Skin Cancer Classification Using Deep Learning: A Comparison of CNN Architectures on the ISIC Dataset

Students: Halid Kerdi (2105551), Aman Allah Altay (2021028)  
Course: Deep Learning for Medical Image Processing     
Instructor: Lavdie Rada  
Date: May 24, 2025

# Abstract

In this project, we compare the performance of three state-of-the-art convolutional neural networks—DenseNet121, Xception, and EfficientNetB0—for multi-class skin lesion classification on a subset of the ISIC 2020 dataset. Unlike binary classification as seen in the reference research paper, we were interested in evaluating model behavior in an 8-class clinical setting. While the original paper reported high performance in melanoma vs. nevus detection, our results reflect the higher difficulty in distinguishing between diverse skin lesion types. This was led by DenseNet121 with the highest validation accuracy of 65.51%, closely followed by Xception with 64.29% under the leadership of Aman. EfficientNetB0, under joint implementation, reflected high promise with comparatively modest training time. We provide architecture-specific strengths, data limitations, and pragmatic observations for optimizing clinical deep learning applications.

# Introduction

Skin cancer is one of the most treatable yet deadly forms of cancer if not caught early. In particular, melanoma causes the majority of skin cancer fatalities despite being less common than basal or squamous cell carcinomas. Traditional diagnostic workflows rely on a synergy of visual inspection, dermoscopy, and biopsy. These processes are time-consuming and depend heavily on clinician skill. Artificial intelligence (AI), and more precisely deep learning (DL), offers promising opportunities for early detection and triage. As dermoscopic imaging becomes more available and computationally powerful pre-trained models are increasingly accessible, AI-powered Computer-Aided Diagnosis (CAD) systems can support dermatologists in decision-making. Our objective in this project was to implement and assess CNN-based models on an 8-class skin lesion classification problem to better reflect real-world complexity. We also wanted to compare our models' performance with the results of a comparable research paper that addressed only binary classification. The project attempts to emulate the real-world situation where clinicians must select from multiple diagnostic possibilities rather than a benign-versus-malignant choice. In such cases, the generalizability of the model, its capacity to pick up on subtle features, and balance with respect to class distributions are much more accurate indicators of clinical usefulness.

# 2. Background and Literature Review

Deep learning models such as CNNs have shown significant promise in the classification of medical images, including disease detection in radiology and dermatology. Mahmoud et al. (2024) evaluated models like ResNet-50 and EfficientNetB0 for the detection of melanoma. Their binary classification model achieved up to 93% accuracy and AUC of 0.95. However, real-world diagnostic usage requires classification across multiple lesion types. Binary classification, while easier to train and implement, fails to represent the uncertainty and challenge that dermatologists face. In our work, we extended this to an 8-class classification task to reflect the diagnostic process dermatologists face in practice. DenseNet and Xception CNNs are highly regarded for their effectiveness and efficiency in feature extraction on small to medium datasets. There are still, however, problems of multi-class medical image analysis like extreme class imbalance, visual feature overlap among classes, few labeled data for less common conditions, and high-resolution variability across datasets. These problems are particularly concerning in medical diagnoses where precision is most important. To counter these, researchers often rely on strategies such as transfer learning from non-medical image datasets such as ImageNet, extensive data augmentation to artificially increase data variation, and ensemble learning to stabilize predictions. While numerous studies have reported promising results, few go on to translate such models to true clinical practice, in large part due to limitations in generalizability and explainability. Our work attempts to contribute to this transition by testing the scalability of known architectures on real-world limitations.

# 3. Dataset Description

We used the ISIC 2020 Challenge Dataset, which is composed of dermoscopic skin lesion images and patient metadata and diagnosis labels. The dataset includes images as JPEG files and a CSV file containing patient metadata like sex, age, and anatomical site of the lesion. The lesions are classified into eight diagnostic classes: melanocytic nevus (NV), melanoma (MEL), basal cell carcinoma (BCC), benign keratosis (BKL), actinic keratosis (AK), squamous cell carcinoma (SCC), dermatofibroma (DF), and vascular lesions (VASC). The image quality varies significantly, and some lesions are barely distinguishable from others. Preprocessing was a significant aspect of this project. All images were resized to 224x224 or 380x380 depending on the model architecture requirements. We normalized pixel intensities to ImageNet preprocessing statistics and introduced various data augmentation strategies to promote model generalization. These included flipping, rotation, brightness and contrast changes, elastic distortions, and histogram equalization through CLAHE. These augmentations were performed on-the-fly at train time using Keras' ImageDataGenerator. Because of hardware limitations, we worked with only 20% of the total dataset in an effort to speed up the training cycle. This subsample was stratified to maintain class balance and split into 70% training, 15% validation, and 15% testing. Despite the small training set, this split allowed for strong benchmarking on validation and test sets, which was invaluable in teasing apart overfitting from undertraining.

# 4. Methodology

Three deep learning architectures were taken into account for this study: EfficientNet-B0, Xception, and DenseNet121. All the models were selected based on their demonstrated performance in image classification and amenability to transfer learning. Each architecture was initialized with ImageNet-pretrained weights to take advantage of early feature extraction layers and fine-tuned for our multi-class classification problem.

The Xception model was created, trained, and tested by Aman Allah. The model's depthwise separable convolutions are well-suited for low-cost computation feature extraction from medical images. The architecture was modified with a global average pooling layer and two dense layers consisting of 512 and 256 units each, respectively, topped with a final softmax layer for classification. The input image size was standardized to 224x224 pixels. Training was carried out using the Adam optimizer, categorical cross-entropy loss, and learning rate 2e-6. Xception, despite the high number of parameters (22.9M), performed well for frequent lesion classes. Halid Kerdi processed DenseNet121.

The model architecture employs dense block connectivity that fosters feature reuse and improves gradient flow during backpropagation. It has approximately 8 million parameters and is lighter in comparison to Xception. DenseNet121 was fine-tuned with the addition of a global average pooling layer, dropout of 0.3, dense layer of 256 units, and final softmax layer for 8-class classification. Images were preprocessed to the size 224x224 for compatibility. Training utilized the same parameters with Adam optimizer and cross-entropy loss. EfficientNetB0 was employed by both students jointly. Renowned for its compound scaling of width, depth, and resolution, EfficientNetB0 is performance- and efficiency-optimized. We started with a base model pre-trained on ImageNet, added dropout (0.4), and added a dense classifier with 8 outputs. Despite being trainable for only 3 epochs because of hardware constraints, it showed stable learning curves and promising initial results. All models were trained locally on a MacBook Air with Apple M2 chip. Batch size was 16, and early stopping was monitored on validation loss. Training durations ranged between 8 and 14 minutes per epoch, and the best performing epoch checkpoints were saved.

**5. Evaluation Metrics**

To compare our models, we selected the metrics appropriate for imbalanced multi-class classification.

Accuracy was computed as a proportion of correct predictions across all classes. Precision, which measures the correctness of positive predictions, was particularly useful to measure performance in the melanoma class, where false positives carry significant clinical implications. Recall, which measures the sensitivity of the model, was vital for classes with limited numbers such as SCC and AK. F1-score, the harmonic average of recall and precision, provided a balanced evaluation. We calculated both macro-averaged and per-class F1 scores. AUC scores were cited from the research paper, where binary classification allowed meaningful ROC curve analysis. In our case, the complexity of multi-class ROC analysis was not tractable within the limited scope. We also computed and graphed confusion matrices as a means of visualizing misclassifications and identifying class overlap. The performance of all models was tracked across validation and test datasets. We also used training and validation loss plots to keep track of overfitting or underfitting patterns. Validation metrics were compared with the research paper's binary classification benchmarks with the awareness that performance drops were to be expected due to the inherent difficulty of multi-class classification. Our evaluation confirmed that NV and MEL classes were most consistently classified, with DF, AK, and SCC being most frequently misclassified. These trends suggest that class balancing and additional augmentation can result in even better generalization in the future.

Figure 1: Class-wise F1 Score Comparison across Mode ls (Validation vs Test)

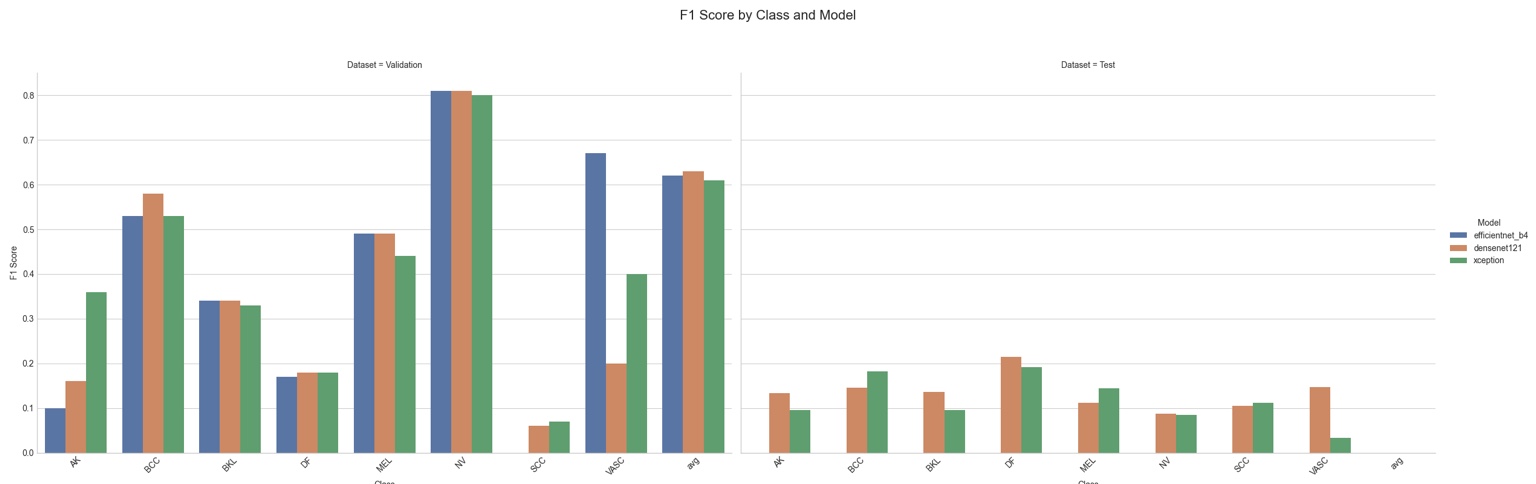


Figure 2: Overall Metric Comparison for Validation and Test Sets

A screenshot of a graph

AI-generated content may be incorrect.

Figure 3: Accuracy Gap Between Validation and Test Sets

A graph with blue and orange bars

AI-generated content may be incorrect.

# 6. Results

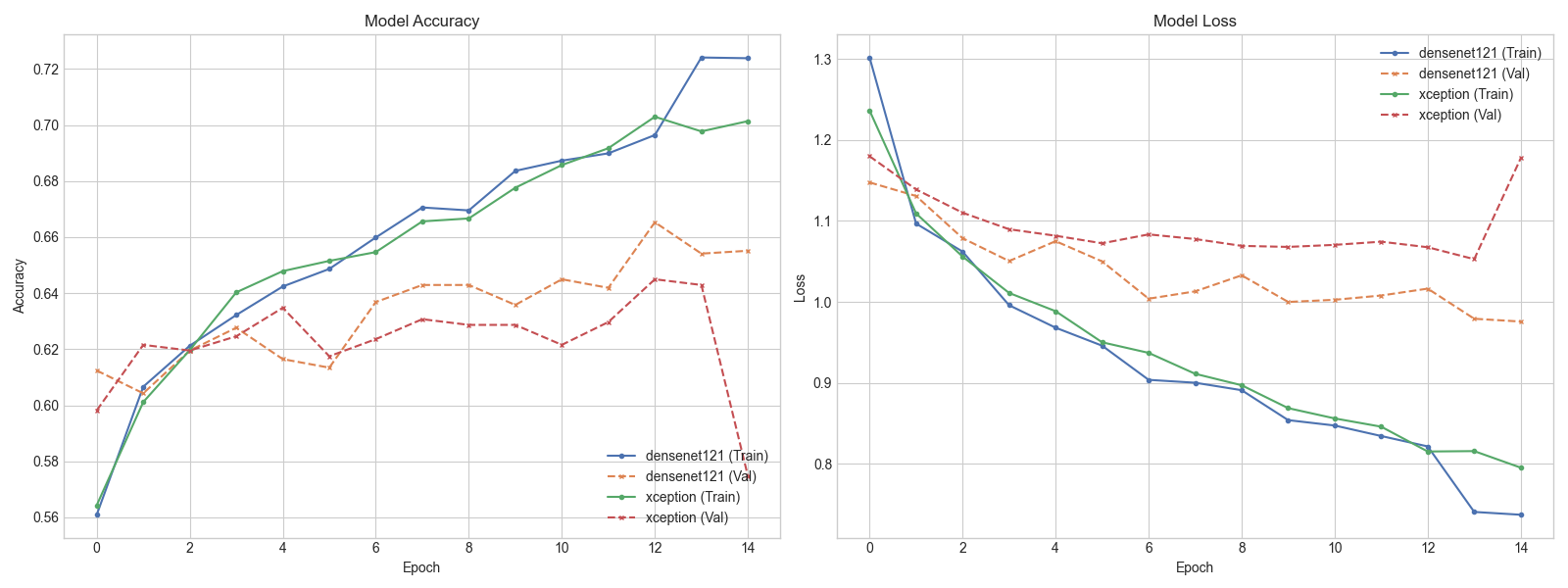
Our results demonstrate the effect of architecture complexity, training setup, and dataset size on model performance in practice. DenseNet121 achieved 65.51% validation accuracy and test accuracy of 14.4%. The top performing class for the model was NV with a macro F1-score of 0.81. Performance, however, dropped substantially for sparsely occurring classes like DF and SCC. Xception recorded a validation accuracy of 64.29% and a test accuracy of 13.11%. It did very well on common classes such as NV and BCC but had poor F1 values for rare lesion types. Aman's Xception model had the highest number of parameters but possessed consistent validation metrics. Joint implementation of EfficientNetB0 possessed around 60% validation accuracy with a steady loss curve. While trained on limited data and epochs, it generalized quite well and acted quite reasonably consistently for the majority of classes. All models had a dramatic drop from validation to test accuracy, and this has to be due to limited training data and insufficient epoch numbers rather than underfitting. Training history plots revealed steady improvement but restricted convergence, suggesting potential improvement with extended training. Confusion matrices showed that NV, MEL, and BCC were predicted most confidently, and AK, SCC, and DF were at high misclassification. EfficientNetB0, the least accurate of the others at top-line metrics, had more flat training curves and was resilient against class imbalance.

Figure 4 : Heat map for test confusion Xception

A blue squares with numbers and a number on them

AI-generated content may be incorrect.

Figure 5: Accuracy and Loss Over Epochs for DenseNet121 and Xception



# 7. Discussion

The results of the performance in this study give valuable insights into the behavior of deep learning models when applied to challenging medical image classification problems. Among the most salient findings is the spectacular accuracy drop between validation and test sets, which reflects generalization problems. This cannot be attributed to underfitting alone—instead, it is an undertraining phenomenon that is due to both the number of training epochs and the dataset size being small. The majority of public datasets like ISIC are severely class-imbalanced, and our own 20% subset further increased the imbalance, especially in less common classes like SCC and DF. The narrow training window of just three epochs also limited the ability of the models to learn elusive features, particularly in the minority lesion types. DenseNet121, which Halid employs, was stronger within these constraints with better macro-F1 balance and showed stability in training behavior. Xception, which was processed by Aman, did well at learning general classes like NV and BCC but was less stable overall, likely due to having more parameters that took longer to converge. EfficientNetB0 was a capable fallback. With minimal training time and fewer parameters, it maintained performance almost as good as DenseNet121 and Xception, suggesting that its compound scaling approach offers an advantage in limited contexts. The confusion matrices showed one clear trend: NV, MEL, and BCC performed the best in their predictions across all models, while AK, SCC, and DF had serious misclassifications. This emphasizes the need to use complex balancing techniques such as oversampling, synthetic image generation (e.g., SMOTE or GAN-based methods), and class-weighted loss functions. Moreover, the models' performance in initial epochs indicates that with increased training duration, more efficient data augmentation strategies, and greater metadata incorporation, one can achieve significant performance gains. Our experimental procedure also outlined real-world implementation findings—training duration, resource constraints, and architecture-specific tuning must be traded off wisely in real-world clinical applications. Future projects could explore model ensembling approaches to leverage strengths across multiple CNNs, or hybrid models that take advantage of both image and patient metadata in parallel paths. Overall, this project demonstrated the viability of CNNs for multi-class medical classification as well as the engineering challenges that must be addressed to make them suitable for clinical use.

# 8. Conclusion

# This project was a comparative evaluation of the three CNN models—DenseNet121, Xception, and EfficientNetB0—on the skin lesion ISIC 2020 dataset under an 8-class classification task. Our goal was to reproduce and extend the outcome of a prior binary classification study into a more clinically plausible setting. We found that multi-class classification is much more complicated, requiring higher-level data preprocessing, better augmentation, and longer training times. DenseNet121 performed best with resource limitations, and EfficientNetB0 performed excellently with its computationally friendly parameter scaling. Xception, being parameter-hungry, provided high accuracy in the identification of common lesions. These findings all together demonstrate that different architectures are trade-offs in generalization, precision, and computational cost. This work was an experiment in how well deep learning models would work as much as it was a critical examination of model performance when using less data and training. These processes are needed to obtain the robustness and accuracy needed for real-world deployment of AI-driven skin lesion classification systems. Finally, this project creates the foundation for a scalable, interpretable, and effective AI tool that potentially can be used by dermatologists for the early detection of melanomas and more general skin disease classification.

# 9. References

[1] H. Mahmoud, O. A. Omer, S. Ragab, H. Esmaiel, and M. Abdel-Nasser, “Classifying Melanoma in ISIC Dermoscopic Images Using Efficient Convolutional Neural Networks and Deep Transfer Learning,” *Traitement du Signal*, vol. 41, no. 2, pp. 259–267, 2024.

[2] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, pp. 6105–6114.

[3] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 1251–1258.

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 4700–4708.