The Effect of Image Similarity on Melanoma Classification

Hongyuan Xie
Manchester Metropolitan University, United Kingdom

Yanlong Zhang
Manchester Metropolitan University, United Kingdom

ABSTRACT

Skin cancer is one of the most common types of cancer, with research now increasingly focused on the use of deep learning algorithms to perform diagnosis in experimental settings. Deep neural networks can be used to assist early detection; however, accuracy can be highly reliant on aspects such as dataset quality and class distribution. This study investigates the impact on melanoma classification when using images that are visually similar from the publicly available ISIC 2019 dataset. The negative effect of image duplication is well known in deep learning; however, the effect of image similarity is an under-researched topic. In this work, we used an open source image similarity algorithm to identify similar images in the ISIC 2019 dataset. We identify groups of similar images at different similarity thresholds and investigate the effect of removing each threshold on a classification model. We then evaluate the best performing model on the ISIC 2019 datatest. Our results show that the best performing model was DenseNet201 when trained using the 100% similarity threshold images, and InceptionResNetV2 when trained using the 95% similarity threshold images. These results indicate that highly similar images present in the ISIC 2019 training set result in performance degrading bias, and that their removal shows in a boost to model performance.

CCS CONCEPTS

• **Human-centered computing**; • **HCI theory, concepts and models.**;

KEYWORDS

Skin Cancer, Deep Learning, Classification, Image Similarity, Data Quality

ACM Reference Format:
Hongyuan Xie and Yanlong Zhang. 2023. The Effect of Image Similarity on Melanoma Classification. In *2023 7th International Conference on Cloud and Big Data Computing (ICCBDC 2023), August 17–19, 2023, Manchester, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3616131.3616144

1 INTRODUCTION

Melanoma is considered to be one of the most serious types of cancer , with early and accurate diagnosis being crucial to ensuring that treatments are most effective in terms of patient survival [1]. Classification with big data has become one of the latest trends when talking about learning from the available information. [2].

ICCBDC 2023, August 17–19, 2023, Manchester, United Kingdom

ACM ISBN 979-8-4007-0733-9/23/08.
https://doi.org/10.1145/3616131.3616144

Deep learning neural networks have attracted significant attention in recent years as a promising tool for early detection and diagnosis of dermatological conditions, including skin diseases [3]. The effectiveness of deep learning neural networks for skin disease diagnosis largely depends on the quality and diversity of the dataset used for training the network [4]. The International Skin Imaging Collaboration (ISIC) [5] datasets are the largest and most popular source of skin lesion images used by researchers to train deep learning models. One issue that can negatively impact the performance of trained models is image duplication. This can occur when binary identical images are included in training, testing, and across both training and testing sets. This can lead to the introduction of bias within the network as certain features become over- represented. The impact of image duplication is further compounded in large-scale datasets such as ImageNet [6], which is often used for pretrained weights in deep learning models. The inclusion of duplicate images in datasets can have significant negative consequences for the performance of trained models [7][8] and is therefore important to carefully curate datasets to avoid such issues. To address these issues, and to improve the performance of deep learning models researchers have explored pre-processing techniques. For instance, studies have proposed using clustering, outlier detection, and dimensionality reduction to mitigate the effects of feature biases. This paper aims to investigate the impact of image similarity on publicly available skin lesion images from the ISIC2019 dataset when training convolutional neural networks (CNNs) for multi-class classification.

To accomplish this, we used an open source tool called dupeGuru [9] that uses a fuzzy algorithm to locate visually similar images. We then assess the performance of a selection of classification models using the processed training set after removing images at different similarity ranges. To provide further context on our research, Section 2 will review and discuss related literature, while Section 3 will provide a brief introduction to the ISIC2019 dataset. Section 4 will expound upon our research methodology, including the approaches we used to collect and process our data. Section 6 will present and analyze our experimental findings, and final section will conclude the paper by identifying potential areas for future research.

2 RELATED WORKS

Most currently research paper published at Jan/2023 by Gilani et al, Skin Cancer Classification Using Deep Spiking Neural Network, they employed deep spiking neural networks using the surrogate gradient descent method to classify 3670 melanoma and 3323 non-melanoma images from the ISIC 2019 dataset, achieved an accuracy of 89.57% and an F1 score of 90.07% using the proposed spiking VGG-13 model [10]. This work using ISIC2019 Dataset, they focus on the neural network. Villaruz et al. [11] use deep convolutional neural network feature extraction for berry trees classification, achieved very good results for plants by using pretrained model from ImageNet. Cassidy et al. [12] proposed a strategy for removing

Table 1: Dataset composition of the ISIC2019 dataset.

	Training and Test Sets		
1	Train images	25,331	Labeled
2	Val images (Splitting)	5,066	Labeled
3	Test images	8,238	Unlabeled
	Lesion Class		
1	Melanoma	MEL	4522
2	Melanocytic nevus	NV	12,875
3	Basal cell carcinoma	BCC	3323
4	Actinic Keratosis	AK	867
5	Benign keratosis	BKL	2624
6	Dermatofibroma	DF	239
7	Vascular Lesion	VASC	253
8	Squamous cell carcinoma	SCC	628
9	None of the others	UNK	0

duplicate image files from the ISIC 2017 - 2020 dataset as a means of reducing bias in deep learning models trained on these dataset. They presented results from a range of commonly used CNN classification architectures trained on a curated balanced dataset which indicated excellent class distribution and improved performance measures. This work performed preliminary image similarity experiments using ImageHash, mean squared error, structural similarity index measure, and cosine similarity, however, detailed results were not reported.

Dipto et al. [8] conducted a series of multi-class classification experiments using the Diabetic

Foot Ulcer 2021 dataset [13] to ascertain which similarity thresholds had the biggest effect on validation and test metrics. They found that the training set with 80% similarity threshold images removed achieved the best performance using InceptionResNetV2. The experiments demonstrated improvements in test results for F1-score, precision, and recall of 0.023, 0.029,and 0.013, respectively.

3 DATASET

The ISIC datasets [14] are a leading repository for researchers in deep 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 and in some cases additional metadata. The yearly ISIC challenges have resulted in significant contributions to the field [12]. The dataset used in our experiments, ISIC-2019 (see Table 1), contains 25,331 images and comprises 9 different classes which includes and unknown category. The test dataset consists of 8,238 images whose labels are not publicly available. The test dataset includes an additional class that is not contained in the training dataset. Predictions on the ISIC2019 test dataset are assessed using an automatic online evaluation system. The goal of the ISIC2019 challenge is to classify dermoscopic images among nine different diagnostic categories:

4 METHODOLOGY

In this study, we aimed to investigate the impact of similar images on the performance of multi-class classification models trained using ISIC-2019 dataset. To achieve this, we employed a strategy that gradually removes groups of similar images from the training dataset and evaluated the perfor- mance of the models on the modified dataset.

We create a validation set by splitting 20% of the dataset that falls under the threshold of 60%. Before doing the split, we perform cross-image similarity and duplicated file checks to ensure that there are no similar images between the training set and the validation set. This ensures that the validation set is entirely distinct from the training set, preventing any overlapping or duplication in the images between the two sets. Each set of similar images are identified using the dupeGuru [9] Windows application. This open source application implements a fuzzy search algorithm capable of identifying visually similar images. The dupeGuru search results can be filtered by similarity in percentage, so that similar images in the ISIC2019 dataset can be identified. For each threshold level (60-100%, where 100% represents binary identical images), we identified groups of similar images and saved the filenames of the images to a CSV file. The CSV files contained the group ID and the image filenames, which we later used to remove similar images from the training dataset to produce a set of new training sets, with each representing a similarity threshold. The group ID indicates a collection of similar images within the dupeGuru results.

Table 2 shows a summary of each similarity threshold together with the number of images removed and the total number of images remaining in each new training set.

4.1 Fuzzy Algorithm

The dupeGuru application uses a fuzzy algorithm in its image search functionality, which was first introduced by Lotfi Zadeh in 1965 [15]. The algorithm used by dupeGuru operates in several steps. First, each image is read in RGB bitmap mode and divided into blocks. For each block, the average color is computed and stored in a cache database, resulting in a grid of average colors for each image.

Table 2: Summary number of similar images removed for dataset

Total Images	Threshold	Total removed	Total remaining
25,331	100%	50	25,281
	95%	87	25,244
	90%	199	25,132
	85%	469	24,862
	80%	1064	24,267
	75%	2256	23,075
	70%	4006	21,325
	65%	5902	19,429
	60%	7845	17,486

This process helps to reduce the size of the data being compared, making the algorithm more efficient. Next, the algorithm compares the corresponding grids of average colors between the two images being compared. To do so, it computes the difference between the red, green, and blue values of each corresponding pair of grid tiles and sums them. This produces a score that reflects the similarity of the two images. Finally, the RGB differences are added together to obtain a score for the two images. If the score is below the threshold set by the user, the algorithm indicates that the images are similar. If the user defines a threshold of 100, the algorithm adds an additional constraint that requires the two images to contain identical binary data. This algorithm compares the average color of small regions of the image, rather than the entire image itself. This strategy allows the algorithm to detect similar image that have undergone slight modifications, such as resizing or cropping. The use of a threshold also allows the user to customize the sensitivity of the algorithm and avoid false positives.

4.2 Similar Images Identification

To identify similar images in the ISIC 2019 dataset, we utilized the dupeGuru application and set similarity threshold levels at 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, and 100%, where 100% represents binary identical images. The threshold level indicates the minimum percentage of similarity required for the application to flag two images as similar. For example, if the threshold level is set at 75% the algorithm will identify any two images that are 75% or more similar to each other.

We used the dupeGuru application fuzzing algorithm on both the training and testing sets, as well as the combined train and test set to identify similar images present in each set separately and in combination. The resulting output is summarized in Table 2, which shows the number of similar images detected by the dupeGuru fuzzy algorithm for each threshold level on the training set, the test set, and the combined train and test sets.

4.3 Similar Image Removal

The dupeGuru fuzzy algorithm was used to identify each threshold group of similar images in the 60% to 100% similarity threshold ranges. After running the algorithm, the filenames of the similar images in each similarity threshold are saved to a CSV file, along with a group ID assigned by the algorithm to each group of similar images. The CSV file output from dupeGuru is then merged with the file containing the ground truth labels for the training set based on the image filename. This step is used to ensure that only images from the training set are used for training, and not images from the test set. Next, for each group of similar images, all images in the group except the first image are removed. This means that only one example image from each similarity group ID is kept for inclusion in each new similarity threshold training dataset. By doing this, we ensure that each training dataset contains only distinct examples according to the similarity threshold. Finally, the number of unique images in each similarity threshold training dataset is calculated, and the results are presented in Table 2.

4.4 Similarity Threshold Examples

In this section, we analyse a selection of images from each similarity threshold returned by the dupeGuru fuzzy algorithm prior to training the multi-class classification models.

4.4.1 Training Set Similarity. Figure 1 shows an example of two similar images within the 60% similarity threshold in the original training set. We observe that these images contain dark corner artifacts introduced by the use of a dermascope, and that these dark features will be part of the similar image features. Figure 2 shows an example of similar images within the 65% similarity threshold within the original training set. These two images are clearly captured from different lesions, and we observe that much of the image similarity results from the texture of the surrounding skin, although areas of the surrounding lesion also exhibit similar features. Figure 3 shows an example of training set images that fall within the 70% similarity threshold. As per figure 2, a large part of the similar features appear to come from the surrounding skin, with a smaller degree of similarity present in the actual lesion. Figure 4 shows an example of similar training set images within the 75% similarity threshold. These examples are clearly taken from the same lesion at different intervals, with skin and lesion features varying subtly. Figure 5 shows an example of similar training set images within the 80% similarity threshold. These images are also examples of the same lesion taken at different time frames, with very subtle differences in the actual skin and lesion areas, with the later case exhibiting a darker lesion area.

4.4.2 Inter-class Similarity. Figure 6 shows two training set images that exhibit inter-class similarity in the 60% similarity threshold, where (a) is NV and (b) is MEL. We observe that although the images

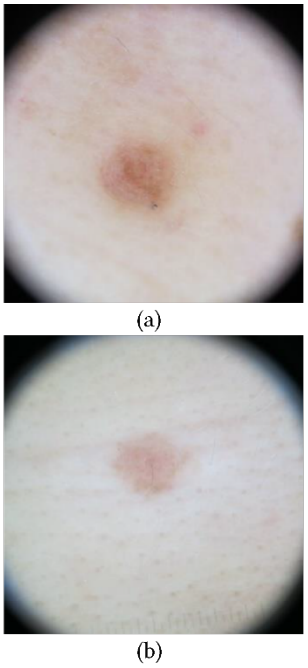

Figure 1: Illustration of two training set images identified by the dupeGuru fuzzy algorithm in the 60% similarity threshold.

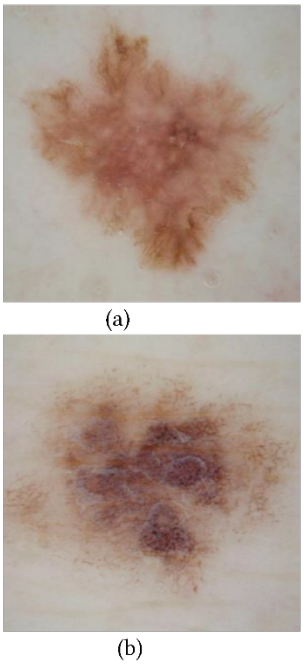

Figure 2: Illustration of two training set images identified by the dupeGuru fuzzy algorithm with a similarity threshold of 65%.

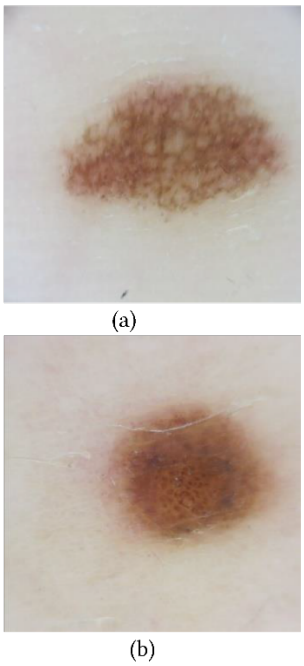

Figure 3: Illustration of two training set images identified by the dupeGuru fuzzy algorithm with a similarity threshold of 70%.

are similar, the melanoma example shows a much darker lesion area.

4.4.3 Train & Test Similarity. Figure 7 shows an image from the training set and an image from the test set in the 60% similarity threshold. The ISIC2019 test set ground truth is not publicly available so are not able to indicate if the two lesions are of the same class.

5 EXPERIMENTAL SETUP

The hardware configuration used in our experiments was as follows: an Intel Core i7-6700K CPU @ 4.00GHz, NVIDIA TITAN X (Pascal) 12GB GPU, 32GB RAM. The software configuration was as follows: Python3.9.12, and TensorFlow2.4.1-GPU for the multi-class classification experiments, we trained five model architectures - DenseNet201, InceptionResNetV2, ResNet50, VGG16, and Vit-b32. We used a batch size of 32, a learning rate of 0.0002, and the Adam optimizer with early stopping.

6 RESULTS AND DISCUSSION

Table 3 shows the accuracy between the baseline validation results and the best performing thresholds for the five models. The DenseNet201 model had the highest baseline validation accuracy of 70.61%, with the 100% binary identical threshold increasing accuracy to 72.45%. The VGG16 model showed the lowest baseline validation accuracy of 67.12%, with the 70% threshold model showing a minor increase to 67.17%. We observe that the lowest performing models,

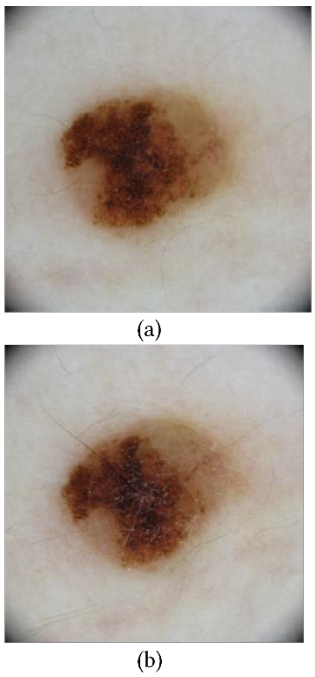

Figure 4: Illustration of two training set images identified by the dupeGuru fuzzy algorithm with a similarity threshold of 75%.

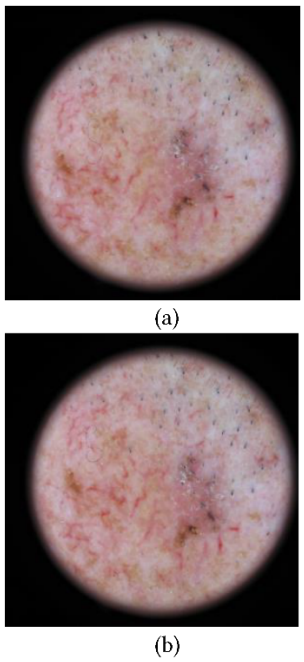

Figure 5: Illustration of two training set images identified by the dupeGuru fuzzy algorithm with a similarity threshold of 80%.

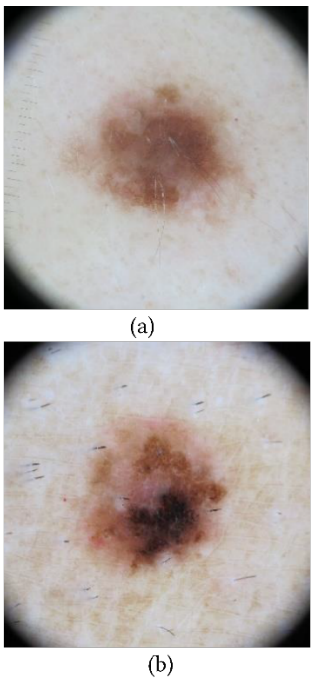

Figure 6: Illustration of two training set images identified by the dupeGuru fuzzy algorithm with a similarity threshold of 60% which exhibit inter-class similarity, where (a) is NV, and (b) is MEL.

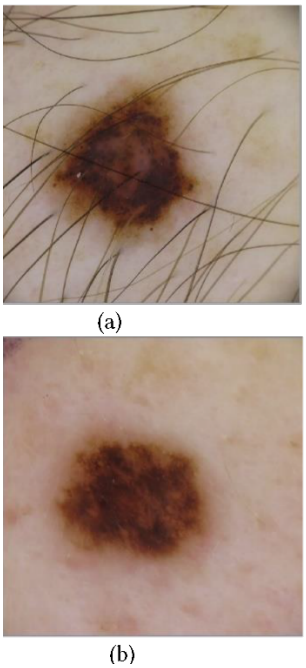

Figure 7: Illustration of two images identified by the dupeGuru fuzzy algorithm with a similarity threshold of 60% where (a) is from the training set, and (b) is from the test set.

Table 3: Comparison of baseline validation accuracy results for 5 best performing classification models trained on similarity threshold training sets ranging from 75% to 100%.

Model	Baseline Val Acc	Best Threshold	Best Threshold Val Acc
DenseNet201	70.61%	100%	72.45%
InceptionResNetV2	70.45%	95%	71.52%
ResNet50	68.42%	95%	68.93 %
VGG16	67.12%	70%	67.17%
Vit-b32	67.64%	80%	69.30%

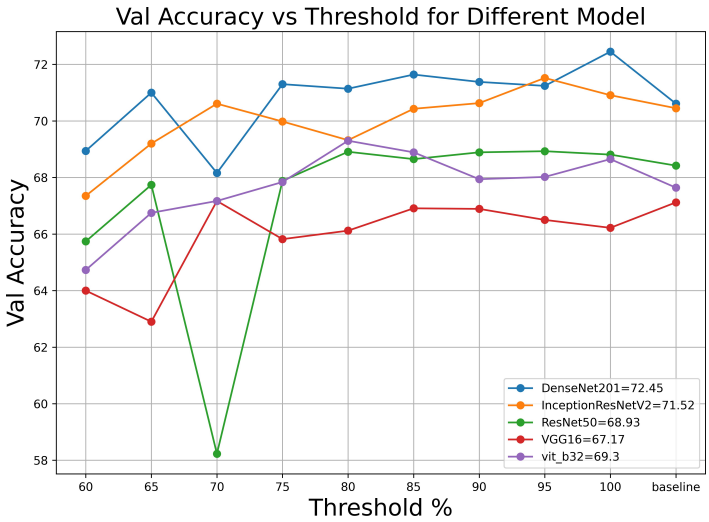

Figure 8: Comparison of the val accuracy of different model architectures trained in our experiments.

in terms of baseline validation accuracy, show a reduced performance from the 95% threshold onwards . Our results indicate that different model architectures respond to visually similar images in different ways. Some benefit from the elimination of redundant images, while others could experience a performance decline if too many are removed. It is necessary to gauge the impact of similarity and establish an optimum threshold for each model and dataset for the most successful outcomes.

A summary of validation results for each trained model when compared to the best baseline result is shown in Figure 8.

Our tests on the ISIC2019 dataset show that the DenseNet201 model had the best overall validation performance among the different similarity thresholds. In terms of binary identical images, where the similarity threshold is 100%, our best model demonstrated a validation accuracy of 72.45%. For images that were not binary identical, the 85% threshold yielded the best validation accuracy of 71.64%. In terms of non-identical images, our results correspond with those in a previous study which investigated the effect of image similarity on diabetic foot ulcer image classification [8]. The prior work showed that the 80% similarity threshold was the best performing threshold, which is close to our best performing non-identical threshold result of 85%, Table 4. DenseNet201 utilises an Inception module with residual connections which allows for the detection of multiple scale features enabling the network to counter the vanishing gradient problem. Our findings corroborate with other classification experiment results in other fields. However, despite the superior performance of DenseNet201 in our experiments, we should note that our analysis was limited by the lack of available ground truth labels for the ISIC2019 test set. Future work could expand on our initial findings by testing on other ISIC datasets and reporting on test set metrics where test set data is available. Results from such experiments should be compared to results of similar experiments in other domains to ascertain if there is a correlation of similarity thresholds which yield to best results.

7 CONCLUSIONS

In this paper, we sought to explore the effect of image similarity on a range of popular deep learning multi-class classification models, which were trained and validated on the ISIC2019 dataset. We hypothesised that the performance of these models would degrade due to biases that visually similar images may introduce. Thus, we ran several experiments by removing groups of increasingly similar images from the training set. The results of our study demonstrated that classification models may be negatively affected if too many or too few similar images are removed from the training set. Our findings highlighted the importance of identifying bias within a large dataset and how such challenges may be overcome in terms of the removal of similar images, advancements in cloud and big data computing, the benefits of such techniques in terms of storage

Table 4: Training and validation results for the DenseNet201 model for similarity thresholds between 60% to 100% and baseline.

Threshold	Best Epoch	Train Acc	Val Acc
60%	11	97.03%	68.94%
65%	20	98.63%	71.00%
70%	13	97.51%	70.61%
75%	17	98.19%	71.30%
80%	24	98.12%	71.14%
85%	28	98.66%	**71.64%**
90%	21	98.36%	71.38%
95%	29	97.26%	71.24%
100%	**29**	**98.87%**	**72.45%**
Baseline	19	97.03%	70.64%

optimization, processing efficiency, accuracy improvement, and overall system scalability. The aim of this research is to highlight the negative effects that image similarity may present in the use of a deep learning dataset, so that , this is preliminary research just based on ISIC skin cancer dataset , for further work, experiment on using less amount of data based on crossed domain dataset, can also obtain better accuracy.

REFERENCES

[1] Melanoma UK. 2020 melanoma skin cancer report. Online, 2020.

[2] Victoria López, Sara del Río, José Manuel Benítez, and Francisco Herrera. Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. *Fuzzy Sets and Systems*, 258:5–38, 2015. Special issue: Uncertainty in Learning from Big Data.

[3] Manu Goyal, Amanda Oakley, Priyanka Bansal, Darren Dancey, and Moi Hoon Yap. Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. *IEEE Access*, 8:4171–4181, 2019.

[4] Titus J. Brinker, Achim Hekler, Alexander H. Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Stefan Frohling, Jochen S. Utikal, and Christof von Kalle. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 2019.

[5] International Skin Imaging Collaboration. ISIC melanoma project dataset v1.0. https://doi.org/10.7910/DVN/DBW86T, 2018.

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, *et al.* Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[7] Joanna Jaworek-Korjakowska, Andrzej Brodzicki, Bill Cassidy, Connah Kendrick, and Moi Hoon Yap. Interpretability of a deep learning based approach for the classification of skin lesions into main anatomic body sites. *Cancers*, 13(23), 2021.

[8] Imran Chowdhury Dipto, Bill Cassidy, Connah Kendrick, Neil D Reeves, Joseph M Pappachan, Vishnu Chandrabalan, and Moi Hoon Yap. Quantifying the effect of image similarity on diabetic foot ulcer classification. In *Diabetic Foot Ulcers Grand Challenge: Third Challenge, DFUC 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pages 1–18. Springer, 2023.

[9] dupeGuru, 2018. [Online] Available from: https://dupeguru.voltaicideas.net/ [Accessed: 7th June, 2022].

[10] Syed Qasim Gilani, Tehreem Syed, Muhammad Umair, and Oge Marques. Skin cancer classification using deep spiking neural network. *Journal of Digital Imaging*, pages 1–11, 2023.

[11] Jolitte A Villaruz. Deep convolutional neural network feature extraction for berry trees classification. *Journal of Advances in Information Technology Vol*, 12(3), 2021.

[12] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the isic image datasets: usage, benchmarks and recommendations. *Medical image analysis*, 75:102305, 2022.

[13] Moi Hoon Yap, Bill Cassidy, Joseph M. Pappachan, Claire O'Shea, David Gillespie, and Neil D. Reeves. Analysis towards classification of infection and ischaemia of diabetic foot ulcers. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4, 2021.

[14] Josef Steppan and Sten Hanke. Analysis of skin lesion images with deep learning. *arXiv preprint arXiv:2101.03814*, 2021.

[15] L. A. Zadeh. Fuzzy sets. In *Information and Control*, volume 8, pages 338–353, 1965.