Deep Learning for Skin Cancer Detection: A Practical Framework for Automated Skin Lesion Classification

Debabrata Pruseth
AI Architect & Applied AI Researcher
Singapore

Author Note

This article is a research-style companion version of the author's blog post "How I built a beginner-friendly skin-cancer detector" and the associated GitHub project "HAM10000_SkinCancer_AI"

Abstract

Automated skin lesion classification has become an important research area within medical artificial intelligence due to the increasing availability of annotated dermatoscopic image datasets and advances in convolutional neural networks. Early detection of malignant skin lesions, particularly melanoma, can improve clinical outcomes; however, the development of reliable automated systems requires careful attention to dataset quality, class imbalance, validation design, model selection, and responsible-use boundaries. This paper presents a practical deep learning framework for automated skin lesion classification using the HAM10000 dataset and transfer learning with EfficientNetB0. The objective is not to propose a clinically deployable diagnostic system, but to demonstrate a reproducible, beginner-accessible, and methodologically disciplined workflow for medical image classification.

The framework is based on the author's blog article, "How I built a beginner-friendly skin-cancer detector," and the accompanying annotated GitHub notebook. The implementation uses Google Colab, TensorFlow, GPU acceleration, EfficientNetB0 pretrained on ImageNet, lesion-aware train-validation splitting, image preprocessing, data augmentation, class weighting, and supervised multi-class classification across seven lesion categories. The HAM10000 dataset contains 10,015 dermatoscopic images of common pigmented skin lesions and was introduced to address the limited availability of large public datasets for training neural networks in dermatology imaging.

The proposed framework emphasizes several practices that are especially important in medical AI education: exploratory metadata analysis, explicit handling of class imbalance, prevention of data leakage through lesion-level grouping, use of transfer learning to reduce training complexity, and cautious interpretation of performance metrics. EfficientNetB0 is

selected because EfficientNet models were designed around compound scaling of network depth, width, and resolution, providing a practical balance between accuracy and computational efficiency for image classification tasks.

The study concludes that beginner-friendly medical AI projects should go beyond model training and include methodological safeguards, ethical boundaries, and transparent limitations. The framework provides a practical foundation for learners to understand how automated lesion classification systems are built, while reinforcing that such systems require external validation, explainability, calibration, bias assessment, and clinical governance before any medical use.

---

1. Introduction

Skin cancer is a major public health concern, and visual examination of skin lesions remains central to early recognition and referral. In dermatology, dermatoscopic imaging provides enhanced visualization of pigmented lesions and has become an important modality for both clinical assessment and machine learning research. The increasing availability of public dermatoscopic datasets has made it possible for students, researchers, and practitioners to experiment with automated classification models. However, the apparent simplicity of image classification can be misleading. Medical image classification introduces challenges that are not always present in general computer vision tasks, including class imbalance, uncertain labels, patient-level leakage, demographic bias, image acquisition variability, and the risk of inappropriate clinical interpretation.

The author’s project was developed to make this learning pathway more accessible. The blog explains the implementation as a beginner-friendly skin cancer detector that classifies skin lesion images into seven categories using a trained deep learning model. It explicitly states that the tool is educational and not a substitute for medical advice. This distinction is important because automated lesion classification systems can appear authoritative to non-specialists even when they are only experimental prototypes.

This paper reframes the project as a structured applied research study. The central aim is to describe a practical framework for building a supervised deep learning classifier for skin lesion images using the HAM10000 dataset. The emphasis is not only on model construction, but also on the surrounding methodological decisions that determine whether an experiment is meaningful. These include how the dataset is inspected, how labels are encoded, how images are split into training and validation sets, how class imbalance is addressed, how transfer learning is applied, and how model outputs should be interpreted.

The study is motivated by a broader concern in AI education. Many beginner tutorials demonstrate how to train an image classifier, but fewer explain the special responsibilities associated with medical datasets. A model that performs reasonably on a validation split may still be unsuitable for clinical use if the validation split contains leakage, if minority classes are poorly detected, if the model is not externally validated, or if performance is not examined across patient subgroups. Therefore, the proposed framework is intentionally educational but also methodologically cautious.

2. Related Work

The HAM10000 dataset is one of the most widely used public datasets for pigmented skin lesion classification. Tschandl, Rosendahl, and Kittler introduced HAM10000 as a collection of 10,015 dermatoscopic images acquired from multiple sources and covering seven diagnostic categories of pigmented lesions. The dataset was designed to support machine learning research and comparisons with human experts, and the original paper notes that more than half of the lesions were confirmed by histopathology, while the remaining cases were supported by follow-up, expert consensus, or confocal microscopy.

Deep learning has been widely applied to this dataset and related dermoscopy benchmarks. Earlier approaches often used convolutional neural networks such as ResNet, DenseNet, Inception, and EfficientNet variants. These models learn hierarchical visual features from images, beginning with low-level edges and textures and progressing toward higher-level lesion structures. In the context of limited medical datasets, transfer learning has become a common strategy because models pretrained on large natural-image datasets can provide useful feature representations that are adapted to medical domains through fine-tuning.

EfficientNet is particularly relevant for resource-aware learning environments. Tan and Le proposed EfficientNet as a family of convolutional neural networks based on compound scaling, where network depth, width, and image resolution are scaled together in a principled manner rather than independently. This design produced strong accuracy-efficiency tradeoffs and made EfficientNet architectures attractive for transfer learning tasks where computational resources are limited.

The present study differs from performance-maximizing papers that aim to achieve state-of-the-art accuracy on HAM10000. Its primary contribution is pedagogical and methodological: it demonstrates how a learner can build an end-to-end lesion classification pipeline while being introduced to core medical AI concepts such as leakage prevention, imbalance handling, validation design, and responsible communication. In this sense, the work is closer to an applied framework than a benchmark competition entry.

3. Dataset

The study uses the HAM10000 dataset, formally titled *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*. The dataset contains 10,015 dermatoscopic images and seven diagnostic categories. Its public availability has made it a common benchmark for educational and research experiments in automated skin lesion classification.

The seven diagnostic labels used in the implementation are actinic keratoses and intraepithelial carcinoma (`akiec`), basal cell carcinoma (`bcc`), benign keratosis-like lesions (`bkl`), dermatofibroma (`df`), melanocytic nevi (`nv`), melanoma (`mel`), and vascular lesions (`vasc`). The metadata file contains image identifiers, lesion identifiers, diagnosis labels, diagnostic confirmation type, age, sex, and anatomical localization. These metadata attributes

are valuable not only for training but also for understanding possible dataset biases and subgroup distribution.

A key property of HAM10000 is class imbalance. The melanocytic nevi category dominates the dataset, while categories such as dermatofibroma and vascular lesions contain far fewer examples. This imbalance creates a methodological risk because a model optimized only for overall accuracy may learn to favor the majority class. In a medical context, such behavior is problematic because clinically important minority classes may be missed even when aggregate performance appears acceptable.

Another important property is that multiple images may correspond to the same lesion. If a conventional random image-level split is used, images of the same lesion can appear in both training and validation sets. This creates leakage because the model may be evaluated on visually similar examples that are not independent of the training data. The notebook addresses this risk by using lesion-level grouping during the train-validation split.

---

4. Methodology

The proposed framework follows a supervised multi-class image classification design. The author implemented the pipeline in Google Colab using TensorFlow and GPU acceleration. The workflow begins with loading the dataset from Google Drive, extracting the image folders, merging image files into a single directory, and reading the HAM10000 metadata into a structured dataframe. The implementation then performs exploratory analysis to verify dataset shape, class distribution, demographic attributes, and anatomical localization patterns.

After exploratory analysis, the framework maps diagnostic labels into numeric classes and constructs full image paths for each sample. Images are resized to 224 × 224 pixels to match the expected input dimensions of EfficientNetB0. The preprocessing pipeline decodes image files, converts them into RGB tensors, resizes them, applies EfficientNet-specific preprocessing, and pairs each image with a one-hot encoded class label.

A central methodological decision is the use of lesion-aware splitting. The implementation uses the lesion identifier as a grouping variable so that all images belonging to the same lesion remain in the same split. This reduces the likelihood that the validation set contains near-duplicate information from the training set. In educational terms, this step is especially important because it demonstrates that validation design is not a mechanical afterthought but a core part of scientific model evaluation.

The training pipeline uses TensorFlow's `tf.data` API for efficient batching and prefetching. Training images are augmented through simple transformations such as horizontal flipping, vertical flipping, and brightness variation. These augmentations are intended to improve robustness by exposing the model to plausible variations in image orientation and illumination. Because the task involves medical images, augmentation is intentionally conservative; transformations that could distort clinically relevant lesion structure are avoided.

The model architecture uses EfficientNetB0 pretrained on ImageNet as a feature extractor. The original classification head is removed and replaced with a global average pooling layer, dropout regularization, and a dense softmax layer with seven output units. Initially, the EfficientNetB0 base is frozen so that the newly added classification head can learn dataset-specific mappings without destabilizing the pretrained representation. This is a standard transfer learning approach and is particularly suitable for beginner-friendly experimentation because it reduces training cost and complexity.

Class imbalance is addressed through class weighting. The implementation computes balanced class weights from the training labels so that minority classes contribute more strongly to the loss function. This does not eliminate imbalance, but it reduces the tendency of the model to optimize primarily for the dominant class. The model is compiled using categorical cross-entropy loss and monitored using accuracy, AUC, precision, and recall. The inclusion of recall is important because, in medical screening-related contexts, missed positive cases are often more concerning than false alarms.

Training is managed using callbacks. Model checkpointing saves the best-performing model based on validation AUC. Early stopping prevents unnecessary training after validation performance stops improving. Learning-rate reduction on plateau allows optimization to continue more cautiously when improvement slows. Together, these callbacks provide a disciplined training procedure suitable for an educational medical AI experiment.

---

5. Experimental Design

The experimental design is intentionally practical rather than benchmark-maximizing. The objective is to create a reproducible baseline pipeline that learners can inspect, run, and extend. The implementation choices reflect this goal.

The use of Google Colab lowers the barrier to entry by removing the need for local GPU configuration. The use of EfficientNetB0 balances accuracy and computational cost. The use of `tf.data` introduces learners to scalable input pipelines. The use of class weights introduces the problem of imbalanced medical datasets. The use of lesion-aware splitting introduces the concept of data leakage and patient- or lesion-level independence.

The classification problem is framed as seven-class supervised learning. Given an input dermatoscopic image x, the model estimates a probability distribution over seven lesion classes:

$$p(y \mid x) = \text{softmax}(f_\theta(x))$$

where $f_{\theta(x)}$ theta represents the neural network parameterized by \theta, and y is one of the seven diagnostic classes. The model is trained by minimizing categorical cross-entropy:

$$\mathcal{L} = -\sum_{c=1}^{7} y_c \log(\hat{y}_c)$$

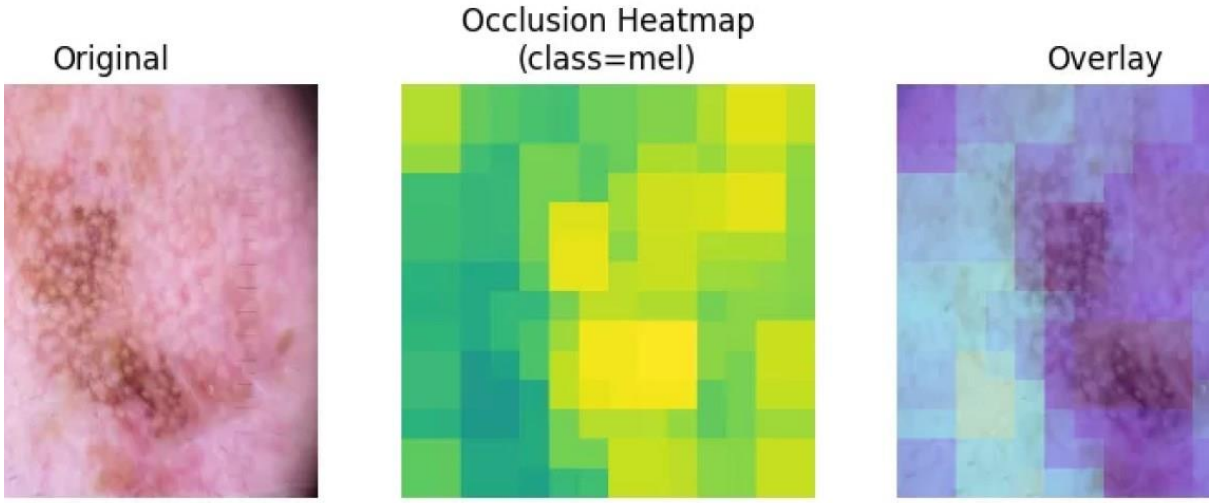
With class weighting, the loss contribution is adjusted so that underrepresented classes receive larger weights:

$$\mathcal{L}_{weighted} = -\sum_{c=1}^{7} w_c y_c \log(\hat{y}_c)$$

This formulation is appropriate for a baseline multi-class classification framework. However, it does not by itself guarantee clinical usefulness. Clinical relevance would require careful threshold selection, calibrated probabilities, external validation, and dermatologist-supervised evaluation.

---

6. Results and Evaluation Approach



The notebook demonstrates that the training process proceeds successfully under GPU acceleration and that the model learns from the dataset. Early training logs show improvement in validation AUC during the initial epochs, indicating that the transfer learning pipeline is able to extract useful predictive signal from the dermatoscopic images. However, the goal of this research-style writeup is not to overstate performance based on internal validation. Instead, the focus is on establishing an evaluation approach appropriate for medical AI learning.

For imbalanced multi-class medical classification, overall accuracy is insufficient. Accuracy can be dominated by the majority class, especially when one category accounts for a large share of the dataset. A model that performs well on melanocytic nevi but poorly on

melanoma or rare lesion classes may still appear acceptable if only accuracy is reported. Therefore, evaluation should include per-class precision, recall, F1-score, AUC, and confusion matrix analysis.

Recall is particularly important for malignant or clinically significant classes because false negatives may delay clinical review. Precision is also important because excessive false positives can increase unnecessary anxiety and clinical workload. AUC provides a threshold-independent measure of separability, but it should be interpreted alongside class-specific metrics. Confusion matrices are useful because they reveal which lesion types are most frequently confused.

The framework should also include calibration analysis. Medical AI systems should not only assign the correct class; their confidence scores should be meaningful. A model that predicts melanoma with 95% confidence should be correct much more often than a model that gives the same confidence indiscriminately. Calibration methods such as reliability diagrams, expected calibration error, and temperature scaling can be incorporated in future iterations.

External validation is also essential. A model evaluated only on a split from the same dataset may not generalize to images from different devices, clinical settings, populations, or acquisition conditions. Since HAM10000 consists of dermatoscopic images, performance on ordinary smartphone images should not be assumed. This is one reason the model must remain clearly labeled as an educational tool rather than a diagnostic product.

---

7. Discussion

The project demonstrates that a beginner-friendly implementation can still incorporate important elements of rigorous medical AI practice. The strongest aspect of the framework is that it does not treat model training as the only objective. Instead, it introduces learners to the broader experimental discipline required for medical image classification.

The lesion-aware split is especially important. In many beginner projects, data is split randomly at the image level. This can inflate validation results when correlated images from the same lesion appear in both training and validation sets. By grouping samples by lesion identifier, the implementation moves closer to a realistic evaluation design. This choice reflects an understanding that the independence of validation data matters.

The use of class weighting is also appropriate. HAM10000 is heavily imbalanced, and the majority class can dominate optimization. Weighting the loss function does not fully solve imbalance, but it is a practical first step. More advanced methods such as focal loss, balanced batch sampling, oversampling, targeted augmentation, and threshold tuning could be explored in later versions.

EfficientNetB0 is a reasonable architecture for the project's objectives. It is computationally efficient, well supported in TensorFlow, and suitable for transfer learning. The original EfficientNet work showed that compound scaling can improve accuracy-efficiency tradeoffs, which is helpful for Colab-based experimentation and beginner learning environments.

The project also illustrates a broader principle: educational medical AI should teach caution, not just capability. A classifier that produces a label for a skin lesion can easily be misunderstood by users as a diagnostic assistant. For this reason, the author's disclaimer that the model is not medical advice is not a minor note; it is part of the responsible design of the project. Automated lesion classification should be communicated as a learning exercise unless validated through appropriate clinical, regulatory, and operational processes.

---

8. Limitations

First, the implementation uses a public dermatoscopic dataset, and dermatoscopic images differ from ordinary consumer photographs. The model should not be assumed to work on smartphone images, images taken under uncontrolled lighting, or images outside the HAM10000 distribution.

Second, the validation is internal to the available dataset. Although lesion-aware splitting reduces leakage, it does not provide the same evidence as external validation on independent datasets from different clinical sources.

Third, the dataset is imbalanced. Class weighting helps but does not guarantee strong performance on minority classes. Future work should report detailed per-class metrics and examine whether clinically important minority categories are reliably detected.

Fourth, the current framework is primarily image-based. Metadata such as age, sex, and lesion localization is analyzed but not deeply integrated into the predictive model. Future multimodal approaches could combine image and metadata features, but this would require additional care around bias, missingness, and interpretability.

Fifth, the model is not calibrated or clinically validated. Probability outputs from neural networks may not reflect true clinical risk. Any use beyond education would require calibration, threshold analysis, prospective evaluation, expert review, and regulatory assessment.

Sixth, explainability is not yet fully integrated into the reported workflow. Although the notebook includes libraries that could support explainability, future work should generate and analyze Grad-CAM or similar heatmaps to determine whether the model focuses on lesion regions rather than artifacts such as rulers, color charts, borders, or imaging noise.

---

9. Future Work

Future work should extend the framework in five directions. First, the model should be evaluated with a full per-class classification report, confusion matrix, macro-averaged F1-score, balanced accuracy, and class-specific recall for melanoma and other clinically significant classes. Second, external validation should be performed using an independent dermatoscopic dataset to evaluate generalization. Third, explainability methods such as Grad-CAM should be added to visualize the image regions that contribute most to predictions.

Fourth, calibration should be assessed so that predicted probabilities can be interpreted more reliably. Fifth, the framework should explore additional architectures and imbalance strategies, including EfficientNet variants, DenseNet, MobileNet, focal loss, and balanced sampling.

For educational use, future versions could also include an interactive notebook section that allows learners to upload a sample image while displaying a strong disclaimer that the output is not diagnostic. Such a demo should avoid medical claims and should encourage users to consult a qualified clinician for any suspicious lesion.

---

10. Conclusion

This paper presented a practical deep learning framework for automated skin lesion classification using the HAM10000 dataset and EfficientNetB0 transfer learning. The framework was developed as an educational implementation based on the author's blog and annotated GitHub notebook. It demonstrates how learners can build a medical image classification pipeline while being introduced to important research practices such as exploratory data analysis, lesion-aware validation splitting, class imbalance handling, transfer learning, data augmentation, and responsible-use communication.

The main contribution of the work is not a claim of clinical-grade diagnostic performance. Rather, it is a structured and reproducible learning framework that shows how medical AI experimentation should be approached with methodological caution. The project reinforces that successful medical AI development requires more than training a model. It requires careful data design, validation discipline, clinically meaningful evaluation, transparency, bias awareness, and explicit boundaries around use.

The study concludes that beginner-friendly AI education can and should incorporate research-grade habits. By presenting a practical yet cautious framework for skin lesion classification, the project helps bridge the gap between introductory deep learning tutorials and responsible medical AI experimentation.

References

Pruseth, D. (2025). *How I built a beginner-friendly skin-cancer detector*. Blog article.

Pruseth, D. (2025). *HAM10000 SkinCancer AI: Annotated Notebook*. GitHub notebook.

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.

Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. Proceedings of the 36th International Conference on Machine Learning.

Suggested Citation

Pruseth, D. (2025). *Deep Learning for Skin Cancer Detection: A Practical Framework for Automated Skin Lesion Classification*.