantization was performed in all skin-like pixels of each image to detect the main color cluster center using a K-Means strategy (Scikit-learn Version 0.23.1). This extracted color cluster centers were considered to be the dominant skin color in each image. Fitzpatrick Skin tone classification (Type I-VI) was finally generated by analyzing value channel of the dominant color such that: Type I = [214-255], Type II = [171-213], Type III = [128-170], Type IV = [85-127], Type V = [43-84], Type VI = [0-42]. Using this automatically generated classification, images were placed in independent folders for non-expert initial visual grading inspection (L.S.), with subsequent expert-grading revision (R.S.). A total of 153 images with incorrect Fitzpatrick grading were identified in the initial visual inspection (L.S.) and changed accordingly. Only 12 images additional images were changed in grading during single-expert revision (R.S.). The distribution of the Fitzpatrick skin types in the generated database can be seen in table S2.

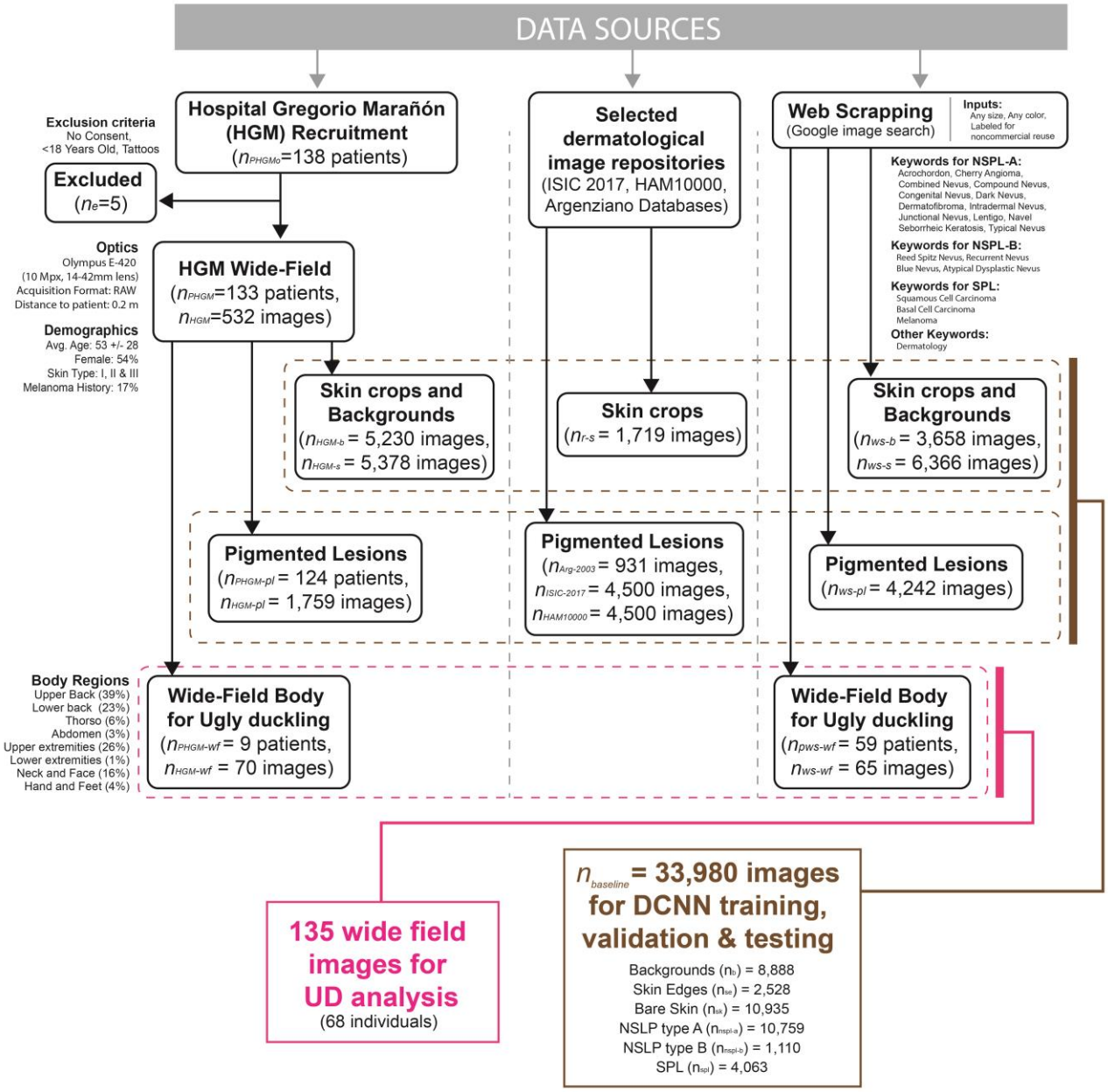
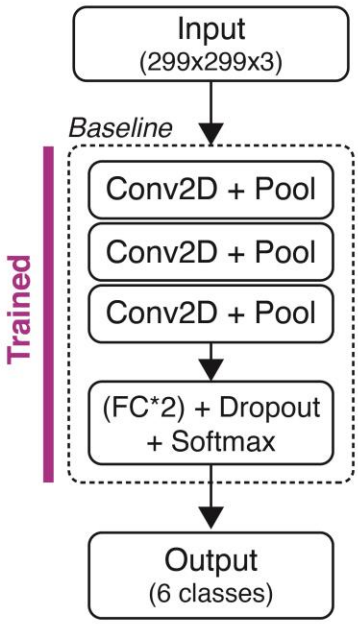


Fig. S1. STARD diagram of data aggregation. Data sources presented in the STAndards for the Reporting of Diagnostic (STARD) diagram are divided into three main components: 1) data collected from the Hospital Gregorio Marañón (Madrid, Spain); 2) data retrieved from selected open-source dermatological repositories such as the ISIC - 2017 (*60*), HAM10000 - 2018 (*61*) and Argenziano - 2003 (*59*) databases; and 3) data obtained through web scrapping on open resources for non-commercial reuse. Images collected from Hospital Gregorio Marañón (HGM) were acquired using an Olympus E-420 (10 Mega Pixel, 14-42mm lens) camera, in RAW format, at a distance of 0.2 m from the patient. Basic demographics of HGM recruited population are also provided. Image collection hardware for other data sources was not controlled. All three data sources were used for aggregation of pigmented lesions, skin and background crops, whereas wide-field body images for ugly duckling (UD) analysis were obtained only from HGM and web

scrapping. Only a non-overlapping randomized subset of 4,500 images was collected from each HAM10000 and ISIC datasets for this work. Images with assessment of low-resolution, substantial pixelation or blurriness by L.S., B.M., J.B., J.T. or J.A, were not added during data aggregation. The total number of aggregated images for deep convolutional neural network (DCNN) model training, validation and testing was 33,980 images. The total number of wide-field images for ugly duckling analysis was 135, spanning 68 different individuals. The distribution of observable body regions on these images is also specified, with the highest proportion (39%) for the upper back. NSPL= Non-suspicious pigmented lesions, SPL= Suspicious pigmented lesions.

A



B

Baseline Architecture

Layer (type)	Output Shape	Param #
Input (Image)	(299, 299, 3)	-
conv2d_1 (Conv2D)	(297, 297, 32)	896
activation_1 (Activation)	(297, 297, 32)	-
max_pooling2d_1 (MaxPooling2)	(148, 148, 32)	-
conv2d_2 (Conv2D)	(146, 146, 32)	9248
activation_2 (Activation)	(146, 146, 32)	-
max_pooling2d_2 (MaxPooling2)	(73, 73, 32)	-
conv2d_3 (Conv2D)	(71, 71, 64)	18496
activation_3 (Activation)	(71, 71, 64)	-
max_pooling2d_3 (MaxPooling2)	(35, 35, 64)	-
flatten_1 (Flatten)	(78400)	-
dense_1 (Dense)	(64)	5017664
activation_4 (Activation)	(64)	-
dropout_1 (Dropout)	(64)	-
dense_2 (Dense)	(6)	390

Total params: 5,046,694
Total Output classes: 6

Fig. S2. Baseline DCNN model architecture. (A) Block diagram of the three-layered baseline convolutional neural network model. An image input size of 299x299x3 was chosen to accommodate for comparison with transfer learning models with maximum typical image input size of 299x299x3 (Xception), as well as models with smaller input sizes in the 150x150x3 (VGG16). **(B)** Details of baseline network with layer input-out sizes as well as parameter count "Param #."

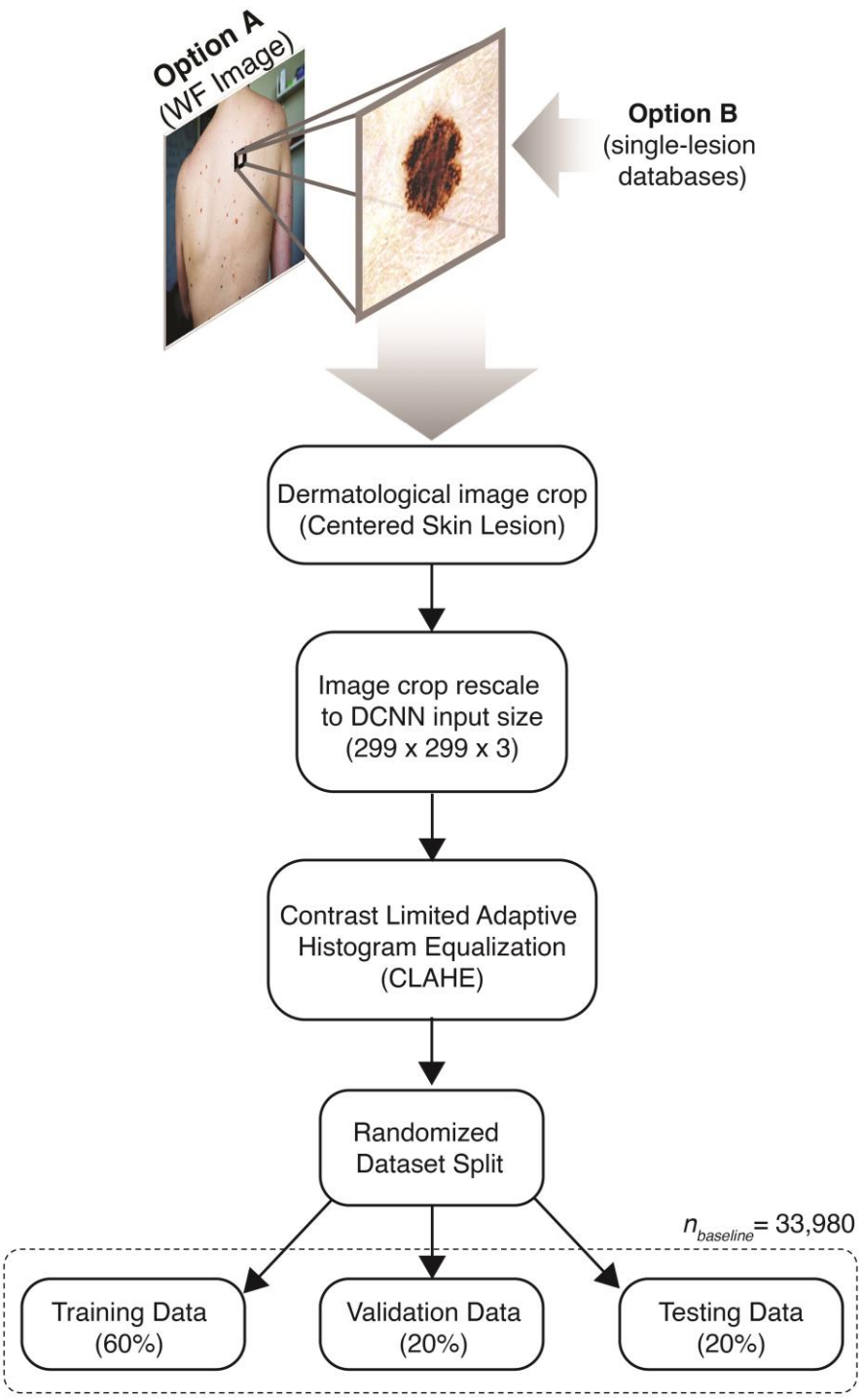


Fig. S3. Preprocessing and splitting of the base database into training, validation, and testing sets. Randomized dataset splits were done before any data augmentation and at single-lesion crop level. WF=wide-field.

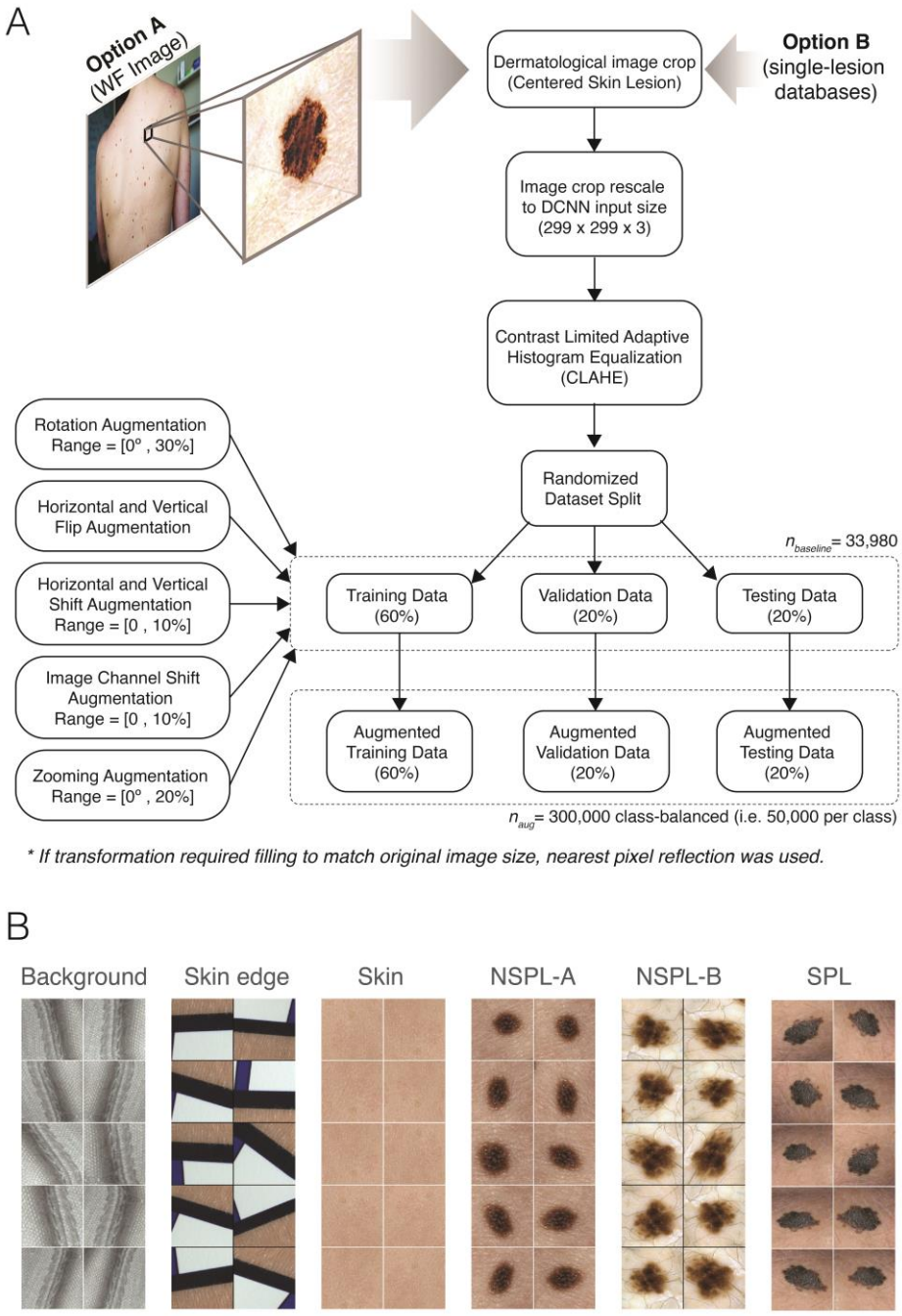


Fig. S4. Data augmentation strategy. (A) Dataset pre-processing yielding $n_{baseline}$=33,980 images are split into training (60%), validation (20%) and testing (20%) sets. Randomized dataset splits were done before any data augmentation and at single-lesion crop level. Then each split is augmented approximately 10-fold to generate n_{aug}=300,000 non-overlapping images across training, validation or testing sets, but exhibiting balanced classes (50,000 images per class). The augmentation strategy considers five basic types of data augmentation randomly selected from the provided ranges. A random combination of transformations was also allowed. **(B)** Example augmentation outputs for a single image per class.

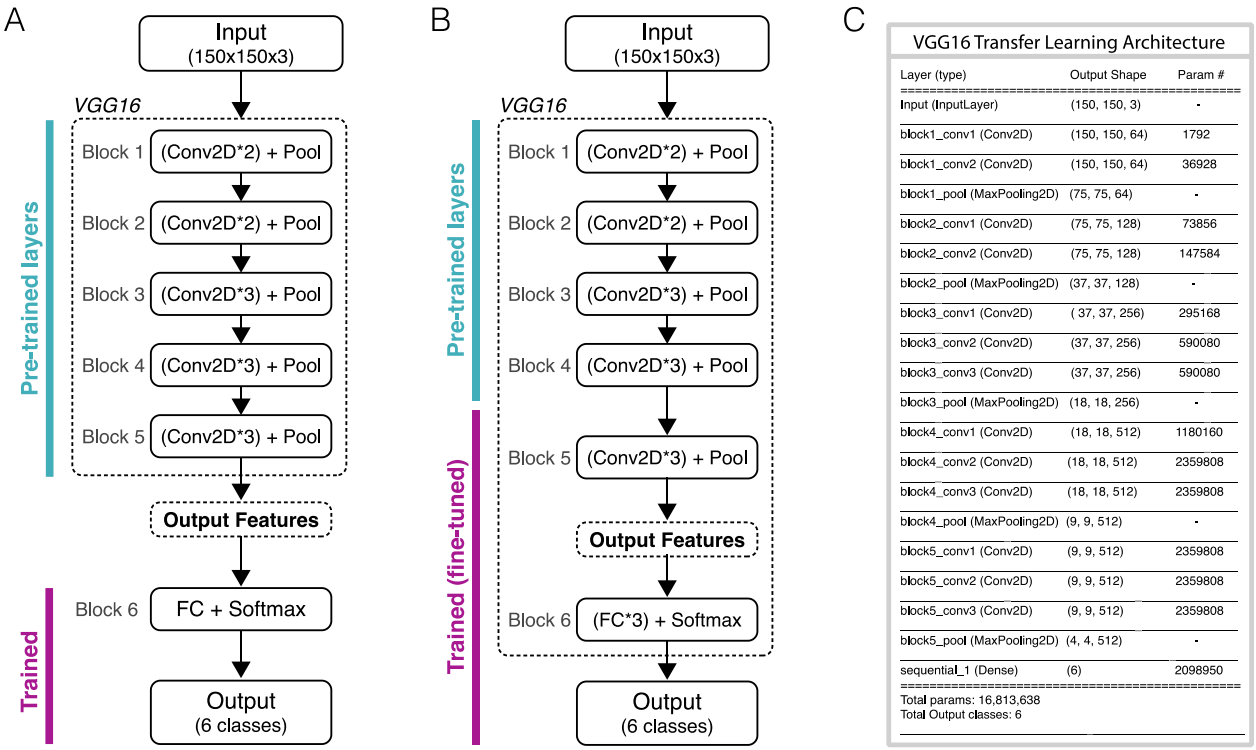


Fig. S5. Transfer learning DCNN model architecture based on VGG16. (A) Block diagram of VGG16 convolutional neural network model with the indication of bottleneck trained sections corresponding to the last fully connected (FC) and activation (SoftMax) layers. An image input size of 150x150x3 was used. **(B)** Block diagram of VGG16 convolutional neural network model with the indication of fine-tuned block sections corresponding to Block 5 and 6. **(C)** Details of bottleneck VGG16 transfer learning network with layer input-out sizes as well as parameter count "Param #."

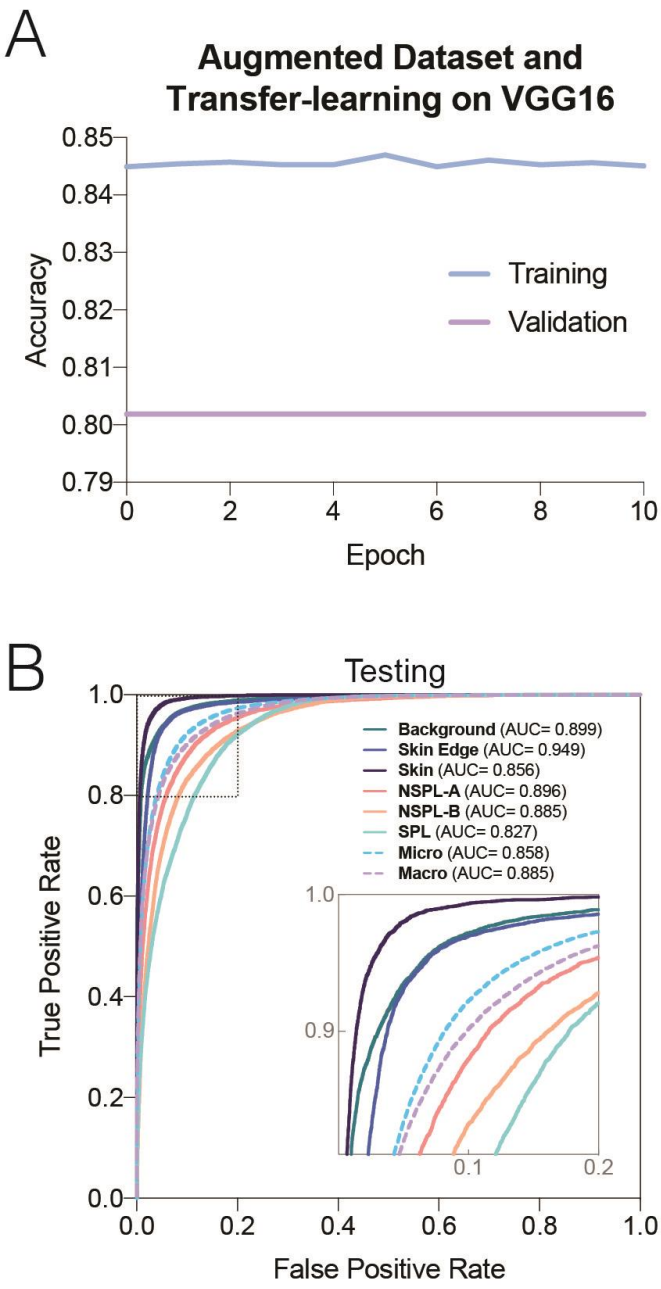


Fig. S6. Training, validation, and testing of fine-tuned VGG16 DCNN model. (A) Recorded accuracy on training and validation sets per Epoch for the VGG16 architecture model using transfer learning and fine-tuning on the 10x augmented dataset. **(B)** Multi-class ROCs for the VGG16 architecture model using transfer learning and fine-tuning on the augmented dataset.

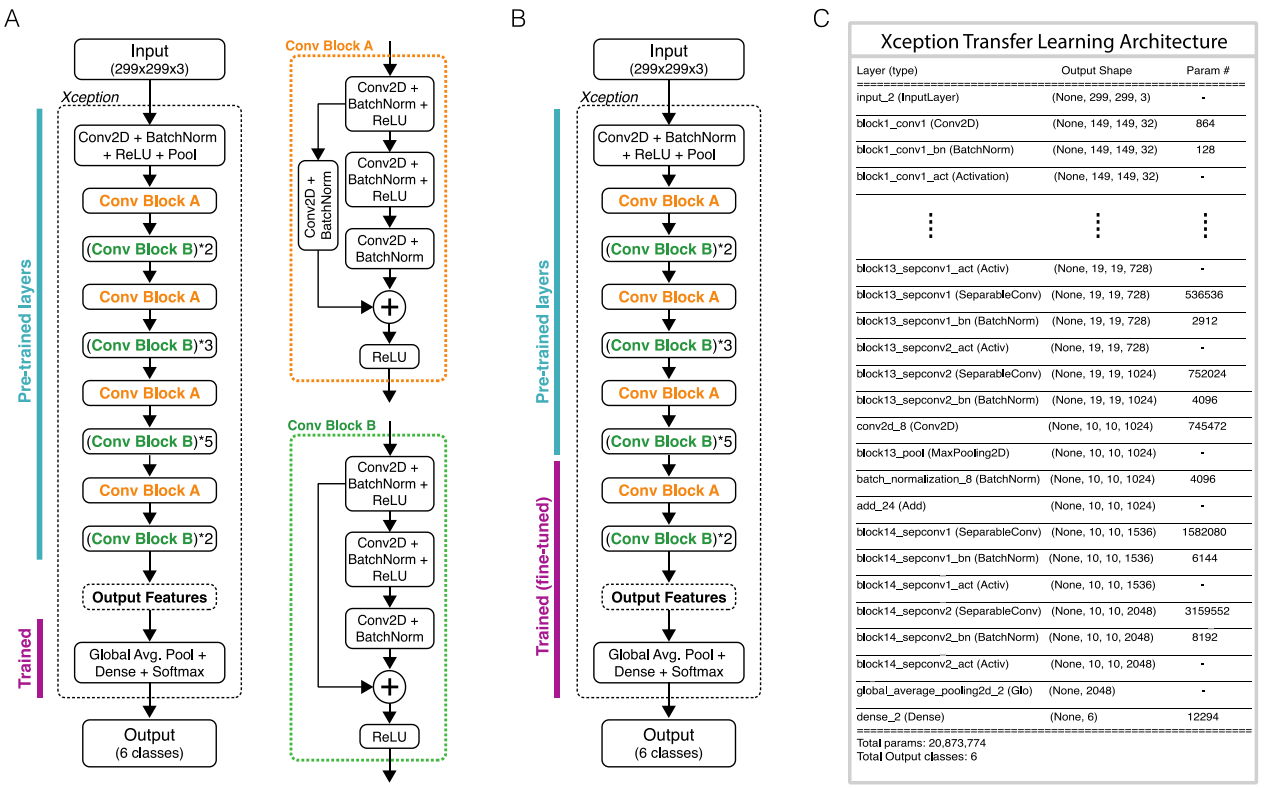


Fig. S7. Transfer learning DCNN model architecture based on Xception. (A) Block diagram of Xception convolutional neural network model with the indication of bottleneck trained sections corresponding to the last global average max-pooling (Pool), dense and activation (SoftMax) layers. An image input size of 299x299x3 was used. **(B)** Block diagram of Xception convolutional neural network model with the indication of fine-tuned block sections corresponding to the last three blocks of the network (last block A and last two block Bs., global average max-pooling (Pool), dense and activation layers). **(C)** Details of bottleneck Xception transfer learning network with layer input-out sizes as well as parameter count "Param #," vertical triple dots indicate intermediate layers not shown in the table.

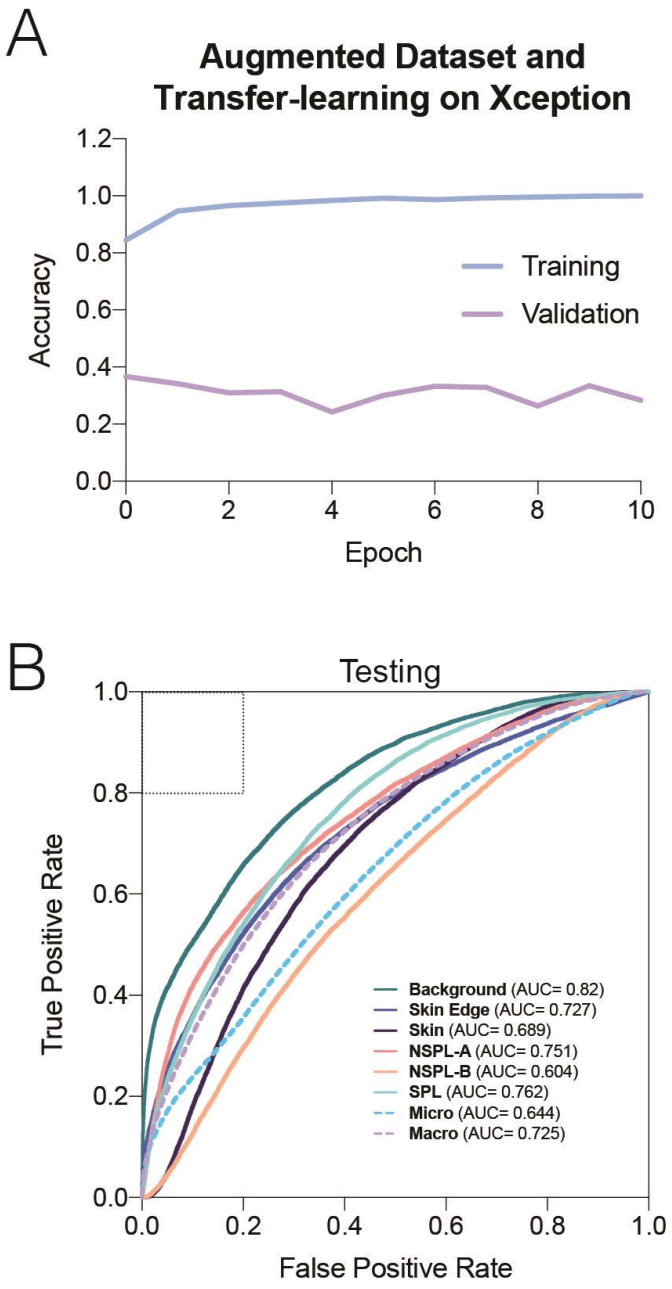


Fig. S8. Training, validation, and testing of fine-tuned Xception DCNN model. (A) Recorded accuracy on training and validation sets per Epoch for the Xception architecture model using transfer learning and fine-tuning on the 10x augmented dataset. **(B)** Multi-class ROCs for the Xception architecture model using transfer learning and fine-tuning on the augmented dataset.

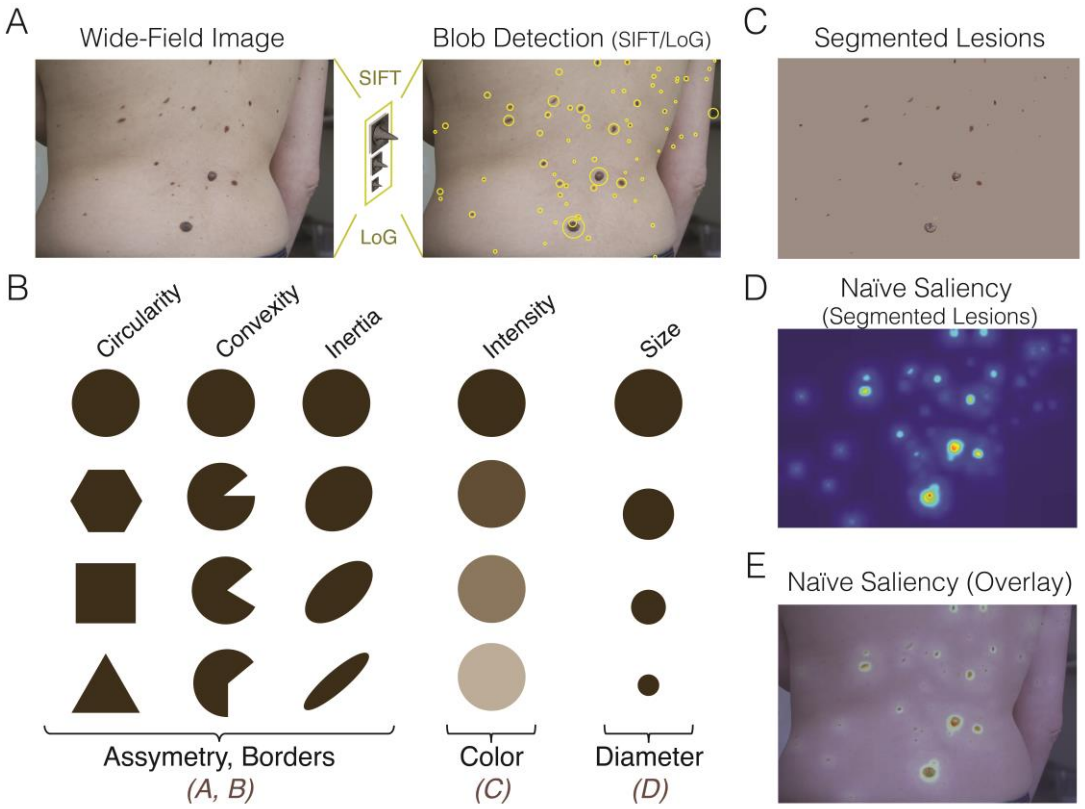


Fig. S9. Blob detection and naïve saliency calculation. (A) Wide-field image analyzed for blob detection for both, DCNN analysis and naïve feature-based saliency using a standard SIFT-LoG algorithm based on geometrically defined features. **(B)** Geometrically defined features in blobs (i.e., circularity, convexity, inertia, intensity, and size) used for naïve detection of pigmented lesions in wide-field images with correspondence to the ABCD criteria (i.e., asymmetry, borders, color, and diameter). The detected blob-like key points were calculated using the OpenCV Computer Vision Library and the SimpleBlobDetector function (OpenCV.org) on grayscale versions of the analyzed wide-field images. The standard filter parameters used in this function included: Threshold filter (minThreshold = 0, maxThreshold = 255); Area filter (minArea = 10x10 pixels, maxArea = image height * image width); Circularity filter (minCircularity = 0.1); Convexity filter (minConvexity = 0.1); and Inertia filter (minInertiaRatio = 0.1). This blob detected output was used as starting point for both DCNN analysis and naïve saliency calculation. **(C)** Sample of lesion segmentation from the detected blob-like regions using grayscale thresholding. Segmented masks are overlaid over a synthetically averaged monochrome wide-field image base on the input field. **(D)** The output of saliency-based visual attention based on the Itti et al. method (*63*), as an alternative saliency method independent for comparison with the deep learning feature method used primarily in this work. In this specific saliency algorithm, a visual attention mechanism inspired by the behavior of the early primate visual system is used. **(E)** Multiscale image features are combined into a single topographical with the non-DCNN saliency map created through lesion segments collaged into an inconspicuous (non-salient) synthetic background created by averaging the original wide-field dermatological image. In this naïve, but computationally efficient approach, saliency maps can be generated; unfortunately, this algorithm also appears to be sensitive to the presence of fabrics, backgrounds, and other outstanding features drawing attention to non-lesions (see data files S1 and S2).

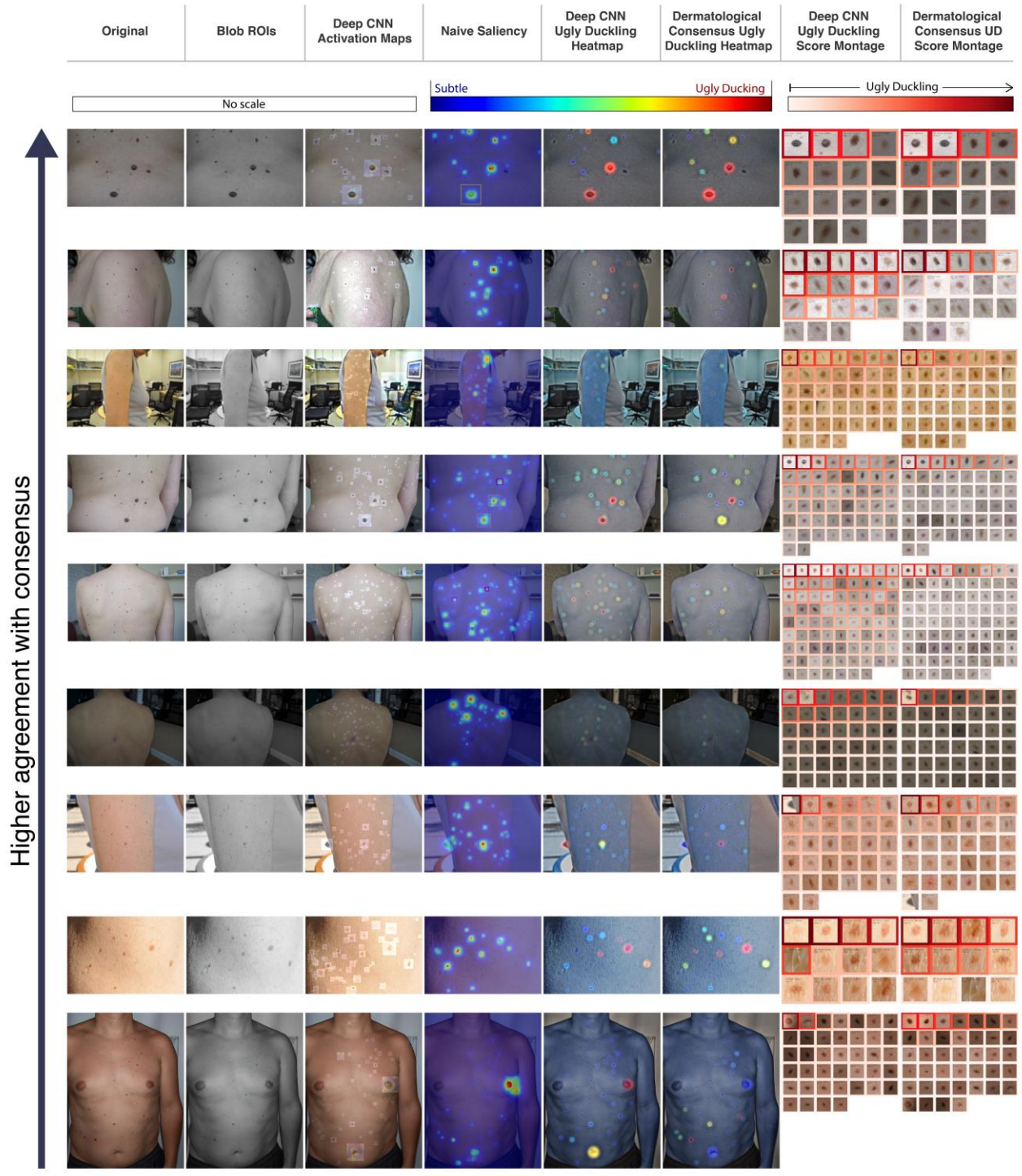


Fig. S10. Selected samples of DCNN ugly duckling outputs as compared to naïve saliency and dermatological consensus. Side-to-side ugly-duckling heatmaps and montages are shown with color-coded ranking. Varying degrees of success can be appreciated in this selection, from situations where belly button and nipples are automatically analyzed as suspicious pigmented lesions to events with a high agreement with dermatological consensus. The greater richness in evaluations can be seen in most images when analyzed using our DCNN system as compared with the top-10 dermatological consensus.

Table S1. Taxonomy of pigmented lesions included in our study's baseline dataset. All classification labels (NSLP-A, NSPL-B and SPL) were generated by expert evaluation and majority consensus from R.S. M.M. and C.K. Pigmented lesions in our baseline dataset that differed in classification among all expert reviewers (n_d=15) during primary evaluation were resolved by follow-up revision among all reviewers. No images from the baseline dataset were removed from analysis due lack of consensus. The total number of instances includes dermoscopy and non-dermoscopy images. The number of dermoscopy images from these subsets is also included evincing a more substantial proportion of non-dermoscopy images for all classes.

Lesion Type	Management	Class/ Taxonomy	Immediate Referral required	Priority	Count (n)	Total (N)	Dermoscopy
Junctional Nevus	Nothing	NSPL-A	NO	Low	3334	10,759	91
Combined Nevus	Nothing	NSPL-A	NO	Low	203		
Congenital Nevus	Nothing	NSPL-A	NO	Low	178		
Dermal Nevus	Nothing	NSPL-A	NO	Low	4509		
Dermatofibroma	Nothing	NSPL-A	NO	Low	184		
Lentigo	Nothing	NSPL-A	NO	Low	177		
Seborrheic keratosis	Nothing	NSPL-A	NO	Low	754		
Acrochordons	Nothing	NSPL-A	NO	Low	782		
Cherry angiomas	Nothing	NSPL-A	NO	Low	638		
Atypical Nevus (i.e. Dysplastic, Clark)	Follow	NSPL-B	NO	Medium	960	1,110	191
Blue Nevus	Follow	NSPL-B	NO	Medium	118		
Recurrent Nevus	Follow	NSPL-B	NO	Medium	7		
Reed Spitz Nevus (Pigmented Spindle Cell Nevus of Reed)	Follow	NSPL-B	NO	Medium	7		
Miscellaneous (Other non-cancer)	Follow	NSPL-B	NO	Medium	18		
Basal Cell Carcinoma	Excision	SPL	YES	High	589	4,063	398
Squamous Cell Carcinoma	Excision	SPL	YES	High	568		
Melanoma (Stage 0 - IV)	Excision	SPL	YES	High	2906		

Table S2. Distribution of Fitzpatrick skin tones along all skin-relevant classes in the base dataset. Skin Type III is the most represented (36.17%), whereas Type VI is the least represented (0.46%).

	Skin Edge		Skin		NSPL-A		NSPL-B		SPL		ALL	
Fitzpatrick	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Type I	218	8.62%	887	8.11%	817	7.59%	455	40.99%	1322	32.54%	3699	12.58%
Type II	533	21.08%	3197	29.24%	2872	26.69%	325	29.28%	1469	36.16%	8396	28.56%
Type III	739	29.23%	4192	38.34%	4473	41.57%	254	22.88%	975	24.00%	10633	36.17%
Type IV	527	20.85%	2484	22.72%	2235	20.77%	72	6.49%	220	5.41%	5538	18.84%
Type V	388	15.35%	174	1.59%	357	3.32%	3	0.27%	72	1.77%	994	3.38%
Type VI	123	4.87%	1	0.01%	5	0.05%	1	0.09%	5	0.12%	135	0.46%
Total:	**2528**		**10935**		**10759**		**1110**		**4063**		**29395**	

Data file S1. Montage of analysis outputs for wide-field images, numbers 1 to 35.

Data file S2. Montage of analysis outputs for wide-field images, numbers 36 to 70.

Data file S3. Montage of analysis outputs for wide-field images, numbers 71 to 105.

Data file S4. Montage of analysis outputs for wide-field images, numbers 106 to 135.