Skin Cancer Detection with Deep Learning

Antonio Cantillo Molina
antoocantillo@correo.ugr.es

Jaime Corzo Galdó
jaimecrz04@correo.ugr.es

Leandro Jorge Fernández Vega
leandrofdez@correo.ugr.es

Mario Líndez Martínez
mariolindez@correo.ugr.es

Abstract

Early detection of Melanoma is a critical factor in reducing skin cancer mortality rates. While Deep Learning has shown promise in automated diagnosis, achieving a balance between high sensitivity and specificity remains a challenge due to class imbalance and the visual similarity between malignant and benign lesions. In this work, we present a comprehensive comparative study between Convolutional Neural Network models (EfficientNetB4, Inception-ResNet-v2) and Transformer-based architectures (ViT, ViT+DINO) using the ISIC 2017 dataset. We employ a rigorous training protocol utilizing the One Cycle Policy and data augmentation. Our experiments reveal a distinct trade-off: while ViT-based models demonstrate superior global performance (Accuracy: 76.67%, ROC-AUC: 0.8268) and robustness against class imbalance, traditional CNNs like Inception-ResNet-v2 achieve higher raw sensitivity (Recall: 0.70) at the cost of increased false positive rates. Furthermore, we analyze the impact of self-supervised learning (DINO) and decision thresholds, providing insights into the optimal operating points for clinical decision support.

1. Introduction

Skin cancer is a major public health concern worldwide, with Melanoma being the most dangerous form. Early diagnosis is crucial for patient survival; when detected in its initial stages, Melanoma is highly treatable. However, late-stage diagnosis significantly reduces the chances of effective treatment. Therefore, developing reliable and accessible tools for early detection is a priority in the medical community.

Traditionally, the diagnosis relies on visual inspection by dermatologists, often assisted by dermoscopy. While effective, this process is subjective and depends heavily on the clinician's training and experience. Distinguishing Melanoma from benign skin lesions is visually challenging because they often share similar colors, shapes, and textures. This visual similarity can lead to diagnostic errors or unnecessary biopsies.

In recent years, Deep Learning has revolutionized the field of Computer Vision. Convolutional Neural Networks (CNNs) have shown remarkable success in medical image analysis, becoming the standard approach for automated skin lesion classification. These models can learn complex representations from images, potentially matching or exceeding the accuracy of human experts.

Despite the success of Deep Learning, building a robust detection system remains difficult due to issues like image quality variations and the vast diversity of skin lesion appearances. Besides, the overrepresentation of lighter skin tones in the literature limits the generalizability of current DL models.

In this paper, we address these challenges by proposing a Deep Learning framework for the accurate detection of Melanoma. Our goal is to provide an effective tool that can assist dermatologists in making better diagnostic decisions.

Consequently, a critical open question remains: To what extent do emerging Vision Transformers (ViTs) offer a distinct advantage over CNN structures in balancing the sensitivity-specificity trade-off required for clinical melanoma screening?

In this paper, we address this question by proposing a comprehensive comparative framework. We evaluate whether the global context modeling of Transformers translates into more robust decision-making compared to the local feature extraction of CNNs. Our goal is to provide an effective tool that can assist dermatologists in making better diagnostic decisions.

Our main contributions can be summarized as follows:

- We propose a range of Deep Learning models optimized for skin lesion classification.
- We offer a comparison between these models and decide which one would be suitable for real-world scenario.
- We offer an insight to explainable AI (XAI) applied to

this topic and draw conclusions.

2. Background

Skin cancers are among the most common and notorious diseases nowadays. This scientific paper focuses on the classification of three types of skin lesions that present significant visual similarities. Because of these similarities, they pose a major challenge for Deep Learning classification tasks intended for the introduction of these types of models into medical fields.

To begin with, Dermoscopy is a specialized imaging technique that allows for the examination of skin lesions with magnification and reduced surface reflection. By using a dermatoscope, clinicians can observe deep morphological structures such as the pigment network, dots, globules, and blue-white veils.

The dataset used for this classification problem includes high-resolution dermoscopic images, which provide the necessary visual features for deep learning models to identify patterns that are often imperceptible to the naked eye.

The three main skin cancer lesions we focus on are:

- **Melanoma:** Recognized by the World Health Organization (WHO) as the most aggressive and lethal form of skin cancer. It originates in the melanocytes, which are the cells responsible for producing melanin. This lesion represents a high risk of metastasis if not caught in its early stages [31].

- **Seborrheic Keratosis:** According to the Skin Cancer Foundation, Seborrheic Keratosis is one of the most common non-cancerous (benign) skin growths in older adults. It is characterized by a waxy or scaly skin texture and a color range from light tan to black [19]. Why is this type of non-cancerous lesion considered for a skin cancer classification problem? Its importance lies in its morphological similarity to Melanoma, which frequently leads to diagnostic confusion.

- **Nevus:** Commonly known as a mole, a Nevus is a benign proliferation of melanocytes. The International Agency for Research on Cancer (IARC) classifies them as stable neoplasms that typically appear during childhood and adolescence. While most nevi remain benign throughout a patient's life, the distinction between a common Nevus and an early-stage Melanoma becomes critical for prevention [31].

3. Related Works

The field of skin lesion analysis has evolved significantly, transitioning from traditional Computer-Aided Diagnosis (CAD) systems relying on handcrafted feature extraction to robust, end-to-end Deep Learning (DL) pipelines. Currently, Convolutional Neural Networks (CNNs) represent the standard widely adopted in the literature due to their translation invariance and local processing capabilities. However, the landscape is shifting, with Transformer-based architectures rapidly emerging as the new state-of-the-art by offering superior modeling of global contexts.

3.1. CNN-Based Approaches

For years, CNNs have been the dominant paradigm for Melanoma detection due to their ability to learn hierarchical representations from dermoscopic images [1, 14]. Extensive experimental studies, such as the one conducted by Perez *et al*. [16], have benchmarked various architectures, specifically highlighting that while deeper networks like DenseNet201 and ResNet152 require higher computational resources, they offer superior sensitivity in detecting subtle pigment network irregularities compared to shallower architectures like VGG19.

To further improve performance and bridge the semantic gap, researchers have focused on hybrid and fusion strategies. Almaraz-Damian *et al*. [5] demonstrated that fusing deep features with handcrafted ones—specifically those derived from the clinical ABCD rule (Asymmetry, Border, Color, Diameter)—using mutual information measures can significantly boost classification accuracy. Similarly, techniques integrating segmentation with classification have proven effective in isolating the lesion of interest. For instance, the ResBCU-Net proposed by Badshah and Ahmad [6] integrates Bi-directional ConvLSTMs within a U-Net architecture to capture spatiotemporal-like correlations in the spatial domain, thereby refining the boundary delineation of lesions. In a parallel development, Zhang and Chaudhary [32] recently established a new benchmark with a hybrid framework that effectively combines diverse feature extractors in a multi-stream fashion, validating that composite architectures can capture multi-scale representations better than single-stream models.

Addressing the specific challenges of dermoscopy, such as artifacts and class imbalance, has also been a priority. Suiçmez *et al*. [22] emphasized the critical role of preprocessing, utilizing Wavelet Transforms to decompose images and filter high-frequency noise components caused by hair artifacts, which significantly enhances the signal-to-noise ratio before feature extraction. Regarding data imbalance, Adepu *et al*. [2] proposed a Knowledge Distillation framework where a larger "Teacher" network guides a smaller "Student" network. This method effectively transfers inductive biases, improving the Student's ability to generalize on minority classes like Melanoma without being overwhelmed by the majority class (nevi). Furthermore, Spo-

laor *et al*. [21] and Varma *et al*. [28] validated the efficacy of transfer learning, demonstrating that fine-tuning specific deep layers rather than full retraining is crucial when working with limited clinical datasets to prevent overfitting.

Efficiency remains a critical research vector for deploying these models in clinical settings. Aldhyani *et al*. [4] proposed lightweight dynamic kernel networks, achieving high diagnostic accuracy with reduced parameter counts. This aligns with the industry's move towards efficient architectures, similar to the principles of EfficientNet, balancing accuracy with the low computational cost required for handheld diagnostic devices.

3.2. The Shift to Transformers and Attention Mechanisms

Despite the success of CNNs, they are inherently limited by their local receptive fields, which may fail to capture the whole structure of large or disseminated lesions. Recent literature suggests a paradigm shift towards Vision Transformers (ViT) and attention mechanisms, which utilize self-attention to capture long-range global dependencies essential for distinguishing subtle lesion patterns.

Venugopal [30] critically analyzed the potential of Vision Transformers in oncology, concluding that while ViTs lack the inherent inductive bias of CNNs (translation invariance), they offer superior feature abstraction capabilities when trained on large-scale datasets. The study suggests that ViTs are particularly adept at ignoring background noise by focusing attention dynamically on relevant pathological regions. This capability is supported by Heroza *et al*. [12], who introduced a self-attention fusion approach. Similarly, Omeroglu *et al*. [15] utilized soft attention-based multi-modal frameworks, allowing the model to weigh different input modalities dynamically to enhance multi-label classification performance.

Most recently, comparative studies by Aksoy *et al*. [3] have evaluated advanced models like Swin Transformers and ConvNeXt against traditional baselines. Their findings indicate that Swin Transformers, which employ a shifted windowing mechanism to calculate self-attention locally while maintaining cross-window connections, solve the computational bottleneck of standard ViTs on high-resolution medical images. Consequently, these attention-based models often outperform standard CNNs in complex diagnostic tasks where context is key. These findings motivate the inclusion of Transformer-based architectures in modern benchmarks alongside robust CNNs.

In this work, we align with these recent trends by evaluating a diverse set of architectures—ranging from established CNNs (EfficientNetB4, Inception-ResNet-v2) to modern Transformers (ViT)—to provide a comprehensive perspective on the current landscape of Melanoma detection.

4. Methods

We will cover two tasks in this study:

- Multiclass classification performance.
- Melanoma v.s. non-Melanoma classification performance.

Medically speaking, classifying skin cancer versus non-skin cancer is the most important aspect to be taken into consideration.

In this section, we describe the Deep Learning architectures selected for our study. These models offer innovative and balanced performance according to the recent literature. To the best of our knowledge, there are few or no approaches that apply this specific combination of architectures to the problem of Melanoma detection. We evaluate each model's performance to illustrate their adaptability to skin lesion classification.

- **ViT Transformer**: The Vision Transformer (ViT) [29] [10] represents a paradigm shift from traditional Convolutional Neural Networks (CNNs). Instead of processing pixels using local receptive fields, ViT treats the input image as a sequence of fixed-size patches, similar to how tokens are processed in Natural Language Processing. This mechanism allows the model to capture long-range dependencies and global context across the entire image from the very first layer, which is crucial for identifying structural patterns in skin lesions.
- **ViT Transformer + DINO**: DINO (Self-distillation with no labels) [7] is a novel self-supervised learning approach applied to Vision Transformers [29] [10]. Unlike standard supervised training, DINO learns visual representations by distilling knowledge from a "teacher" network to a "student" network without relying on explicit class labels. This architecture is particularly relevant for medical imaging, as it can leverage unlabelled data to learn robust features before being fine-tuned for the specific task of Melanoma detection.
- **Inception-ResNet-v2**: The Inception-ResNet-v2 [11] [24] [25] [23] architecture combines two powerful concepts in Deep Learning: the Inception modules and Residual connections. The Inception blocks allow the network to perform multi-scale feature extraction by running convolutions of different sizes in parallel, while the residual connections accelerate training and allow the network to be significantly deeper. This hybrid approach enables the model to capture both fine-grained textures and high-level abstract features required for distinguishing malignant lesions.

- **EfficientNetB4**: EfficientNet [27] [26] [18] [13] [17] introduces a compound scaling method that uniformly scales the network's width, depth, and resolution. Rather than arbitrarily increasing one dimension, this family of models optimizes all three to achieve maximum accuracy with minimal computational cost. We specifically employ the EfficientNetB4 variant, which offers a strong balance between model size and performance, making it highly effective for handling the high variability found in dermoscopic images.

5. Experiments

5.1. Dataset

The classification task is performed using a subset of the ISIC 2017 Challenge dataset, originally introduced by Codella et al. (2017) [9] and later expanded in the works summarized by Codella et al. [8]. These data were sourced from the curated repository hosted on Kaggle (https://www.kaggle.com/datasets/wanderdust/skin-lesion-analysis-toward-Melanoma-detection). This dataset includes high-resolution dermoscopic images, which provide the necessary visual features for deep learning models to identify patterns that are often imperceptible to the naked eye.

The dataset is already organized in three folders:

- **Training**: 374 Melanoma images, 1372 Nevus images and 254 Seborrheic Keratosis images.
- **Validation**: 30 Melanoma images, 78 Nevus images and 42 Seborrheic Keratosis images.
- **Testing**: 117 Melanoma images, 393 Nevus images and 90 Seborrheic Keratosis images.

5.2. Experimental Protocol

To ensure the reproducibility and robustness of our results, all experiments were conducted using the *PyTorch* framework. The training and evaluation pipeline follows a rigorous standardization process.

Data Preprocessing. Input images are resized to a fixed resolution of 224×224 pixels for ViT models , and 384×384 pixels for the others, using bilinear interpolation. We apply distinct normalization to the input data using the mean and standard deviation statistics from the *ImageNet* dataset. This step is crucial to facilitate model convergence and transfer learning stability.

Regularization and Optimization. To mitigate the risk of overfitting, a common challenge in medical imaging due to limited data, we employ a comprehensive *data augmentation* strategy. This includes random geometric transformations (such as horizontal flipping and rotation up to $10°$) and photometric distortions. Regarding the training dynamics, we adopt the *One Cycle Policy* proposed by Smith [20]. This scheduling technique dynamically modulates the learning rate during the training cycle, increasing it during a warm-up phase to a maximum value before annealing it, while inversely adjusting the momentum. This approach promotes *super-convergence*, allowing the optimizer to traverse saddle points effectively and settle into flatter, more robust minima. Additionally, we utilize *weight decay* to penalize large weights and further improve generalization.

Evaluation Logic. The models are evaluated on a strict hold-out test set comprising 600 images, ensuring no data leakage from the training phase.

5.3. Evaluation Metrics

5.3.1 Multiclass Metrics

We report:

- Accuracy on the test set.
- Per-class Precision / Recall / F1-score, plus Macro average (uniform across classes) and Weighted average (weighted by support).
- Two imbalance-aware global metrics computed from the full confusion matrix: MCC (Matthews Correlation Coefficient) and Cohen's Kappa score.

5.3.2 Binary Melanoma vs Rest Metrics

To align with clinical decision-making, we also evaluate Melanoma vs non-Melanoma, using the model's Melanoma probability as a score. The test prevalence of Melanoma is 19.5% (117/600). We report:

- ROC curve and ROC-AUC score.
- Precision–Recall curve and Average Precision (AP).
- Operating points (threshold analysis):
  - True Positive Rate. $TPR = \frac{TP}{TP+FN}$
  - True Negative Rate. $TNR = \frac{TN}{TN+FP}$
  - Positive Predictive Value. $PPV = \frac{TP}{TP+FP}$
  - Negative Predictive Value. $NPV = \frac{TN}{TN+FN}$

Thresholds are selected to satisfy either a minimum TPR (detect at least X% Melanomas) or a minimum TNR (reject at least X% non-Melanomas), choosing among valid thresholds the one that optimizes the complementary objective (maximize TNR when fixing TPR, and maximize TPR when fixing TNR).

5.4. Multiclass Results and Discussion

Model	Acc	Macro-F1	W-F1	Mel. P	Mel. R	MCC
ViT	0.7667	0.6885	0.7640	0.6277	0.5043	0.5420
ViT+DINO	0.7667	0.6815	0.7606	0.6591	0.4957	0.5280
EfficientNetB4	0.6667	0.6100	0.6845	0.4823	0.5812	0.4491
Inception-ResNet-v2	0.6167	0.5871	0.6381	0.4385	0.7009	0.4440

Table 1. Multiclass test results (600 images). Mel.=Melanoma.

The multiclass evaluation shows that both ViT-based models achieve the best overall performance (Acc = 0.7667) and the strongest imbalance-aware global agreement (MCC = 0.542 / 0.528 and Kappa = 0.541 / 0.525 for ViT and ViT+DINO, respectively). Their strongest class is Nevus (F1 ≈ 0.85), which is consistent with it being the majority category.

However, the clinically critical class Melanoma remains challenging for all models. The two ViT models obtain Melanoma recall ≈ 0.50 (TP=59 and TP=58 out of 117), meaning roughly half of Melanomas are missed under an *argmax* decision rule. The transfer-learning CNNs shift toward higher Melanoma sensitivity: EfficientNetB4 increases Melanoma recall to 0.581 (TP=68), and Inception-ResNet-v2 reaches the highest recall (0.701, TP=82), but at the cost of clearly lower overall accuracy and macro-F1, indicating a stronger redistribution of errors across classes.

Overall, the results reveal a trade-off: ViTs provide the most consistent global multiclass behavior (best accuracy and robust agreement metrics), while the CNN transfer-learning models increase Melanoma recall but degrade global performance.

5.5. Binary Results

5.5.1 ROC-AUC and AP

Model	ROC-AUC	AP
ViT	0.8268	0.6039
ViT+DINO	0.8239	0.6130
Inception-ResNet-v2	0.7976	0.5285
EfficientNetB4	0.7911	0.5521

Table 2. Binary Melanoma-vs-rest results (test prevalence 19.5%).

The ViT models lead in discrimination under threshold-independent metrics (highest ROC-AUC, and highest AP values close to each other), indicating better separation between Melanoma and non-Melanoma across thresholds.

5.5.2 Operating Points

Fixing high sensitivity (TPR constraint). When enforcing a very high sensitivity requirement (TPR ≥ 0.95), all models reach the same Melanoma detection level ($TP = 112$, $FN = 5$). Therefore, the comparison is mainly driven by the number of false positives. In this extreme regime, Inception-ResNet-v2 achieves the best false-positive control ($TNR = 0.329$, $FP = 324$, $PPV = 0.257$), while precision remains low for all approaches (approximately 0.24–0.26), reflecting the high cost in false alarms required to guarantee near-complete Melanoma detection.

Relaxing the constraint to TPR ≥ 0.90 reduces false positives and improves precision. Among the evaluated operating points, ViT+DINO provides the most balanced trade-off, reaching $TPR = 0.906$ ($FN = 11$) with $TNR = 0.522$ ($FP = 231$) and $PPV = 0.315$.

Fixing false-positive control (TNR constraint). When constraining the specificity to TNR ≥ 0.90, the Vision Transformer models preserve Melanoma detection better than the transfer-learning CNNs. In particular, ViT achieves $TPR = 0.581$ ($TP = 68$, $FN = 49$) with $PPV = 0.607$ and $FP = 44$, while ViT+DINO attains a very similar sensitivity ($TPR = 0.581$) with $PPV = 0.591$ and $FP = 47$.

With a stricter constraint of TNR ≥ 0.92, the same pattern holds: ViT-based models remain the strongest options, with $TPR = 0.538$ (ViT) and $TPR = 0.521$ (ViT+DINO), maintaining PPV around 0.62–0.63 and low false-positive counts ($FP = 38$ and $FP = 36$, respectively).

Under even stricter specificity requirements (TNR $\geq$ 0.95–0.97), all methods become increasingly conservative, leading to a sharp drop in sensitivity. In this high-specificity regime, ViT+DINO offers the best compromise at TNR ≥ 0.95, obtaining $TPR = 0.393$ ($TP = 46$, $FN = 71$) with $PPV = 0.667$ and $FP = 23$.

These results confirm that the decision threshold strongly determines clinical behavior: enforcing extremely high sensitivity yields many false alarms, whereas enforcing extremely high specificity increases missed Melanomas. In moderate operating regimes (e.g., $TPR \approx 0.90$–0.80 or $TNR \approx 0.90$–0.92), the Vision Transformer models provide the most consistent overall trade-off.

5.6. False Negative Analysis

To better understand the limitations of the evaluated classifiers, we analyze Melanoma *false negatives* (Melanoma samples predicted as non-Melanoma) under the standard multiclass *argmax* decision rule. The number of Melanoma false negatives for each model is:

- **ViT**: 58 FNs
- **ViT+DINO**: 59 FNs
- **EfficientNetB4**: 49 FNs

- **Inception-ResNet-v2**: 35 FNs

A consistent pattern emerges across models. The ViT-based approaches tend to miss Melanomas primarily by confusing them with Nevus, whereas the transfer-learning CNNs (EfficientNetB4 and Inception-ResNet-v2) more frequently misclassify Melanomas as Seborrheic Keratosis. Importantly, a subset of false negatives occur with high confidence and very low predicted Melanoma probability, indicating that these errors are not necessarily borderline cases. Qualitative inspection suggests that recurrent factors among missed lesions include the presence of acquisition artifacts (e.g., hair occlusions, rulers or marks) as well as small, low-contrast, or visually ambiguous lesions. These observations highlight the need for careful operating-point selection and support the use of these models as decision-support tools rather than fully autonomous diagnostic systems.

6. Conclusions

In this study, we addressed the challenge of automated Melanoma detection by evaluating the transition from traditional CNN architectures to modern Vision Transformers. Our comparative analysis yields several critical insights for the deployment of Deep Learning in dermatological settings.

Technically, we observed that Vision Transformers (ViT and ViT+DINO) offer the most robust overall performance, achieving the highest accuracy and ROC-AUC scores. They demonstrate a superior ability to model global context, resulting in better separation between classes and fewer false positives compared to CNNs. However, CNNs (specifically Inception-ResNet-v2) exhibited higher intrinsic sensitivity for Melanoma detection. This suggests that while Transformers are better general classifiers, CNNs may currently be more naturally inclined towards "screening" behavior where missing a positive case is penalized more heavily than a false alarm.

From a clinical perspective, our threshold analysis confirms that no single model is perfect for all scenarios. For a screening tool, operating points must be adjusted to maximize sensitivity (TPR), where Inception-ResNet-v2 excels despite lower precision. Conversely, for a diagnostic support tool intended to reduce unnecessary biopsies, the ViT-based models offer a better trade-off between sensitivity and specificity (TNR).

Furthermore, the deployment of such systems in clinical practice relies heavily on trust and interpretability. We emphasize that high quantitative metrics alone are insufficient; the integration of Explainable AI (XAI) techniques is essential to allow clinicians to verify that the model focuses on medically relevant features rather than confounding artifacts. Ultimately, these models must always be utilized under the supervision of a qualified medical professional, serving as an adjunct to human expertise rather than a replacement, ensuring that the final diagnostic responsibility remains with the dermatologist.

Despite these promising results, limitations remain. The overrepresentation of lighter skin tones in the literature limits the generalizability of current DL models to diverse populations. Future work must prioritize dataset diversification and demographic transparency to avoid algorithmic bias. Additionally, ensuring reproducibility and standardization in evaluating model performance across different datasets is critical for advancing the field. Finally, the exploration of ensemble methods combining the sensitivity of CNNs with the robustness of Transformers represents a promising avenue for future research.

References

[1] A.A. Adegun and S. Viriri. Deep learning-based system for automatic melanoma detection. *IEEE Access*, 8:7160–7172, 2019. 2

[2] A.K. Adepu, S. Sahayam, U. Jayaraman, and R. Arramraju. Melanoma classification from dermatoscopy images using knowledge distillation for highly imbalanced data. *Computers in Biology and Medicine*, 154:106571, 2023. 2

[3] S. Aksoy, P. Demircioglu, and I. Bogrekci. Enhancing melanoma diagnosis with advanced deep learning models focusing on vision transformer, swin transformer, and convnext. *Dermatopathology*, 11(3):239–252, 2024. 3

[4] T.H. Aldhyani, A. Verma, M.H. Al-Adhaileh, and D. Koundal. Multi-class skin lesion classification using a lightweight dynamic kernel deep-learning-based convolutional neural network. *Diagnostics*, 12(9):2048, 2022. 3

[5] J.A. Almaraz-Damian, V. Ponomaryov, S. Sadovnychiy, and H. Castillejos-Fernandez. Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures. *Entropy*, 22(4):484, 2020. 2

[6] N. Badshah and A. Ahmad. Resbcu-net: deep learning approach for segmentation of skin images. *Biomedical Signal Processing and Control*, 71:103137, 2022. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 3

[8] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Recognition Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019. 4

[9] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the inter-

national skin imaging collaboration (ISIC). *arXiv preprint arXiv:1710.05006*, 2017. 4

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2016. 3

[12] R.I. Heroza, J.Q. Gan, and H. Raza. Enhancing skin lesion classification: a self-attention fusion approach with vision transformer. In *Annual Conference on Medical Image Understanding and Analysis*, pages 309–322. Springer, 2024. 3

[13] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018. arXiv:1709.01507. 4

[14] I. Iqbal, M. Younus, K. Walayat, M.U. Kakar, and J. Ma. Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. *Computerized Medical Imaging and Graphics*, 88:101843, 2021. 2

[15] A.N. Omeroglu, H.M. Mohammed, E.A. Oral, and S. Aydin. A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification. *Engineering Applications of Artificial Intelligence*, 120:105897, 2023. 3

[16] E. Perez, O. Reyes, and S. Ventura. Convolutional neural networks for the automatic diagnosis of melanoma: an extensive experimental study. *Medical Image Analysis*, 67:101858, 2021. 2

[17] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. arXiv:1710.05941. 4

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. arXiv:1801.04381. 4

[19] Skin Cancer Foundation. Seborrheic keratosis, 2023. Accessed: 2024-05-20. 2

[20] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *Proceedings of SPIE*, 2019. Vol. 11006, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. 4

[21] N. Spolaor, H.D. Lee, A.I. Mendes, C.V. Nogueira, A.R.S. Parmezan, W.S.R. Takaki, and R. Fonseca-Pinto. Fine-tuning pre-trained neural networks for medical image classification in small clinical datasets. *Multimedia Tools and Applications*, 83(9):27305–27329, 2024. 3

[22] Ç. Suiçmez, H.T. Kahraman, A. Suiçmez, C. Yılmaz, and F. Balcı. Detection of melanoma with hybrid learning method by removing hair from dermoscopic images using image processing techniques and wavelet transform. *Biomedical Signal Processing and Control*, 84:104729, 2023. 2

[23] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2017. 3

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2015. 3

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2016. 3

[26] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2820–2828, 2019. arXiv:1807.11626. 4

[27] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 4

[28] P.B.S. Varma, S. Paturu, S. Mishra, B.S. Rao, P.M. Kumar, and N.V. Krishna. Sldcnet: skin lesion detection and classification using full resolution convolutional network-based deep learning cnn with transfer learning. *Expert Systems*, 39(9):e12944, 2022. 3

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3

[30] S. Mannumadam Venugopal. Can vision transformers be the next state-of-the-art model for oncology medical image analysis? *AI in Precision Oncology*, 1(6):286–305, 2024. 3

[31] World Health Organization. Radiation: Ultraviolet (UV) radiation and skin cancer, 2023. Fact sheets. Accessed: 2024-05-20. 2

[32] Peng Zhang and Divya Chaudhary. Hybrid deep learning framework for enhanced melanoma detection. *IEEE Transactions on Computational Biology and Bioinformatics*, 2025. Epub ahead of print. 2