applied sciences

MDPI

Article

Computer-Aided Diagnosis of Skin Diseases using Deep Neural Networks

Muhammad Naseer Bajwa [1,2,*], Kaoru Muta [3], Muhammad Imran Malik [4,5], Shoaib Ahmed Siddiqui [1,2], Stephan Alexander Braun [6,7], Bernhard Homey [7], Andreas Dengel [1,2] and Sheraz Ahmed [2]

1 Fachbereich Informatik, Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany; shoaib_ahmed.siddiqui@dfki.de (S.A.S.); andreas.dengel@dfki.de (A.D.)
2 German Research Center for Artificial Intelligence GmbH (DFKI), 67663 Kaiserslautern, Germany; sheraz.ahmed@dfki.de
3 College of Engineering, Osaka Prefecture University, Naka, Sakai, Osaka 599-8531, Japan; muta@m.cs.osakafu-u.ac.jp
4 School of Electrical Engineering and Computer Science, National University of Sciences and Technology (NUST), H-12, Islamabad 46000, Pakistan; malik.imran@seecs.edu.pk
5 Deep Learning Laboratory, National Center of Artificial Intelligence, Islamabad 46000, Pakistan
6 Department of Dermatology, University Hospital Münster, Von-Esmarch-Str 58, 48149 Münster, Germany; stephanalexander.braun@ukmuenster.de
7 Department of Dermatology, University Hospital Düsseldorf, Moorenstr. 5, 40225 Düsseldorf, Germany; bernhard.homey@med.uni-duesseldorf.de
* Correspondence: naseer.bajwa@dfki.de

Received: 5 March 2020; Accepted: 29 March 2020; Published: 4 April 2020

Abstract: Propensity of skin diseases to manifest in a variety of forms, lack and maldistribution of qualified dermatologists, and exigency of timely and accurate diagnosis call for automated Computer-Aided Diagnosis (CAD). This study aims at extending previous works on CAD for dermatology by exploring the potential of Deep Learning to classify hundreds of skin diseases, improving classification performance, and utilizing disease taxonomy. We trained state-of-the-art Deep Neural Networks on two of the largest publicly available skin image datasets, namely DermNet and ISIC Archive, and also leveraged disease taxonomy, where available, to improve classification performance of these models. On DermNet we establish new state-of-the-art with 80% accuracy and 98% Area Under the Curve (AUC) for classification of 23 diseases. We also set precedence for classifying all 622 unique sub-classes in this dataset and achieved 67% accuracy and 98% AUC. On ISIC Archive we classified all 7 diseases with 93% average accuracy and 99% AUC. This study shows that Deep Learning has great potential to classify a vast array of skin diseases with near-human accuracy and far better reproducibility. It can have a promising role in practical real-time skin disease diagnosis by assisting physicians in large-scale screening using clinical or dermoscopic images.

Keywords: artificial intelligence in dermatology; automated skin disease diagnosis; computer-aided diagnosis; medical image analysis

1. Introduction

Deep Learning (DL) [1] is a branch of Artificial Intelligence (AI) in which a computer algorithm analyses raw data and automatically learns discriminatory features needed for recognizing hidden patterns in them. Over the last decade, this field has witnessed striking advances in the ability of DL-based algorithms to analyse various types of data, especially images [2] and natural language [3]. The most common DL models are trained using supervised learning, in which datasets are composed of

inputs (e.g., dermoscopic images of skin diseases) and corresponding target output labels (e.g., diagnoses or skin disease classes such as 'benign' or 'malignant'). Healthcare and medicine can greatly benefit from recent advances in image classification and object detection [4], particularly those medical disciplines in which diagnoses are primarily based on detection of morphologic changes such as pathology, radiology, ophthalmology and dermatology etc. In such medical domains, digital images are captured and provided to DL algorithms for Computer-Aided Diagnosis (CAD). These advance algorithms have already made their mark on automated detection of tuberculosis [5], breast malignancy [6], glaucoma [7], diabetic retinopathy [8] and serious brain findings such as stroke, haemorrhage, and mass effects [9].

Large scale manual screening for diseases is exhaustively laborious, extremely protracted, and severely susceptible to human predisposition and fatigue. Since manual diagnosis may also be affected by physicians' level of experience and different dermoscopic algorithms in which they are formally trained, multiple experts might disagree on their diagnosis for a certain condition [10,11]. Additionally, due to physicians' subjective judgements, manual diagnosis is hardly reproducible [12]. On the other hand, CAD can provide swift, reliable and standardized diagnosis of various diseases with consistency and accuracy. CAD can also afford the opportunity of efficient and cost-effective screening and prevention of advanced tumour diseases to people living in rural or remote areas where expert dermatologists are not readily available.

Most of publicly available datasets for clinical or dermoscopic images like Interactive Atlas of Dermoscopy [13], Dermofit Image Library [14], Global Skin Atlas, MED-NODE [15] and PH2 [16] etc. contain only a few hundreds to a couple of thousand images. Ali et al. [17] reported that around 78% of the studies they surveyed used datasets smaller than 1000 images and the study using the largest dataset had 2430 images. Therefore, most of existing works on CAD of skin diseases use either private or very small publicly available datasets. Additionally, these studies usually render overwhelming focus on only binary or ternary classification of skin diseases and not much attention is paid to multi-class classification to explore the full potential of DL. Therefore, such studies act merely as a proof-of-concept for the efficacy of AI in dermatology.

In this work, we extend previous works by showing that DL model are fairly capable of recognising hundreds of skin lesions, and therefore should be capitalized to their full extent. We trained many state-of-the-art DL models for classification of skin diseases using two of the largest publicly available datasets, namely DermNet and ISIC Archive (2018 version). We also employed non-visual data in the form of disease taxonomy to improve our classification results and show that DL can process and utilize multi-model input for better classification performance.

Related Work

Convolutional Neural Networks (CNNs) are computer models inspired by biological visual cortex. These models have been proven to be very efficient, accurate and reliable in image classification. They have already achieved near-human performance in many challenging natural image stratification tasks [18–21] and have also been used to classify diseases from medical images [4].

Towards automated skin disease classification, Kawahara et al. [22] employed CNNs to extract features and trained a linear classifier on them using 1300 images of Dermofit Image Library to perform 10-ary classification. Similar approach was used by Ge et al. [23] on MoleMap dataset to do 15-ary classification. Esteva et al. [24] used a pre-trained Inception v3 on around 130,000 images. Although their results for two binary-classification tasks are merely "on par with all tested experts", yet this work was the first credible proof-of-concept based on a large dataset that DL can make a practical contribution in real-world diagnosis. Following their steps, Haenssle et al. [25] pitched their fine-tuned Inception v4 model against 58 dermatologist after evaluating binary classification performance of their model on two test sets of size 100 and 300 only. The sensitivity and specificity of their Deep Neural Network (DNN) model is certainly higher than that of dermatologists' mean performance on two private test sets, however, their performance on publicly available International Symposium on Biomedical Imaging (ISBI) 2016 Challenge [26] test data is below the performance of first two winning entries in that challenge.

To address the scarcity of available data for tracking and detecting skin diseases, Li et al. [27] developed a domain-specific data augmentation technique by merging individual lesions with full body images to generate large volume of synthetic data. Li and Shen [28] also used DNN to segment lesions, extract their dermoscopic features and classify them.

2. Materials and Methods

2.1. Datasets

DermNet is a freely available dataset of around 23,000 images gathered and labelled by Dermnet Skin Disease Atlas. We were able to download 22,501 images because the links for rest of them appeared to be inactive. This dataset provides diagnosis for 23 super-classes of diseases which are taxonomically divided into 642 sub-classes. However, there were some duplicate, empty and irrelevant sub-classes in the data. After pruning, 21,844 images in 622 sub-classes remained. Distribution of DermNet dataset used in this work is given in Table 1.

Table 1. Overview of DermNet Dataset and Distribution of Classes.

Class Label	Abbreviation	Super-Class Name	Np. of Images	No. of Sub-Classes
0	ACROS	Acne and Rosacea	912	21
1	AKBCC	Actinic Keratosis, Basal Cell Carcinoma, and other Malignant Lesions	1437	60
2	ATO	Atopic Dermatitis	807	11
3	BUL	Bullous Diseases	561	12
4	CEL	Cellulitis, Impetigo, and other Bacterial Infections	361	25
5	ECZ	Eczema Photos	1950	47
6	WXA	Exanthems and Drug Eruptions	497	18
7	ALO	Alopecia and other Hair Diseases	195	23
8	HER	Herpes, Genetal Warts and other STIs	554	15
9	PIG	Pigmentation Disorder	711	32
10	LUPUS	Lupus and other Connective Tissue diseases	517	20
11	MEL	Melanoma and Melanocytic Nevi	635	15
12	NAIL	Nail Fungus and other Nail Disease	1541	48
13	POI	Poison Ivy and other Contact Dermatitis	373	12
14	PSO	Psoriasis Lichen Planus and related diseases	2112	39
15	SCA	Scabies Lyme Disease and other Infestations and Bites	611	25
16	SEB	Seborrheic Keratoses and other Benign Tumors	2397	50
17	SYS	Systemic Disease	816	43
18	TIN	Tinea Candidiasis and other Fungal Infections	1871	36
19	URT	Urticaria	265	9
20	VASCT	Vascular Tumors	603	18
21	VASCP	Vasculitis	569	17
22	WARTS	Common Warts, Mollusca Contagiosa and other	1549	26
		Total	**21844**	**622**

The second dataset is an online archive of around 24,000 images divided into seven classes maintained by The International Skin Imaging Collaboration (ISIC). Their growing archive of high quality clinical and dermoscopic images is manually labelled. Distribution of images in ISIC Archive-2018 dataset can be found in Table 2.

Table 2. Overview of ISIC Archive Dataset and Distribution of Classes.

Class Label	Abbreviation	Class	Np. of Images
0	AKIEC	Bowen Disease	334
1	BCC	Basal Cell Carcinoma	583
2	BKL	Benign Keratosis-like Lesions	1674
3	DF	Dermatofibroma	122
4	MEL	Melanoma	2177
5	NV	Melanocytic Nevi	18,618
6	VASC	Vascular Lesions	157
		Total	**23,665**

2.2. Experimental Setup

We used various state-of-the-art DNN architectures developed in the recent years like residual networks, inception networks, densely connected networks, and frameworks facilitating architecture search. To cope up with never-ending appetite of deep CNNs for data, we used these models pre-trained on ImageNet, which is a large dataset of around 1.5 million natural scene images divided into 1000 classes. We fine-tuned these models on dermatology datasets to leverage the benefits of transfer learning. From various CNN architectures tried for this task, we eventually selected ResNet-152 [29], DenseNet-161 [30], SE-ResNeXt-101 [31], and NASNet [32] for their better performance. To report the final results, we combined the potential of all of these biologically inspired neural networks by taking ensemble of their individual predictions. For ensemble we used average of individual predictions of four best performing CNNs to output final prediction.

It is important to note here that comparing researches that use different datasets, different subsets or train/test splits of the same dataset is not scientifically correct. Since neither of the two datasets used in this work provided instructions on dividing the data into train and test sets, we used stratified k-fold cross validation ($k = 5$ in this work) so that any future research can be compared with our work at least. The k-fold cross validation is a statistical method to ensure that the classifier's performance is less biased towards a randomly taken train/test split. The k-fold cross validation is performed by dividing the whole dataset into k, possibly equal, portions or folds. During a training iteration, one of these folds is kept aside for validation and rest of $k - 1$ folds are used for training the model. In next training iteration a different fold is kept aside for validation and remaining $k - 1$ are used for training. This way, the train and test sets in each iteration are completely mutually exclusive. This process is repeated k times such that each of the k-folds is used for validation exactly once. This cross-validation approach provides a more realistic generalization approximation. For training, we randomly cropped the images with scale probability ranging between 0.7 and 1.0 while maintaining the aspect ratio. These cropped images are then resized to 224×224 pixels (for NASNet the input is resized to 331×331) before feeding them to the network. The images are also randomly flipped horizontally with flip probability 0.5. During testing, an image is cropped into four corners (top left, top right, bottom left, and bottom right) and one central crop of required size. These cropped images are given to the classifier for inference and ensemble of five predictions is taken to provide final output. Initial learning rate is set to 10^{-4} and is halved every five epochs. The networks are trained for 20 epoch and 10 epochs for DermNet and ISIC Archive, respectively. The number of training epochs for each dataset and initial learning rate were determined empirically. To handle class imbalance, we used weighted loss where the weight for a certain class equals reciprocal of that class's ratio in the dataset.

3. Results

3.1. Results on DermNet

As DermNet provides the opportunity to leverage taxonomical relationship among various diseases, therefore, for 23-ary classification we conducted our experiments in two ways. In the first experiment (Exp-1), we trained our networks on 23 classes and inferred on 23 classes. This is the most prevalent approach. We achieved $77.53 \pm 0.64\%$ Top-1 accuracy and $93.87 \pm 0.37\%$ Top-5 accuracy with $97.60 \pm 0.15\%$ Area Under the Curve (AUC) using ensemble of four best models. In second experiment (Exp-2) we made use of additionally given ontology in the dataset. We trained our network on 622 classes but inferred on 23 classes only. The use of this disease ontology information translates into incorporation of expert knowledge into the network. We implemented this by summing the predictions of all sub-classes to calculate the prediction of respective super-class. This approach gave us noticeable boost in our classifiers' performance. We got $79.94 \pm 0.45\%$ Top-1 accuracy, $95.02 \pm 0.15\%$ Top-5 accuracy and $98.07 \pm 0.07\%$ AUC using ensemble.

Top-N accuracy indicates the capability of a classifier to predict correct class in first N attempts. This metric gives a deeper insight into the classifier's learning and discriminating ability. Our results,

of Exp-2 for example, show that the model was able to predict the correct diagnosis out of 23 possible diseases in first attempt with almost 80% accuracy. However, when allowed to make 5 most probable predictions about a given image, the classifier achieved more than 95% accuracy. This means that even when the first prediction of the classifier is wrong, the would-be correct prediction is high on the list of next four predictions. Table 3 shows detailed performance metrics of 23-ary classification in both experiments. Accuracies and AUC scores of individual classifiers for Exp-1 and Exp-2 are given in Table A1 in Appendix A.

Table 3. Performance Metrics for 23-ary Classification of DermNet using Ensemble. Exp-1: Training on 23 classes and testing on 23 classes without using disease ontology. Exp-2: Training on 622 classes and testing on 23 classes using disease ontology. Refer to Table 1 for full-form of class abbreviations.

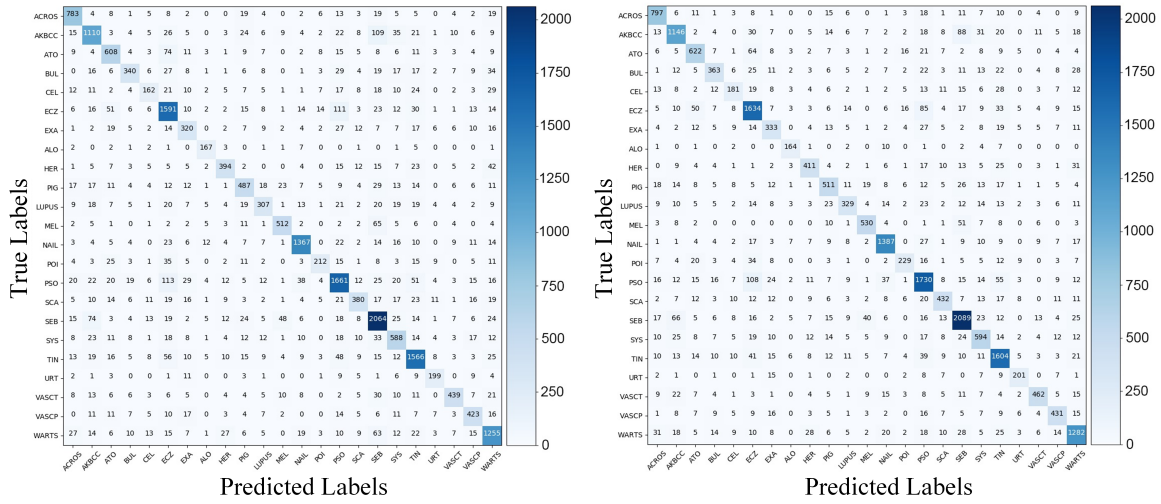
Class	Precision (%)		Sensitivity (%)		Specificity (%)		F-1 Score (%)	
	Exp-1	Exp-2	Exp-1	Exp-2	Exp-1	Exp-2	Exp-1	Exp-2
ACROS	81.39	81.66	85.86	87.39	98.90	98.94	83.56	84.43
AKBCC	79.17	81.45	77.24	79.75	98.19	98.43	78.20	80.59
ATO	71.95	75.76	75.34	77.08	98.57	98.83	73.61	76.41
BUL	75.72	74.08	60.61	64.71	99.35	99.26	67.33	69.08
CEL	61.60	64.18	44.88	50.14	99.40	99.42	51.92	56.30
ECZ	75.19	78.41	81.59	83.79	96.69	97.24	78.26	81.01
WXA	62.99	65.17	64.39	67.00	98.88	98.97	63.68	66.07
ALO	76.96	81.19	85.64	84.10	99.70	99.78	81.07	82.62
HER	77.87	77.99	71.12	74.19	99.33	99.32	74.34	76.04
PIG	69.57	73.31	68.50	71.87	98.72	98.91	69.03	72.59
LUPUS	69.61	74.60	59.38	63.64	99.20	99.35	64.09	68.68
MEL	82.85	83.46	80.63	83.46	99.36	99.38	81.72	83.43
NAIL	89.64	89.08	88.71	90.01	99.00	98.95	89.17	89.53
POI	76.81	75.33	56.84	61.39	99.62	99.57	65.33	67.65
PSO	78.39	79.61	78.65	81.91	97.09	97.26	78.52	80.75
SCA	74.51	77.42	62.19	70.70	99.22	99.27	67.80	73.91
SEB	79.14	85.16	86.10	87.15	96.47	97.69	82.48	86.14
SYS	68.61	72.35	72.06	72.79	98.38	98.67	70.29	72.57
TIN	80.97	80.97	83.70	85.73	97.66	97.68	82.31	83.28
URT	75.67	78.21	75.09	75.85	99.62	99.68	78.38	77.01
VASCT	83.30	84.77	72.80	76.62	99.47	99.51	77.70	80.49
VASCP	72.43	77.24	74.34	75.75	99.03	99.26	73.37	76.49
WARTS	77.76	81.97	81.02	82.76	97.76	98.29	79.36	82.36
Weighted Average	**71.81**	**79.82**	**77.53**	**79.94**	**98.14**	**98.40**	**77.34**	**79.80**
Standard Deviation	06.46	05.89	11.20	09.83	00.95	00.75	08.42	07.72

Figure 1 shows that many reciprocatory misclassifications in Exp-1, like between Eczema (Abbreviated as ECZ in Figure 1) and Psoriasis Lichen Planus (PSO) and between Actinic Keratosis BCC (AKBCC) and Seborrheic Keratosis (SEB), are corrected to a large extent in Exp-2 by utilizing taxonomical relationship among diseases.

We not only performed classification for 23 super-classes but also took a step forward and tried to classify all 622 unique sub-classes as well. We obtained $66.74 \pm 0.64\%$ Top-1 accuracy and $86.26 \pm 0.54\%$ Top-5 accuracy with $98.34 \pm 0.09\%$ AUC. Small values of standard deviation in all of these results signify the stability and consistency of our classifier's performance.

Previous works on DermNet have generally opted for a subset of 23 super-classes for classification. However, Haofu Liao [33] chose to classify all 23 classes and reported best Top-1 accuracy of 73.1% and Top-5 accuracy of 91% on 1000 randomly chosen test images. Cícero et al. [34] reported Top-1 accuracy of 60% on 24 classes (they split "Melanoma and Melanocytic Nevi" into malignant and benign classes). They picked only 100 examples of each class for their test set. To the best of our knowledge, previously the classification task with highest number of classes using DermNet has been performed by Prabhu et al. [35]. They performed 200-ary classification and obtained highest Mean Class Accuracy

(MCA) around 51%. Classification accuracy and AUC of individual models for 622-ary classification are given in Table A2 in Appendix A.



(**a**) Confusion Matrix for Exp-1 on DermNet (**b**) Confusion Matrix for Exp-2 on DermNet

Figure 1. Accumulated confusion matrix of 23-ary classification of DermNet dataset.

3.2. Results on ISIC Archive-2018

ISIC Archive consists of high resolution clinical and dermoscopic images. It does not provide any ontology information about the diseases. Therefore, the approach used in Exp-2 for DermNet cannot be applied here. We achieved Top-1 accuracy of 93.06% $\pm$ 0.31% and Top-2 accuracy of 98.18% $\pm$ 0.06% with 99.23% $\pm$ 0.02% AUC using ensemble approach. Since this dataset has only seven classes, we restricted ourselves to Top-2 accuracy. Table 4 shows that the ensemble of four classifiers was able to achieve high precision of over 80% for all classes except Vascular Lesions that can be justified by small number of images (157 only) in this class. Confusion matrix showing number of correctly classified and misclassified images per class in this dataset is shown in Figure 2. Table A3 in Appendix A presents accuracy and AUC scores of individual classifiers.

Table 4. Performance Metrics of ISIC Archive-2018 using Ensemble.

Class	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Bowen Disease (AKIEC)	80.43	78.74	99.71	79.58
Basal Cell Carcinoma (BCC)	91.85	86.96	99.79	89.34
Benign Keratosis-like Lesions (BKL)	85.55	77.48	98.95	81.32
Dermatofibroma (DF)	91.67	81.15	99.96	86.09
Melanoma (MEL)	84.64	66.05	98.75	74.20
Melanocytic Nevi (NV)	94.90	98.30	79.09	96.57
Vascular Lesions (VASC)	66.10	74.52	99.73	70.06
Weighted Average	85.02	80.46	96.57	82.45
Standard Deviation	09.10	09.38	07.15	08.38

The ISIC Challenges of 2016 [26] and 2017 [36] have focused on binary classification of skin lesions whereas ISIC Challenge 2018 [37] included seven classes. However, as shown in our experiments, DL has enormous capacity to discern far many diseases with high sensitivity and specificity if given enough data. While reliable and accurate detection of melanoma is of utmost importance because of its lethality, it might also be of interest for dermatologists to use CAD to detect other non-lethal skin diseases.

Figure 3 shows some examples of correct and misclassified images. We observed that some of these misclassified images had very high correlation with other classes. For example, there is significantly small inter-class variance between Figure 3a and Figure 3e and between Figure 3d and

Figure 3h. Therefore, CAD had really hard time classifying those classes. ROC AUC curves for all experiments are depicted in Figure A1 in Appendix A.

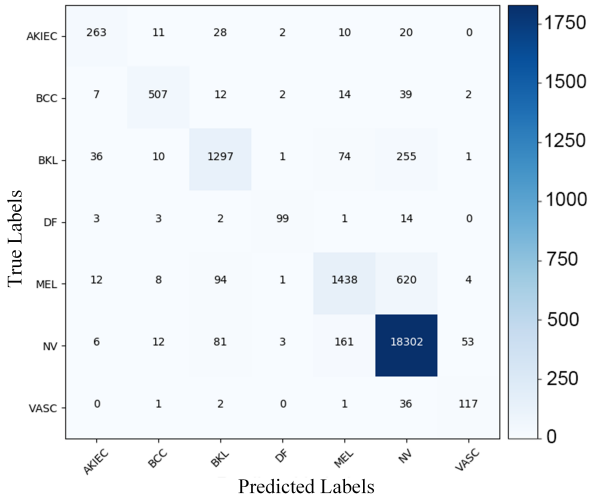
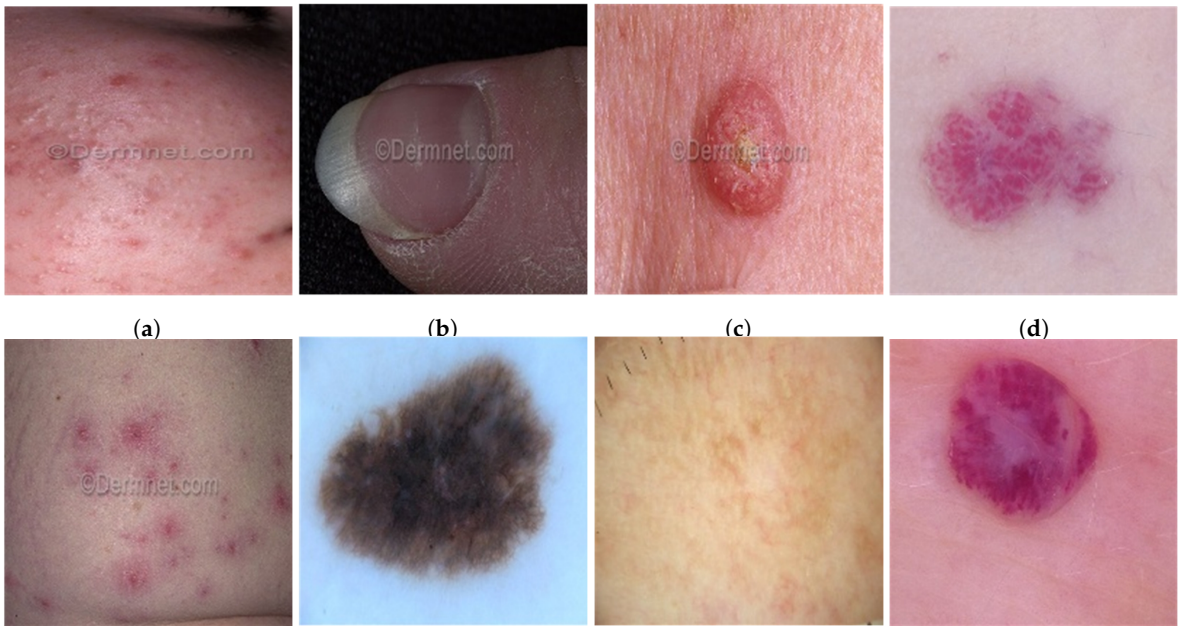


Figure 2. Confusion Matrix showing number of correctly classified and misclassified images per class in ISIC Archive-2018.



(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

Figure 3. Examples of correctly and incorrectly classifies diseases. (**a**) Correctly classified ACROS in DermNet (**b**) Correctly Classified NAIL in DermNet (**c**) Correctly Classified SEB in DermNet (**d**) Correctly Classified VASC in ISIC (**e**) CEL Misclassified as ACROS in DermNet (**f**) Correctly Classified AKIEC in ISIC (**g**) BKL Miscalssified as MEL in ISIC (**h**) NV Misclassified as VASC in ISIC. All Images are resized to fit in square windows.

4. Discussion

Automated diagnosis of skin diseases has enjoyed much attention from researchers for quite some time now. However, most of these researches confine themselves to only binary or ternary classification [38–43] even when large number of classes are available [44]. The importance of early detection of melanoma is understandable given the growing risk it poses to the patient's survival with every passing day. However, there are thousands of other skin diseases [24] that might not be as fatal

as melanoma but have an enormous impact on a patient's quality of life. DL is extremely competent to take on hundreds of classes simultaneously, as evident by our results. We believe that this is right time to harvest the potential of DL to its full extent and start conducting real impactful research that can actually translate into industry standard solution for automated skin disease diagnosis on a larger scale. These solutions can have a far-reaching social impact by not only helping dermatologist with their diagnosis in a clinical setup but also providing an economical and efficient initial screening for underprivileged people in both developed and developing countries.

Another consideration in terms of application of DL in dermatology is that many researchers either use private datasets or public datasets with their own choice of train/test splits (although randomly taken) and number of classes. For this reason, there is little common ground, and often times no ground at all, to compare various classification methods—as also noted by Brinker et al. [45]. This issue of non-comparability can be resolved by collecting and maintaining a standardized publicly available large dataset with explicitly specified train/test splits and standard performance metrics for benchmarking. Notwithstanding that some public datasets, like ISIC Challenges datasets, do provide this beforehand train/test split but their size is normally small and task is usually restricted to binary or ternary classification. Any research on such small datasets cannot be reliably generalized and although the results are publishable, they cannot be used as stepping stone for practical applications of AI in real-world diagnosis. On the other hand, large public datasets normally have a lot of noise, images with disgracefully low resolution or are watermarked. Significant useful information required for fine-grained classification of seemingly similar diseases is lost in such low resolution or watermarked images. Additionally, non-visual metadata, like medical history, is not usually available with medical image datasets. However, this additional information could be pivotal for confident and accurate diagnosis. We were able to utilize disease taxonomy for DermNet dataset and improve our results by 2.5% (refer to Table A1). If multi-model datasets are curated and provided publicly, AI can surely leverage additional information to improve its classification performance.

While understanding and interpreting results of any AI-based classifier it is important to realize that accuracy, or even sensitivity and specificity, might not portray the complete picture of a model's performance. That is why Area Under Receive Operating Characteristic (ROC) Curve (AUC) is also reported along with other performance metrics. From AI point of view, we might argue that achieving around 80% average sensitivity with 1.6% average false positive rate (Table 3, Exp-2) for 23-ary classification task using highly unbalanced datasets of low-resolution and watermarked images is a reasonable achievement. Nevertheless, the actual performance of any AI-based classifier can be significantly different in practical clinical setup as noted by Navarrete-Dechent et al. [46]. They found that the classifier developed by Han et al. [47] did not generalize well when presented with data from an archive of different demography than the one which was used to train the classifier. For a dermatologist it is certainly a cause of concern. However, Han et al. advocated in their response [48] that a classifier should not be judged merely on the bases of sensitivity and specificity. The ROC curves indicate the true ability of a classifier to perform under a wide range of operating points or thresholds while making a diagnosis prediction for a given image. Varying this threshold from 0 to 1 on model's output can change the trade-off between sensitivity and specificity and yield different accuracy. Therefore, higher AUC values ensure that the model has the ability to correctly predict a certain disease, for examples melanoma, with minimum chance of classifying any other disease as that particular disorder.

5. Conclusions

In this paper we have build on previous works on CAD for dermatology and exhibited that DNNs are fairly competent to identify hundreds of skin lesions, and therefore, should be exploited to their full potential instead of employing them to classify only a handful of diseases. We have also set new state-of-the-art result for 23-ary classification on DermNet. Non-visual metadata is not normally available with most of medical image datasets. However, if such additional information is available, DNNs are capable of utilizing it and improving their classification performance as is evident from our experiment with using disease taxonomy to noticeably improve our classification accuracy.

Author Contributions: Conceptualization, M.N.B. and S.A.S.; Formal analysis, S.A.S. and S.A.B.; Investigation, K.M.; Methodology, M.N.B. and K.M.; Project administration, M.I.M. and S.A.; Supervision, B.H., A.D. and S.A.; Writing—original draft, M.N.B.; Writing—review & editing, K.M., M.I.M., S.A.B., B.H. and S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly funded by National University of Science and Technology (NUST) Pakistan (0972/F008/HRD/FDP), BMBF project DeFuseNN (01IW17002) and BMBF project ExplAINN (01IS19074).

Acknowledgments: The authors would like to extend their gratitude to Dieter Metze and Kerstin Steinbrink for providing insightful feedback and valuable suggestions to improve the draft. M. N. Bajwa is also thankful to Arbab Naila for helping with validation of results.

Conflicts of Interest: The authors have no conflict of interest to declare.

Abbreviations

The following abbreviations are used in this manuscript:

DL	Deep Learning
AI	Artificial Intelligence
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
DNN	Deep Neural Network
ISBI	International Symposium on Biomedical Imaging
ISIC	International Skin Imaging Collaboration
AUC	Area Under the Curve
MCA	Mean Class Accuracy
ROC	Receiver Operating Characteristic

Appendix A

This section presents classification accuracy and AUC for individual classifiers and their ensembles for both DermNet and ISIC Archive-2018 datasets.

Table A1. Detailed results of 23-ary classification for individual classifiers and their ensemble on DermNet.

Model	Top-1 Accuracy (%)		Top-5 Accuracy (%)		AUC (%)	
	Exp-1	Exp-2	Exp-1	Exp-2	Exp-1	Exp-2
Resnet-152	70.13 ± 0.89	75.09 ± 0.40	91.17 ± 0.61	93.12 ± 0.31	96.15 ± 0.27	97.31 ± 0.11
Densenet-161	73.34 ± 0.68	77.21 ± 0.40	92.16 ± 0.36	93.91 ± 0.35	96.61 ± 0.15	97.66 ± 0.06
SE_ResNeXt-101	74.46 ± 0.29	77.28 ± 0.60	92.59 ± 0.95	94.07 ± 0.25	96.84 ± 0.22	97.56 ± 0.05
NASNet	72.78 ± 0.73	77.21 ± 0.48	91.68 ± 0.58	92.57 ± 0.32	96.19 ± 0.34	96.79 ± 0.15
Ensemble	77.53 ± 0.64	79.94 ± 0.45	93.87 ± 0.37	95.02 ± 0.15	97.60 ± 0.15	98.11 ± 0.07

Table A2. Detailed results of 622-ary classification for individual classifiers and their ensemble on DermNet.

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)	AUC (%)
Resnet-152	60.82 ± 0.51	82.16 ± 0.43	98.50 ± 0.10
Densenet-161	63.51 ± 0.68	84.46 ± 0.46	98.49 ± 0.06
SE_ResNeXt-101	64.03 ± 0.77	84.26 ± 0.66	98.48 ± 0.08
NASNet	60.69 ± 0.72	81.09 ± 0.61	97.90 ± 0.03
Ensemble	66.74 ± 0.64	86.26 ± 0.54	98.77 ± 0.07

Table A3. Detailed results of 7-ary classification for individual classifiers and their ensemble on ISIC Archive-2018.

Model	Top-1 Accuracy (%)	Top-2 Accuracy (%)	AUC (%)
Resnet-152	89.79 ± 0.29	97.30 ± 0.24	98.97 ± 0.02
Densenet-161	91.27 ± 0.35	97.46 ± 0.21	99.04 ± 0.03
SE_ResNeXt-101	91.63 ± 0.17	97.77 ± 0.21	99.07 ± 0.03
NASNet	91.52 ± 0.38	97.57 ± 0.28	98.97 ± 0.05
Ensemble	93.06 ± 0.31	98.18 ± 0.06	99.23 ± 0.02

Figure A1 shows ROC curves and Area under these ROC curves for all experiments conducted and reported above.

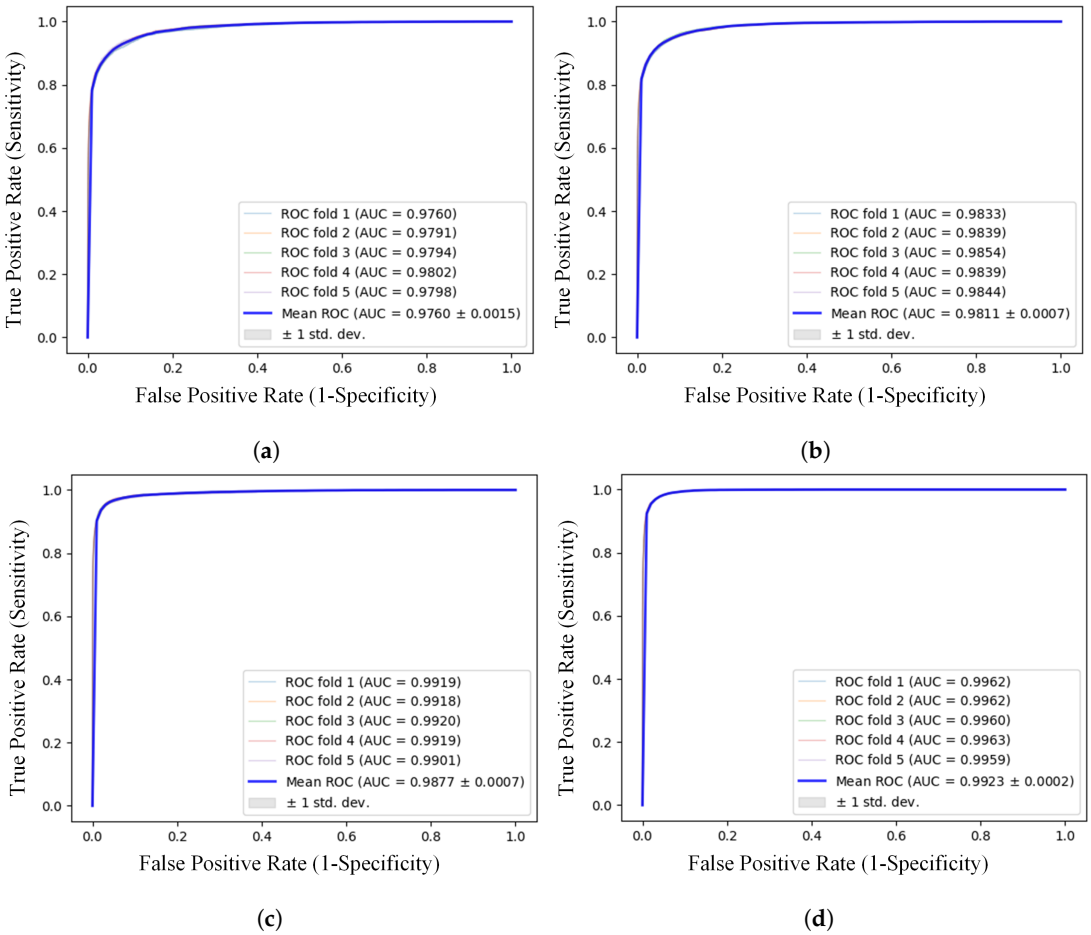


Figure A1. Receiver Operating Characteristics (ROC) curves for DermNet and ISIC Archive-2018 datasets. (**a**) ROC curve for 23-ary classification of DermNet without using ontology. (**b**) ROC curve for 23-ary classification of DermNet with using ontology (**c**) ROC curve for 622-ary classification of DermNet. (**d**) ROC curve for 622-ary classification of ISIC Archive.

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
2. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
3. Hirschberg, J.; Manning, C.D. Advances in natural language processing. *Science* **2015**, *349*, 261–266. [CrossRef] [PubMed]

4. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
5. Wen, L.T.; Tanaka, A.; Nonoyama, M. Identification of Marek's disease virus nuclear antigen in latently infected lymphoblastoid cells. *J. Virol.* **1988**, *62*, 3764–3771. [CrossRef]
6. Teare, P.; Fishman, M.; Benzaquen, O.; Toledano, E.; Elnekave, E. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. *J. Dig. Imaging* **2017**, *30*, 499–505. [CrossRef]
7. Bajwa, M.N.; Malik, M.I.; Siddiqui, S.A.; Dengel, A.; Shafait, F.; Neumeier, W.; Ahmed, S. Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 136.
8. Bajwa, M.N.; Taniguchi, Y.; Malik, M.I.; Neumeier, W.; Dengel, A.; Ahmed, S. Combining Fine-and Coarse-Grained Classifiers for Diabetic Retinopathy Detection. In Proceedings of the Annual Conference on Medical Image Understanding and Analysis, Liverpool, UK, 24–26 July 2019; pp. 242–253.
9. Prevedello, L.M.; Erdal, B.S.; Ryu, J.L.; Little, K.J.; Demirer, M.; Qian, S.; White, R.D. Automated critical test findings identification and online notification system using artificial intelligence in imaging. *Radiology* **2017**, *285*, 923–931. [CrossRef]
10. Carli, P.; Quercioli, E.; Sestini, S.; Stante, M.; Ricci, L.; Brunasso, G.; De Giorgi, V. Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. *Br. J. Dermatol.* **2003**, *148*, 981–984. [CrossRef]
11. Carrera, C.; Marchetti, M.A.; Dusza, S.W.; Argenziano, G.; Braun, R.P.; Halpern, A.C.; Jaimes, N.; Kittler, H.J.; Malvehy, J.; Menzies, S.W.; et al. Validity and reliability of dermoscopic criteria used to differentiate nevi from melanoma: A web-based international dermoscopy society study. *JAMA Dermatol.* **2016**, *152*, 798–806. [CrossRef]
12. Masood, A.; Ali Al-Jumaily, A. Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms. *Int. J. Biomed. Imaging* **2013**, *2013*. [CrossRef] [PubMed]
13. Argenziano, G.; Soyer, H.; De Giorgi, V.; Piccolo, D.; Carli, P.; Delfino, M. Interactive Atlas of Dermoscopy (Book and CD-ROM). 2000. Availabel onine: http://www.dermoscopy.org/atlas/order_cd.asp (accessed on 5 March 2020).
14. Ballerini, L.; Fisher, R.B.; Aldridge, B.; Rees, J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*; Springer: Berlin, Germany, 2013; pp. 63–86.
15. Giotis, I.; Molders, N.; Land, S.; Biehl, M.; Jonkman, M.F.; Petkov, N. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Syst. Appl.* **2015**, *42*, 6578–6585. [CrossRef]
16. Mendonça, T.; Ferreira, P.M.; Marques, J.S.; Marcal, A.R.; Rozeira, J. PH 2-A dermoscopic image database for research and benchmarking. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 5437–5440.
17. Ali, A.R.A.; Deserno, T.M. A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data. Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment. *Int. Soc. Opt. Photon.* **2012**, *8318*, 83181I.
18. Sato, I.; Nishimura, H.; Yokoi, K. Apac: Augmented pattern classification with neural networks. *arXiv* **2015**, arXiv:1505.03229.
19. Graham, B. Fractional max-pooling. *arXiv* **2014**, arXiv:1412.6071.
20. Lee, C.Y.; Gallagher, P.W.; Tu, Z. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. *Artif. Intell. Stat.* **2016**, 464–472.
21. Liu, C.; Zoph, B.; Neumann, M.; Shlens, J.; Hua, W.; Li, L.J.; Fei-Fei, L.; Yuille, A.; Huang, J.; Murphy, K. Progressive neural architecture search. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 19–34.
22. Kawahara, J.; BenTaieb, A.; Hamarneh, G. Deep features to classify skin lesions. In Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague, Czech Republic, 13–16 April 2016; pp. 1397–1400.

23. Ge, Z.; Demyanov, S.; Bozorgtabar, B.; Abedini, M.; Chakravorty, R.; Bowling, A.; Garnavi, R. Exploiting local and generic features for accurate skin lesions classification using clinical and dermoscopy imaging. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, Australia, 18–21 April 2017; pp. 986–990.
24. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
25. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842. [CrossRef]
26. Gutman, D.; Codella, N.C.; Celebi, E.; Helba, B.; Marchetti, M.; Mishra, N.; Halpern, A. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). *arXiv* **2016**, arXiv:1605.01397.
27. Li, Y.; Esteva, A.; Kuprel, B.; Novoa, R.; Ko, J.; Thrun, S. Skin cancer detection and tracking using data synthesis and deep learning. In Proceedings of the Workshops at the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–5 February 2017.
28. Li, Y.; Shen, L. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors* **2018**, *18*, 556. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
32. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8697–8710.
33. Liao, H. *A Deep Learning Approach to Universal Skin Disease Classification*; University of Rochester Department of Computer Science, CSC: Rochester, NY, USA, 2016.
34. Cıcero, F.M.; Oliveira, A.H.M.; Botelho, G.M.; da Computaçao, C.d.C. Deep learning and convolutional neural networks in the aid of the classification of melanoma. In Proceedings of the Conference on Graphics, Patterns and Images, SIBGRAPI, Sao Paulo, Brazil, 4–7 October 2016.
35. Prabhu, V.; Kannan, A.; Ravuri, M.; Chablani, M.; Sontag, D.; Amatriain, X. Prototypical Clustering Networks for Dermatological Disease Diagnosis. *arXiv* **2018**, arXiv:1811.03066.
36. Codella, N.C.; Gutman, D.; Celebi, M.E.; Helba, B.; Marchetti, M.A.; Dusza, S.W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 168–172.
37. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* **2019**, arXiv:1902.03368.
38. Menegola, A.; Fornaciali, M.; Pires, R.; Bittencourt, F.V.; Avila, S.; Valle, E. Knowledge transfer for melanoma screening with deep learning. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, Australia, 18–21 April 2017; pp. 297–300.
39. Nasr-Esfahani, E.; Samavi, S.; Karimi, N.; Soroushmehr, S.M.R.; Jafari, M.H.; Ward, K.; Najarian, K. Melanoma detection by analysis of clinical images using convolutional neural network. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 17–20 August 2016; pp. 1373–1376.
40. Yu, L.; Chen, H.; Dou, Q.; Qin, J.; Heng, P.A. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans. Med. Imaging* **2016**, *36*, 994–1004. [CrossRef] [PubMed]

41. Codella, N.; Cai, J.; Abedini, M.; Garnavi, R.; Halpern, A.; Smith, J.R. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In Proceedings of the International Workshop on Machine Learning in Medical Imaging, Munich, Germany, 5 October 2015; pp. 118–126.
42. Ma, Z.; Tavares, J.M.R. Effective features to classify skin lesions in dermoscopic images. *Expert Syst. Appl.* **2017**, *84*, 92–101. [CrossRef]
43. Lopez, A.R.; Giro-i Nieto, X.; Burdick, J.; Marques, O. Skin lesion classification from dermoscopic images using deep learning techniques. In Proceedings of the 2017 13th IASTED International Conference on Biomedical Engineering (BioMed), Innsbruck, Austria, 20–21 February 2017; pp. 49–54.
44. Shoieb, D.A.; Youssef, S.M.; Aly, W.M. Computer-aided model for skin diagnosis using deep learning. *J. Image Graph.* **2016**, *4*, 122–129. [CrossRef]
45. Brinker, T.J.; Hekler, A.; Utikal, J.S.; Grabe, N.; Schadendorf, D.; Klode, J.; Berking, C.; Steeb, T.; Enk, A.H.; von Kalle, C. Skin cancer classification using convolutional neural networks: Systematic review. *J. Med. Internet Res.* **2018**, *20*, e11936. [CrossRef]
46. Navarrete-Dechent, C.; Dusza, S.W.; Liopyris, K.; Marghoob, A.A.; Halpern, A.C.; Marchetti, M.A. Automated dermatological diagnosis: Hype or reality? *J. Invest. Dermatol.* **2018**, *138*, 2277–2279. [CrossRef]
47. Han, S.S.; Kim, M.S.; Lim, W.; Park, G.H.; Park, I.; Chang, S.E. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J. Invest. Dermatol.* **2018**, *138*, 1529–1538. [CrossRef]
48. Han, S.S.; Lim, W.; Kim, M.S.; Park, I.; Park, G.H.; Chang, S.E. Interpretation of the Outputs of a Deep Learning Model Trained with a Skin Cancer Dataset. *J. Invest. Dermatol.* **2018**, *138*, 2275. [CrossRef]