Analysis of the ISIC image datasets: Usage, benchmarks and recommendations

Bill Cassidy [a], Connah Kendrick [a], Andrzej Brodzicki [b], Joanna Jaworek-Korjakowska [b], Moi Hoon Yap [a,*]

[a] *Manchester Metropolitan University, John Dalton Building, Chester Street, Manchester M1 5GD, UK*
[b] *AGH University of Science and Technology, Al Mickiewicza 30, 30–059 Krakow, Poland*

A B S T R A C T

The International Skin Imaging Collaboration (ISIC) datasets have become a leading repository for researchers in machine learning for medical image analysis, especially in the field of skin cancer detection and malignancy assessment. They contain tens of thousands of dermoscopic photographs together with gold-standard lesion diagnosis metadata. The associated yearly challenges have resulted in major contributions to the field, with papers reporting measures well in excess of human experts. Skin cancers can be divided into two major groups - melanoma and non-melanoma. Although less prevalent, melanoma is considered to be more serious as it can quickly spread to other organs if not treated at an early stage. In this paper, we summarise the usage of the ISIC dataset images and present an analysis of yearly releases over a period of 2016 - 2020. Our analysis found a significant number of duplicate images, both within and between the datasets. Additionally, we also noted duplicates spread across testing and training sets. Due to these irregularities, we propose a duplicate removal strategy and recommend a curated dataset for researchers to use when working on ISIC datasets. Given that ISIC 2020 focused on melanoma classification, we conduct experiments to provide benchmark results on the ISIC 2020 test set, with additional analysis on the smaller ISIC 2017 test set. Testing was completed following the application of our duplicate removal strategy and an additional data balancing step. As a result of removing 14,310 duplicate images from the training set, our benchmark results show good levels of melanoma prediction with an AUC of 0.80 for the best performing model. As our aim was not to maximise network performance, we did not include additional steps in our experiments. Finally, we provide recommendations for future research by highlighting irregularities that may present research challenges. A list of image files with reference to the original ISIC dataset sources for the recommended curated training set will be shared on our GitHub repository (available at www.github.com/mmu-dermatology-research/isic_duplicate_removal_strategy).

1. Introduction

Skin cancer is the most common of all cancers, with more people being diagnosed with the condition each year than all other cancers combined. There are 9500 new cases being diagnosed every day in the US (Skin Cancer Foundation, 2017). Melanoma, the deadliest form of skin cancer, is projected to reach almost half a million cases by 2040. This represents a 62% increase since 2018. One person dies of skin cancer every 4 minutes. As such, the rise of skin cancer incidence is seen by many dermatologists as a global epidemic (Melanoma UK, 2020). Early intervention for skin cancer, melanoma in particular, is essential to ensure high survival rates in the face of an ever growing number of cases (Thörn et al., 1994; Cormier et al., 2015). The main identifiable cause of skin cancer is excessive exposure to ultraviolet (UV) radiation (NHS, 2020a). This may be due to exposure to natural sunlight (Cancer Research UK), or from other UV sources such as indoor tanning devices (World Health Organization, 2017). Depleted ozone levels lead to a rise in ground-level UV radiation which can increase the risk of exposure in natural sunlight (Department for Environment Food & Rural Affairs, 2020). There is also evidence of increased incidence of non-melanoma skin cancer in populations living in lower latitude regions where UV radiation levels are high (Henriksen et al., 1989). Other modifiable risk factors may also include poor diet (Sarnoff and Gerome, 2017), alcohol consumption (Ruiz, 2018; American Institute for Cancer Research, 2018) and smoking (De Hertog et al., 2001).

---

* Corresponding author.
    E-mail address: M.Yap@mmu.ac.uk (M.H. Yap).

Dermoscopy is a widely used imaging technique that enables the skin surface to be visualised by light-amplification using immersion fluid (Kittler et al., 2002), however, its diagnostic accuracy is highly dependant on the experience of dermatologists (Brinker et al., 2019b; 2019c; Haenssle et al., 2018). Scarcity of expert resources in poorer countries can significantly impact timely treatment for skin cancers. Many of the publicly available statistics relating to skin cancer are thought to be underestimates due to issues such as non-melanoma cases not being tracked by cancer registries, incomplete registrations due to successful treatment or poorer countries not having cancer registries (American Institute for Cancer Research, 2018).

Due to increased demands that skin cancer cases are incurring on global healthcare services, the need for remote automated diagnosis solutions is becoming increasingly important. This is particularly pertinent in poorer countries where patients do not have access to the latest medical equipment and expertise required for accurate diagnosis. Skin lesion classification has become a popular field of research in recent years following the growing adoption of deep learning techniques in the field of medical image analysis. However, as the majority of the state-of-the-art solutions are data-driven, the reliability and the consistency of open datasets are key factors for algorithm development. Therefore, in this paper we analyse images from the largest dermoscopic open datasets released over the past five years - the International Skin Imaging Collaboration (ISIC) datasets. The main contributions of this paper are as follows:

1. We analyse the usage of ISIC image datasets with a selection of well-cited research papers from the past 3–4 years and identify related issues.
2. We propose a duplicate removal strategy to curate the datasets. By removing the duplicate images (overlap images between and within the test and training sets), we produced a cleaned (non-duplicate) dataset and a balanced dataset.
3. We benchmark the curated balanced training set using 19 state-of-the-art deep learning architectures for melanoma recognition. We evaluate the performance of our benchmark algorithms on the ISIC 2020 testing set (on Kaggle) for binary classification (melanoma and non-melanoma), with additional analysis on ISIC 2017 testing set.
4. We provide recommendations for future research and share our research findings on our GitHub repository (available at www.github.com/mmu-dermatology-research/isic_duplicate_removal_strategy).

2. Related work

This section outlines the usage of dermoscopic datasets, focusing on the ISIC image datasets and issues relating to their use, including research discussing duplicate images, class imbalance, image resolution and label noise. A growing number of studies have demonstrated that CNNs are just as capable as human experts in the diagnosis of malignant and benign skin lesions, and in some cases are able to out-perform them (Esteva et al., 2017b; 2017a; Brinker et al., 2019b; 2019a; Fujisawa et al., 2019; Pham et al., 2020; Jinnai et al., 2020). A shortage of experienced dermatologists in some countries, combined with high observer variability, presents an opportunity to develop solutions that address this problem.

2.1. Usage of ISIC datasets

We conducted a survey on the usage of ISIC datasets for research purposes. As ISIC datasets are widely used, it is not possible to provide an exhaustive list, however, we have selected some of the more prominent well-cited papers from the past 3–4 years for our analysis of usage that demonstrate significant contributions in the field. Table 1 shows a summary of the 35 papers we surveyed. From these publications, only 5 implemented some sort of duplicate removal. The remaining 30 papers did not mention any form of duplicate removal. Additionally, we observe that most of the recent research used multiple datasets, where the number of papers indicating the use of single ISIC datasets was 13, and 22 papers indicated the use of multiple ISIC datasets. Given the large number of duplicates across the ISIC datasets, we observe that experiments that utilise multiple ISIC datasets that do not implement some form of duplicate removal may exhibit bias in their results. Therefore, we propose to perform an analysis to verify the existence of such biases.

The usage of ISIC datasets is broad, with the majority of tasks focusing on classification and segmentation. The most popular research involves binary classification, as these challenges provide more images to train the algorithms. With the introduction of ISIC 2018 and ISIC 2019, researchers started to explore multi-class classification, the majority of which used the ISIC 2020 dataset. However, the ISIC 2020 challenge focused on melanoma detection, therefore further additional binary classification papers are expected. Segmentation tasks appear to be not as popular as lesion diagnosis as ISIC did not continue this challenge type beyond 2019. Only the ISIC 2016–2018 datasets provided delineated segmentation masks, and are relatively few in number compared to those found in the classification tasks. Other usage of ISIC datasets include a study of the effect of colour constancy (Ng et al., 2019) and data augmentation using generative adversarial networks (Kendrick et al., 2020).

2.2. Related issues

In this section we present the issues related to the usage of ISIC dataset images, supported by the findings of recent state-of-the-art research.

Image duplication and images exhibiting high similarity within datasets used to train CNNs may introduce unwanted bias in the resulting models. To address this problem, researchers have investigated methods of identifying visually similar images within large datasets. Hu et al. (2018) proposed the use of a deep constrained siamese hash coding network with binary constrained regularization to detect near duplicate images. They tested their network on three datasets and demonstrated an additional load balancing method that was shown to further increase performance in terms of accuracy and speed. Zhang (2018) adopted a different approach to test for image similarity. They implemented a deep CNN using a double-channel architecture. Such architectures might prove useful in the balancing of deep learning datasets, especially those where a high number of samples have been sourced from a low number of participants. More recently, a number of researchers (Sucholutsky and Schonlau, 2020) have suggested that the number of unique features contribute more to network performance as opposed to simply increasing the number of input images. Although data augmentation is widely used in the majority of previous research (as shown in Table 1), it is unclear if this helps in increasing the number of unique features.

Brinker et al. (2018) and Gessert et al. (2020) showed that there was a clear benefit to network performance from the inclusion of patient metadata when training CNNs. Rezvantalab et al. (2018) conducted experiments on two public dermoscopy skin cancer datasets using four CNNs pretrained on ImageNet in the classification of eight skin lesion types, including melanoma. These networks were trained using the HAM10000 and PH^2 datasets, the former comprising a large part of the ISIC datasets, namely ISIC 2018 - 2020. This work concluded

Table 1

A summary of a subset of research papers that use the ISIC image datasets (non-exhaustive list). *Multi-class classification divided into seven classes: (1) actinic keratosis, (2) basal cell carcinoma, (3) benign keratosis, (4) dermatofibroma, (5) melanocytic nevi, (6) melanoma, (7) vascular skin lesion. **Multi-class classification into eight classes: (1) melanoma, (2) melanocytic nevi, (3) basal cell carcinoma, (4) benign keratosis, (5) actinic keratosis and intraepithelial carcinoma, (6) dermatofibroma, (7) vascular lesions, (8) atypical nevi. ***(keratinocyte carcinomas vs benign seborrheic keratosis and malignant melanomas vs benign nevi ****Three classes: (1) melanoma, (2) nevi, and (3) benign.

Publication	Dupl. Removal	Datasets	No. of Images	DA	Usage
Le et al. (2020)	Yes	HAM10000	7470	Yes	Multi-class (7 classes*)
Hekler et al. (2020)	Yes	HAM10000, ISIC (not specified)	804	No	Binary classification
Rotemberg et al. (2020)	Yes	ISIC 2020	33,126	No	Binary classification
Bisla et al. (2019a)	Yes	ISIC 2017, ISIC 2018, PH^2, Dermofit	1875	Yes	Binary classification
Bissoto et al. (2019)	Yes	Atlas, ISIC 2018	3466	Yes	Binary classification
Hasan et al. (2021)	No	ISIC 2016, ISIC 2017, ISIC 2018	14,044	Yes	Binary classification
Xie et al. (2021)	No	ISIC 2017/2018	6344	Yes	Segmentation
Brinker et al. (2019a)	No	train ISIC (not specified), test PH^2	4204	No	Binary classification
Acosta et al. (2021)	No	ISIC 2017, PH^2 malignant	2,742	Yes	Binary classification
Adegun and Viriri (2020a)	No	ISIC 2017, PH^2	2860	Yes	Binary classification
Hasan et al. (2020)	No	ISIC 2017, PH^2	2750	Yes	Semantic segmentation (3-class ISIC, binary PH^2)
Xie et al. (2020)	No	ISIC 2017, PH^2	3520	Yes	Segmentation and binary classification
Nahata and Singh (2020)	No	ISIC 2018/2019	35,348	Yes	Binary classification
Ha et al. (2020)	No	ISIC 2018, ISIC 2019,ISIC 2020	33,000+	Yes	Binary classification (with 9-class output)
Gessert et al. (2020)	No	ISIC 2019, 7-point, In-house	27,665	Yes	Multi-class (8 classes**)
Adegun and Viriri (2020b)	No	ISIC 2018, ISIC 2019	-	-	State-of-the-art survey
Goyal et al. (2019)	No	ISIC 2017, PH^2	3520	Yes	Segmentation
Hekler et al. (2019)	No	HAM10000, ISIC supplement	11,444	Yes	5-class and binary classification
Brinker et al. (2019b)	No	ISIC 2016, HAM10000	12,378	No	Binary classification
Mahbod et al. (2019)	No	ISIC 2016/2017	2787	Yes	Binary classification + seborrheic keratosis
Hosny et al. (2019)	No	MED-NODE, Derm (IS & Quest), ISIC 2017	2376	Yes	Binary classification and Multi-class (3 classes)
Bisla et al. (2019b)	No	ISIC 2017/2018, Dermofit, PH^2	16,270	Yes GAN	Segmentation and binary classification
Tang et al. (2019)	No	ISIC 2016/2017, PH^2	4079	Yes	Segmentation
Rezvantalab et al. (2018)	No	HAM10000, PH^2	10,135	No	Multi-class (8 classes**)
Carcagn et al. (2019)	No	HAM10000	10,015	Yes	Multi-class (7 classes*)
Tschandl et al. (2019)	No	HAM10000	10,015	-	Multi-class (7 classes*), State-of-the-art survey
Sagar and Dheeba (2020)	No	ISIC (not specified)	3600	Yes	Binary classification
Brinker et al. (2019c)	No	ISIC 2019	13,737	No	Binary classification
Kassem et al. (2020)	No	ISIC 2019	25,331	Yes	Multi-class (8 classes**)
Ratul et al. (2020)	No	HAM10000	10,015	Yes	Multi-class (7 classes*)
Majtner et al. (2019)	No	ISIC 2016	1279	No	Binary classification
Almaraz-Damian et al. (2020)	No	HAM10000	10,015	Yes SMOTE	Binary classification
Barbosa and Baleiras (2019)	No	ISIC 2017	2,750	Yes	Binary classification
Gessert et al. (2018)	No	ISIC 2018	13,500	Yes	Multi-class (7 classes*)
Al-antari et al. (2018)	No	ISIC 2018	11,720	Yes	Multi-class (7 classes*)

Dupl. Removal–authors mention removal of duplicate images, DA–Data augmentation, ISIC (not specified)–year not stated in the paper

that all models had difficulty discerning between melanoma and melanocytic nevi. Additionally, all 4 networks performed poorly when classifying actinic keratosis and benign keratosis.

Class imbalance has been shown to significantly impact model performance (Tschandl et al., 2019), with data augmentation being used in the training of CNNs as a means of addressing this problem. Hosny et al. (2019) conducted six classification experiments using AlexNet to achieve >95% accuracy. They performed two sets of experiments on three datasets - ISIC 2017, MED-NODE and dermatology information system (DermIS). This work showed that various data augmentation techniques combined with adjustments to Softmax contributed to significant improvements in output measures for networks trained on all three datasets. This work was also notable for using a dataset with known low quality images, found in the DermIS dataset, which may have contributed to the robustness of the classification model.

Le et al. (2020) devised an ensemble of ResNet50 networks that utilised class-weighting with a focal loss function to mitigate the inherent class imbalance in the HAM10000 dataset, which they used as training data. They experimented using a pre-processing stage where lesions were segmented. However, this approach re-

sulted in a reduction in accuracy, suggesting that the area of skin surrounding the lesion provides a vital contribution to the discerning features learned by the network.

With the recent development of EfficientNet (Tan and Le, 2020), Gessert et al. (2020) found that EfficientNet models trained on higher resolution images from the ISIC 2019 dataset improved network performance. This is likely due to the scaling functionality inherent within the EfficientNet architecture, where model width and depth are scaled uniformly to input size. They also found that addressing class imbalance using loss balancing improved network performance.

Hekler et al. (2020) investigated the effects of label noise on CNNs for skin cancer classification. This research noted that many skin cancer classification studies used non-biopsy-verified training images, and that such imperfect ground truth could introduce systematic error. They observed a correlation between models trained with diagnosis from several dermatologists and high quality results on a test set whose labels had been produced by dermatologists. They found that CNNs could identify the features that dermatologists also identified, but that the CNNs also learned sources of errors in dermatological decisions. They also observed that if

Table 2
Summary of the ISIC 2016 - 2020 datasets. Note that image counts do not include mask and superpixel images.

Dataset	Train	Test	Total
ISIC 2016	900	379	1279
ISIC 2017	2000	600	2600
ISIC 2018	10,015	1512	11,527
ISIC 2019	25,331	8238	33,569
ISIC 2020	33,126	10,982	44,108

Table 3
Class distribution within the ISIC 2017 - 2020 training sets. Note that all unknown cases for ISIC 2020 are diagnosed as benign.

Class	2017	2018	2019	2020
Melanoma	374	1113	4522	584
Atypical melanocytic proliferation	-	-	-	1
Cafe-au-lait macule	-	-	-	1
Lentigo NOS	-	-	-	44
Lichenoid keratosis	-	-	-	37
Nevus	-	-	-	5193
Seborrheic keratosis	254	-	-	135
Solar lentigo	-	-	-	7
Melanocytic nevus	-	6705	12,875	-
Basal cell carcinoma	-	514	3323	-
Actinic keratosis	-	327	867	-
Benign keratosis	-	1099	2624	-
Dermatofibroma	-	115	239	-
Vascular lesion	-	142	253	-
Squamous cell carcinoma	-	-	628	-
Other / Unknown	1372	-	-	27,124
Total	2000	10,015	25,331	33,126

a CNN trained with majority decisions was tested on a biopsy-verified ground truth, there was a significant decrease in performance, with accuracy dropping from 75.03% to 64.24%. However, there were several limitations in this study, namely (1) they used only 804 test and training images; (2) they tested only one deep learning architecture (ResNet50 pretrained on ImageNet); (3) all lesions assessed were biopsied, which are naturally more difficult to classify, and therefore represent edge cases, with the authors noting that the introduction of simpler cases would likely increase network accuracy.

Researchers in medical image analysis of skin cancer who use dermoscopic image datasets for the early detection of skin cancer and malignancy assessment are focused on developing new computer algorithms. However, issues inherent within the datasets used are often overlooked or under researched. In the following section we analyse the largest and most widely used dermoscopic datasets, namely, the ISIC datasets.

3. Datasets

The ISIC challenges have become a driving force for research into melanoma classification. They provide biopsy-proven digital high resolution skin lesion image datasets, with expert annotations and metadata from around the world. The aim is to promote research in the field, which will lead to the development of automated Computer Aided Diagnosis (CAD) solutions for the diagnosis of melanoma and other cancers. This community also organises yearly skin lesion challenges to attract wider participation of researchers to improve the diagnosis of CAD algorithms and to spread awareness of the growing problem that skin cancer represents (Codella et al., 2018b). Table 2 shows a summary of the number of images within the ISIC datasets (2016–2020). We note that the number of images has increased substantially every year since its introduction.

The ISIC 2016 dataset (Gutman et al., 2016) contains 900 training images and 379 test images, a total of 1279 images. Ground truth data is provided for both training and test sets, indicating if each lesion is malignant or benign. This dataset has limited future use, as in clinical practice dermatologists often identify the specific types of malignancy and benignancy.

The ISIC 2017 dataset (Codella et al., 2017) contains 2000 training images and 600 test images, a total of 2600 images. Ground truth and patient metadata are provided for both training and test sets, indicating if the lesion is one of four class groups: (1) melanoma; (2) nevus or seborrheic keratosis; (3) seborrheic keratosis; or (4) melanoma or nevus. The patient's approximate age and gender are also provided as additional metadata. Table 3 shows the detailed split of class distribution for ISIC 2017 - 2020.

In 2018, ISIC shared a more substantial dataset (Codella et al., 2018a; Tschandl, 2018) which contains 10,015 training images and 1512 test images, a total of 11,527 images. Ground truth data is provided for the training set only, which includes more detailed lesion type labels, including melanoma, melanocytic nevus, basal

cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma and vascular lesions.

In the following year, the ISIC 2019 dataset (Tschandl, 2018; Codella et al., 2017; Combalia et al., 2019) was released. This dataset contains 25,331 training images and 8238 test images, a total of 33,569 images. Similar to ISIC 2018, ground truth data is provided for the training set only, indicating the following classes: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesions and squamous cell carcinoma. The testing set consists of 9 classes, 8 classes as in the training set plus an additional unknown class. Patient metadata is provided for both training and testing sets. The training metadata indicates the patient's approximate age, anatomical site, lesion ID and gender. Lesion ID is specified for 23,247 images, and unspecified for 2084 images, with 11,848 unique IDs from a total of 25,331 images. The testing metadata indicates the patient's approximate age, anatomical site and gender. The ISIC 2019 dataset is also notable for including multiplets of single lesions which feature the same lesion at different zoom levels which may provide important unique features at different levels of magnification.

In 2020, the largest ISIC dataset was released (Rotemberg et al., 2020) which contains 33,126 training images and 10,982 test images, a total of 44,108 images. Similar to the previous year, ground truth data is provided for the training set only, indicating patient ID, lesion ID, gender, approximate age, anatomical site, diagnosis (see Table 3) and benign or malignant status. Of the 33,126 images in the training set, there are 2056 unique patient IDs and 32,701 unique lesion IDs. This would suggest that a large number of lesion images have been sourced from a relatively small pool of patients at different intervals. The test set also includes patient metadata indicating patient ID, patient approximate age, anatomical site and gender.

The ISIC datasets (2016–2020) comprise of 18 underlying sub-datasets. A summary of these sub-datasets is shown in Table 4. We obtained these figures from the ISIC Archive Gallery (ISIC, 2020).

The first observation on overlap images can be seen in the ISIC 2018 - 2020 datasets, which include the HAM10000 dataset, comprising 10,015 training images and 1511 test images, a total of 11,526 images. Moreover, the ISIC 2019 and 2020 datasets include the BCN20000 dataset, a total of 19,424 images, which includes lesions found in hard to diagnose locations such as nails and mucosa. Note that we excluded segmentation masks and super-pixel images from our experiments and analysis.

As shown in Table 3, there are a total of 16 classes and 70,472 images for ISIC training sets 2017 - 2020. We note that although

Table 4
Composition of the ISIC datasets (2016–2020).

Sub-dataset	No. of Images
2018 JID Editorial Images	100
BCN 20,000	12,413
BCN 2020 Challenge	7311
Brisbane ISIC Challenge 2020	8449
Dermoscopedia (CC-BY)	5
HAM10000	10,015
ISIC 2020 Challenge - MSKCC contribution	11,108
ISIC 2020 Vienna (part 1)	2231
ISIC 2020 Vienna (part 2)	2143
MSK-1	1100
MSK-2	1535
MSK-3	225
MSK-4	947
MSK-5	111
SONIC	9251
Sydney (MIA/SMDC) 2020 ISIC challenge contribution	1884
UDA-1	557
UDA-2	60
Total	69,445

Table 5
Train and test splits for individual ISIC datasets associated with classification tasks.

Dataset	Task No.	Train	Test	Total
2016	3	900	379	1279
2017	3	2000	600	2600
2018	3	10,015	1512	11,527
2019	1	25,331	8238	33,569
2020	-	33,126	10,982	44,108
Split Total	-	71,372	21,711	93,083

Table 6
Summary of binary identical image files within individual ISIC datasets. Note that these figures do not include downsampled duplicates.

Dataset	Train	Test	Train & Test
2016	1	0	3
2017	0	2	2
2018	2	0	0
2019	50	0	0
2020	433	78	0

Table 7
Summary of downsampled duplicate image files where the ISIC code in the downsampled file name is the same as a non-downsampled file name. Note that all downsampled image files are part of the 2019 training set.

	2016	2017	2018	2019	2020
Train	291	1283	0	0	0
Test	95	594	0	0	0

the total number of images has doubled from 2019 to 2020, the dataset remains imbalanced, with deficiency in actinic keratosis, dermatofibroma, vascular and squamous cell carcinoma. We also note the significant reduction in the number of melanoma examples between the 2019 and 2020 training sets, in addition to the large number of unknown cases present in the 2020 training set.

To analyse and compare the datasets, we downloaded the ISIC datasets from 2017 to 2020. The following section describes the approaches that we used to analyse these datasets.

4. Method

This section details the following: (1) the implementation of a proposed duplicate removal strategy to address class imbalance within and across the ISIC datasets; (2) following the implementation of our proposed duplicate removal strategy, we curated a new cleaned and balanced dataset (henceforth curated dataset), using images from the ISIC 2017 - 2020 datasets (ISIC 2016 was excluded due to missing labels of the type melanoma and non-melanoma); and (3) we train a selection of the most widely used pretrained deep CNNs using our curated dataset and report on the benchmark results.

4.1. Duplicate removal strategy

As an initial preprocessing stage, we removed all 2000 super-pixel images contained in the ISIC 2017 training dataset and all 600 superpixel images contained in the ISIC 2017 test dataset. A summary of all datasets following the removal of all superpixel image files is shown in Table 5. Task number refers to the task number category on the ISIC dataset website, as some datasets are split into tasks for each year. We only used datasets from classification tasks, where comma-separated value (CSV) ground truth labelling was available for the corresponding training set. Table 6 shows a

summary of binary identical image files present in each dataset. The total number of binary identical duplicate image files found across all training sets is 12,039. This includes duplicates found both within individual training sets and across training sets. The total number of binary identical duplicate image files across all test sets is 1,592, which includes duplicates found both within individual test sets and across test sets. The total number of binary identical image files across all training and test sets is 13,976 which includes duplicates found both within individual training and test sets, and across training and test sets.

The main aim of our experiments was to remove duplicates from the training sets only, as we would be evaluating our baseline results on the ISIC 2020 challenge website. Duplicates were removed in year order, e.g. remove all 2016 training set duplicates, then remove all 2017 training set duplicates, etc. The 2019 training set contains a subset of 2074 downsampled image files, denoted by the filename suffix "_downsampled". These are images that have been reduced in size (height and width) so would not be identified by algorithms that check for identical binary data. There were no images containing the "_downsampled" suffix in any other training or testing set. No formal description of the downsampled images is provided on the ISIC 2019 challenge website or in the associated challenge papers. We identified a total of 2263 duplicate image files in the downsampled set where the ISIC code in the downsampled image file name is the same as the ISIC code in a non-downsampled image file name. A summary of the downsampled duplicates is shown in Table 7. Given that duplicates may exist within training sets, within testing sets and across training and testing sets, we devised a duplicate removal strategy, comprising the following stages:

1. Delete all image files from the 2019 training set where the following criteria are satisfied: (i) the filename contains the suffix "_downsampled"; and (ii) the filename contains the ISIC code found in any other image file in any of the testing sets.
2. Delete all image files from the 2019 training set where the following criteria are satisfied: (i) the filename contains the suffix "_downsampled"; and (ii) the filename contains the ISIC code found in any other image file in any of the training sets.
3. Delete all duplicate binary identical image files across all training sets (2016–2020).
4. Delete all image files from each individual training set where a duplicate is found in any of the test sets.

Table 8
Number of image files deleted from each ISIC training set after applying our duplicate removal strategy. Note that figures include both binary identical duplicates and downsampled duplicates in the 2019 training set.

Year	Task No.	Images Removed	Images Remaining
2016	3	826	74
2017	3	801	1199
2018	3	10,015	0
2019	1	2235	23,096
2020	-	433	32,693
Total	-	14,310	57,062[1]

1 Note: this total differs from the total number of images we used in our combined training set as we did not use the 2016 dataset.
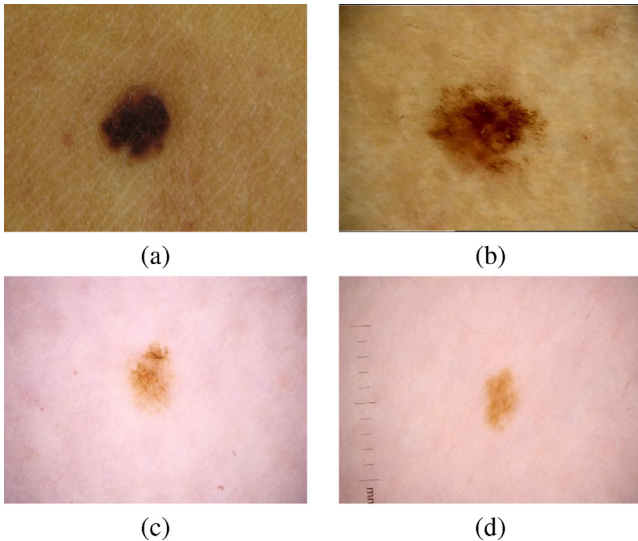
Fig. 1. Illustration of images identified by ImageHash with high similarity, cutoff = 5 (a and b); cutoff = 1 (c and d).

Fig. 2. Illustration of images identified by MSE with high similarity (MSE < 100).

Fig. 3. Illustration of images identified by SSIM with high similarity (SSIM > 0.8).

Fig. 4. Illustration of images identified by cosine similarity with high similarity (> 0.9).

The total number of duplicate image files deleted from all training sets is 14,310, of which 1927 downsampled duplicate image files were deleted from the 2019 training set. Table 8 shows a summary of image files deleted from each training set, and the number remaining, after applying our duplicate removal strategy. We note that the deletion of all of the 2018 training set is due to the 2018 training data comprising the HAM10000 dataset, which was used in its entirety in subsequent ISIC datasets. Additionally, we do not count multiplets as duplicates, given that they represent lesions at different levels of magnification with slight variations in angle and lighting. As a final checking stage, we counted the number of binary identical files across all training sets, with a total of zero duplicate image files found.

Following the completion of stages 1–4 of our duplicate removal strategy, we conducted experiments using four image similarity algorithms to determine if there were any other examples of downsampled images that had not yet been identified. First, we tested the ImageHash Python library which uses multiple image hash algorithms (average, perceptual, difference and wavelet) to analyse the image structure on luminance without colour information. The colour hash algorithm analyses the colour distribution and black and gray fractions without position information (Buchner, 2020). We tested this method with cutoff values set to 5 and then to 1 on a random selection of training images for 72 hours. No exact matches were found within this time frame, however, we report on two examples of the false positives identified by the algorithm (see Fig. 1). Note that lower values indicate closer similarity.
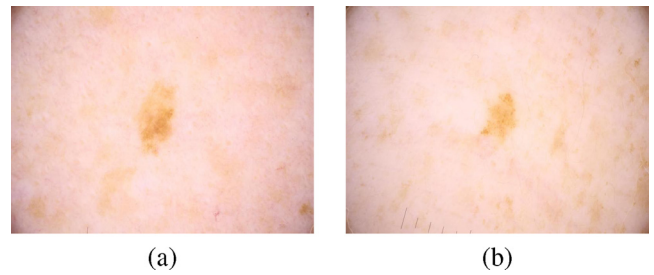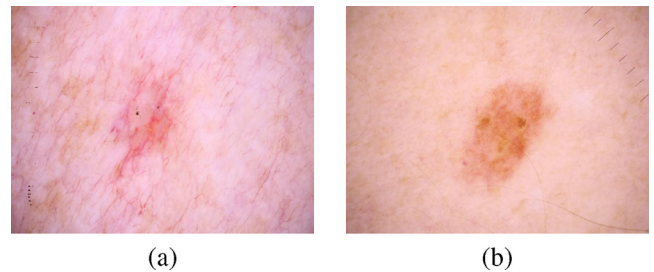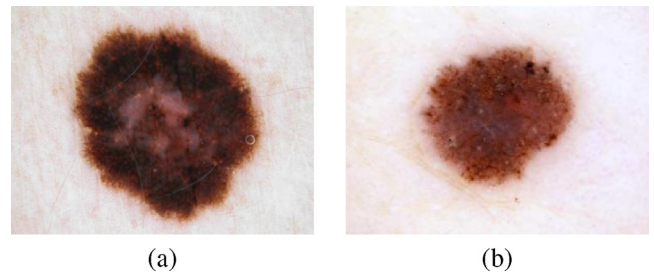
The second image similarity method we tested was mean squared error (MSE) which tests for image similarity by calculating the sum of the squared difference between the two images, resulting in an estimate of the perceived errors. An MSE value of zero indicates perfect similarity, with larger values indicating reduced similarity. Although this was faster than the ImageHash algorithm, the results were all false positives when tested over a 72 h period on a random selection of training images. See Fig. 2 for two such examples. Note that the closer the value is to zero, the closer the similarity.

The third image similarity method we tested was the Structural Similarity Index Measure (SSIM) which models the perceived change in the structural information of the image (Zhou Wang et al., 2004). After testing for 72 hours on a random selection of the training images, this method resulted in only false positive results (see Fig. 3). Note that a value of 1 indicates perfect similarity.

For the fourth image similarity method, we tested cosine similarity. This method measures the similarity between two vectors of an inner product space using the cosine of the angle between two vectors, and determines whether the two vectors are pointing in roughly the same direction (Han et al., 2012). After testing for 72 hours on a random selection of training images, this method also resulted in only false positive results (see Fig. 4). Note that the closer the value is to 1, the closer the similarity.

Although none of the image similarity methods we tested were able to identify any identical images, the application of image similarity techniques might be employed in future studies in order to reduce the over-representation of features within datasets. This
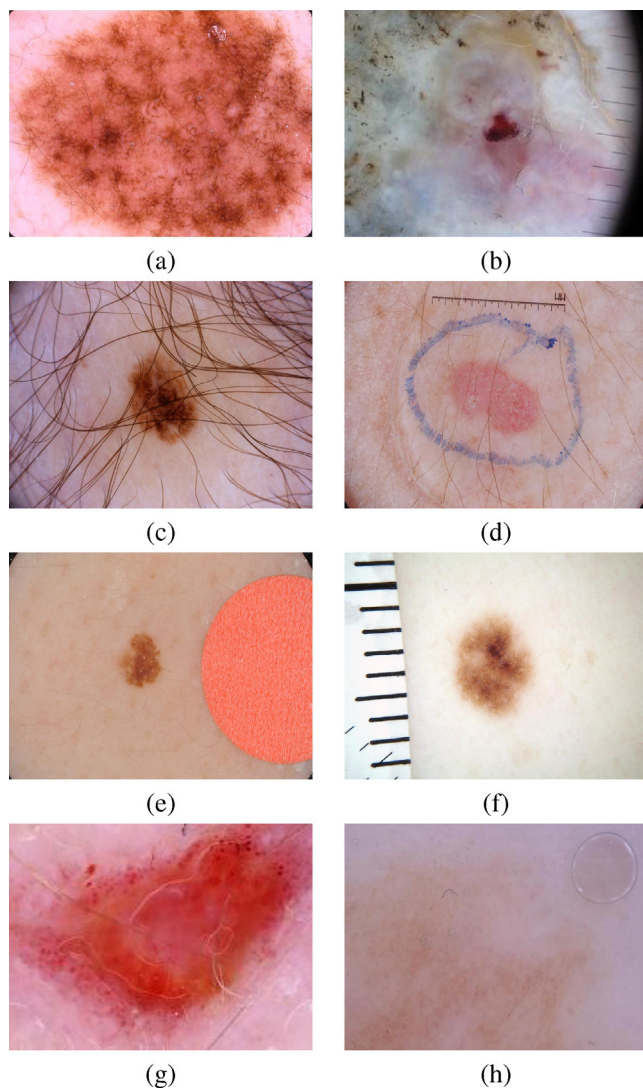
(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 5. Illustration of (a) cropped lesion; (b) obfuscation of lesion by dermoscope measurement overlay; (c) lesion obfuscated by hair; (d) presence of clinical markings; (e) presence of circular size reference stickers; (f) presence of physical ruler; (g) presence of immersion fluid causing distortion of lesion; (h) presence of immersion fluid air pocket.

(a)

(b)

(c)

(d)

Fig. 6. Illustration of (a) duplicate image with different filenames (ISIC_0016018.jpg and ISIC_0012271.jpg) found in the 2017 training and testing sets; (b) duplicate image with the same filename (ISIC_0029847.jpg) found in the 2018 and 2019 training sets; (c) duplicate image with the same filename (ISIC_0011132.jpg) found in two training sets (2017, 2019) and one testing set (2016); (d) duplicate image with different filenames (ISIC_5448850.jpg and ISIC_9881235.jpg) found in the 2020 training set.

- Images may contain clinical markings around the lesion (see Fig. 5(d)).
- Images may contain size reference stickers placed close to the lesion (see Fig. 5(e)).
- Images may contain physical rulers placed close to the lesion (see Fig. 5(f)).
- Images may show pockets of air, a result of the application of immersion fluid used during a dermoscopic examination (see Fig. 5(g) and (h)).

Fig. 6 shows examples of images with duplicated file names and duplicate images with different file names found both within individual datasets and across multiple datasets. The ISIC 2019 training set contains 2074 image files with the suffix "_downsampled". We observed that although the image dimensions for many of these files had been reduced compared to the non-resized originals found in other training sets, the file sizes were often more than double that of the original non-resized images. This is most likely a side-effect of using a lower compression rate when the images were resized in order to avoid introducing additional compression artefacts. Fig. 7 shows two examples of downsampled images that fall into this category.

We observe that there may be edge cases where our duplicate removal strategy may have missed some duplicates. E.g. an image may have been resized, but does not contain the "_downsampled" suffix in the filename, or duplicates (of a different image size) may also have different filenames. However, we believe that our strategy will at least provide a basis for removing a large number of duplicates, which could help to eliminate bias and to enable networks trained on the ISIC datasets to better generalise to new data.

We note that for the ISIC 2016 dataset, the ground truth data does not indicate if a lesion is of melanoma type. Only the malignant and benign status is defined. Given that not all malignant skin cancers are melanomas (NHS, 2020b), we did not include the ISIC 2016 data in any of our experiments reported in this paper.

For researchers who do not intend on uploading their results to the ISIC competition website, we recommend an additional step for duplicate removal. This step would involve the removal of all duplicates found across all test sets. For our study, we did not per-

may help to improve a network's ability to generalise. However, such an approach would need to be carefully considered, as the removal of too many images may result in the opposite desired effect, causing a reduction in the model's ability to generalise.

All stages of our duplicate removal strategy were completed using the Linux application FSlint, created by Brady (2014), which uses rigorous file comparison techniques that compare file size, hardlinks, Message Digest 5 (MD5) and Secure Hash Algorithm 1 (SHA-1). MD5 is used to check both the first 4 kilobytes of a file and the entire file, whereas SHA-1 is used to check the entire file.

Other notable observations of images found within the training sets which may impede model performance include:

- Images may be heavily cropped - removing large amounts of the lesion and/or normal skin boundary regions (see Fig. 5(a)).
- Images may exhibit dermascope measurement overlays, sometimes obfuscating the lesion or lesion boundary (see Fig. 5(b)).
- Images may contain varying amounts of hair, which has been shown to impede model performance (Le et al., 2020) (see Fig. 5(c)).
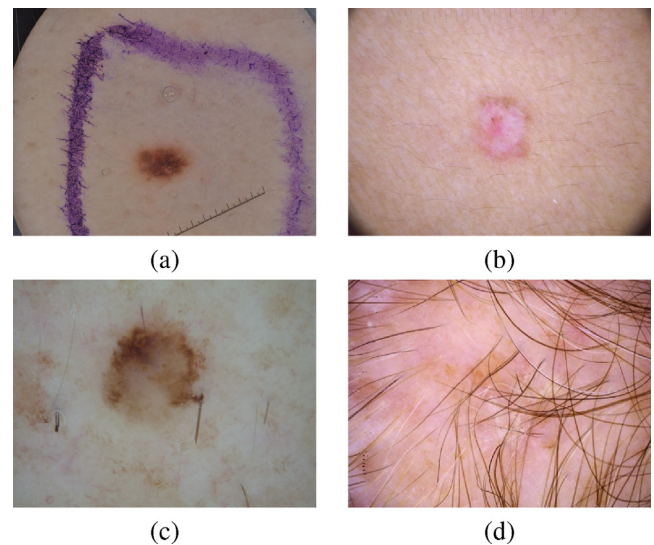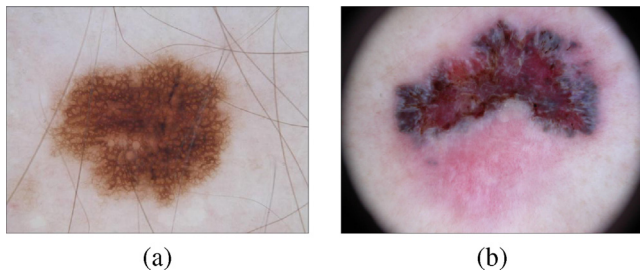
(a)   (b)

Fig. 7. Illustration of downsampled image files that have a larger file size compared to the non-resized original files: (a) duplicate image found in the 2016 (ISIC_0000019.jpg; 1504x1129; 107.6KB), 2017 (ISIC_0000019.jpg; 1504x1129; 107.6KB) and 2019 (ISIC_0000019_downsampled.jpg; 1024x768; 211.7KB) training sets; (b) duplicate image found in the 2016 (ISIC_0000030.jpg; 1503x1129; 95.1KB), 2017 (ISIC_0000030.jpg; 1503x1129; 95.1KB) and 2019 (ISIC_0000030_downsampled.jpg; 1024x769; 195.6KB) training sets; Both examples exhibit a file size more than double that of the original image file.

form this final step as our intention was to submit our results to the ISIC competition website.

Our initial experiments showed that the significant class imbalance in the curated dataset resulted in model over-fitting, despite the use of categorical crossentropy as a means to address the imbalance. For our subsequent experiments, we balanced the dataset using undersampling by removing images from the majority class (non-melanoma). We refer to this dataset as the curated balanced dataset.

The ISIC 2020 dataset includes patient ID for each image. As a further dataset cleaning step to make the data more heterogeneous, it might be appropriate to include only images from unique patient IDs. We observed that from a total of 33,126 images in the ISIC 2020 training set, there are 3078 unique patient IDs and 30,048 duplicates. With duplicate patient IDs removed, there are a total of 428 melanoma and 2650 non-melanoma cases. In the ISIC 2020 test set, there are a total of 10,982 images with 690 unique patient IDs and 10,292 duplicates. With such a limited number of patient examples, and the very low number of unique melanoma cases, this may represent a bias in the dataset which could affect the robustness of models trained exclusively on this dataset, as limited unique cases might not represent the general public from different skin types. However, patients may have presented the same lesion over a period of time during different clinical visits, thus the lesions may provide important unique indicators at different stages of development. Additionally, a patient may present more than one lesion.

4.2. Recommendations of curated datasets for training

Our duplicate removal strategy resulted in a curated dataset with a total of 45,590 image files in the training set (3,924 melanoma and 41,666 other) and a total of 11,397 image files in the validation set (981 melanoma and 10,416 other). The total number of image files is 56,987 (4,905 melanoma and 52,082 other). Users may consider using this curated dataset if they have high performance machines and additional datasets to balance the classes.

Due to the high level of class imbalance (with a ratio of 1:10.62 for melanoma versus others), we recommend a cleaned and balanced dataset (curated balanced dataset). This resulted in a total of 7848 image files in the training set (3924 melanoma and 3924 others) and 1962 image files in the validation set (981 melanoma and 981 others). The total number of image files is 9810 (4905 melanoma and 4905 other) with a 1:1 ratio. For this study, since we are focusing on ISIC datasets only, we will analyse and experiment using the curated balanced dataset.

To study the distribution of the curated balanced dataset of 9810 training images, we perform statistical analysis and analyse its distribution using Unified Uniform Manifold Approximation and Projection (UMAP), devised by McInnes et al. (2018).

UMAP is a dimensional reduction tool based on manifold learning. McInnes et al. (2018) demonstrated that UMAP is a competitive tool when compared to t-SNE, which has been shown to be less computationally expensive. Fig. 8 shows the UMAP visualisation data feature distributions (input, EfficientNetB0, top dropout layer, dense layer and output), where blue regions represent non-melanoma and orange regions represent melanoma. It is noted that after training with EfficientNetB0, the two classes become separable, which is further supported by the statistic metrics in Table 9. On the input distribution, the intra-class and inter-class values show high similarity.

We can observe that moving from the input dataset through the feature vector of EfficientNetB0 to the dense layer that both intra-class and inter-class values increase, which indicates a better separability of the dataset. This is further confirmed by higher values of the silhouette score and Calinski-Harabasz index. We observe that the feature distribution of EfficientNetB0 shows that the inter-class distance (18.6252) is larger than the intra-class distances (7.6235 for melanoma and 8.6028 for non-melanoma).

Recent developments in the field have shown a growing emphasis on models capable of multi-class predictions (Codella et al., 2018b; Kassem et al., 2020). Due to the growing importance of multi-class CNNs in this domain, we conduct further analysis on multi-class using our curated dataset, and evaluate the performance on the ISIC 2018 test set (Task 3: multi-class lesion diagnosis) (Codella et al., 2018b). Task 3 ISIC 2018 consists of 7 different types of skin lesions, including Melanoma (MEL), Nevi (NV), Basal cell carcinoma (BCC), Actinic keratosis / Bowens disease (intraepithelial carcinoma) (AKIEC), Benign keratosis (BKL), Dermatofibroma (DF) and Vascular (VASC). The curated dataset consists of 4905 MEL, 11,421 NV, 3316 BCC, 859 AKIEC, 2520 BKL, 239 DF and 253 VASC, respectively. Fig. 9 illustrates the UMAP visualisation data feature distribution of multi-class. Due to factors such as class imbalance and image similarity, the composition of the dataset presents a significant challenge for skin lesion diagnosis, with overlaps on every stage during training. It is noted that the value of Calinski-Harabasz, silhouette score and Davies-Bouldin index indicate low separability for distribution, with some improvement from input distribution (411.8463, -0.0613 and 12.3489) after training with EfficientNetB0 (928.5820, -0.1254 and 3.9409).

To allow for reproducibility of this work, the list of image files and corresponding labels can be downloaded from our GitHub repository (available at www.github.com/mmu-dermatology-research/isic_duplicate_removal_strategy).

4.3. Benchmarks

For baseline experiments, we trained 19 of the most widely used deep learning architectures: DenseNet121, DenseNet169, DenseNet201, EfficientNetB0 - B4, InceptionResNetV2, InceptionV3, ResNet50, ResNet50V2, ResNet101, ResNet101V2, ResNet152, ResNet152V2, VGG16, VGG19 and Xception. For training data, we used an 80:20 split for training and validation using our curated balanced dataset based on images from the ISIC 2017 - 2020 datasets. Transfer learning was not used for any of the experiments, as the purpose of this paper is to provide baseline results without additional strategies.

We trained each of the 19 networks for 50 epochs with a batch size of 32 using stochastic gradient descent with an initial learning rate of 0.01 and momentum of 0.9. We implemented early stopping, until each network converged, determined by a patience of 10 epochs. For pre-processing, all images were resized, with the
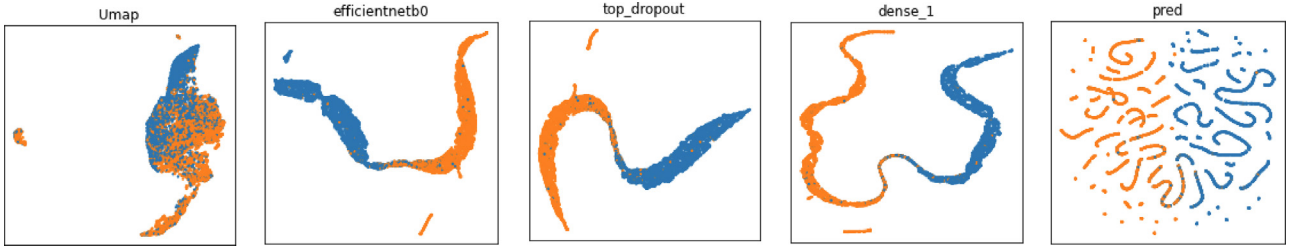
Fig. 8. UMAP visualisation of the curated balanced training set, where orange regions represent melanoma and blue regions represent others. From left to right: input, feature distributions extracted with EfficientNetB0, the top dropout layer, dense layer and output. These graphs visually illustrate an increase in the separability of melanoma versus non-melanoma in the deep learning architecture. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 9

Statistical analysis on the separability of melanoma versus non-melanoma on the curated balanced dataset. Intra: average distance between samples of the same class; Inter-class: average distance between samples of different classes; Si: silhouette score; CH: Calinski-Harabasz index; DB: Davies-Bouldin index.

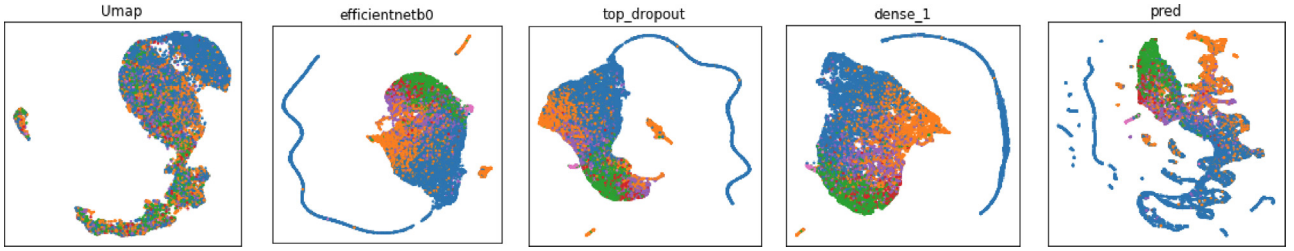| Method	Metrics					
	Intra (melanoma)	Intra (non-melanoma)	Inter-class	CH	Si	DB
Input	5.5740	6.2936	6.8420	1284.7040	0.1041	2.2834
EfficientNetB0	7.6235	8.6028	18.6252	14421.3350	0.5399	0.7032
Dropout layer	7.5033	8.2194	17.5380	13994.5712	0.5223	0.7334
Dense layer	10.8150	13.4581	19.3595	4755.8860	0.3618	1.2904

Fig. 9. UMAP visualisation of the curated multi-class training set, where orange regions represent MEL, blue regions represent NV, green regions represent BCC, red regions represent AKIEC, purple regions represent BKL, brown regions represent DF and pink regions represent VASC. From left to right: input, feature distributions extracted with EfficientNetB0, the top dropout layer, dense layer and output. These graphs visually illustrate the low separability of multi-class skin lesions within the deep learning architecture. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

shortest side reduced to 224 pixels and center-cropping (224 × 224 pixels) using the high-quality downsampling filter found in the Python Image Library, created by Lund and Clark (2013). To address the limited size of the training set, we applied several data augmentation techniques, including random rotations, random zooms, random width and height shift, shearing and horizontal and vertical flipping.

The hardware configuration used to train the networks was an Intel i7-7700 Quad Core 3.60GHz CPU with 64GB DDR4 2400MHz Duel Channel RAM and a GTX1080 Ti 11GB GPU. The software configuration used Tensorflow GPU 2.4.1 and Keras 2.3.1 running on Windows 10.

We evaluate the performance of the baseline models on Kaggle, however, given that we do not have access to the ground truth labels for the ISIC 2020 test set, we cannot perform further analysis on this dataset. In order to discuss the performance of the baseline models, we use the ISIC 2017 test set, which is fully exclusive from our curated balanced training set.

To produce baselines for our curated multi-class dataset, we use four popular deep learning models (VGG19, DenseNet121, ResNet101 and EfficientNetB0) and evaluate the results on the ISIC 2020 live leaderboard. Our experiments set the maximum epoch to 200 and adopt an early stopping strategy. We save the best model that maximised validation accuracy, with early stopping im-plemented when the categorical accuracy of validation did not increase after 8 epochs. The initial learning rate was set to 0.001, which was reduced by a factor of 0.1 when the validation score did not increase after 5 epochs.

5. Results

Table 10 presents benchmark results using our curated balanced dataset for 19 deep learning architectures of their best epochs on the ISIC 2020 test set. Note that the ground truth for ISIC 2020 is not publicly available, therefore the metric provided by the organiser on Kaggle is used - area under the Receiver Operating Characteristics Curve (AUC).

For the ISIC 2020 test results, the highest performing network was VGG19 with an AUC of 0.80, indicating an increase of 0.03 over the next best performing networks (VGG16 and DenseNet121). The next best performing networks were ResNet101, ResNet50, ResNet50V2, EfficientNetB2, EfficientNetB3, Xception, EfficientNetB0, EfficientNetB1, ResNet101V2, VGG16 and DenseNet121, reporting an AUC in the range of 0.70 - 0.77. InceptionV3 was shown to be the lowest performing network with an AUC of 0.5. In addition to InceptionV3, the other lowest performing networks were DenseNet201, InceptionResNetV2, ResNet152V2, EfficientNetB4, DenseNet169 and ResNet152, all reporting an AUC in

Table 10

A performance comparison of the baseline models on the ISIC 2020 testing set without the use of a pre-trained model, results are reported on their best epoch.

Model	Params	Best epoch	AUC
DenseNet121	7,039,554	25	0.77
DenseNet169	12,646,210	47	0.66
DenseNet201	18,325,826	5	0.63
EfficientNetB0	4,052,133	37	0.75
EfficientNetB1	6,577,801	30	0.76
EfficientNetB2	7,771,387	67	0.73
EfficientNetB3	10,786,609	8	0.75
EfficientNetB4	17,677,409	45	0.65
InceptionResNetV2	54,339,810	10	0.64
InceptionV3	21,806,882	47	0.50
ResNet50	23,591,810	27	0.71
ResNet50V2	23,568,898	41	0.73
ResNet101	42,662,274	21	0.70
ResNet101V2	42,630,658	37	0.76
ResNet152	58,375,042	18	0.67
ResNet152V2	58,335,746	25	0.65
VGG16	134,268,738	21	0.77
VGG19	139,578,434	30	**0.80**
Xception	20,865,578	16	0.75

Table 11

More detailed performance measures on the ISIC 2017 test set. Note that these results can not be compared with the ISIC 2017 leaderboard as these results are based on binary classification, while the ISIC 2017 leaderboard is based on 3-class.

Method	Accuracy	Precision	Recall	F1	AUC
DenseNet121	0.41	0.20	0.68	0.31	0.50
DenseNet169	0.46	0.16	0.42	0.23	0.46
DenseNet201	0.36	0.21	0.83	0.33	**0.56**
EfficientNetB0	0.55	0.19	0.41	0.26	0.51
EfficientNetB1	0.46	0.20	0.60	0.30	0.53
EfficientNetB2	0.41	0.21	0.71	0.32	0.53
EfficientNetB3	0.53	**0.22**	0.58	0.32	0.54
EfficientNetB4	0.52	0.17	0.39	0.24	0.49
InceptionResNetV2	0.40	0.20	0.67	0.30	0.49
InceptionV3	0.30	0.21	**0.94**	**0.34**	0.53
ResNet50	0.40	0.20	0.71	0.32	0.52
ResNet50V2	0.44	0.20	0.64	0.31	0.53
ResNet101	0.45	0.21	0.66	0.32	0.54
ResNet101V2	0.38	0.20	0.73	0.31	0.54
ResNet152	0.32	0.21	0.86	0.33	0.54
ResNet152V2	0.40	0.20	0.67	0.30	0.51
VGG16	0.48	0.20	0.57	0.30	0.54
VGG19	**0.56**	0.20	0.41	0.26	0.51
Xception	0.44	0.19	0.59	0.29	0.50

the range of 0.63 to 0.67. For all reported networks, the mean average for AUC was 0.704, with a standard deviation of 0.071. The poor performance of EfficientNetB4 compared to EfficientNetB0 - B3 may be due to a disparity between the large size of the network architecture and the relatively small size of the training set images. We note that the best performing network (VGG19) also has the highest number of parameters, however, the poorest performing network (InceptionV3) did not have the lowest number of parameters.

Table 11 shows the benchmark results using our curated balanced dataset for 19 deep learning architectures of their best epochs on the ISIC 2017 test set. For the ISIC 2017 test results, VGG19 demonstrated the highest accuracy at 0.56, with InceptionV3 having the lowest accuracy of 0.30. For precision, EfficientNetB3 showed the highest result at 0.22, while DenseNet169 reported the lowest result at 0.16. InceptionV3 showed the highest recall, with a result of 0.94, indicating that the network over-classified melanoma cases. Conversely, EfficientNetB4 showed the lowest recall at 0.39, which is comparable to its low performance on the 2020 test set. For AUC, DenseNet201 showed the highest

result at 0.56, with DenseNet169 reporting the lowest result of 0.46. Measures for all networks using the ISIC 2017 test set demonstrated poor performance compared to those returned by the ISIC 2020 test set experiment.

F1-score is the best indicator of overall network performance, indicating the harmonic mean between precision and recall. Fig. 10 shows six examples of test images from the ISIC 2017 test set where noise affected the performance of three predictions. Fig. 11 shows a selection of heatmaps for test images from the ISIC 2020 test set, compared against original test images. Given that the ground truth data is not publicly available for the ISIC 2020 test set, we present these results to demonstrate that the trained networks are clearly focusing on noise present within the dataset. However, in the case of the ISIC 2017 test results, noise would appear to not always affect the accuracy of the prediction.

Table 13 shows the benchmark results of multi-class classification evaluated on Task 3 of the ISIC 2018 testing set. As the ground truth of this dataset is not publicly available, we evaluate our results on the live leaderboard and report the Balanced Multi-class Accuracy (Codella et al., 2018b). The deep learning models achieve better accuracy with pretrained models based on ImageNet. We observe that the best baseline accuracy of our curated dataset is 0.621, achieved by EfficientNetB0 with a pretrained model.

Since we do not have access to ground truth data for the ISIC 2018 test set, we conduct further analysis of our best baseline model on the ISIC 2017 classification dataset, which consists of 3 classes: MEL, NV and seborrheic keratosis (SK). Fig. 13 illustrates the Grad-CAM heatmap visualisation of correct predictions (on the left: a), c) and e)) and incorrect predictions (on the right: b), d) and f)). We note that although the majority of the network focused on regions for correct predictions for skin lesions, some cropped regions and areas of noise were also included. Our findings on multi-class classification are consistent with our binary classification results, where noise appears to not always affect the accuracy of the prediction.

We further analyse the curated dataset by annotating features that enable the network to further inflate its results, e.g. clinical pen markings. We categorise these non-lesion features into 7 separate class labels: (1) dermoscope ruler; (2) light and dark hair; (3) clinical pen marking; (4) size reference sticker; (5) air pocket; (6) dermoscope borders; (7) Other. The last category contains artefact types where there were too few examples to warrant creating new categories for, including images with dates printed onto them and images that were extremely blurry. Next, we train the network with all the non-lesion classes removed, similar to a cross-fold validation. This experiment shows how non-lesion features (noise) affect model accuracy. We train each model by removing all non-lesion images in the dataset, then rebalance by removing images from the majority class, with 20% reserved for validation.

Table 14 demonstrates the diversity of images within the ISIC datasets, including within our curated dataset. Furthermore, it highlights how certain features could give bias to melanoma, such as the dermoscopic borders which are much less present in melanoma cases, which gives a minor accuracy increase. Similarly, dermoscope ruler overlays have a slight bias towards non-melanoma. However, the best performing network uses the full dataset (None removed), which is in contrast to Table 15 where none of the best scores come from networks trained on the full dataset. Training with dermoscope ruler artefacts removed, DenseNet 201 and InceptionV3 received improved accuracy. Similarly, VGG19 with air pocket artefacts removed had the best performance overall. These results show that some models are susceptible to noise from the artefacts that disrupt performance, in some cases significantly. Future work could involve the use of ensemble networks trained on the removed class to overcome these obstacles.
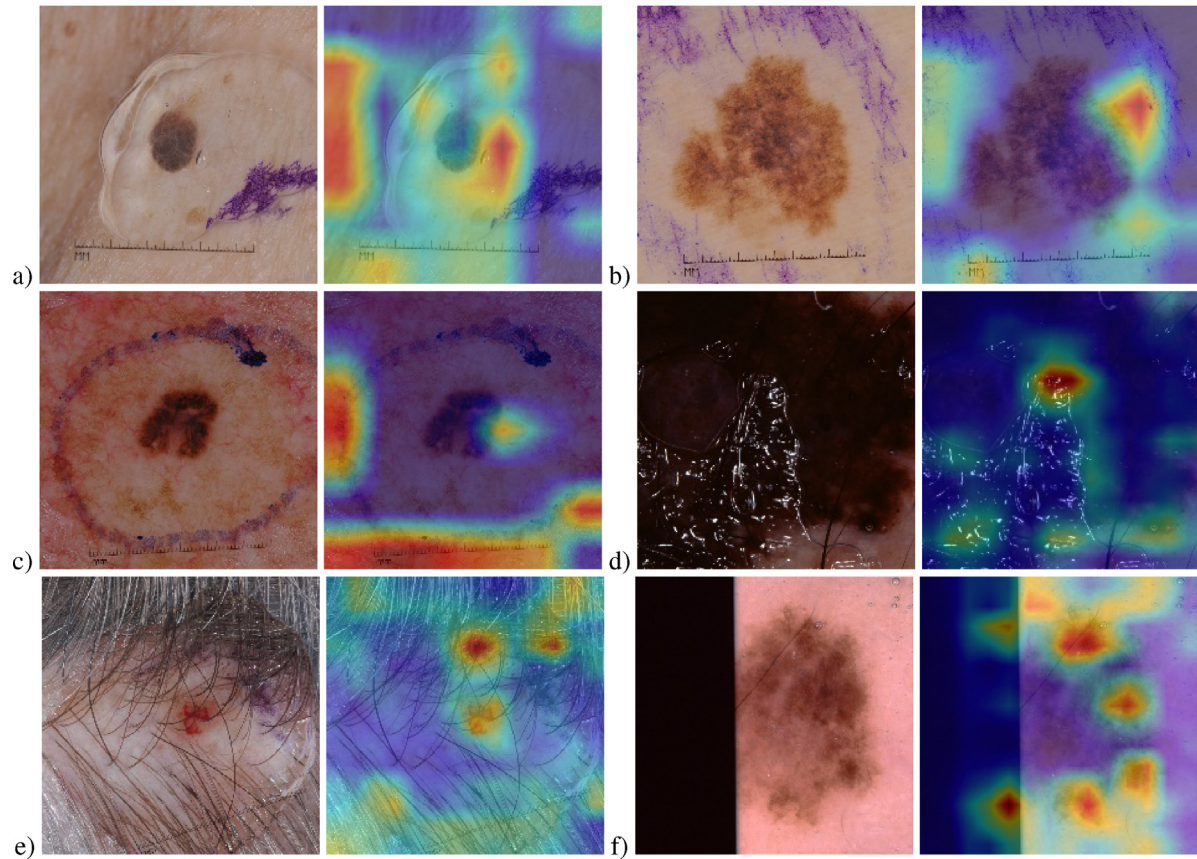
Fig. 10. Grad-CAM heatmap visualisation for the ISIC 2017 test results: a) network focused only on areas surrounding the lesion, including immersion fluid (DenseNet201) - prediction: melanoma; ground truth: other, b) network focused only on areas around the lesion, including clinical pen markings and dermascope measurement overlay (DenseNet201) - prediction: melanoma; ground truth: melanoma, c) network focused mainly on areas surrounding the lesion, including clinical pen markings and dermascope measurement overlay (DenseNet201) - prediction: melanoma; ground truth: melanoma, d) network focused mainly on immersion fluid (VGG16) - prediction: melanoma; ground truth: melanoma, e) network focused mainly on clinical pen markings and hair (VGG19) - prediction: melanoma; ground truth: other, f) network focused on lesion and cropped image area (VGG19) - prediction: melanoma; ground truth: other.
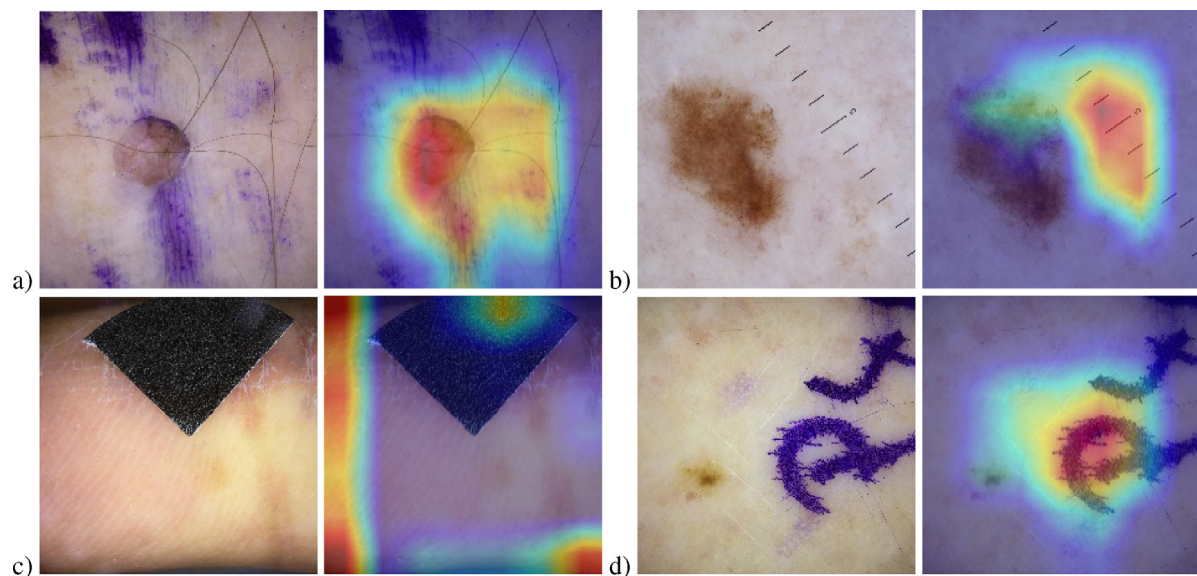

Fig. 11. Grad-CAM heatmap visualisation for the ISIC 2020 test results using DenseNet121: a) network focused on both lesion and clinical pen markings, b) network focused primarily on dermoscope measurement overlay, c) network focused on surrounding skin and wound dressing, d) network focused primarily on clinical pen markings.
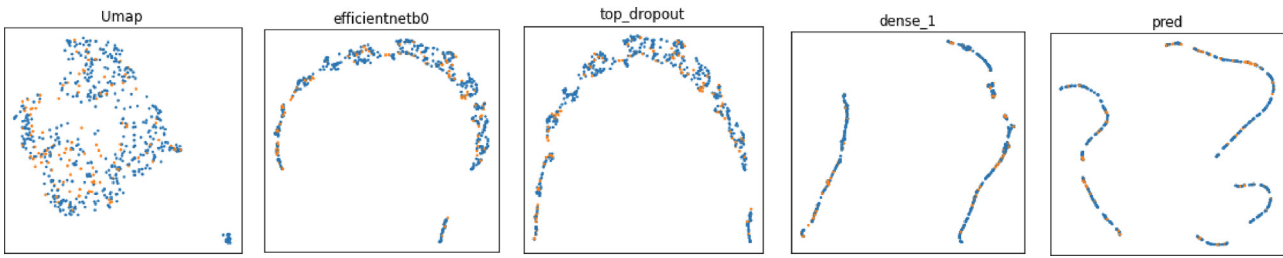
Fig. 12. UMAP visualisation of EfficientNetB0 on the ISIC 2017 test set, where orange regions represent melanoma and blue regions represent others. From left to right: input, feature distributions extracted with EfficientNetB0, the top dropout layer, dense layer and output. These graphs visually illustrate the separability of melanoma versus non-melanoma. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 12

Statistical analysis on the separability of melanoma versus non-melanoma on the ISIC 2017 testing set. Intra: average distance between samples of the same class; Inter-class: average distance between samples of different classes; Si: silhouette score; CH: Calinski-Harabasz index; DB: Davies-Bouldin index.

Method	Metrics					
	Intra (melanoma)	Intra (non-melanoma)	Inter-class	CH	Si	DB
Input	3.8030	3.2105	3.6288	10.3804	-0.0238	5.1868
EfficientNetB0	8.7547	9.0875	9.0051	1.49800	0.0206	15.3246
Dropout layer	8.8238	9.1935	9.0970	1.8324	0.0217	13.7170
Dense layer	13.7832	13.8319	13.9366	1.9848	0.0065	13.4191

6. Discussion

All network architectures we trained using our curated balanced dataset provided comparable results. However, we note that the ISIC 2020 testing set we used to evaluate our models is significantly larger than our training set - 7848 training examples vs 10,982 test examples. Performance may be further impacted by possible class imbalance in the ISIC 2020 testing set. We also note that the ISIC 2020 test set contains 78 duplicate image files, identified using FSlint. Exact details of the ISIC 2020 test set are not currently publicly available, therefore we can only speculate on its composition and possible effects when obtaining evaluation metrics.

Given the comparatively small dataset size of our curated balanced dataset, and the lack of any additional fine-tuning of the trained networks, we would regard the test results from the ISIC 2020 test set to be good for the best performing networks. However, the results for the experiment performed on the ISIC 2017 test set were poor for all measures for all networks. As shown in the UMAP visualisation (using the same settings for the balanced training set) in Fig. 12 and statistical analysis in Table 12, the ISIC 2017 test set is less separable, with inter-similarities and intra-dissimilarities between the two classes. We identify four possible causes for this: (1) the number of duplicates within the ISIC 2017 test set; (2) class imbalance in the ISIC 2017 test set; (3) the amount of noise present in both the training and testing sets; and (4) the relatively small size of our curated balanced training set. We identify noise as cases where lesions may be obfuscated by hair follicles, hair, air pockets resulting from the application of immersion fluid, size reference stickers, rulers, dermascope measurement overlays and clinical pen markings. Ju et al. (2021) noted that medical datasets tend to have asymmetric (class-dependent) noise and suffer from high observer variability. Rolnick et al. (2018) showed that deep learning models trained on large supervised datasets are capable of generalising from training data where true labels are massively outnumbered by incorrect labels. However, this was only demonstrated on MNIST, CIFAR and ImageNet datasets and requires a significant increase in dataset size that is related to the factor by which correct labels have been diluted. Our results may indicate the importance of transfer learning and dataset size in this domain.

Table 13

Benchmarking the performance of the curated dataset on multi-class classification evaluated on the ISIC 2018 test set. The primary metric value for the live leaderboard is Balanced Multi-class Accuracy. Pretrained indicates that the model is using a pretrained model based on ImageNet.

Model	Settings		Metric
	Pretrained	Best epoch	Accuracy
VGG19	×	20	0.279
VGG19	✓	31	0.322
DenseNet121	×	26	0.436
DenseNet121	✓	18	0.565
ResNet101	×	14	0.410
ResNet101	✓	18	0.495
EfficientNetB0	×	12	0.368
EfficientNetB0	✓	17	**0.621**

We balanced our curated dataset using undersampling, which involved the removal of images from the majority class (non-melanoma). However, it may be useful for future research to compare the results of this approach with other balancing techniques such as image augmentation of the minority class (melanoma), or weight balancing such as the implementation of a focal loss function, as per Lin et al. (2017).

For multi-class classification, we provide baseline results with four popular deep learning models on the ISIC 2018 Task 3 lesion diagnosis test set. The curated dataset is imbalanced and requires additional strategies to improve the performance of networks trained using this dataset. For future improvement, we recommend the use of data augmentation methods and/or the inclusion of external non-ISIC datasets to balance the classes, particularly on AKIEC, DF, VASC and SCC.

Future work might focus on the effect of the large number of visually similar images on trained models that use the ISIC datasets. We tested four image similarity methods on a limited set of data over a short period of time. However, other techniques such as those employing feature extraction, may be worth investigating given that recent works, such as Sucholutsky and Schonlau (2020), suggest that unique features are more important than a large number of training images when training deep CNNs. We also note the importance of colour space in the processing of medical image data (Barata et al., 2014). This could contribute to future im-
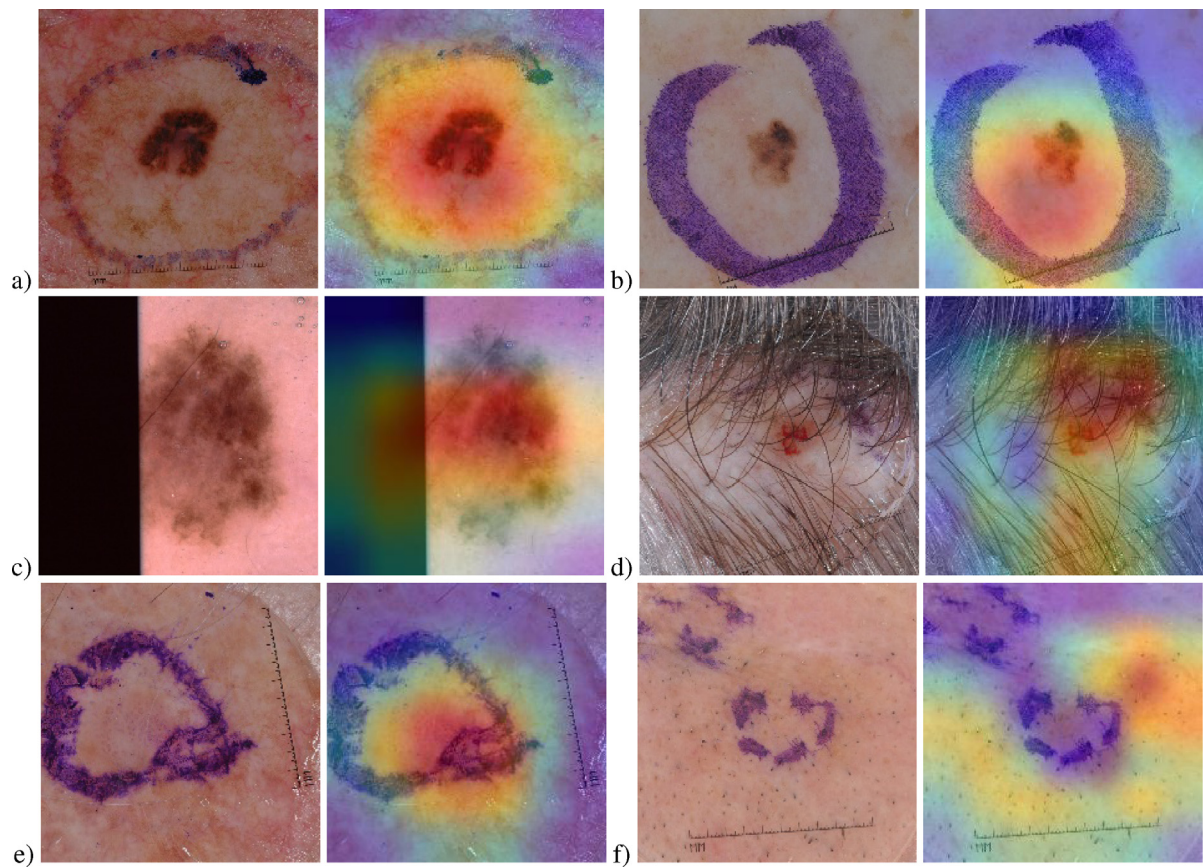
Fig. 13. Grad-CAM heatmap visualisation for multi-class classification on the ISIC 2017 test results: a) network focused mainly on the lesion - prediction: melanoma; ground truth: melanoma, b) network focused on the lesion and clinical pen marking - prediction: nevus; ground truth: melanoma, c) network focused on lesion and cropped area - prediction: nevus; ground truth: nevus, d) network focused mainly on clinical pen markings and hair regions - prediction: melanoma; ground truth: nevus, e) network focused on the lesion and clinical pen markings - prediction: seborrheic keratosis; ground truth: seborrheic keratosis, f) network focused entirely on the surrounding skin - prediction: Seborrheic keratosis; ground truth: melanoma.

Table 14

Results on the validation set for artefacts removed. Note: when removing artefacts we re-balance the dataset to the class with the lowest number of remaining images, then take 20% from both classes for validation.

Class Removed	Images Mel	Image Oth	Total train	DenseNet 201	Inception V3	VGG 19	EfficentNet B3
None removed	4905	4905	7848	**0.7900**	**0.7935**	**0.7956**	**0.7848**
Dermoscope ruler	3760	4163	6016	0.7267	0.7434	0.7407	0.7314
Hair	2265	2241	3586	0.6462	0.6406	0.5826	0.6395
Clinical markings	4793	4831	7670	0.6584	0.6923	0.6208	0.6563
Size reference sticker	4893	4890	7824	0.6752	0.6445	0.6639	0.6747
Air pockets	4067	4744	6508	0.6981	0.6994	0.6975	0.6444
Dermoscope borders	2564	4426	4104	0.7441	0.7354	0.7305	0.7080
Other	4577	4833	7324	0.6990	0.6957	0.6864	0.6831

Table 15

Results of individual artefact class removal on the ISIC 2020 testing set.

Class removed	DenseNet201	InceptionV3	VGG19	EfficentNetB3
None Removed	0.4514	0.6480	0.7988	0.7079
Dermoscope ruler	**0.7411**	**0.7377**	07967	0.7060
Hair	0.7293	0.6524	0.6588	0.7239
Clinical markings	0.6477	0.6500	0.6065	0.6801
Size reference sticker	0.6900	0.6369	0.7884	**0.7363**
Air pockets	0.7208	0.6229	**0.8067**	0.6700
Dermoscope borders	0.6976	0.5933	0.7210	0.6464
Other	0.7368	0.6362	0.7925	0.6913

provements to algorithm design for skin lesion diagnosis, and will be explored further in future work.

As our paper focuses on skin lesions classification, we did not include duplicate analysis on skin lesion segmentation datasets.

Whilst classification tasks provide the diagnosis of the lesions, lesion segmentation, such as in ISIC 2018 Task 1 on lesion boundary segmentation, provides better localisation of the lesions. This could be used in future studies for the overlap of the computer generated heatmap and the dermatologist's annotation.

7. Conclusion

In this work, we propose a strategy for removing duplicate image files from the ISIC 2017 - 2020 datasets as a means of reducing bias in deep learning models trained on these datasets. We present results from a variety of commonly used CNN architectures trained on a curated balanced dataset which indicates excellent class distribution and good performance measures. The aim of this work is to highlight the potential biases of the usage of duplicate images of ISIC datasets, and other numerous issues, such as noise,

present within the ISIC datasets and to better understand their effects on deep CNNs. This work is not intended to maximise the performance of the CNNs, therefore we did not include any additional steps such as transfer learning with different pretrained models, fine-tuning or adjustments to network configurations. The effects of noise inherent within the ISIC datasets, in addition to a relatively small training set size, were shown to contribute to a significant reduction in network performance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Bill Cassidy: Conceptualization, Data curation, Formal analysis, Writing – original draft. **Connah Kendrick:** Conceptualization, Formal analysis, Writing – original draft. **Andrzej Brodzicki:** Conceptualization, Formal analysis, Writing – original draft. **Joanna Jaworek-Korjakowska:** Conceptualization, Formal analysis, Writing – original draft. **Moi Hoon Yap:** Conceptualization, Data curation, Formal analysis, Writing – original draft.

Acknowledgment

References

Acosta, M.F.J., Tovar, L.Y.C., Garcia-Zapirain, M.B., Percybrooks, W., 2021. Melanoma diagnosis using deep learning techniques on dermatoscopic images. BMC Med Imaging 21.

Adegun, A.A., Viriri, S., 2020. Deep learning-based system for automatic melanoma detection. IEEE Access 8, 7160–7172.

Adegun, A.A., Viriri, S., 2020. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. Artif Intell Rev 1–31.

Al-antari, M.A., Rivera, P., Al-masni, M., Valarezo Aazco, E., Gi, G., Kim, T.-Y., Park, H., Kim, T.-S., 2018. An automatic recognition of multi-class skin lesions via deep learning convolutional neural networks.

Almaraz-Damian, J.-A., Ponomaryov, V., Sadovnychiy, S., Castillejos-Fernandez, H., 2020. Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures. Entropy 22 (4). doi:10.3390/e22040484.

American Institute for Cancer Research, 2018. Skin cancer statistics. Online. https://www.wcrf.org/dietandcancer/cancer-trends/skin-cancer-statistics.

Barata, C., Celebi, M.E., Marques, J.S., 2014. Improving dermoscopy image classification using color constancy. IEEE J Biomed Health Inform 19 (3), 1146–1152.

Barbosa, J., Baleiras, M., 2019. Melanoma detection using deep learning methods.

Bisla, D., Choromanska, A., Berman, R., Stein, J., Polsky, D., 2019. Towards automated melanoma detection with deep learning: data purification and augmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2720–2728.

Bisla, D., Choromanska, A., Stein, J., Polsky, D., Berman, R., 2019. Skin lesion segmentation and classification with deep learning system. ArXiv abs/1902.06061.

Bissoto, A., Fornaciali, M., Valle, E., Avila, S., 2019. (de)constructing bias on skin lesion datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Brady, P., 2014. Fslint. Online. http://www.pixelbeat.org/fslint/.

Brinker, T.J., Hekler, A., Enk, A.H., Berking, C., Haferkamp, S., Hauschild, A., Weichenthal, M., Klode, J., Schadendorf, D., Holland-Letz, T., von Kalle, C., Frhling, S., Schilling, B., Utikal, J.S., 2019. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer 119, 11–17. doi:10.1016/j.ejca.2019.05.023. http://www.sciencedirect.com/science/article/pii/S0959804919303491

Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Frhling, S., Utikal, J.S., von Kalle, C., 2019. 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. doi:10.1016/j.ejca.2019.02.005. http://www.sciencedirect.com/science/article/pii/S0959804919301443

Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., Utikal, J.S., von Kalle, C., 2019. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer 113, 47–54. doi:10.1016/j.ejca.2019.04.001. http://www.sciencedirect.com/science/article/pii/S0959804919302217

Brinker, T.J., Hekler, A., Utikal, J.S., Grabe, N., Schadendorf, D., Klode, J., Berking, C., Steeb, T., Enk, A.H., von Kalle, C., 2018. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 20 (10), e11936. doi:10.2196/11936.

Cancer Research UK, 2019. How does the sun and uv cause cancer? Online. https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/sun-uv-and-cancer/how-does-the-sun-and-uv-cause-cancer.

Buchner, J., 2020. Imagehash. Online. https://www.pypi.org/project/ImageHash/.

Carcagn, P., Leo, M., Cuna, A., Mazzeo, P.L., Spagnolo, P., Celeste, G., Distante, C., 2019. Classification of Skin Lesions by Combining Multilevel Learnings in a DenseNet Architecture, pp. 335–344.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A., 2018a. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). 1902.03368.

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., 2018. 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: Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on. IEEE, pp. 168–172.

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A., 2017. 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). 1710.05006.

Combalia, M., Codella, N. C. F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A. C., Puig, S., Malvehy, J., 2019. Bcn20000: Dermoscopic lesions in the wild. 1908.02288.

Cormier, J., Voss, R., Woods, T., Cromwell, K., Nelson, K., 2015. Improving outcomes in patients with melanoma: strategies to ensure an early diagnosis. Patient Relat Outcome Meas 6, 229. doi:10.2147/PROM.S69351.

De Hertog, S.A., Wensveen, C.A., Bastiaens, M.T., Kielich, C.J., Berkhout, M.J., Westendorp, R.G., Vermeer, B.J., Bavinck, J.N.B., 2001. Relation between smoking and skin cancer. Journal of Clinical Oncology 19.

Department for Environment Food & Rural Affairs, 2020. Depletion of the ozone layer leading to an increase in ground-level ultraviolet radiation. Online. https://www.uk-air.defra.gov.uk/research/ozone-uv/moreinfo?view=increase-uv-radiation.

Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542. doi:10.1038/nature21056.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115–118.

Fujisawa, Y., Otomo, Y., Ogata, Y., Nakamura, Y., Fujita, R., Ishitsuka, Y., Watanabe, R., Okiyama, N., Ohara, K., Fujimoto, M., 2019. 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 (2), 373–381. doi:10.1111/bjd.16924.

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A., 2020. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. MethodsX 7, 100864. doi:10.1016/j.mex.2020.100864.

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

Goyal, M., Oakley, A., Bansal, P., Dancey, D., Yap, M.H., 2019. Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. IEEE Access 8, 4171–4181.

Gutman, D., Codella, N. C. F., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A., 2016. 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). 1605.01397.

Ha, Q., Liu, B., Liu, F., 2020. Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge. 2010.05351.

Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., Uhlmann, L., 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology 29 (8), 1836–1842. doi:10.1093/annonc/mdy166. Immune-related pathologic response criteria

Han, J., Kamber, M., Pei, J., 2012. Getting to Know Your Data. In: Han, J., Kamber, M., Pei, J. (Eds.), Data Mining: Concepts and Techniques. Morgan Kaufmann, Boston, pp. 39–82. doi:10.1016/B978-0-12-381479-1.00002-2.

Hasan, M., Dahal, L., Samarakoon, P., Tushar, F.I., Marly, R.M., 2020. Dsnet: automatic dermoscopic skin lesion segmentation. Comput. Biol. Med. 120, 103738.

Hasan, M., Elahi, M.E., Alam, M.A., 2021. Dermoexpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation. medRxiv.

Hekler, A., Kather, J.N., Krieghoff-Henning, E., Utikal, J.S., Meier, F., Gellrich, F.F., Upmeier zu Belzen, J., French, L., Schlager, J.G., Ghoreschi, K., Wilhelm, T., Kutzner, H., Berking, C., Heppt, M.V., Haferkamp, S., Sondermann, W., Schadendorf, D., Schilling, B., Izar, B., Maron, R., Schmitt, M., Frhling, S., Lipka, D.B., Brinker, T.J., 2020. Effects of label noise on deep learning-based skin cancer classification. Front Med (Lausanne) 7, 177. doi:10.3389/fmed.2020.00177.

Hekler, A., Utikal, J.S., Enk, A.H., Hauschild, A., Weichenthal, M., Maron, R.C., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., Schilling, B., Holland-Letz, T., Izar, B., von Kalle, C., Frhling, S., Brinker, T.J., 2019. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer 120, 114–121. doi:10.1016/j.ejca.2019.07.019. https://www.sciencedirect.com/science/article/pii/S0959804919304277

Henriksen, K., Stamnes, K., Volden, G., Falk, E., 1989. Ultraviolet radiation at high latitudes and the risk of skin cancer. Photodermatology, Photoimmunology and Photomedicine 6.

Hosny, K., Kassem, M., Fouad, M., 2019. Classification of skin lesions using transfer learning and augmentation with alex-net. PLoS ONE 14, e0217293. doi:10.1371/journal.pone.0217293.

Hu, W., Fan, Y., Xing, J., Sun, L., Cai, Z., Maybank, S., 2018. Deep constrained siamese hash coding network and load-balanced locality-sensitive hashing for near duplicate image detection. IEEE Trans. Image Process. 27 (9), 4452–4464. doi:10.1109/TIP.2018.2839886.

ISIC, 2020. Isic archive gallery. Online. https://www.isic-archive.com.

Jinnai, S., Yamazaki, N., Hirano, Y., Sugawara, Y., Ohe, Y., Hamamoto, R., 2020. The development of a skin cancer classification system for pigmented skin lesions using deep learning. Biomolecules 10 (8).

Ju, L., Wang, X., Wang, L., Mahapatra, D., Zhao, X., Harandi, M., Drummond, T., Liu, T., Ge, Z., 2021. Improving medical image classification with label noise using dual-uncertainty estimation. 2103.00528.

Kassem, M.A., Hosny, K.M., Fouad, M.M., 2020. Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning. IEEE Access 8, 114822–114832. doi:10.1109/ACCESS.2020.3003890.

Kendrick, C., Gillespie, D., Yap, M.H., 2020. Anysize gan: a solution to the im-age-warping problem. arXiv preprint arXiv:2003.03233.

Kittler, H., Pehamberger, H., Wolff, K., Binder, M., 2002. Diagnostic accuracy of dermoscopy. The lancet oncology 3 (3), 159–165.

Le, D. N. T., Le, H. X., Ngo, L. T., Ngo, H. T., 2020. Transfer learning with class-weighted and focal loss function for automatic skin cancer classification. 2009.05977.

Lin, T., Goyal, P., Girshick, R., He, K., Dollr, P., 2017. Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007. doi:10.1109/ICCV.2017.324.

Lund, F., Clark, A., 2013. Pillow. https://www.github.com/python-pillow/Pillow.

Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Pitiot, A., Wang, C., 2019. Fusing fine–tuned deep features for skin lesion classification. Comput Med Imaging Graph 71, 19–29.

Majtner, T., Yildirim Yayilgan, S., Hardeberg, J., 2019. Optimised deep learning features for improved melanoma detection. Multimed Tools Appl 78, 11883–11903. doi:10.1007/s11042-018-6734-6.

McInnes, L., Healy, J., Melville, J., 2018. Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Melanoma UK, 2020. 2020 melanoma skin cancer report. Online. https://www.melanomauk.org.uk/2020-melanoma-skin-cancer-report.

Nahata, H., Singh, S., 2020. Deep Learning Solutions for Skin Cancer Detection and Diagnosis, pp. 159–182.

Ng, J.H., Goyal, M., Hewitt, B., Yap, M.H., 2019. The effect of color constancy algorithms on semantic segmentation of skin lesions. In: Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, Vol. 10953. International Society for Optics and Photonics, p. 109530R.

NHS, 2020a. How does the sun and uv cause cancer?Online. https://www.nhs.uk/conditions/melanoma-skin-cancer/causes/.

NHS, 2020b. Skin cancer (non-melanoma). Online. https://www.nhs.uk/conditions/non-melanoma-skin-cancer/.

Pham, T.C., Hoang, V.D., Tran, C.T., Luu, M.S.K., Mai, D.A., Doucet, A., Luong, C.M., 2020. Improving binary skin cancer classification based on best model selection method combined with optimizing full connected layers of deep cnn. In: 2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), pp. 1–6. doi:10.1109/MAPR49794.2020.9237778.

Ratul, A.R., Hamed Mozaffari, M., Lee, W.-S., Parimbelli, E., 2020. Skin lesions classification using deep learning based on dilated convolution. bioRxiv doi:10.1101/860700.

Rezvantalab, A., Safigholi, H., Karimijeshni, S., 2018. Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. arXiv preprint arXiv:1810.10348.

Rolnick, D., Veit, A., Belongie, S., Shavit, N., 2018. Deep learning is robust to massive label noise. 1705.10694.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J., Soyer, H. P., 2020. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. 2008.07360.

Ruiz, E. S., 2018. Is there a link between alcohol and skin cancer?Online. https://www.health.harvard.edu/blog/loose-link-alcohol-skin-cancer-2017120812861.

Sagar, A., Dheeba, J., 2020. Convolutional neural networks for classifying melanoma images. bioRxiv doi:10.1101/2020.05.22.110973.

Sarnoff, D., Gerome, D., 2017. Can your diet help prevent skin cancer?Online. https://www.skincancer.org/blog/can-your-diet-help-prevent-skin-cancer/.

Skin Cancer Foundation, 2017. Skin cancer facts and statistics. Online. https://www.skincancer.org/skin-cancer-information/skin-cancer-facts.

Sucholutsky, I., Schonlau, M., 2020. 'less than one'-shot learning: Learning n classes from m<n samples. 2009.08449.

Tan, M., Le, Q. V., 2020. Efficientnet: Rethinking model scaling for convolutional neural networks. 1905.11946.

Tang, P., Liang, Q., Yan, X., Xiang, S., Sun, W., Zhang, D., Coppola, G., 2019. Efficient skin lesion segmentation using separable-unet with stochastic weight averaging. Comput Methods Programs Biomed 178, 289–301.

Thörn, M., Pontén, F., Bergström, R., Sparén, P., Adami, H., 1994. Clinical and histopathologic predictors of survival in patients with malignant melanoma: a population-based study in sweden. J. Natl. Cancer Inst. 86 10, 761–769.

Tschandl, P., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. 10.7910/DVN/DBW86T

Tschandl, P., Codella, N., Akay, B., Argenziano, G., Braun, R., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., Lallas, A., Lapins, J., Longo, C., Malvehy, J., Marchetti, M., Marghoob, A., Menzies, S., Oakley, A., Paoli, J., Puig, S., Rinner, C., Rosendahl, C., Scope, A., Sinz, C., Soyer, P., Thomas, L., Zalaudek, I., Kittler, H., 2019. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol. 20, 938–947. doi:10.1016/S1470-2045(19)30333-X.

World Health Organization, 2017. More can be done to restrict sunbeds to prevent increasing rates of skin cancer. Online. https://www.who.int/phe/news/sunbeds-skin-cancer/en/.

Xie, Y., Zhang, J., Lu, H., Shen, C., Xia, Y., 2021. Sesv: accurate medical image segmentation by predicting and correcting errors. IEEE Trans Med Imaging 40, 286–296.

Xie, Y., Zhang, J., Xia, Y., Shen, C., 2020. A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE Trans Med Imaging 39, 2482–2493.

Zhang, Y., 2018. Learning near duplicate image pairs using convolutional neural networks. International Journal of Performability Engineering 14. doi:10.23940/ijpe.18.01.p18.168177.

Zhou Wang, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612. doi:10.1109/TIP.2003.819861.