Imperial College London

AI Health

UKRI Centre for Doctoral Training in Artificial Intelligence for Healthcare

Imperial College London

Written Report

Dermatologist level classification of skin cancer with deep neural networks

Esteva et al., 2017

Leo Huang

February 1, 2024

Module: 70152 - Research Tutorial - AI and Machine Learning for Healthcare

1 Paper Summary

Title: Dermatologist level classification of skin cancer with deep neural networks

Authors: Andre Esteva1, Brett Kuprel1, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, Sebastian Thrun

Publication details: Published in Nature - Letters, Volume 542, 2017

Reference: [9]

1.1 Background

Skin cancer is one of the most common cancers worldwide, with 5.4 million new diagnoses in the US alone each year [22]. Early detection of skin cancer is critical for timely recovery; in the case of melanoma, commonly cited as the deadliest form of skin cancer, five-year survival rates can drop from 99% for early-stage diagnoses down to a mere 5% when diagnosed at the late stage of the disease.

Skin cancers normally manifest as localised lesions on the surface of the skin. As such, diagnosis is typically performed via visual screening with the naked eye and dermoscopy, followed by histopathological confirmation of a biopsied skin sample. Such clinical assessment, however, can be both costly and difficult to access. The task is further complicated by the existence of visual similarities between malignant and benign skin lesions.

In an attempt to alleviate these challenges, deep convolutional neural networks (CNNs) have been proposed as an effective diagnostic tool for skin cancer. Specifically, CNNs have the potential to aid clinical diagnosis via the automated classification of skin lesions taken from both digital photographs and dermoscopy images. This paper aims to explore the applicability of CNNs in such contexts, demonstrating that deep neural networks have the capacity to match the skin cancer classification capabilities of experienced dermatologists.

1.2 Methods

1.2.1 Dataset

This study utilised a dataset of 129,450 images and corresponding dermatologist-verified disease labels. The dataset comprised 126,076 digital photographs and 3,374 dermoscopy images, with a total 2,032 individual cancerous and non-cancerous skin diseases represented. Images were collected in collaboration with Stanford University Medical Center as well as from online public repositories [2]. 98% of the dataset was used to train and validate the neural network, while the remaining 2% was held out for subsequent testing. Note that only the test image disease labels were biopsy-verified.

1.2.2 Disease Partitioning Algorithm

To enhance the network training process, this paper proposes a novel recursive tree-based algorithm that merges individual diseases based on clinical and visual similarity. By employing this taxonomical algorithm, the original 2,032 individual diseases were grouped into 757 disease classes, a subset of which is shown in Fig. 1.

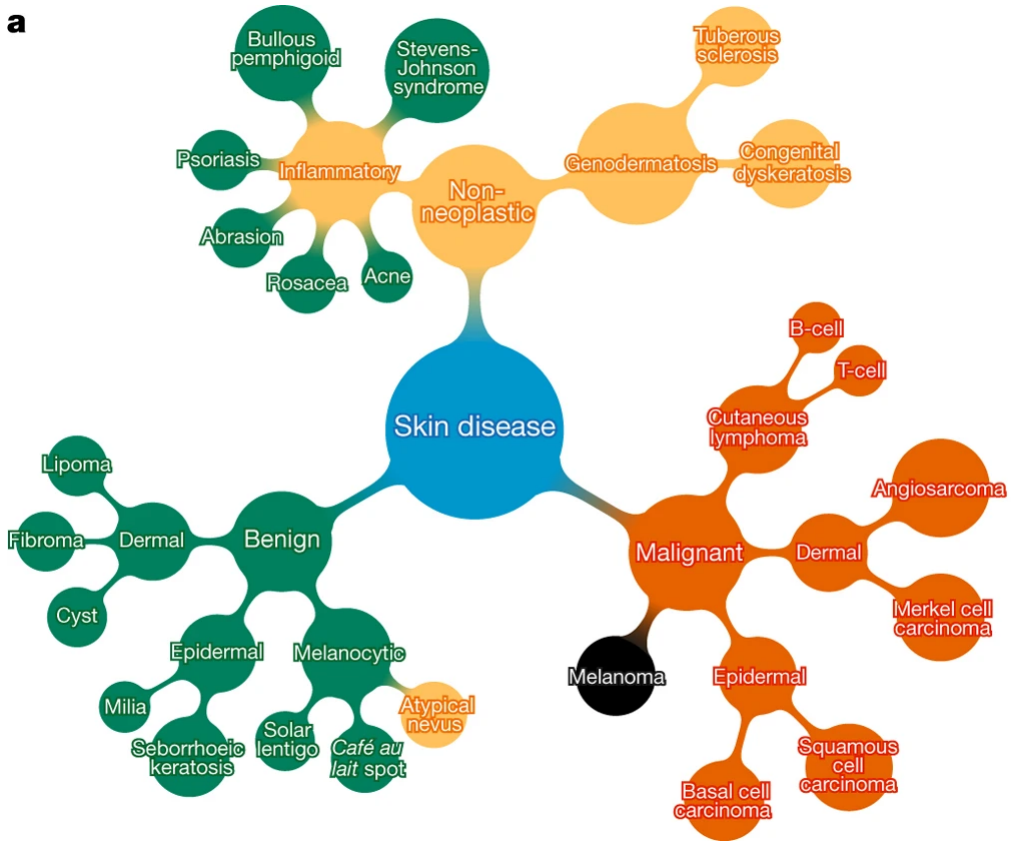

Figure 1: Subset of the top of the algorithm-generated skin disease taxonomy. Green nodes are benign diseases, red nodes are malignant, and orange nodes can be either benign or malignant. The first two levels are used in validation testing. [9]

1.2.3 Training

The network utilised a pre-trained GoogLeNet Inception v3 CNN architecture [26] (for discussion on the progress of Inception v3 since the publication of this paper, see Section 2.4). Training was conducted via transfer learning, whereby the final classification layer of the pre-trained network was removed and re-trained using the partitioned disease labels. Importantly, these training images were resized to 299 x 299 pixels in order to leverage the natural features learned by the pre-trained network. Model hyperparameters were optimised via standard backpropagation.

1.2.4 Validation

Nine-fold cross-validation was employed on the model using the partitioned skin disease classes. Specifically, the ability of the CNN to correctly classify images based on the three first-level disease classes in Fig. 1 was compared to that of two dermatologists. The same experiment was repeated using the nine second-level disease classes in Fig. 1.

1.2.5 Testing

Testing was performed in the form of three medically important use cases: 1) malignant versus benign carcinoma photographs (representing the most common skin cancer), 2) malignant versus benign melanoma

photographs (representing the deadliest skin cancer), and 3) malignant versus benign melanoma dermoscopy images. For each experiment, CNN classification performance was compared to the individual and average performance of 21-25 dermatologists via metrics of sensitivity (true malignancy rate) and specificity (true benignity rate). In the case of the CNN, these metrics were computed by choosing a threshold value t for each malignancy probability P and defining the prediction $\hat{y}$ as $\hat{y} = P \geq t$.

1.3 Results

1.3.1 Skin Disease Classification Performance

Validation results indicated superior skin disease classification performance of the CNN compared to that of the two dermatologists for both three-way and nine-way disease partitions, as shown in Table 1.

Table 1: Validation testing results for three-way and nine-way disease classification. CNN = network directly trained on the three and nine classes. CNN - PA = network trained via partitioning algorithm. CNN accuracies shown as mean ± standard deviation. [9]

a.

Classifier	Three-way accuracy
Dermatologist 1	65.6%
Dermatologist 2	66.0%
CNN	69.4 ± 0.8%
CNN - PA	**72.1 ± 0.9%**

b.

Classifier	Nine-way accuracy
Dermatologist 1	53.3%
Dermatologist 2	55.0%
CNN	48.9 ± 1.9%
CNN - PA	**55.4 ± 1.7%**

1.3.2 Skin Cancer Classification Performance

For all three test cases, the CNN outperforms the average dermatologist at skin cancer classification using a subset of the testing set, as shown by the specificity-sensitivity curves in Fig. 2. Notably, similar results were achieved when the CNN was re-tested using the entirety of the testing set.

2 Discussion and Critique

2.1 Paper Analysis

This study compared the performance of a CNN trained on partitioned disease classes versus that of a group of experienced dermatologists in both classification of skin diseases and differentiation between malignant and benign skin cancers from digital images.

Results from validation testing demonstrate that the CNN was able to learn relevant information during network training, as shown by its higher skin disease classification accuracy compared to the two dermatologists. Specifically, the CNN achieved an accuracy of 72.1±0.9% with the partitioning algorithm for three-way disease classification compared to 65.6% and 66.0% for the dermatologists. The CNN also performed slightly better in classifying finer disease partitions (nine-way), achieving an accuracy of 55.4±1.7% compared to 53.3% and 55.0% for the same two dermatologists. These metrics, however, are inconclusive for verifying absolute classification ability, given that the ground truth disease labels for the validation images were not biopsy-proven.

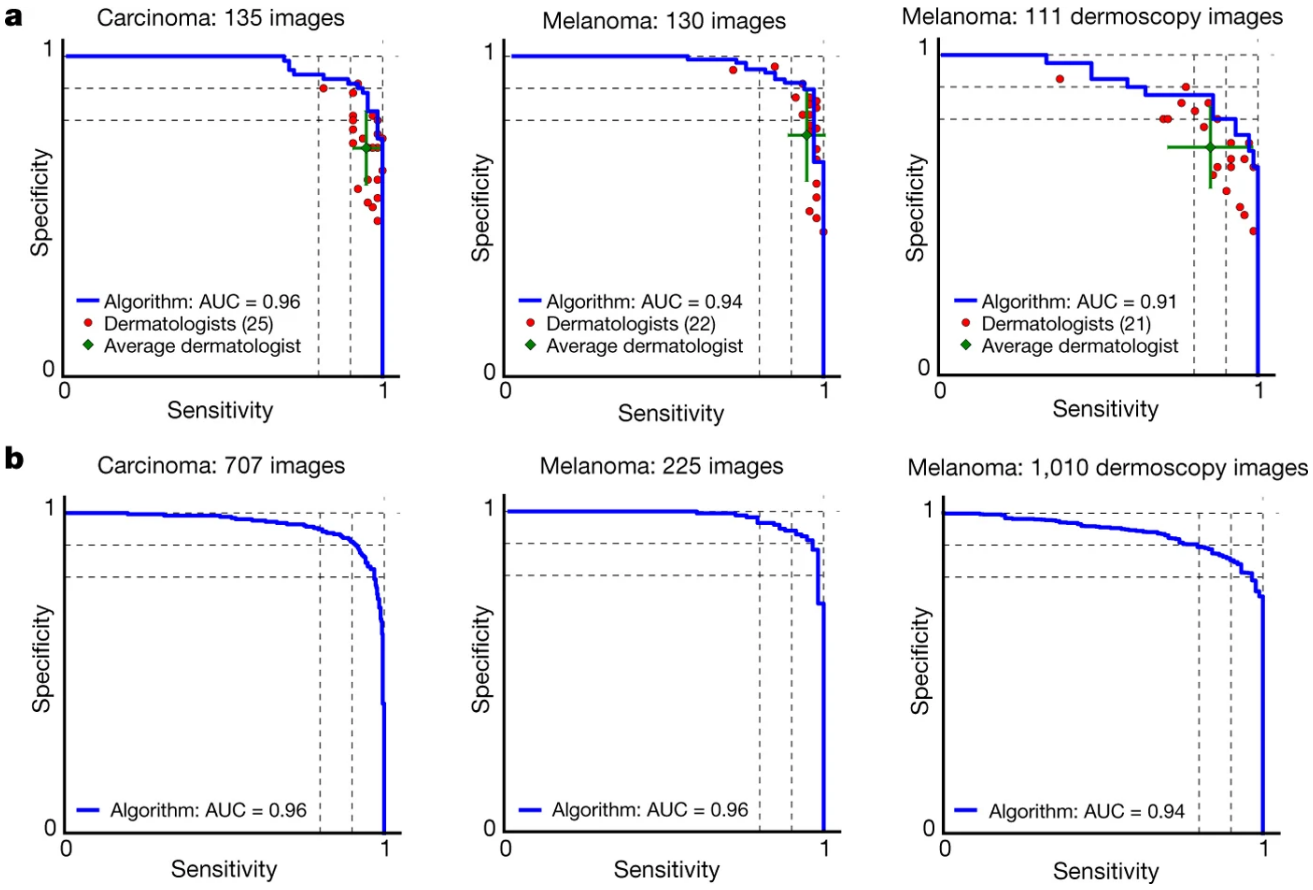

Figure 2: Specificity-sensitivity curves showing CNN skin cancer (malignant versus benign) classification performance for a range of threshold probabilities. The CNN achieves superior performance to a given dermatologist if its red point lies below the blue curve. AUC = area under the curve (higher the better). **a**, Test case results (CNN versus dermatologists) for carcinoma photographs, melanoma photographs, and melanoma dermoscopy images using a subset of the testing set. **b**, Test case results (CNN only) using the entire testing set. [9]

Analysis of the confusion matrices generated from validation testing (Fig. 3) highlights that both the CNN and dermatologists tend to misclassify similar skin diseases. Inflammatory conditions were commonly mistaken for skin lesions, for example, while it was often difficult for both groups to distinguish between malignant and benign melanocytic lesions. In the latter case, dermatologists tended to err on the side of malignant classification, suggesting the greater importance of minimising false negative over false positive classifications in such cases. Finally, malignant and benign dermal tumours were frequently misclassified, likely due to their often indistinguishable appearance as small nodules underneath the skin surface.

The CNN was shown to outperform the average dermatologist in all three test cases for this chosen dataset, demonstrating the CNN's equal level of competence for the chosen skin cancer classification tasks. While the CNN was largely able to equal or outperform most individual dermatologists on unseen data, one dermatologist in each of the melanoma photograph and dermoscopy image cases was able to achieve superior performance.

Interestingly, there was a slight drop in skin cancer classification performance for dermoscopy images for both the CNN and dermatologists in comparison to the photographic images, with the CNN achieving

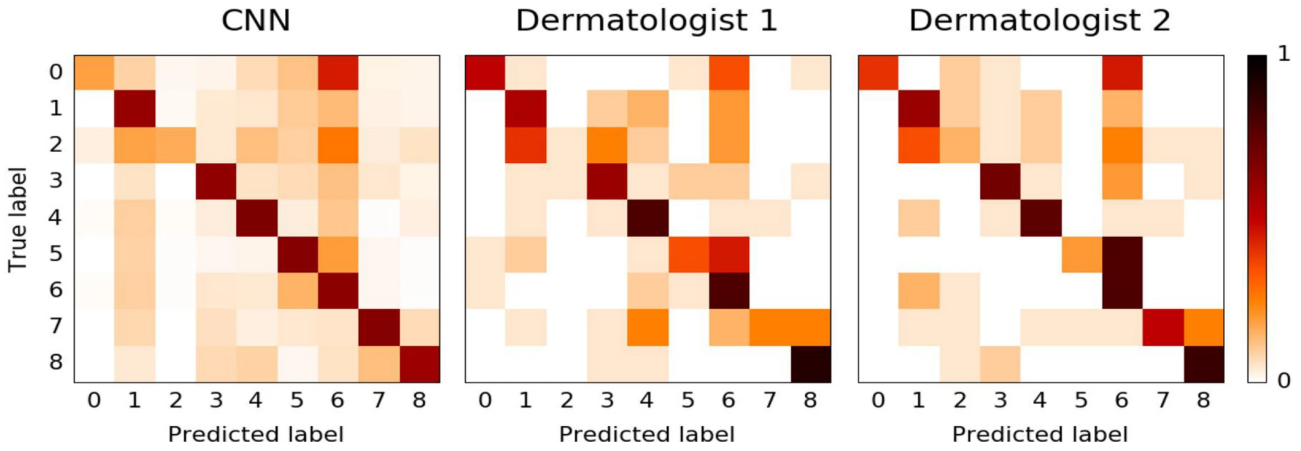

Figure 3: Comparison of confusion matrices between the CNN and two dermatologists for the nine-way skin disease classification during validation testing. For each confusion matrix, element (i, j) is the probability of predicting class j given a true label of class i. Class 0 = cutaneous lymphoma, class 1 = benign dermal, class 2 = malignant dermal, class 3 = benign epidermal, class 4 = malignant epidermal, class 5 = genodermatoses, class 6 = inflammatory, class 7 = benign melanocytic, class 8 = malignant melanoma. [9]

an AUC of 0.91 for the melanoma dermoscopy images compared to 0.96 and 0.94 for the carcinoma and melanoma photographs, respectively.. This trend likely reflects the greater visual challenge and higher classification difficulty of dermoscopy images, rather than being an indication of their diagnostic accuracy.

In summary, this paper was one of the first to demonstrate the effective application of artificial intelligence to the classification of both general skin conditions and specific skin cancers, matching or outperforming dermatologist performances across three critical diagnostic tasks. In terms of clinical implications, the authors concluded that fitting mobile devices with such CNNs could enhance proactive and early diagnostic care outside of clinical visits, as well as augment clinical decision-making by providing complementary information to contextual patient factors used in real-world diagnosis. Subsequent developments in the field are discussed in Section 2.4.

2.2 Paper Strengths

This study showed numerous strengths and demonstrated a strong proof-of-concept for the increased usage of deep neural networks in dermatological diagnosis to improve automation and reduce costs at a time when the field was not as developed, reflected in the paper's high citation count (see Section 2.4 for related works). For example, the dataset used to train and test the model was two orders of magnitude larger than previous similar studies [3][11][18][20]. Moreover, the network incorporated lower-quality images (i.e., perturbed by zoom and blurriness) during training to reflect the expected quality of patient input images and improve robustness to potential variability in the real world. The use of biopsy-proven ground truth labels during testing further provided an empirical baseline for subsequent comparison.

The approach of using transfer learning via the Inception v3 network was beneficial in this context owing to the lack of large real-world dermatological datasets for model training at the time of publication. In this way, the model can efficiently combine the basic learned features from the pre-trained model with the domain-specific skin lesion features from the retrained layers.

Notably, the paper's usage of its taxonomical algorithm was beneficial in generating both medically relevant and deep learning-suitable training classes. Specifically, the algorithm was able to find a harmony between having too many fine-grained classes that lack sufficient data to be learned properly, as well as

too few coarse classes that are too data-abundant and likely to introduce bias. Furthermore, the algorithm outputs a probability metric for each class during inference, allowing for the model uncertainty over a given prediction to be expressed and utilised during clinical practice.

2.3 Paper Limitations

The most apparent limitation of the paper was likely the lack of real-world generalisability of the CNN-based model. The study did not report its dataset's distribution of population demographics, which may have an impact on the appearance of skin cancer and hence impact the model's classification ability. A high imbalance was also present between the number of photographic and dermoscopic images in favour of the former, meaning that the algorithm had much fewer dermoscopy features to learn from. Furthermore, the authors state that, due to the difficulty of obtaining such images, the full spectrum of lesions encountered in typical clinical practice was not covered within the dataset.

To enhance clinical applicability of the model described, further test cases outside of the three reported with a significantly greater number of dermatologists are required to determine whether the classification algorithm is scalable to other lesion types and skin diseases, as well as the inclusion of additional contextual factors as model features to better replicate real-world clinical practice. Both these points reinforce the importance of data-driven diagnostic approaches. Moreover, a detailed discussion of the computational complexity and inference times was lacking, which would have better supported the authors' claim of the algorithm's compatibility with mobile devices.

In addition to generalisability, model accuracy was potentially compromised due to the possibility of images being incorrectly labelled by dermatologists and causing error propagation during the training process. Outside of dataset limitations, the interpretability of the algorithm and subsequent results is lowered by the black box nature of the CNN employed, limiting the model's usage outside of academic research settings. This is further compounded by the lack of source code published by the authors to verify their experimental results. In terms of the metrics used, further detail on the false positive and false negative rates of the CNN predictions would have been beneficial for evaluating the model's reliability and accuracy. Additionally, the skin cancer classification task was a binary one that only considered malignancy and not the stage of the disease, and thus earlier diagnosis could not be measured as a clinical outcome of the study.

The use of 9-fold cross-validation was an interesting methodological choice that was unfortunately not justified by the authors. In general, k-fold cross-validation with $k = 10$ is more commonly utilised and has been cited to provide sufficient variance in the training data to enable learning while also balancing computational cost in most cases [12]. Leave-one-out cross-validation is also a common option when robust estimates of model performance are more important than computational efficiency. Lower values of k may increase the bias of performance estimates, hence some form of quantitative comparison would have been useful in helping to rationalise the experimental design.

2.4 Subsequent Works

Since the publication of this seminal paper in 2017, the field of deep learning-based skin cancer classification from digital images has seen tremendous advancements. Prior to this paper's publication, CNNs had already shown exceptional performance in computer vision tasks outside of the dermatological domain, and their continued prevalence for skin cancer classification in the proceeding years is described in a number of recent reviews [7][21]. Naqvi et al. [21] report the following, in chronological order of development, as the most commonly used CNN architectures in skin cancer analysis: AlexNet (2012) [19], VGG (2014) [24], Inception v3 (2015) [26], ResNet (2015) [13], DenseNet (2017) [16], and MobileNet (2017) [15].

It is interesting to note that Inception v3 remains more commonly used for skin cancer analysis than

its subsequent iterations Inception v4 and Inception-ResNet, despite Szegedy et al. [25] reporting a superior performance of 3.08% top-5 error on the test set of the ImageNet classification challenge using an ensemble of three residual and one Inception v4. One example of such usage is the work of Emara et al. [8], who utilised a single Inception v4 model in which long residual connections allowed the concatenation of features extracted from earlier layers with high-level layers to improve classification performance. The authors achieved an accuracy of 94.7% using the International Skin Imaging Collaboration (ISIC) 2018 Challenge dataset [4][27].

More recently, the idea of deep learning-based ensemble methods has been gaining traction for classifying skin cancers. Kausar et al. [17] proposed an ensemble of five network architectures, namely ResNet, Inception v3, DenseNet, Inception-ResNet v2, and VGG-19. Using majority voting and weighted majority methods, the authors achieved accuracies of 98% and 98.6%, respectively, on the ISIC Archive dataset (`https://www.isic-archive.com`, accessed on 1 February 2024), which were higher than that of any individually trained model. Deep ensembles have also played a role in model selection for skin cancer diagnosis via uncertainty quantification. For example, Abdar et al. [1] integrated deep ensembles and ensemble Monte Carlo dropout methods in their classification models (ResNet152 v2, MobileNet v2, DenseNet201, and Inception-ResNet v2), achieving an accuracy of 88.95% on a subset of the ISIC Archive dataset.

Outside of advancements in the algorithms employed, the number of publicly available skin cancer image datasets has also seen substantial development since 2017. For example, in addition to the ISIC Archive mentioned above, annual ISIC challenges between 2017 and 2020 [4][5][6][23][27] have prompted the curation of datasets containing a total of 72,957 images and corresponding ground truths. Dataset sizes remain a limiting factor in clinical deep learning applications, however, when compared to the orders of magnitude present in non-medical imaging datasets, while images containing a greater representation of skin colours also remain insufficient for addressing skin colour bias.

While Esteva et al. [9] have since largely expanded their scope of research to other clinical applications of deep learning, a perspective article published by Esteva and Topol in 2019 [10] discussed the barriers towards clinical translation of deep learning systems for skin cancer diagnosis from a more practical perspective, in which the topic of humans and artificial intelligence working alongside one another was raised. This issue was further discussed by Hekler et al. [14], who found that combining human and artificial intelligence for the classification of skin cancer images yielded superior results (accuracy of 82.95%) compared to that achieved by artificial and human intelligence alone (accuracies of 81.59% and 42.94%, respectively), supporting the recent increase in acceptance and adoption of artificial intelligence into real-world clinical practice.

3 Peer Feedback and Reflection

3.1 Tutorial Discussion

Following presentation of the paper, a number of discussion and feedback points were raised by colleagues. For example, it was noted that the model made use of an architecture that had been pre-trained on a variety of non-domain-specific images, instead of using only skin disease images relevant to the task. While the practical benefits of using such a pre-trained network remain apparent, a discussion on the extent to which domain-specific networks have been utilised in the ensuing years since this paper's publication was not included in the presentation, hence its subsequent inclusion in Section 2.4 of this report.

One common point of confusion among colleagues was the motivation and usage of the taxonomical algorithm to generate partitioned disease classes. The specific method used to group certain diseases together was not explicitly detailed in the paper, yet the algorithm was presented as one of the main unique contributions of the study, which may have prevented its utility from being fully understood and

verified.

3.2 Personal Reflection

Overall, the synthesis of information detailed in this paper and analysis of the main findings presented a difficult but highly enjoyable task. I chose to present on the topic of deep neural networks applied to the task of image analysis in dermatology due to its direct relevance to my current research project. With this paper being one of the earliest works in the field, I was able to better appreciate the context from which subsequent algorithmic and dataset developments have evolved. I believe my critical analysis skills have improved as a result, which will be of great benefit for the remainder of my PhD.

The main challenge I found when compiling the presentation was determining how much background theory, in terms of both medicine and artificial intelligence, should be included, as well as choosing which results would be most relevant to the main message of the paper. To improve for next time, I will aim to keep my findings more concise and ensure that I present the results in an appropriate and logical sequence. Furthermore, given that I was unable to answer some of the more technical questions about the methodology, for example, regarding the disease partitioning algorithm, I will aim to spend more time understanding the specific methods employed, which may include reading the supplementary material in more detail or further references suggested by the authors. I will also take note of when the paper was published and give context on the work that has been done in the field since the paper's publication to include in the presentation, something which was subsequently added in the writing of this report.

4 Acknowledgements

This project was supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1].

References

[1] M. Abdar, M. Samami, S. D. Mahmoodabad, T. Doan, B. Mazoure, R. Hashemifesharaki, L. Liu, A. Khosravi, U. R. Acharya, V. Makarenkov, et al. Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning. *Computers in biology and medicine*, 135:104418, 2021.

[2] L. Ballerini, R. Fisher, B. Aldridge, and J. Rees. *A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-melanoma Skin Lesions*, volume 6, pages 63–86. 01 2013.

[3] M. Binder, H. Kittler, A. Seeber, A. Steiner, H. Pehamberger, and K. Wolff. Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network. *Melanoma research*, 8(3):261–266, 1998.

[4] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019.

[5] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern. 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), 2018.

[6] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

[7] M. Dildar, S. Akram, M. Irfan, H. U. Khan, M. Ramzan, A. R. Mahmood, S. A. Alsaiari, A. H. M. Saeed, M. O. Alraddadi, and M. H. Mahnashi. Skin cancer detection: a review using deep learning techniques. *International journal of environmental research and public health*, 18(10):5479, 2021.

[8] T. Emara, H. M. Afify, F. H. Ismail, and A. E. Hassanien. A modified inception-v4 for imbalanced skin cancer classification dataset. In *2019 14th International Conference on Computer Engineering and Systems (ICCES)*, pages 28–33, 2019.

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

[10] A. Esteva and E. Topol. Can skin cancer diagnosis be transformed by ai? *The Lancet*, 394(10211):1795, 2019.

[11] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern. 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), 2016.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[14] A. Hekler, J. S. Utikal, A. H. Enk, A. Hauschild, M. Weichenthal, R. C. Maron, C. Berking, S. Haferkamp, J. Klode, D. Schadendorf, et al. Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer*, 120:114–121, 2019.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[17] N. Kausar, A. Hameed, M. Sattar, R. Ashraf, A. S. Imran, M. Z. u. Abidin, and A. Ali. Multi-class skin cancer classification using ensemble of fine-tuned deep learning models. *Applied Sciences*, 11(22):10593, 2021.

[18] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder. Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3(3):159–165, 2002.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[20] A. Masood and A. Al-Jumaily. Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms. *International journal of biomedical imaging*, 2013:323268, 01 2013.

[21] M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, and H.-C. Kim. Skin cancer detection using deep learning—a review. *Diagnostics*, 13(11):1911, 2023.

[22] H. W. Rogers, M. A. Weinstock, S. R. Feldman, and B. M. Coldiron. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the US Population, 2012. *JAMA Dermatology*, 151(10):1081–1086, 10 2015.

[23] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34, 2021.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision, 2015.

[27] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.