Skin Cancer Classification via Machine Learning

Claire Rong

Received March 20, 2024
Accepted September 20, 2024
Electronic access October 15, 2024

As technology becomes more and more advanced, it allows many fields that traditionally require in-person interaction to turn virtual. For dermatology in particular, the rise of telemedicine is forcing dermatologists to consider implementing artificial intelligence into their clinics. The use of machine learning in the domain of skin cancer, particularly for skin cancer classification, has become popular. This research presents a comparative analysis of three different machine learning algorithms to evaluate their performance in skin cancer classification. We found that machine learning algorithms successfully classified about 80% of skin lesions, surpassing their human counterparts. This research also explores various areas that require improvement before AI can be realistically implemented in clinical settings.

Keywords: CNN, dermatology, image classification, machine learning, skin cancer.

Introduction

The COVID-19 pandemic has significantly limited access to healthcare providers, leading to a surge in the popularity of telemedicine, particularly in the field of image-based classification[1]. One area where this approach shows great promise is in skin cancer detection. Skin cancer is the most common cancer in the United States, with one in five Americans expected to develop it at some point in their lives[2]. Early detection is critical for improving survival rates. For example, the survival rate for melanoma, one of the deadliest forms of skin cancer, is nearly 99% with early detection, but it plummets to 23% if treatment is delayed[3].

Traditionally, dermatologists diagnose skin cancer through visual examination, a challenging process that requires careful consideration of numerous visual cues such as lesion size, shape, texture, and color. This complexity leaves room for human error[4]. Image-based skin cancer classification offers the potential not only to enhance diagnostic accuracy but also to provide a more convenient and flexible option for patients by enabling remote diagnosis. This could lead to earlier detection and treatment[5].

Given these advantages, ongoing research is exploring ways to make image-based classification a viable solution in the near future. In this study, we will evaluate various machine learning algorithms to determine which models achieve the highest accuracy in image classification. Our goal is to assess whether machine learning algorithms can outperform human dermatologists and to identify the necessary improvements required before these models can be implemented in clinical settings.

Background

Skin cancer is the most common cancer in the United States, affecting one in five Americans over their lifetime[2]. While it is treatable, successful recovery hinges on early detection and prompt treatment. Delays in treatment can lead to significantly higher mortality rates and increased risk of metastasis. Traditionally, skin cancer is diagnosed through visual examination or biopsy. However, these methods can be time-consuming and are heavily dependent on the skill and experience of the examiner, often resulting in inconsistent outcomes. This has spurred growing interest in machine learning algorithms, particularly deep learning techniques, to enhance diagnostic accuracy and efficiency.

Among these, convolutional neural networks (CNNs) have shown remarkable promise in image classification tasks, making them a popular focus in the field of skin cancer detection. This review examines recent literature on the application of deep learning, especially CNNs, in skin cancer classification.

Numerous studies have investigated the potential of CNNs for skin cancer classification. Naqvi et al. (2023) discuss various deep learning architectures, comparing their performance and computational cost[6]. Dildar et al. (2021) provide a comprehensive overview of skin cancer detection using deep learning techniques, offering a detailed analysis of the latest advancements and methodologies in the field[7]. Similarly, Brinker et al. (2018) discuss the application of CNNs for skin cancer classification, offering insights into potential future research directions[8].

Recent research has shown that deep learning methods have achieved high levels of success in skin cancer detection. For instance, Tschandl et al. (2018) developed a CNN model for classi-

fying skin lesions into benign or malignant categories, achieving an Area Under the Curve (AUC) score of 0.94[9]. Gajera et al. (2023) reported a CNN with an accuracy of 98.33% and an F1 score of 96%[10]. Hossain et al. (2023) utilized the max voting ensemble technique combined with advanced pre-trained deep learning models, achieving an AUC of 0.932. This study also emphasized the importance of diverse datasets and addressed challenges related to clinical utility and integration[11]. Other studies have demonstrated that specific CNN models, such as DenseNet-121 and Xception, achieve higher accuracy than others[12]. For example, Codella et al. (2018) compared the performance of Inception-V3, ResNet50, and DenseNet-121, finding DenseNet-121 to have the highest classification accuracy with an area under the receiver operating characteristic curve (AUC-ROC) of 0.91[13,14].

Despite the promising potential of CNNs in skin cancer classification, several challenges remain in the field. Li et al. discussed the difficulties in applying deep learning to skin disease diagnosis, including data limitations and interpretability issues, providing a balanced perspective on the obstacles that researchers and clinicians must overcome[15]. Naqvi et al. (2023) also highlighted the importance of diverse datasets and address challenges related to the lack of large-scale skin cancer datasets and skin color bias in existing datasets[6].

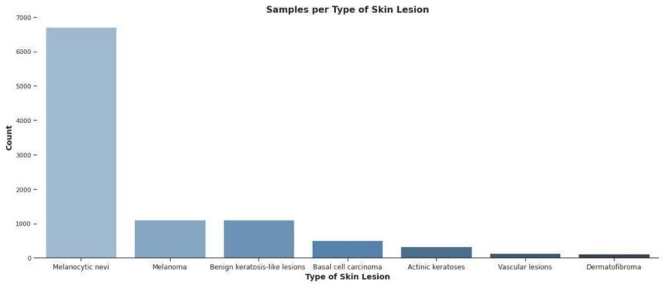

Fig. 1 Original Distribution of Skin Lesions in HAM10000

A significant issue in skin cancer classification through machine learning is the lack of diversity in existing datasets. Current datasets represent only a limited range of skin lesions—HAM10000, for instance, covers only seven different types of lesions—raising concerns about model performance on less common lesion types. Additionally, dataset biases, particularly the overrepresentation of certain lesion types as noted by Tschandl et al. (2018), can skew model performance[9]. As seen in Figure 1 and Table 1, in the HAM10000 dataset, melanocytic nevi accounts for approximately 70% of the lesions, while it represents only about 20% of actual skin lesions in clinical settings. This significant imbalance indicates a potential bias that could affect the model's generalizability and performance in real-world applications.

To mitigate these issues, CNN models should be trained on datasets with a more balanced distribution of lesion types and a more diverse portfolio of skin lesions. Addressing these challenges is crucial for advancing the reliability and clinical applicability of machine learning models in skin cancer detection.

Lesion Type	**Percentage of lesions in HAM10000**	**Percentage of lesions in real life**
Actinic Keratosis	3%	20%
Basal Cell Carcinoma	5%	5%
Benign Keratosis	11%	30%
Dermatofibroma	1.5%	10%
Melanocytic Nevi	67%	20%
Melanoma	11%	1%
Vascular Lesions	1.5%	15%

Table 1 Lesion distribution in HAM10000 vs real life

Methodology

The dataset used in this research is the HAM10000 ("Human Against Machine with 10000 training images"). The dataset is publicly available through the ISIC archive and consists of 10,015 dermatoscopic images collected from various ages, genders, and localizations of skin lesions. The dataset was collected over a period of 20 years from two different sites: The Department of Dermatology at the Medical University of Vienna, Austria and the skin cancer practice of Cliff Rosendahl in Queensland, Australia[9]. There are seven different skin disease types including Melanocytic nevi, Melanoma, Benign keratosis-like lesions, Basal cell carcinoma, Actinic keratoses, Vascular lesions, and Dermatofibroma, as shown in Table 1.

We selected three distinct CNN architectures—ResNet50, DenseNet121, and Xception—based on their proven effectiveness in image classification tasks.

ResNet50

ResNet50, a variant of the ResNet architecture, consists of 50 layers and is distinguished by its use of residual connections. Figure 2 shows a high-level diagram of the ResNet50 Architecture. These connections help mitigate the vanishing gradient problem by enabling gradients to flow directly through the network, bypassing certain layers. This design allows the network to be trained deeper and more effectively, leading to improved accuracy in image classification tasks[11].

DenseNet121

DenseNet121 is a variant of the DenseNet architecture and consists of 121 layers. It has alternating convolutional and pooling

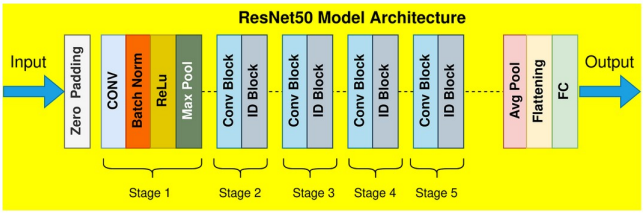

Fig. 2 ResNet50 Model Architecture (Adapted from S. Mukherjee, 2022 [16])

layers, followed by a global average pooling layer and a fully connected layer for classification. Figure 3 shows the diagram of the architecture. Its distinctive dense connectivity structure, where each layer is connected to every other layer in a feed-forward fashion, enables the model to learn complex features more efficiently with fewer parameters compared to other deep learning models. This architecture makes DenseNet121 particularly well-suited for image classification tasks [11].

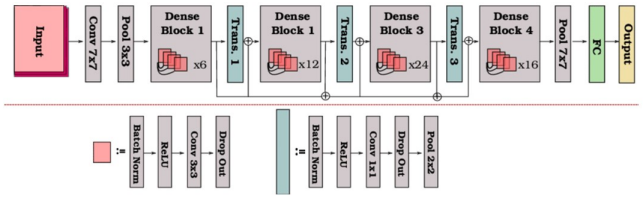

Fig. 3 DenseNet121 Model Architecture (Adapted from N. Radwan, 2019 [17])

Xception

Xception shares similarities with InceptionV3 and has demonstrated promising results in image classification tasks. Its unique architecture shown in Figure 4 is designed to learn highly discriminative features with fewer parameters than other neural networks, leading to improved efficiency and reduced computational resource requirements. This makes Xception an attractive choice for applications where both performance and computational efficiency are critical. [11].

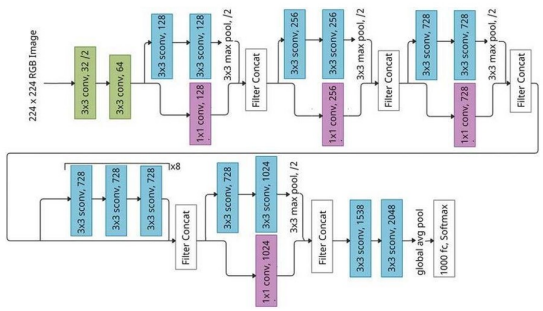

Fig. 4 Xception Model Architecture (Adapted from K. Srinivasan, 2021 [18])

The images were preprocessed by converting them to grayscale, randomly cropping and resizing them to 71 x 71 pixels, and applying various augmentations including random rotations, horizontal and vertical zooming, and flipping. These steps were implemented to enhance the robustness of the models and improve their ability to generalize across different variations in the dataset.

The HAM10000 dataset used in this study exhibits significant class imbalance, with some skin lesion categories being underrepresented compared to others. To address this imbalance and prevent the model from being biased towards the more prevalent classes, we incorporated class weights into our training process. Class weights were computed based on the inverse frequency of each class in the training dataset. This approach ensures that minority classes are assigned higher weights, thereby increasing their impact on the loss function during training. The class weights were calculated as follows:

1. The number of samples for each class in the training dataset was counted.

2. The frequency of each class was determined by dividing the number of samples of that class by the total number of samples in the dataset

3. The real-life frequencies of each skin lesion type were determined using epidemiological data

4. The class weights were calculated by dividing their real-life frequency by their dataset frequency

5. The class weights were normalized so that their sum equals the number of classes, ensuring a balanced influence on the training process.

The formula used for calculating class weights wi for each class i is shown below:

$$w_i = \frac{\text{real-life frequencies}_i}{\text{dataset frequencies}_i}$$

The calculated class weights were integrated into the model training process using the class weight parameter of the fit method in Keras. This parameter adjusts the loss function by multiplying the loss of each sample by its corresponding class weight, thus penalizing the model more for errors on minority classes. We used Adam as the optimizer and Sparse Categorical Cross Entropy as our loss function. The learning rate was initially set to 0.001 and was reduced by a factor of 5 if the validation loss did not improve after 5 epochs, with a minimum learning rate of 0.00001. The models each ran for 50 epochs with early stopping. We developed a separate model for each architecture and fine-tuned all three using the HAM10000 dataset.

To assess the performance of these models, we used metrics including accuracy, recall, precision, and F1 score, providing a comprehensive evaluation of their success.

Results

Metric	DenseNet121	ResNet50	Xception
Accuracy	0.76	0.78	0.80
Precision	0.75	0.77	0.80
Recall	0.76	0.78	0.81
F1 Score	0.75	0.77	0.80

Table 2 Summary of Model Performance Results

The performance metrics for the three CNN models are displayed in Table 2 above. Among our three models, Xception achieved the highest score in all four of our metrics. The accuracy rate of 80% surpasses that of the majority of human dermatologists (65-85%)[19].

To determine if the variations in performance between the three models are statistically significant or simply due to chance, we conducted an ANOVA test followed by Tukey's HSD post-hoc test. The ANOVA test revealed a statistically significant difference in performance between the models ($F(2, N) = 18.10$, $p < 0.001$). Tukey's HSD test further indicated that Xception significantly outperforms both DenseNet121 ($p < 0.001$) and ResNet50 ($p = 0.007$). However, there was no significant difference between DenseNet121 and ResNet50 ($p = 0.056$). Based on these results, the Xception model is the best-performing model among the three, demonstrating statistically superior performance.

The most critical aspect of skin cancer classification is accurately identifying cancerous and high-risk cases. Misclassification in this context can have severe consequences, as it may cause patients to delay seeking the necessary treatment, allowing the cancer to advance to stages where treatment options are more limited and less effective. To evaluate how well our models, perform in this regard, we can examine the confusion matrices, which provide a detailed breakdown of true positives, false positives, true negatives, and false negatives. This analysis helps us understand the model's ability to correctly identify high-risk cases and its potential for misclassification.

A closer examination of the DenseNet121 confusion matrix (Figure 5) reveals that the model frequently misclassifies actinic keratoses, benign keratosis-like lesions, and melanocytic nevi as one another. While such misclassifications are undesirable, the impact is somewhat mitigated by the fact that these are all benign skin lesions. However, a significant point of concern is the 34% misclassification rate of basal cell carcinoma as a benign lesion. This is problematic because basal cell carcinoma is the most common type of skin cancer, with an estimated 2

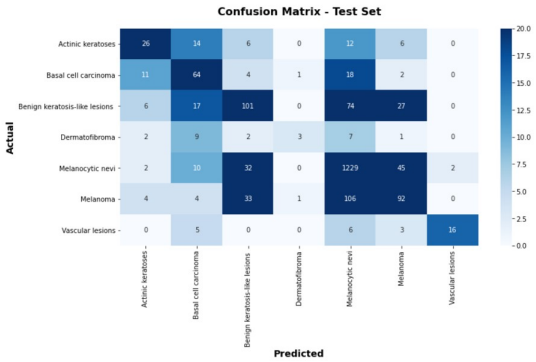

Fig. 5 Confusion Matrix for DenseNet121

to 4 million cases in the United States each year. Even more troubling is the 60% misclassification rate of melanoma as a benign lesion. Given that melanoma is the most dangerous type of skin cancer, this high misclassification rate poses a serious risk[20].

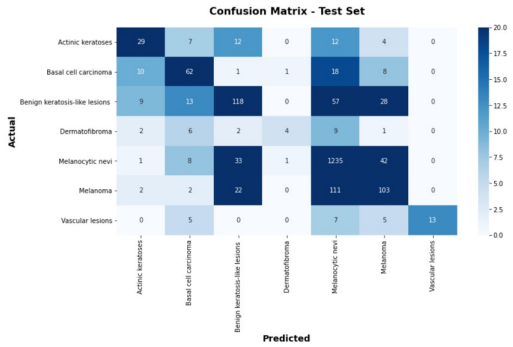

Fig. 6 Confusion Matrix for ResNet50

Similar to the DenseNet121 model, the confusion between melanoma and benign lesion types (57%) is a recurring issue for the ResNet50 model (Figure 6). While the model performs relatively well on basal cell carcinoma, there is still a significant number of misclassifications with other classes (30%). The model also struggles to differentiate between Actinic Keratoses and Benign Keratosis-like Lesions, likely due to their similar appearance.

Compared to the DenseNet121 and ResNet50 models, Xception has a smaller number of misclassifications (Figure 7). As shown in Figure 8, the Xception model has shown improvement in the misclassification rates of melanoma (43%) and basal cell carcinoma (17%) compared to the previous confusion matrix, which is a positive development.

According to a Cleveland Clinic study, compared to patients who were treated within 30 days, patients with stage I melanoma who delayed treatment were 5 percent more likely to die when treated between 30 and 59 days (95% CI 1.01-1.1; P = 0.029), 16 percent more likely to die when treated between 60 and 89

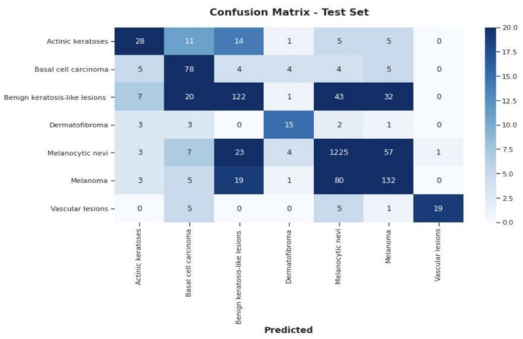

Fig. 7 Confusion Matrix for Xception

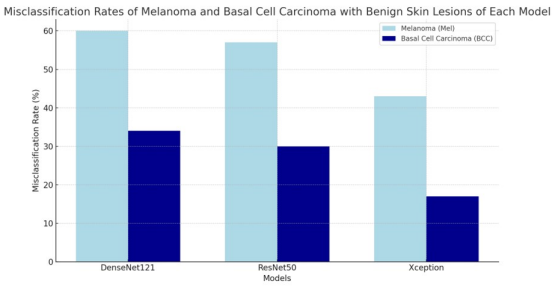

Fig. 8 Comparison of Misclassification Rates for All Models

days (95% CI 1.07-1.25; $P > 0.001$), 29 percent more likely to die when treated between 91 and 120 days (95% CI 1.12-1.48; $P > 0.001$), and 41 percent more likely to die when treated after 120 days (95% CI 1.21-1.65; $P > 0.001$)[20,21].

Xception's melanoma misclassification rate of 43% suggests that in a hypothetical cohort of 1000 patients with melanoma, 430 patients could be incorrectly diagnosed with benign lesions. To understand the potential impact of these misclassifications, consider the survival rates for melanoma at different stages. The 5-year survival rate for early-stage melanoma is approximately 99% but drops drastically to 27% for stage IV melanoma. If 430 patients are misclassified and consequently diagnosed only at a later stage, the survival rate would dramatically decrease. Assuming these patients are diagnosed at stage IV, we could anticipate a mortality rate of 73%. Therefore, out of the 430 misclassified patients, approximately 314 patients could face fatal outcomes due to delayed treatment. This starkly highlights the critical importance of accurate and timely melanoma diagnosis in clinical practice.

The patterns of misclassification were consistent across all three models, with two primary observations: confusion between similar-looking lesions and low F1 scores for minority classes. To address the inherent dataset bias in HAM10000, we applied class weights in our models. While this approach helped in achieving higher accuracy and F1 scores for the CNNs overall, it was insufficient in fully mitigating the issue. Future research could address these challenges through further tuning of class weights, incorporating more diverse and representative training data, or employing advanced techniques such as enhanced data augmentation or model ensembling.

Discussion

In this research, we used three different CNN architectures with the HAM10000 dataset to evaluate model performances in skin cancer classification. All three models achieved relatively high accuracy rates (above 75%), with Xception being the highest with an accuracy rate of 80%.

While CNN models have surpassed human dermatologists in accuracy, significant challenges remain before these models can be reliably implemented in clinical practice.

Current models are limited by dataset issues, including biases and insufficient skin color diversity[19]. To address these challenges, further research is needed to assess CNN performance across diverse skin tones and less common lesions, which are often underrepresented in existing datasets. For example, integrating datasets like HAM10000, which predominantly consists of images from Caucasian patients, with the more racially diverse ISIC archive could help reduce skin color bias. Gathering data from more clinics globally with a standardized data collection procedure and format would ensure a more representative range of skin colors and lesions in training datasets. To mitigate the bias in the HAM10000 dataset, we applied class weighting. Other approaches like under-sampling, over-sampling, and synthetic data generation also warrant exploration to identify the most effective method for handling class imbalance.

In our research, we utilized basic data augmentation techniques, such as rotation, flipping, and grayscaling. Implementing more advanced techniques could further enhance model robustness. It would be valuable to conduct experiments that assess the impact of different data augmentation strategies on model performance.

Ensemble learning, particularly the max voting method, has shown promise in skin cancer classification. Future research should explore whether combining models could enhance performance. A Max Voting Ensemble, weighted by each model's strengths, might optimize performance and reduce misclassifications, particularly false negatives.

Incorporating patient information—such as age, gender, lesion localization, family history, changes in lesions, sun exposure, and symptoms—could more closely mirror the real-life diagnostic process used by dermatologists, potentially improving model performance.

Furthermore, the integration of data segmentation with classification, as demonstrated in the 2018 ISIC Challenge involving HAM10000, presents a promising research avenue. Multi-frame learning networks that simultaneously segment and classify skin lesions could significantly improve diagnostic accuracy.

Developing models with high accuracy is only one aspect of the problem. To deploy these models into the clinical process, there are many other aspects to consider. Before integrating these models into clinical practice, it is essential to conduct rigorous validation and extensive clinical trials across diverse populations. Obtaining regulatory approval from bodies like the FDA or EMA, alongside adherence to ethical standards, is crucial. Ensuring robustness to adversarial attacks involves implementing defense mechanisms and real-time monitoring. Integrating these models seamlessly with clinical workflows and EHR systems while providing decision support will enhance their usability. Developing interpretable models and maintaining transparency in the development process will build trust among clinicians. Continuous learning and feedback mechanisms, along with regular updates, will ensure the model evolves and adapts. Comprehensive training for clinicians on AI tools, fostering collaboration between AI developers and clinicians, and establishing clear accountability guidelines are essential. Long-term monitoring and evaluation, including post-deployment surveillance and impact studies, will ensure ongoing reliability and effectiveness.

Conclusion

From this study, we learn that CNN models used for the purpose of skin cancer classification can outperform the majority of their human counterparts. Of the three models evaluated in this study, Xception had the best overall performance. However, there is still much work to be done before they can be implemented in dermatology clinics. Other than using different technologies to improve model accuracy and reduce false negatives, many other areas need to be covered such as racial fairness, security, clinical workflow incorporation, and continuous improvement. Implementing CNNs in dermatology clinics holds significant potential but requires a cautious and well-regulated approach. By addressing these challenges and uncertainties, we can safely and effectively leverage AI to enhance patient care and support dermatologists in their practice, making a future for machine learning in dermatology a feasible reality rather than a mere possibility.

Acknowledgement

Thank you for the guidance of Nowell Closser from Harvard University in the creation of this research paper.

References

1 P. Webster, *The Lancet*, **395**, 1180–1181.

2 N. Codella, V. Rotemberg, P. Tschandl, M. 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*, `https://arxiv.org/pdf/1902.03368.pdf`, arXiv.

3 A. A. Dermatology, *Skin cancer*, `https://www.aad.org/media/stats-skin-cancer`, Retrieved July 22, 2024, from.

4 C. Clinic, *Skin lesions*, `https://my.clevelandclinic.org/health/diseases/24296-skin-lesions`, Retrieved July 22, 2024, from.

5 M. Giavina-Bianchi, A. Santos and E. Cordioli, *eClinicalMedicine*, **29**, 100641.

6 M. Naqvi, S. Gilani, T. Syed, O. Marques and H. Kim, *Diagnostics*, **13**, year.

7 M. Dildar, S. Akram, M. Irfan, H. Khan, M. Ramzan, A. Mahmood, S. Alsaiari, A. Saeed, M. Alraddadi and M. Mahnashi, *International Journal of Environmental Research and Public Health*, **18**, 5479.

8 T. Brinker, A. Hekler, J. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. Enk and C. Kalle, *Journal of Medical Internet Research*, **20**, year.

9 P. Tschandl, C. Rosendahl and H. Kittler, *Scientific Data*, **5**, 180161.

10 H. Gajera, D. Nayak and M. Zaveri, *Biomedical Signal Processing and Control*, **79**, 104186.

11 M. Hossain, M. Hossain, M. Arefin, F. Akhtar and J. Blake, *Diagnostics*, **14**, 89.

12 A. Ali and T. Deserno, *Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment*, vol. 8318, p. 421–431.

13 N. Codella, D. Gutman, M. Celebi, B. Helba, M. Marchetti, S. Dusza, A. Kalloo, K. Liopyris, N. Mishra and H. Kittler, Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018, p. 168–172.

14 D. Gutman, N. 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*, `https://doi.org/10.48550/arXiv.1605.01397`, arXiv.

15 H. Li, Y. Pan, J. Zhao and L. Zhang, *Neurocomputing*, **464**, 364–393.

16 S. Mukherjee, *Towards Data Science*.

17 N. Radwan, *Leveraging Sparse and Dense Features for Reliable State Estimation in Urban Environments*, `https://doi.org/10.6094/UNIFR/149856`.

18 K. Srinivasan, L. Garg, D. Datta, A. Alaboudi, N. Z.Jhanjhi, R. Agarwal and A. Thomas, *Computers, Materials Continua*, **68**, 1615–1630.

19 D. Wen, S. Khan, A. Xu, H. Ibrahim, L. Smith, J. Caballero, L. Zepeda, C. Perez, A. Denniston, X. Liu and R. Matin, *The Lancet*, **4**, 6.

20 C. Erickson, M. Driscoll and M. Faries, *" Journal of the American Academy of Dermatology*, **78**, 1065–1071.

21 C. Clinic, *Consult QD*.