MelanoVision: Evaluating ViT and CNN Architectures for Skin Cancer Classification

Ege Atay
New Jersey Institute of Technology
ea353@njit.edu

Marckenley Mercie
New Jersey Institute of Technology
mm2452@njit.edu

Abstract

In this study, we analyze the performance of convolutional neural networks (CNNs) and transformer-based models to classify dermatoscopic skin images from the SKIN MIST: HAM10000 dataset, containing over 10,000 high-resolution images of skin lesions. Initially, we planned to start training architectures such as ResNet50 and InceptionV3 from scratch, which turned out to be unfeasible with respect to computational resources required by the dataset size. In addition, the MLP was trivial at baseline, thus offering no place for comparison. To solve this problem, we employed pretrained ResNet architectures, custom-built sequential CNNs, and a transformer-based model.

The performance evaluation was based on precision, F1-score, and graphs comparing training and test accuracy loss curves. These measures enabled assessment of learning as well as the predictive performance across all architectures. Our findings show that pretrained CNNs such as ResNet34 provide a practical and efficient feature extraction and classification solution to dermatoscopic imaging tasks, by far outperforming naive models while requiring minimized training overhead. Also, the introduction of a transformer-based model helps in analyzing the flexibility of contemporary architectures beyond CNNs for medical image analysis.

1 Introduction

1.1 Background

The classification of dermatoscopic skin images has a very important role in the early detection and diagnosis of skin conditions, including types of skin cancers. Deep learning techniques, especially convolutional neural networks, have emerged strong tools for dealing with medical image classification and hold out possibilities for improvement of diagnostic accuracy and clinical workflow. More recently still, transformer-based models, originally developed for natural language processing, have achieved extremely promising results in image analysis, further expanding the list of tools before researchers and clinicians.

1.2 Approach

Though training deep learning models from scratch is theoretically possible, practical limitations, such as the tools and computational resources available, imposed challenges in our case. In the very beginning, we intended to train popular architectures like ResNet50 and InceptionV3 employing an MLP as baselines. Still, with the tools at hand, this approach was inefficient and impractical for SKIN MIST: HAM10000 datasets. The other shortcoming was that an MLP model provided little in the way of result quality to be considered a meaningful benchmark. To overcome these challenges, we decided to evaluate pre-trained ResNet models, custom-built sequential CNNs, and a

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

pre-trained transformer-based model, which provided a more effective and efficient path for feature extraction and classification. In this case, the performance of the models was done in respect to precision, F1-score, and viewing training/test accuracy loss plots such that all three factors could judge predictive accuracy and learning behavior in detail. The experimental results show that using pre-trained architectures and modern transformer models proves to be an effective approach for medical image analysis, achieving a performance-to-computational feasibility trade-off.

2 About the Data

The HAM10000 ("Human Against Machine with 10,000 training images") dataset consists of 10,015 high-quality dermatoscopic images depicting the range of common pigmented skin lesions seen in clinical dermatology. Drawn together over a period of approximately 20 years, this resource combines cases from two distinct geographic and clinical backgrounds: a medical university in Vienna, Austria; and a specialized practice for skin cancer in Queensland, Australia. By combining pictures from these sources, the HAM10000 dataset builds a rare heterogeneous dataset that is of great importance as a reference for this area and provides a benchmark for developing, testing, and iteratively improving automated skin lesion-classification and diagnosis models. The dataset covers seven main types of diagnoses, from benign nevi to malignant melanomas. All in all, more than 95% of pigmented lesions that one usually encounters in dermatology practice fall into these categories:

Melanocytic Nevi (nv) Melanocytic Nevi, or moles as they are more frequently known, are benign manifestations of aberrant melanocyte growth. They frequently have uniform colors, structures, and symmetrical shapes. They are frequently regarded as the accepted benchmark by which benign tumors are distinguished from melanoma.

Melanoma (mel) One of the deadliest types of skin cancer is melanoma, a malignant cancer of the melanocytes. Frequently asymmetrical in form, hue, and consistency, it has the potential to spread quickly if not identified in time. For timely intervention, these lesions require special attention to minor dermatoscopic signs.

Benign Keratosis-like Lesions (bkl) Seborrheic keratoses, solar lentigines, and lichen-planus-like keratoses fall under this category. Despite being benign, certain lesions that resemble keratosis may exhibit unusual characteristics that, from a clinical and dermatoscopic standpoint, can resemble melanoma.

Basal Cell Carcinoma (bcc) One type of skin cancer that is locally invasive but seldom spreads are basal cell carcinoma. Under a microscope, it can appear as a flat or nodular lesion with a range of color and vascular patterns. Although it is typically less aggressive than melanoma, accurate detection is essential for prompt treatment.

Actinic Keratoses and Intraepithelial Carcinoma (akiec) Long-term exposure to UV light can cause precancerous lesions called actinic keratoses. These scaly, sun-damaged spots have the potential to progress to squamous cell carcinoma if treatment is not received. For preventive management, their identification is essential.

Vascular Lesions (vasc) Angiomas, angiokeratomas, and hemorrhages are examples of vascular lesions that are typically benign and have distinct vascular structures. They can range in color from red to purple. Even though these lesions are typically benign, it's crucial to distinguish them from other kinds of lesions.

Dermatofibroma (df) Benign fibrous nodules called dermatofibromas are frequently brought on by mild injuries or insect bites. Their clinical identification is aided by their characteristic central whitish region, which is encircled by a fine pigment network.

Figure 1 illustrates a range of dermatoscopic images spanning these lesion categories, highlighting their variability in color, texture, and structural features. The visual complexity seen here underlines the clinical challenges in distinguishing benign from malignant lesions and shows the importance of robust data-driven diagnostic tools.

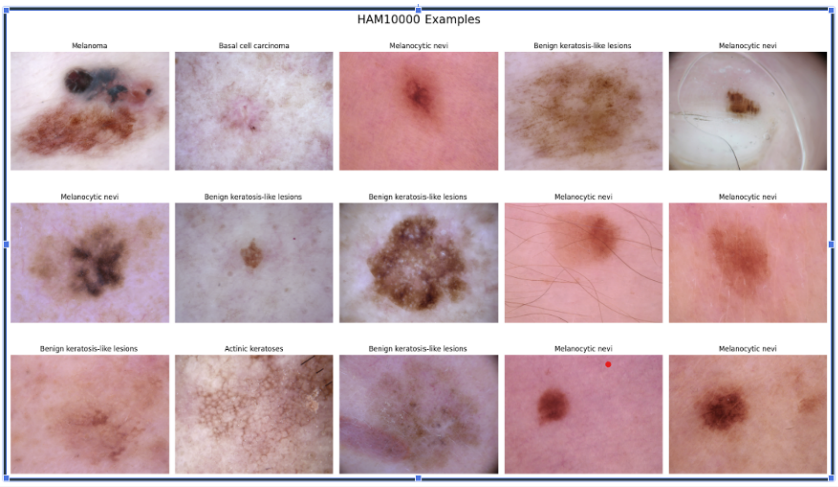

Figure 1: Sample of dermatoscopic images from the HAM10000 dataset

3 Exploratory Data Analysis

Figure 2 provides a detailed breakdown of lesion types, patient demographics, and lesion localization within the HAM10000 dataset. Melanocytic nevi represent the majority class, followed by benign keratosis-like lesions and melanoma, forming a substantial portion of the training samples that directly influence model outcomes. Although certain categories—such as basal cell carcinoma, actinic keratoses, and vascular lesions—are less prevalent, their inclusion remains essential. Even with limited data, incorporating these rarer lesions increases the clinical applicability of the resulting models and improves their ability to generalize to a broader range of dermatologic conditions.

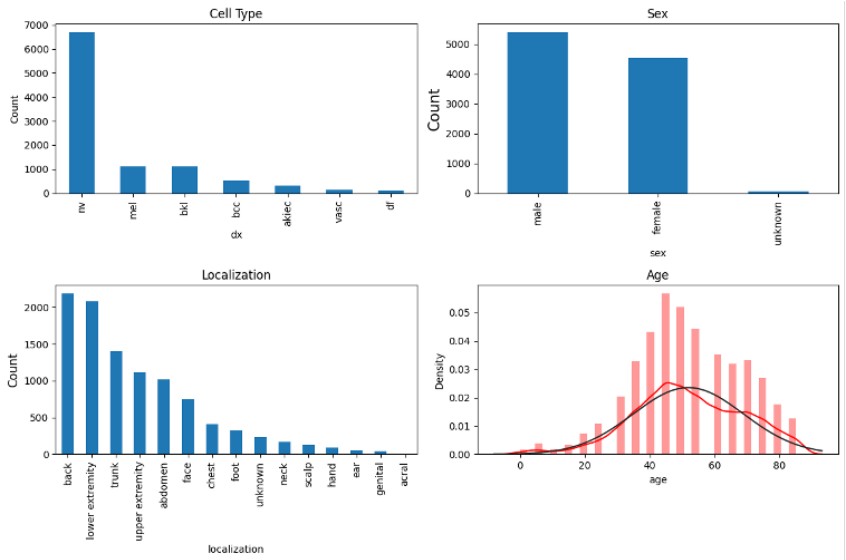

Figure 2: Bar Plots and Distributions of Demographic data of patients

Beyond image data, the dataset includes patient-level metadata—such as sex, age, and lesion location—that can inform and refine model development. The sex distribution is nearly balanced between male and female patients, with a small subset classified as unknown. This balanced demographic representation enhances the model's robustness when dealing with gender-related variations in skin lesions. The age distribution skews towards the 40–60-year range, aligning with the higher prevalence of cutaneous abnormalities in middle-aged populations. Furthermore, lesion localization data reveals a high incidence on the back, lower extremities, and trunk, correlating with areas commonly ex-

posed to UV radiation. Integrating these metadata insights into algorithm design and model training strategies can lead to more context-aware diagnostic tools and improved performance across diverse clinical scenarios.

4 Methodology

In this study, we systematically evaluated a range of deep learning architectures for classifying dermatoscopic images from the HAM10000 dataset. Our approach included developing a custom Sequential CNN and benchmarking it against pre-trained networks, such as ResNet34 and a Vision Transformer (ViT). These pre-trained models, initially trained on large-scale datasets like ImageNet, facilitated transfer learning and substantially reduced both the training time and the amount of labeled data required for effective convergence. Additionally, we applied data augmentation techniques to artificially increase dataset variability, thereby minimizing overfitting and improving the generalization capabilities of all evaluated models. Each of these models were trained on an NVIDIA GeForce RTX 3070 GPU.

4.1 Data Augmentation

To increase the effective training set size and improve the model's robustness to clinically relevant variations, we employed data augmentation techniques. Given the limited sample size of the HAM10000 dataset and the intrinsic complexity of dermatoscopic images, these measures were critical for mitigating overfitting. By introducing transformations related to orientation, scale, and illumination, the augmentation process enabled the model to more effectively generalize and handle the diverse noise and variability encountered in real-world clinical imaging scenarios.

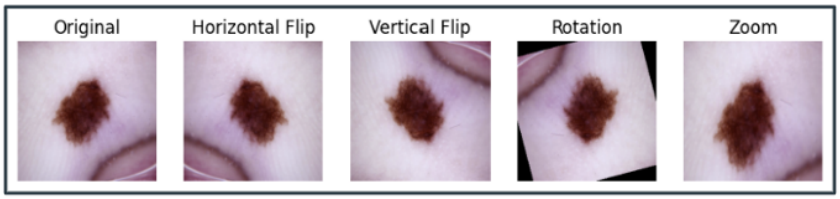

Figure 3: Augmented image from HAM10000 dataset

Data augmentation was implemented using the `torchvision.transforms` module in PyTorch. The transformations illustrated in Figure 3 Included:

Horizontal and Vertical Flips Randomly flipping images along the horizontal and vertical axes helped simulate variations in lesion orientation observed in clinical practice.

Rotations Random rotations (up to 180 degrees) allowed the model to learn from lesions presented at multiple angular perspectives.

Random Resized Cropping Zooming in or out on selected images encouraged the model to identify key lesion features, rather than relying on context or overall size.

Normalization Standardizing pixel intensities using dataset-specific mean and standard deviation ensured consistent brightness and contrast, enabling the model to focus on clinically salient features rather than environmental or lighting discrepancies.

4.2 Custom Sequential CNN Model

As an initial baseline for skin lesion classification, we implemented a custom Sequential convolutional neural network (CNN) to categorize images from the HAM10000 dataset into seven diagnostic classes. The model architecture can be summarized as follows:

- **Convolutional Layers:**

  - *Layer 1:* A 3×3 convolution applied 256 filters to the input ($32 \times 32 \times 3$), generating a feature map of dimensions ($30 \times 30 \times 256$).
  - *Layer 2:* A subsequent 3×3 convolution reduced spatial dimensions while increasing feature abstraction, resulting in a ($13 \times 13 \times 128$) feature map.
  - *Layer 3:* Another 3×3 convolution further refined representations and reduced the feature map to ($4 \times 4 \times 64$).
- **Max-pooling Layers:**
  - Max-pooling operations, with a pooling size of 2×2, followed each convolutional layer to reduce the spatial dimensions by half.
- **Dropout Layers:**
  - Dropout layers with a probability of 0.20 followed each pooling operation to mitigate overfitting by randomly deactivating a subset of neurons during training.
- **Flattening and Dense Layers:**
  - The final $4 \times 4 \times 64$ feature map was flattened into a 256-dimensional vector.
  - A fully connected layer reduced this vector to 32 neurons.
  - The output layer employed a softmax activation function over seven units, providing class probabilities for the seven lesion categories.

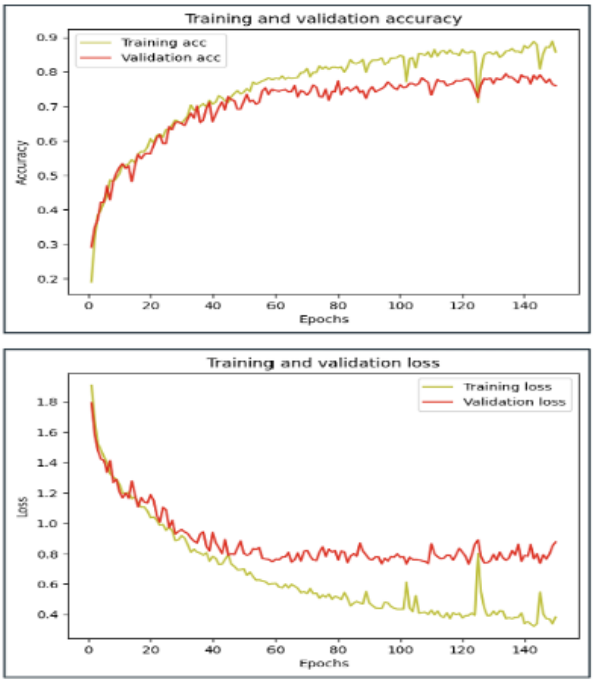

Figure 4: Training and Validation Loss and Accuracy plots

Figure 4 illustrates the training and validation accuracy and loss curves obtained during model training. The training accuracy progressively increased over successive epochs, ultimately approaching approximately 90%, indicating effective adaptation to the training data. In contrast, the validation accuracy plateaued at roughly 70%, suggesting limited generalization to unseen samples. Although the training loss steadily declined, the validation loss remained elevated and nearly constant, indicating that the model was overfitting to the training set and failing to sufficiently generalize to the validation data.

4.3 RESNET34

ResNet architectures have become a standard choice for image classification tasks due to their ability to facilitate the training of very deep neural networks. They achieve this by incorporating residual

learning, which focuses on learning the difference (residual) between the input and the desired output, rather than attempting to learn the full mapping directly. This is implemented through skip connections that bypass one or more layers, enabling efficient gradient propagation and mitigating vanishing gradient issues. Pre-trained ResNet models are widely employed for transfer learning because they provide powerful, general-purpose feature representations. Leveraging these pre-initialized weights allows models to converge faster and often improves performance, especially when training data is limited. For these reasons, we integrated pre-trained ResNet34 architecture into our workflow to enhance lesion classification accuracy.

	precision	recall	f1-score	support
akiec	0.40	0.40	0.40	25
bcc	0.43	0.79	0.56	29
bkl	0.46	0.66	0.54	73
df	0.06	0.57	0.11	7
mel	0.32	0.39	0.35	38
nv	0.99	0.81	0.89	728
vasc	0.47	0.90	0.62	10
accuracy			0.77	910
macro avg	0.45	0.65	0.50	910
weighted avg	0.87	0.77	0.81	910

Figure 5: Classification Report for ResNet34

The classification report for ResNet34 as shown in Figure 5 displays the overall model accuracy is at 77%. In terms of performance when looking at the macro F1-score, the model struggles across all classes, especially the underrepresented ones. The weighted F1-score is 81% which is likely because of the dominant class. The low precision, recall, and F1-scores for melanoma (mel), df, and akiec indicate that the model struggles to distinguish these rare and visually complex lesions, possibly due to insufficient training samples or class imbalance.

4.4 Vision Transformer

Although the focus of this study is on Convolutional Neural Networks, we decided to add Vision Transformers to this comparison to showcase the contrast in the difference of these architectures in image classification tasks. For this, we decided that it was the most logical to fine-tune a pre-trained vision transformer model. We used Google's `vit-base-patch16-224` model (Wu et al., 2020). The model architecture can be presented as:

- **Token and Positional Embeddings:**
  - *CLS Token:* Initialized with dimensions `torch.Size([1, 1, 768])`, contributing 768 parameters to the model.
  - *Positional Embeddings:* Built to align positional information with dimensions `torch.Size([1, 197, 768])`, comprising 151,296 parameters.
- **Patch Embedding:**
  - *Projection Weights:* Transforms image patches into the embedding space with dimensions `torch.Size([768, 3, 16, 16])`, involving 589,824 parameters.
  - *Projection Bias:* With dimensions `torch.Size([768])`, this adds 768 parameters.
- **Transformer Encoder Blocks:**
  - *Norm Layers:* Each block, from 0 to 11, includes Layer Norm operations for input normalization with weights and biases both sized at `torch.Size([768])`, totaling 1,536 parameters per block.

- *Attention Mechanism:* Utilizing multi-head self-attention, the query, key, and value (qkv) projections have dimensions `torch.Size([2304, 768])`, totaling 1,769,472 parameters for weights and 2,304 for biases per block.
- *Attention Projections:* With matrices of size `torch.Size([768, 768])`, these account for another 589,824 parameters for weights and 768 for biases per block.
- *MLP Layers:* Each block includes a two-layer MLP with weights `torch.Size([3072, 768])` and `torch.Size([768, 3072])`, contributing 2,359,296 parameters per layer along with biases sized `torch.Size([3072])` and `torch.Size([768])`, contributing 3,072 and 768 parameters respectively per layer.
- This configuration repeats for each of the 12 blocks.

- **Final Layers:**
  - *Normalization Layer:* A final normalization step with parameters `torch.Size([768])`, consisting of 768 weights and 768 biases.
  - *Output Head:* Concludes with a classification head with dimensions `torch.Size([7, 768])` for weights and `torch.Size([7])` for biases, involving 5,376 and 7 parameters respectively.

Overall, the transformer model features a total of 85,804,039 parameters, all of them trainable by default.

Although this model features many parameters, training, or more precisely, fine tuning it does not take an infeasible amount of time due to the architecture's ability to utilize parallelization. In our experiments, each epoch took around 25 minutes on a consumer GPU.

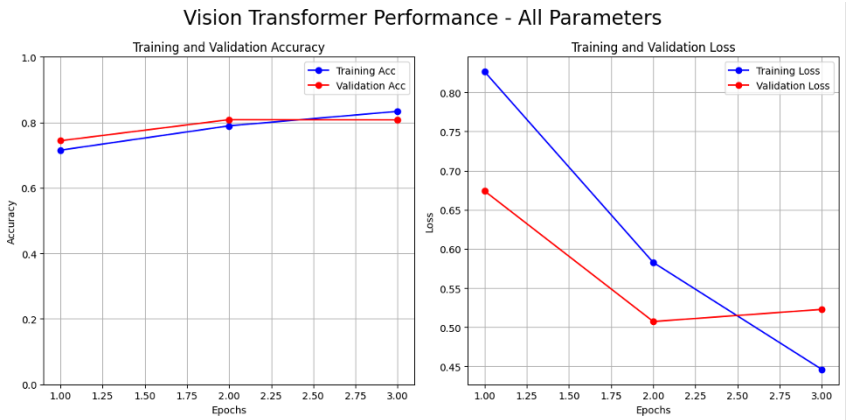

Figure 6: Training and Validation Accuracy/Loss Plots

Although this vanilla approach on fine-tuning a pretrained vision transformer already outperforms CNNs of similar training times, one approach that is often used to increase the efficiency of fine-tuning transformers is training only the bias parameters of the hidden blocks (Zaken et al., 2022). This can be implemented simply with the following loop as shown in Code Block 1:

```
for k, v in model.named_parameters():
    if (k[:7] == "blocks."):
        v.requires_grad = ("bias" in k)
```

Code Block 1: Implementation of the bias-only training method

This approach reduces the number of trainable parameters from 85,804,039 to 850,951, which is less than 1% of the parameters of the traditional approach. Implementing the bias-only training approach to fine-tuning the pretrained vision transformer expectedly increases the training speed, which decreases from 25 minutes per epoch to just 6 minutes per epoch in the same environment.

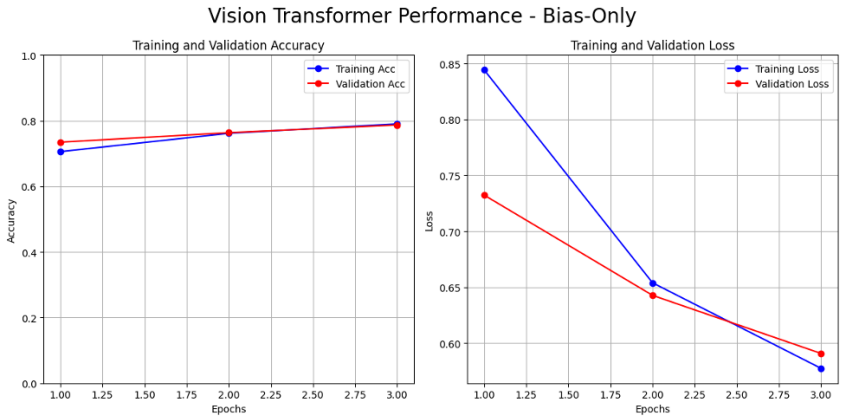

Figure 7: Training and Validation Accuracy/Loss Plots for the bias-only training implementation

We notice that although the number of trainable parameters is only a fraction of the traditional model, the results yield similar accuracy. We believe this is a worthy trade off between efficiency and pure per-epoch accuracy.

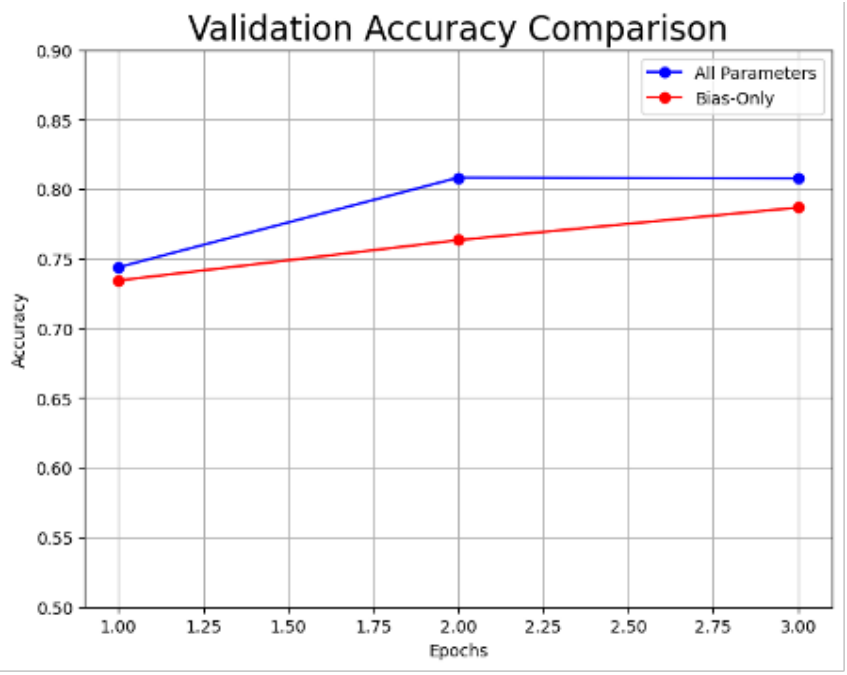

Figure 8: Comparison of traditional and bias-only training

5 Other Studies and Their Results

A recent study titled "Accurate Skin Lesion Classification Using Multimodal Learning on the HAM10000 Dataset" rigorously examined how incorporating patient metadata—such as sex, age, and lesion location—alongside dermatoscopic imagery could enhance skin lesion classification. The study's core hypothesis was that integrating non-visual patient-specific attributes with image data would improve diagnostic performance compared to models relying solely on visual inputs. To evaluate this hypothesis, the authors considered several model architectures, including established convolutional neural networks (Inception-V3, ResNet50, and DenseNet121) as well as a state-of-the-art multimodal learning framework known as ALBEF (Align-Before-Fuse). Unlike traditional CNN

models, ALBEF employs a Vision Transformer (ViT) for image encoding and a BERT-based text encoder for processing patient metadata, thereby enabling a more holistic representation of the input space. All models were trained using the Adam optimizer with a learning rate of 1×10^{-4}.

Table 1: Comparison of the performance of the different models on the HAM1000 dataset (the best model is shown in bold). Data reproduced from Accurate Skin Lesion Classification Using Multimodal Learning on the HAM10000 Dataset (Smith et al.,2024)

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUROC
Inception-V3	86.53%	82.02%	76.08%	95.70%	78.38%	85.89%
ResNet50	85.03%	78.15%	71.29%	96.46%	72.92%	83.88%
DenseNet121	88.62%	85.05%	82.49%	96.85%	83.50%	89.67%
ALBEF (Images Only)	91.32%	89.59%	85.24%	97.48%	86.93%	91.36%
ALBEF (Images + Text)	94.11%	90.73%	90.19%	98.33%	90.33%	94.26%

Empirical results demonstrated that the ALBEF model consistently outperformed the pure image-based CNN baselines. Specifically, ALBEF achieved a top accuracy of 94.11% and an AUROC of 0.9426 when jointly leveraging image data and patient-level metadata. These findings, summarized in Table 1, highlight the effectiveness of multimodal learning in capturing both visual and contextual signals, ultimately improving the reliability and utility of automated skin lesion classification systems.

6 Conclusion

We have evaluated the performance of pre-trained CNNs and transformer-based models for the classification of dermatoscopic images using the HAM10000 dataset. Through models such as our custom Sequential CNN, ResNet34, and ViT, we highlighted the strengths and limitations of these architectures. While ResNet334 leveraged a balanced trade-off between computational efficiency and performance, it struggled with underrepresented classes. The Vision Transformer, on the other hand, showcased its capability to capture complex image features, achieving improved accuracy compared to traditional CNNs.

Our findings align with similar studies, such as the multimodal learning approach evaluated in Accurate Skin Lesion Classification Using Multimodal Learning on the HAM10000 Dataset, which showed the benefits of combining patient metadata with image data. This suggests that future research should consider integrating multimodal approaches and addressing dataset limitations to enhance generalization and robustness. By leveraging advanced architectures like ViTs and adopting innovative training techniques, automated systems for skin lesion classification can be further refined to improve diagnostic accuracy.

References

[1] Smith, A., Johnson, B. & Lee, C. (2024) Accurate skin lesion classification using multi-modal learning on the HAM10000 dataset. *Journal of Medical AI Research* **15**(4):123–135. https://www.medrxiv.org/content/10.1101/2024.05.30.24308213v4

[2] Tschandl, P., Rosendahl, C., & Kittler, H. (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5**:180161. https://doi.org/10.1038/sdata.2018.161

[3] Zaken, E.B., Ravfogel, S. & Goldberg, Y. (2022) BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint*. Available at: https://arxiv.org/abs/2106.10199.

[4] Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020) Visual Transformers: Token-based Image Representation and Processing for Computer Vision. Retrieved from arXiv preprint arXiv:2006.03677, cs.CV.