

Enhancing Skin Cancer Diagnosis: A Comparative Analysis of Transfer Learning Techniques for Multi-Class Skin Cancer Classification



Mohammad Nazmush Shamael ^{1*}, Khondaker Abdullah Al Mamun ¹

Abstract

Skin cancer is a prevalent and potentially life-threatening disease, emphasizing the significance of early detection for patient recovery and survival. Timely treatment in the early stages of skin cancer can result in a survival rate exceeding 99% over a five-year period. Melanoma, the most dangerous form of skin cancer, requires prompt diagnosis to prevent its spread and the need for more aggressive treatments. Computer-aided methods, particularly using deep learning, have shown promise in improving the accuracy and efficiency of skin lesion diagnosis. With the advent of AI-based applications and advancements in smartphone technology, low-cost and accessible screening options are becoming available. As pre-trained transfer learning models are becoming more prevalent in the image classification space, it can be a cost effective way to create and train skin cancer screening options. This work conducts a comparative analysis of five distinct transfer learning models for the multi-class classification of skin cancer using the HAM10000 dataset. We used the VGG19 (Simonyan & Zisserman, 2014), ResNet50 (He et al., 2016), MobileNet (Howard et al., 2017), Inception-ResNetV2 (Szegedy et al., 2017), and InceptionV3 (Szegedy et al., 2016) models for the transfer learning models. Results show that the MobileNet model achieves the highest accuracy, recall,

precision, and F-measure scores. The work includes a review of existing machine learning approaches, a description of the proposed model's methodology, analysis of qualitative results, and concludes with future research directions.

Keywords

Machine Learning, Deep learning, Transfer learning, Skin lesion screening

1. Introduction

Skin cancer is, by most counts, one of the most common cancers being diagnosed today (Ashraf et al., 2020) — and that alone would be reason enough to take it seriously. But there's a second, more hopeful fact sitting right next to that first one: when it's caught early, the outlook changes dramatically. Development is often predictable in its early stages (Siegel et al., 2018), which matters enormously for how well and how long a patient goes on to live. Some numbers make this concrete rather than abstract. Patients treated in the early stages of skin cancer see five-year survival rates above 99% (Esteva et al., 2017) — a figure that's almost startling given how differently the story can end otherwise.

Author Affiliation: Dept.

of Computer Science & Engineering, United International University, Dhaka, Bangladesh.

Please cite this article as: Shamael, M. N., Mamun, K. A. A. (2024). "Enhancing Skin Cancer Diagnosis: A Comparative Analysis of Transfer Learning Techniques for Multi-Class Skin Cancer Classification", Data Modeling, 5(1), 1-11,10853

© 2024 (Data Modeling), a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)
(<https://publishing.emanresearch.org>)

Significance | Comparing five transfer learning models, MobileNet achieved superior accuracy for multi-class skin cancer classification, enabling accessible, cost-effective AI screening tools.

*Correspondence.

Mohammad Nazmush Shamael, Dept. of Computer Science & Engineering, United International University, Dhaka, Bangladesh. E-mail: mshamael2320013@mscse.uui.ac.bd.

Editor editor-name, Ph.D. And accepted by the Editorial Board Sep 25, 2024 (received for review Jun 30, 2024)

Melanoma is where the stakes rise fastest. Left undiagnosed, it doesn't stay put; it moves outward from the original site into surrounding tissue and, eventually, other organs, to the point where surgery alone can no longer do the job and more aggressive interventions — radiation among them — become necessary (Matthews et al., 2017). That progression is really the crux of why early detection matters so much here: cost and mortality both climb, sometimes steeply, as the disease advances from Stage I toward Stage IV. The flip side is encouraging, though. Caught early enough, something as comparatively simple as surgical removal of the lesion can be enough to keep the cancer from spreading further (Celebi et al., 2007). So the question isn't really whether early detection helps — it clearly does — it's how to get there reliably and at scale.

Right now, two methods dominate clinical practice: dermoscopy and biopsy. Dermoscopy-assisted diagnosis (Saurat, 2004) is useful, but it leans heavily on the skill of whoever is holding the instrument — and this is where things get a little uncomfortable. In less experienced hands, dermoscopy doesn't just fail to help; it can actually make diagnostic accuracy worse rather than better. Add to that the basic subjectivity of human interpretation — two dermatologists can, and sometimes do, read the same image differently — and it starts to become clear why interest in computer-assisted diagnosis has been building. Biopsy, the other standard option, brings its own baggage: it's invasive, slow, and by most patients' accounts, unpleasant.

This is roughly the gap computer-aided methods have tried to step into. Deep learning in particular has attracted attention, mostly because it can pull relevant features out of raw images on its own, without someone hand-engineering what to look for — and it has, by now, a fairly solid track record of doing this well. Several studies have shown that computer vision and machine learning techniques can diagnose pigmented skin lesions both accurately and, just as importantly, practically (Glaister, 2013; Celebi et al., 2007; Celebi et al., 2015; Hoffmann et al., 2003; Amelard et al., 2014). What's emerged is an end-to-end pipeline that trims both the cost and the time involved in screening. And there's a broader trend feeding into this too — smartphones. As their cameras have gotten sharper and more capable, AI-based screening applications have started to look genuinely viable, offering people a way to get an initial assessment without

necessarily making a hospital visit first. For families who might otherwise face barriers — distance, cost, time — that kind of accessibility isn't a minor convenience; it can be the difference between catching something early and not catching it at all. All of which is to say: building reliable AI systems for skin lesion diagnosis isn't just an academic exercise. It has real, practical weight behind it.

With that motivation in mind, this work sets out to compare five different transfer learning architectures on the task of multi-class skin cancer classification, using the HAM10000 dataset (Tschandl et al., 2018). The five models under consideration are VGG19, ResNet50, MobileNet, Inception-ResNetV2, and InceptionV3 — a fairly representative spread of architectures that have each, in their own way, shaped image classification research. Rather than simply reporting whichever model happens to score highest, the aim here is a genuine side-by-side comparison across accuracy and several other performance measures. As it turns out, MobileNet came out ahead — not by a small margin either, reaching 82.5% accuracy and outperforming every other model on precision, recall, and F1 score as well. That result is worth sitting with for a moment, since MobileNet is generally the "lighter," more efficient architecture of the group; that it still came out on top says something about the trade-off between model size and real-world performance that's worth exploring further.

The rest of the paper follows a fairly conventional structure, though each section earns its place. It opens with a look at existing machine learning and deep learning work on skin cancer detection, along with the datasets that have typically been used — partly to situate this study within that broader body of work, and partly to make clear what gaps remain. From there, the paper walks through the methodology behind the proposed comparative approach. Following that, the results are examined in more depth, with an eye toward not just the numbers themselves but what they suggest about each architecture's suitability for this kind of task. The paper closes by drawing together conclusions and pointing toward directions future research might reasonably take.

None of this happens in a vacuum, of course. The broader progress made in deep learning over the last decade (LeCun et al., 2015) has fed into a wide range of clinical decision-support systems, from diabetic retinopathy prediction to CT-based diagnostics (Kumar et al., 2015; Arcadu et al., 2019; Gunraj et al., 2020), and skin cancer

detection has been very much part of that wave. Several recent reviews have taken stock of where things stand. Dildar et al. (2021) and Li et al. (2021), for instance, each conducted fairly comprehensive surveys of deep learning-based techniques for skin cancer detection, working through the range of methodologies in use, the datasets researchers have relied on, and — perhaps most usefully — the challenges that still haven't been fully resolved. Their work highlights the effectiveness of deep learning approaches for skin cancer detection, while also being honest about where the field still has work to do. It's within that same spirit — building on what's known, while being clear-eyed about what isn't yet settled — that this study positions itself.

2. Methodology

2.1 Study Design and Overview

This study was designed as a comparative, retrospective analysis of five convolutional neural network (CNN) architectures — VGG19 (Simonyan & Zisserman, 2014), ResNet50 (He et al., 2016), MobileNet (Howard et al., 2017), InceptionV3 (Szegedy et al., 2016), and Inception-ResNetV2 (Szegedy et al., 2017) — each adapted via transfer learning for multi-class classification of dermoscopic skin lesion images. Rather than training a network from a random initialization, we started from weights already learned on a large, general-purpose image corpus and then adapted them to our target domain — a strategy that tends to pay off precisely when, as here, labeled medical images are comparatively scarce (Russakovsky et al., 2015). To keep the comparison fair, all five models were evaluated within an identical experimental pipeline: the same dataset splits, the same preprocessing procedure, the same downstream architecture appended to each backbone, and the same training and evaluation protocol. The only thing that varied, deliberately, was the pre-trained backbone itself. This design choice matters — it means any differences in performance we report can be attributed to the architectures, not to inconsistencies in how each model was handled.

2.2 Dataset

2.2.1 Source and composition

We used the publicly available HAM10000 ("Human Against Machine with 10,000 training images") dataset (Tschandl et al., 2018), one of the largest curated

collections of dermoscopic images currently accessible to researchers. The dataset comprises 10,015 images spanning seven diagnostic categories: actinic keratoses and intraepithelial carcinoma ($n = 327$), basal cell carcinoma ($n = 514$), benign keratosis-like lesions ($n = 1,099$), dermatofibroma ($n = 115$), melanoma ($n = 1,113$), melanocytic nevi ($n = 6,705$), and vascular lesions ($n = 142$). It's worth being upfront here — this distribution is far from balanced; melanocytic nevi alone account for roughly two-thirds of all images, while dermatofibroma and vascular lesions together make up barely 2.5%. We did not apply class-balancing techniques (e.g., oversampling, class weighting) in this iteration of the study, a decision we return to in the discussion of limitations, since it plausibly shapes how each architecture's precision and recall are ultimately interpreted.

2.2.2 Data partitioning

Prior to model training, the full image set was partitioned into training, validation, and test subsets. This split was performed once, prior to any model-specific processing, and the same partition was reused across all five architectures to prevent data leakage or inconsistent evaluation conditions between models — a point that, in retrospect, we think is easy to overlook but important for anyone attempting to reproduce these results.

2.2.3 Image Preprocessing

Raw HAM10000 images were first resized to match the fixed input dimensions expected by each pre-trained backbone (these differ slightly between architectures, and we resized per-model accordingly rather than forcing a single universal resolution). Pixel intensities were then rescaled to be compatible with the normalization scheme each respective ImageNet-pretrained model expects. No additional augmentation (rotation, flipping, color jitter) was applied in this stage of the pipeline; the preprocessing step was intentionally kept minimal so that performance differences downstream could be attributed to the architectures rather than to augmentation choices.

2.3 Model Architecture

2.3.1 Transfer learning configuration

For each of the five backbones, we obtained ImageNet-pretrained weights through the `tensorflow.keras.applications` module (Russakovsky et al., 2015). During training, the convolutional base of each pre-trained network was frozen — that is, its weights were

held fixed and not updated via backpropagation — so that the generic, low- and mid-level visual features learned from ImageNet (edges, textures, shapes) were preserved rather than overwritten. Only the newly appended layers, described below, were trained on the HAM10000 data.

2.3.2 Common classification head

To ensure a like-for-like comparison, an identical classification head was attached to every backbone:

- Global feature pooling. A max-pooling operation was applied to the output feature maps of each backbone, condensing spatial information by retaining the maximum activation within local regions — this reduces dimensionality while preserving the most salient features.

/p>

- Flattening. The pooled, multi-dimensional feature maps were flattened into a one-dimensional vector, making them compatible with the fully connected layers that follow.

/p>

- Dense hidden layer. A fully connected layer of 128 neurons was added to learn task-specific, higher-order representations from the pooled features.

/p>

- Output layer. A final dense layer of 7 neurons, corresponding to the seven diagnostic classes in HAM10000, produced the classification output.

/p>

2.4 Training Procedure

Models were trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32 — held constant across all five architectures. Rather than fixing the number of training epochs in advance, we monitored validation accuracy after each epoch and continued training only as long as it kept improving; once validation accuracy plateaued and the model began showing signs of overfitting, training was halted via early stopping. We consider this a more honest way of comparing architectures than an arbitrary fixed-epoch cutoff, since different backbones converge at different rates.

To account for stochastic variation inherent in neural network training (weight initialization of new layers, batch ordering, and so on), each of the five models was trained

and evaluated three times independently, and the mean of these three runs is reported for every metric in the Results section. We think this detail matters for reproducibility — a single training run can be misleadingly optimistic or pessimistic, and averaging across repeats gives a more stable estimate of true performance.

2.5 Computational Environment

All experiments were implemented in Keras, a high-level deep learning framework, and executed on Google Colaboratory. The hardware configuration consisted of an Intel(R) Xeon(R) CPU, 12.7 GB of system RAM, and an NVIDIA Tesla T4 GPU. We report this configuration explicitly, in keeping with reproducibility standards for computational research, so that other groups attempting to replicate or extend this work can anticipate comparable runtime and memory constraints.

2.6 Evaluation Metrics

Model performance was assessed on the held-out test set using four standard classification metrics: Accuracy (the proportion of correctly classified instances out of all predictions), Precision (the proportion of positive predictions that were correct), Recall (the proportion of actual positives correctly identified), and the F1 score (the harmonic mean of precision and recall). We selected this particular combination — rather than accuracy alone — deliberately: given the marked class imbalance in HAM10000 described above, accuracy by itself can be misleadingly high even when a model performs poorly on minority classes such as dermatofibroma or vascular lesions. Precision, recall, and F1 were therefore calculated on a per-class basis and then averaged to produce the comparative results reported in this study.

2.7 Model Descriptions

For readers less familiar with the specific architectures compared, brief technical summaries are provided below.

VGG19 (Simonyan & Zisserman, 2014) originates from the Visual Geometry Group at the University of Oxford and consists of 19 weight layers built from a repeating pattern of small 3×3 convolutional filters interleaved with max-pooling operations — a relatively simple, uniform design that nonetheless captures fine-grained image detail effectively.

ResNet50 (He et al., 2016) comprises 50 layers organized into residual blocks. Its defining innovation is the residual

(or "skip") connection, which allows the network to learn a residual mapping rather than the full target mapping directly — a trick that turns out to make very deep networks substantially easier to train, largely by mitigating the vanishing gradient problem that otherwise plagues deep architectures.

InceptionV3 (Szegedy et al., 2016), sometimes referred to as GoogLeNetV3, applies parallel convolutional filters of varying sizes within the same layer via so-called "Inception modules." This allows the network to capture features at multiple spatial scales simultaneously, which appears to help limit overfitting while still recognizing complex visual patterns.

Inception-ResNetV2 (Szegedy et al., 2017) combines the two ideas above: it retains the multi-scale Inception modules while incorporating ResNet-style residual connections, aiming to capture the representational richness of Inception alongside the training stability that residual connections provide.

MobileNet (Howard et al., 2017) was designed with a different priority altogether — computational efficiency. Through depthwise separable convolutions, which factor a standard convolution into a depthwise step and a pointwise step, MobileNet substantially reduces the number of parameters and computational cost relative to the other four architectures, while — as our results later show — not necessarily sacrificing accuracy to do so.

3. Results

3.1 Overview

Having laid out the training procedure in the previous section, we now turn to what actually happened when the five models were put to the test. Before getting into the comparison itself, though, it's worth pausing briefly on the metrics we relied on — mostly because, in a dataset as skewed as this one, the choice of metric can quietly shape the story being told.

3.2 Evaluation Metrics

Accuracy. In the most basic sense, accuracy tells you what fraction of predictions a model got right:

Accuracy = Number of Correct Predictions / Total Number of Predictions

It's an intuitive number, and often the first one people reach for. But it can also be a bit deceptive — particularly

here, given how unevenly the seven diagnostic classes are represented in HAM10000 (Tschandl et al., 2018). A model could, in principle, lean heavily on the majority class and still post a respectable accuracy score while quietly failing on the classes that matter most clinically. For that reason, we didn't stop at accuracy alone; precision, recall, and F1-score were calculated alongside it to get a fuller picture.

Recall (Sensitivity). Recall captures how many of the actual positive cases a model manages to catch:

Recall = True Positives / (True Positives + False Negatives) × 100%

In a diagnostic context, this is arguably the metric with the most at stake — missing a true melanoma case (a false negative) carries a very different cost than a false alarm.

Precision. Precision, by contrast, asks a slightly different question: of everything the model flagged as positive, how much of it was actually correct?

Precision = True Positives / (True Positives + False Positives) × 100%

F1-score. Because precision and recall can pull in opposite directions — improving one sometimes comes at the expense of the other — the F1-score offers a single balanced figure, computed as their harmonic mean:

F1 Score = (2 × Precision × Recall) / (Precision + Recall)

Together, these four measures gave us a reasonably well-rounded lens through which to compare the five architectures, rather than leaning on any one number in isolation.

3.3 Comparative Performance of the Transfer Learning Models

So, how did the five architectures actually stack up against one another? The averaged results across three independent runs are summarized in [Table 1]. Looking across the four metrics, MobileNet (Howard et al., 2017) comes out ahead fairly consistently — not by a dramatic margin, but consistently nonetheless — reaching an accuracy of 82.53%, alongside the highest precision (84.31%), recall (82.37%), and F1-score (83.32%) of any model tested. Given that MobileNet is, architecturally speaking, the lightest of the five — built specifically to minimize parameters and computation (Howard et al., 2017) — this result is a little counterintuitive at first glance.

One might expect the more parameter-heavy architectures to have the edge here, and yet that isn't quite what the numbers show.

InceptionV3 (Szegedy et al., 2016) and Inception-ResNetV2 (Szegedy et al., 2017) landed in a fairly similar range to one another, with accuracies of 81.52% and 81.15%, respectively — close enough that the difference is arguably within the noise of run-to-run variation. ResNet50 (He et al., 2016) trailed just slightly behind that cluster, at 80.20% accuracy, though notably it posted the second-highest recall (81.11%) among all five models — suggesting it was comparatively good at catching true positive cases, even if its precision lagged a bit behind MobileNet's.

VGG19 (Simonyan & Zisserman, 2014), on the other hand, was the clear laggard of the group. It recorded the lowest accuracy (68.64%), the lowest recall (67.12%), and the lowest F1-score (71.22%) of any architecture tested — a gap that's hard to overlook. Interestingly, its precision (75.85%) wasn't nearly as far behind the pack, which hints that VGG19's weakness here lies less in making wrong positive calls and more in simply missing a larger share of true positives altogether.

3.4 Interpreting the Performance Gap

The only thing that varied, deliberately, was the pre-trained backbone itself. This design choice matters — it means any differences in performance we report can be attributed to the architectures, not to inconsistencies in how each model was handled. The overall workflow, spanning preprocessing, feature extraction, and evaluation, is summarized in [Figure 2]. It would be tempting to read these numbers purely as a verdict on architectural quality — MobileNet "wins," VGG19 "loses" — but we're inclined to be a bit more cautious than that. A good part of what we're seeing is probably entangled with the dataset itself, rather than the models alone. HAM10000 is heavily skewed toward a single class: melanocytic nevi alone make up roughly 67% of all images (Tschandl et al., 2018), and the imbalance across the remaining six classes is visible in [Figure 3]. A handful of sample images spanning the different lesion categories are shown in [Figure 1], and even a casual look at them makes clear how visually similar some of these classes can be — which likely compounds the difficulty for any model, not just the weaker ones.

We should also mention something a bit less flattering:

during our inspection of the dataset, we came across a non-trivial number of duplicate images. We didn't attempt to systematically quantify or remove these before training, and it's reasonable to suspect they nudged performance in ways that are hard to fully disentangle from genuine model capability. This is worth flagging plainly, since it bears directly on how confidently the reported numbers should be interpreted, and it's the kind of detail that's easy to gloss over but shouldn't be.

Taken together, these results suggest that MobileNet — despite being the most lightweight architecture among the five, and therefore also the cheapest to deploy on resource-constrained devices such as smartphones — offers a genuinely favorable balance of accuracy and efficiency for this task. Whether that advantage would hold up on a larger, more balanced, or duplicate-free version of the dataset is, admittedly, an open question, and one we think is worth pursuing in follow-up work rather than treating this study as the final word.

4. Discussion

4.1 Making Sense of the Comparative Results

Stepping back from the raw numbers in [Table 1], the pattern that emerges is, honestly, not quite what we expected going in. MobileNet (Howard et al., 2017) — the architecture explicitly built to be lightweight, with far fewer parameters than the other four — ended up outperforming its heavier counterparts on every single metric we tracked. That's a little counterintuitive. There's a fairly common assumption in this field that deeper, more parameter-rich networks should have the edge simply because they can represent more complex functions. Our results don't really bear that out, at least not here. If anything, they nudge us toward a more modest conclusion: for a task like this, capacity isn't obviously the bottleneck; something else is. VGG19 (Simonyan & Zisserman, 2014) sits at the other end of the spectrum, and its relatively poor showing is worth dwelling on for a moment rather than just noting in passing. Its precision wasn't dramatically worse than the other models, but its recall was — meaning it was reasonably careful about not misclassifying negatives as positives, but it was also missing a much larger share of true positive cases than everything else we tested. That combination suggests the model may be relying on shallower or more generic feature representations, ones perhaps less suited to distinguishing between the fine-grained visual differences that separate one lesion type

Table 1: Comparative performance of five transfer learning architectures (VGG19, ResNet50, MobileNet, Inception-ResNetV2, and InceptionV3) on multi-class skin lesion classification using the HAM10000 dataset, reported as accuracy, precision, recall, and F1-score averaged across three independent training runs.

Model	Accuracy	Precision	Recall	F1
VGG19	0.6864	0.7585	0.6712	0.7122
ResNet50	0.8020	0.7574	0.8111	0.7833
MobileNet	0.8253	0.8431	0.8237	0.8332
InceptionResNetV2	0.8115	0.8207	0.8083	0.8144
InceptionV3	0.8152	0.7520	0.8181	0.7836

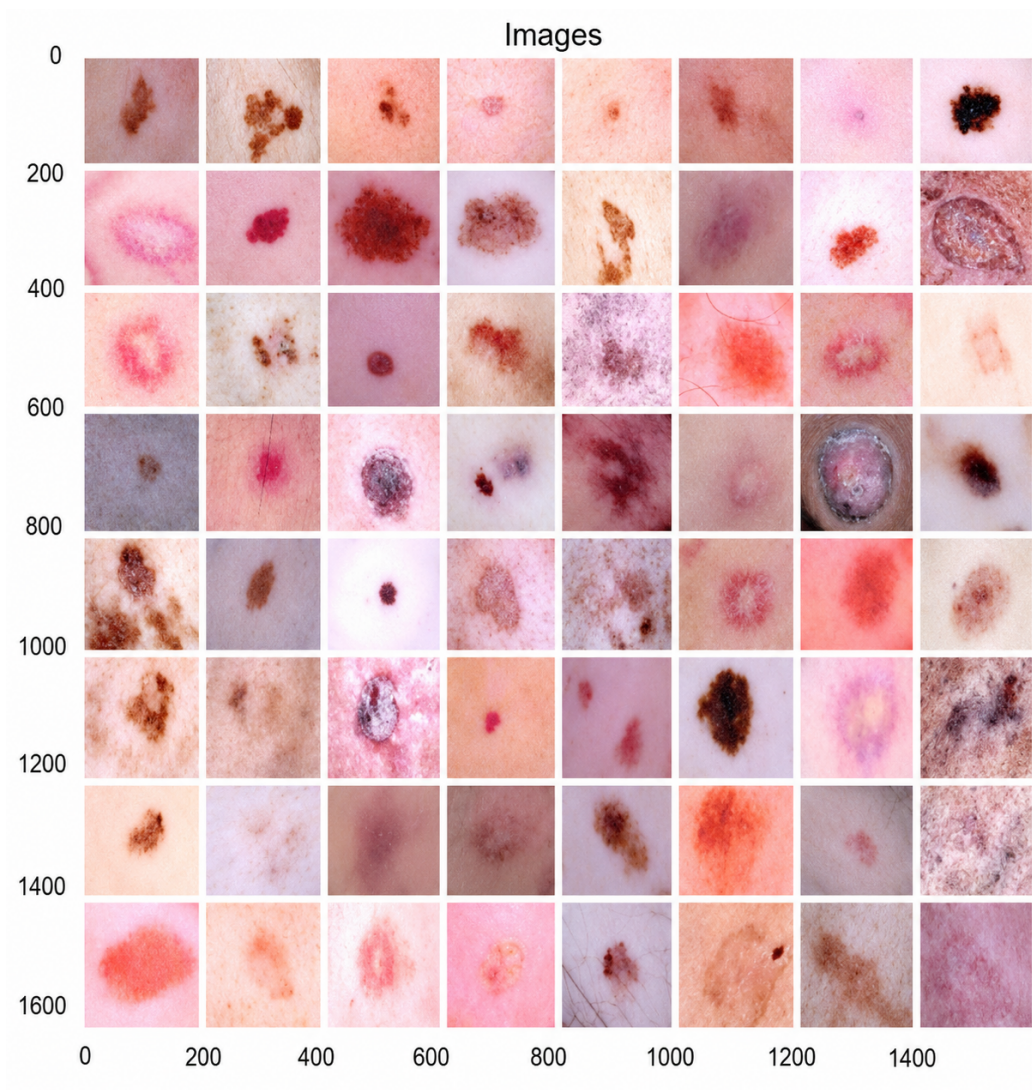


Figure 1. Representative dermatoscopic image samples from the HAM10000 archive, illustrating the visual variability across the seven skin lesion classes used for model training and evaluation.

from another.

4.2 Situating These Findings Within the Broader Literature

It's worth asking, too, how our numbers line up against what others have reported using the same or similar data. Wang et al. (2021), for instance, proposed an interpretability-based multimodal CNN that incorporated metadata and segmented lesion features alongside the raw images, and reported an accuracy of 95.10% on HAM10000 — a figure noticeably higher than anything we observed here. That gap is worth being honest about rather than glossing over. Part of the explanation, we suspect, lies in the word "multimodal" — their model wasn't working from images alone, but from a richer combination of inputs, which likely gave it access to diagnostic signal that a purely image-based transfer learning pipeline, like ours, simply doesn't have. Esteva et al. (2017) reported comparable success using transfer learning at a much larger scale — 129,450 clinical images, dwarfing HAM10000's roughly 10,000 — and achieved performance on par with dermatologists across several diagnostic tasks. Scale, in other words, appears to matter quite a bit, and our results are arguably as much a reflection of dataset size and quality as they are of architectural choice.

Lee et al. (2022), meanwhile, took a somewhat different route — designing Cancer-Net SCa from the ground up, specifically tailored to lean on diagnostically meaningful features rather than incidental visual artifacts, and reported strong performance across multiple datasets. That's a useful counterpoint to our own approach. We relied on off-the-shelf ImageNet-pretrained backbones, which were never designed with skin lesions in mind; a purpose-built architecture, tuned specifically to what actually distinguishes melanoma from a benign nevus, may simply have an inherent advantage that transfer learning from a generic image domain can't fully close. Reviews by Dildar et al. (2021) and Li et al. (2021) reach a broadly similar conclusion — deep learning clearly works for skin cancer detection, but the specific methodology, and not just the raw depth of the network, seems to matter enormously.

4.3 The Dataset Problem

We keep circling back to the dataset, and that's not accidental — we think it's genuinely central to interpreting these results correctly. HAM10000 (Tschandl et al., 2018),

for all its value as one of the more accessible public resources in this space, is still comparatively small next to datasets used in other domains, and it's also skewed quite heavily — melanocytic nevi make up roughly two-thirds of the images, a class imbalance visible in [Figure 3]. This isn't a problem unique to our study. Several other groups working with dermatoscopic datasets, including those relying on the ISIC Archive (ISIC Archive, n.d.; Milton, 2019; Dildar et al., 2021; Albahar, 2019), have run into a related and arguably more troubling issue: these datasets skew heavily toward images of light-skinned patients, with comparatively little representation of darker skin tones. Chabi Adjobo et al. (2022) attempted to address this directly, proposing a skin-tone augmentation method to help rebalance datasets along this dimension. We didn't apply any such correction here, and so it's fair to say our findings — like a fair amount of work in this space — may not generalize cleanly across the full diversity of patients these tools would eventually need to serve. That's a limitation worth naming plainly rather than tucking away.

There's also the more mundane issue we noted earlier: duplicate images within HAM10000 itself. We can't say with precision how much this affected training or the metrics we report, but it's the kind of detail that, left unaddressed, quietly erodes confidence in the numbers — and we'd rather flag it here than let it pass unremarked.

4.4 Why MobileNet's Result Matters Practically

Setting aside architecture-versus-architecture bragging rights for a moment, there's a more practical angle to MobileNet's performance that we think deserves emphasis. Howard et al. (2017) designed MobileNet specifically for deployment on resource-constrained devices — the kind of hardware you'd actually find in a smartphone, not a research lab's GPU cluster. That MobileNet not only kept pace with, but actually outperformed, four considerably heavier architectures in this study is a genuinely encouraging sign for the broader ambition motivating this line of work: building AI-assisted skin cancer screening tools that are cheap enough, and light enough, to run directly on the devices people already carry around. If similar results hold up on larger and more balanced datasets, that has real implications for expanding low-cost, accessible screening — particularly in settings where dermatological expertise or dermoscopy equipment isn't readily available.

4.5 Limitations and What They Mean for Interpretation

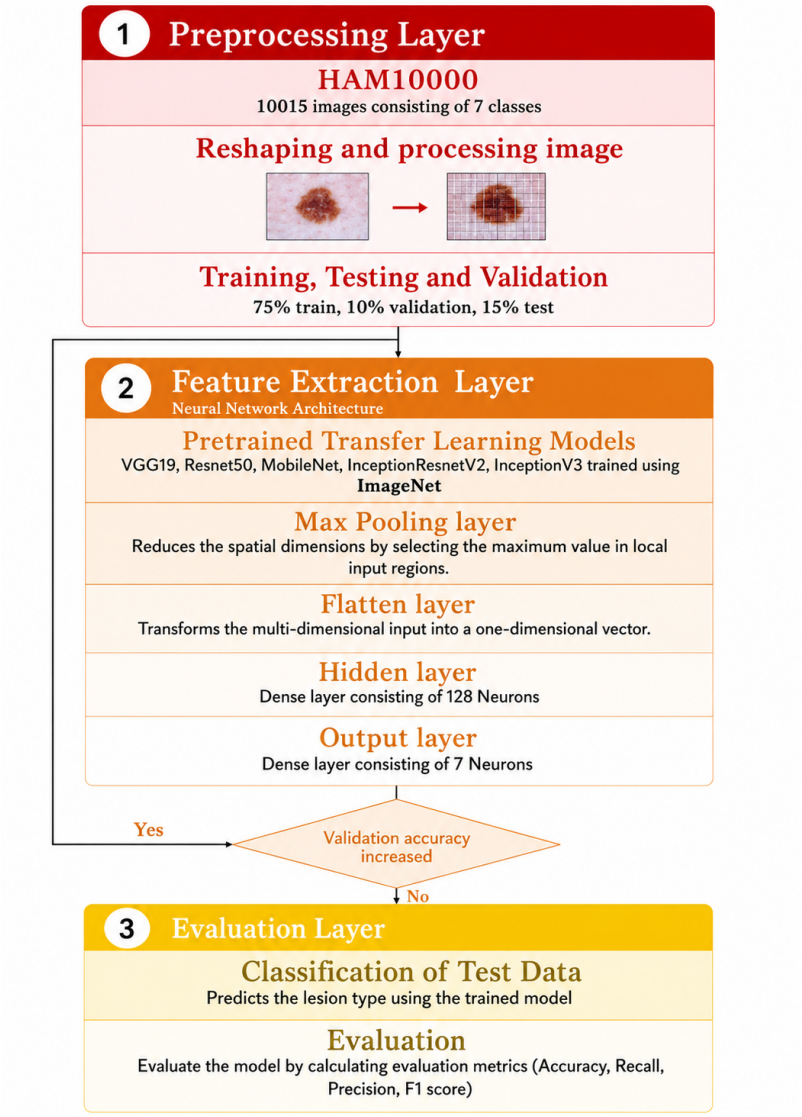


Figure 2. Proposed transfer learning workflow for skin cancer classification, showing the three-stage pipeline — preprocessing, feature extraction via a frozen pre-trained backbone with an appended classification head, and evaluation on the held-out test set.

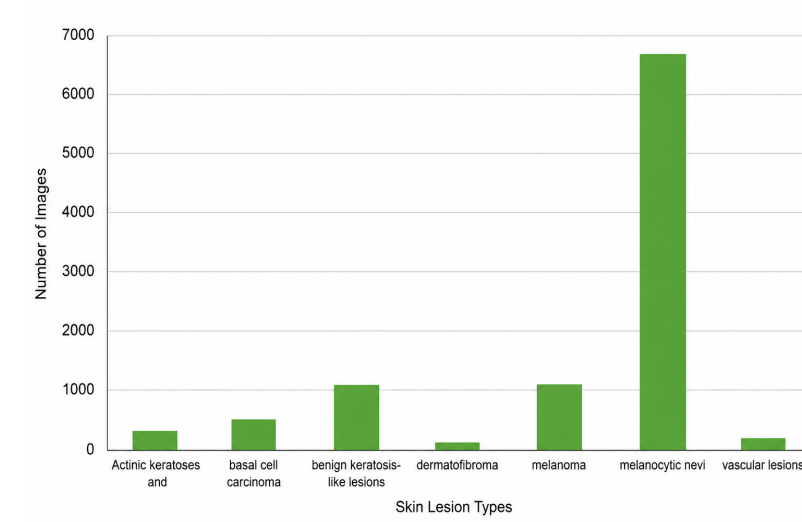


Figure 3. Class distribution of the HAM10000 dataset across the seven diagnostic categories (actinic keratoses/intraepithelial carcinoma, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions), highlighting the marked imbalance toward melanocytic nevi

We'd be doing this work a disservice if we didn't spell out its limitations candidly. First, the dataset's class imbalance and its skew toward lighter skin tones both constrain how confidently these findings can be generalized. Second, the presence of duplicate images, which we did not systematically filter or quantify, introduces some uncertainty into the reported metrics. Third, we relied entirely on frozen, ImageNet-pretrained backbones without further fine-tuning — an approach that's computationally convenient but may understate what these architectures are truly capable of on this task. Taken together, these limitations suggest the comparative rankings reported here should be read as informative rather than definitive — a reasonable starting point for further investigation, not a closing argument.

5. Conclusion

In this comprehensive study, we conducted a thorough comparative analysis of transfer learning models for the vital task of skin cancer classification. Our investigation encompassed the evaluation of five distinct transfer learning models, each demonstrating unique performance characteristics. Remarkably, the MobileNet model emerged as a standout performer, exhibiting robust skin cancer detection capabilities while maintaining an impressive equilibrium between computational efficiency and accuracy. As our planet's ozone layer continues to deplete, the incidence of skin cancer is poised to rise. It is imperative that we develop cost-effective and efficient skin cancer screening tools to address this growing concern. This research underscores the viability of employing transfer learning methodologies for accurate skin cancer detection, offering a compelling alternative to custom-designed CNN models. Moreover, there is potential for further refinement in the future. Exploring additional transfer learning models, fine-tuning existing ones, or incorporating augmented data or newer datasets could potentially elevate the model's accuracy even further.

Author Contribution

M.N.S. conceived and designed the study, implemented and trained the VGG19, ResNet50, MobileNet, Inception-ResNetV2, and InceptionV3 transfer learning models, performed the comparative analysis, and drafted the manuscript. K.A.A.M. supervised the study, provided critical guidance on methodology, and critically reviewed and revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgement

The authors M.N.S. et al., would like to thank the creators of the HAM10000 dataset for making it publicly available for research purposes. The authors also acknowledge the institutional support that facilitated this work.

Competing Financial Interests

The authors M.N.S. et al., declare no competing financial interests.

References

- Albahar, M. A. (2019). Skin lesion classification using convolutional neural network with novel regularizer. *IEEE Access*, 7, 38306–38313.
- Amelard, R., Glaister, J., Wong, A., & Clausi, D. A. (2014). High-level intuitive features (HLIFs) for intuitive skin lesion description. *IEEE Transactions on Biomedical Engineering*, 62(3), 820–831.
- Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Haskova, Z., & Prunotto, M. (2019). Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digital Medicine*, 2(1), 92.
- Ashraf, R., Afzal, S., Ur Rehman, A., Gul, S., Baber, J., Bakhtyar, M., Mehmood, I., Song, O.-Y., & Maqsood, M. (2020). Region-of-interest based transfer learning assisted framework for skin cancer detection. *IEEE Access*, 8, 147858–147871.
- Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., & Moss, R. H. (2007). A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, 31(6), 362–373.
- Celebi, M. E., Wen, Q., Iyatomi, H., Shimizu, K., Zhou, H., & Schaefer, G. (2015). A state-of-the-art survey on lesion border detection in dermoscopy images. In *Dermoscopy Image Analysis* (Vol. 10, No. 1, pp. 97–129).
- Chabi Adjobo, E., Sanda Mahama, A. T., Gouton, P., & Tossa, J. (2022). Towards accurate skin lesion classification across all skin categories using a PCNN fusion-based data augmentation approach. *Computers*, 11(3), 44.
- Dildar, M., Akram, S., Irfan, M., Khan, H. U., Ramzan, M., Mahmood, A. R., Alsaiani, S. A., Saeed, A. H. M., Alraddadi, M. O., & Mahnashi, M. H. (2021). Skin cancer detection: A review using deep learning techniques. *International Journal of Environmental Research and Public Health*, 18(10), 5479.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.
- Glaister, J. L. (2013). Automatic segmentation of skin lesions from dermatological photographs [Master's thesis, University of Waterloo].
- Gunraj, H., Wang, L., & Wong, A. (2020). COVIDNet-CT: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images. *Frontiers in*

- Medicine, 7, 608525.
- Hagerty, J. R., Stanley, R. J., Almubarak, H. A., Lama, N., Kasmi, R., Guo, P., Drugge, R. J., Rabinovitz, H. S., Oliviero, M., & Stoecker, W. V. (2019). Deep learning and handcrafted method fusion: Higher diagnostic accuracy for melanoma dermoscopy images. *IEEE Journal of Biomedical and Health Informatics*, 23(4), 1385–1391.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
- Hoffmann, K., Gambichler, T., Rick, A., Kreutz, M., Anschuetz, M., Grunendick, T., Orlikov, A., Gehlen, S., Perotti, R., Andreassi, L., et al. (2003). Diagnostic and neural analysis of skin cancer (DANAOS): A multicentre study for collection and computer-aided analysis of data from pigmented skin lesions using digital dermoscopy. *British Journal of Dermatology*, 149(4), 801–809.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- International Skin Imaging Collaboration. (n.d.). ISIC Archive. Retrieved from <https://www.isic-archive.com/>
- Kumar, D., Wong, A., & Clausi, D. A. (2015). Lung nodule classification using deep features in CT images. In *2015 12th Conference on Computer and Robot Vision* (pp. 133–138). IEEE.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lee, J. R. H., Pavlova, M., Famouri, M., & Wong, A. (2022). Cancer-Net SCa: Tailored deep neural network designs for detection of skin cancer from dermoscopy images. *BMC Medical Imaging*, 22(1), 1–12.
- Li, H., Pan, Y., Zhao, J., & Zhang, L. (2021). Skin disease diagnosis with deep learning: A review. *Neurocomputing*, 464, 364–393.
- Matthews, N. H., Li, W.-Q., Qureshi, A. A., Weinstock, M. A., & Cho, E. (2017). Epidemiology of melanoma. In *Exon Publications* (pp. 3–22).
- Milton, M. A. A. (2019). Automated skin lesion classification using ensemble of deep neural networks in ISIC 2018: Skin lesion analysis towards melanoma detection challenge. *arXiv preprint arXiv:1901.10802*.
- Rezvantalab, A., Safigholi, H., & Karimijeshni, S. (2018). Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. *arXiv preprint arXiv:1810.10348*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 211–252.
- Saurat, J. H. (2004). Dermoscopy of pigmented lesions: A valuable tool in the diagnosis of melanoma. *Swiss Medical Weekly*, 134(0708), 83–90.
- Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, 68(1), 7–30.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31).
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826).
- 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(1), 1–9.
- Wang, S., Yin, Y., Wang, D., Wang, Y., & Jin, Y. (2021). Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. *IEEE Transactions on Cybernetics*, 52(12), 12623–12637.