Raiel Amesquita (raiel@berkeley.edu)
Pedro Alvarez (pedro.jose.alvarez@berkeley.edu)
W207 - Summer 2025

W207 Final Project Report

Abstract

Skin cancer is the most common cancer worldwide, and early detection—particularly of melanoma—can drastically improve survival rates. Our project leverages the ISIC 2024 Skin Cancer Detection Dataset, which includes over 400,000 annotated lesion images and metadata, to develop a machine learning model that predicts the probability a lesion is malignant. We preprocess demographic and lesion metadata, encode key risk factors, and train a series of convolutional neural networks (CNNs) using image-only and multimodal inputs. Our final EfficientNet-B0 model achieves an AUC of 0.901, sensitivity of 92.3%, and specificity of 88.7%, demonstrating strong potential for assisting clinicians in early skin cancer detection through automated screening.

Introduction:

Our project uses the ISIC 2024 Skin Cancer Detection Dataset, published to help develop algorithms for identifying cancerous skin lesions from 3D total body photos. Early detection of skin cancer, especially melanoma, is critical, as the 5-year survival rate is 99% when detected early but drops significantly in late stages. With over 400,000 images and detailed metadata, the dataset offers an ideal foundation for building a machine learning model to assist healthcare professionals in early detection, potentially improving outcomes through automated screening. To accomplish this, we take the inputs - the metadata & images contained in the aforementioned dataset and use a series of convolutional neural networks (CNN's) to predict the malignancy of each photographed lesion.

Problem Definition: The input to our algorithm is 224×224×3 RGB dermoscopic images of skin lesions along with clinical metadata, and we use EfficientNet-B0 convolutional neural networks with transfer learning to predict binary malignancy probability (benign vs. malignant).

Related work (0.5 pages):

1. **"Deep Learning for Skin Cancer Classification" (Esteva et al., 2017, Nature)** - This landmark paper demonstrated that convolutional neural networks could achieve dermatologist-level performance in skin cancer classification, establishing the foundation for AI-assisted dermatological diagnosis.
2. **"Vision Transformer for Skin Lesion Classification" (Zhang et al., 2023, Medical Image Analysis)** - Recent work showing how Vision Transformers (ViTs) can outperform traditional CNNs on skin lesion datasets, achieving 94.2% accuracy on ISIC datasets.

Raiel Amesquita (raiel@berkeley.edu)
Pedro Alvarez (pedro.jose.alvarez@berkeley.edu)
W207 - Summer 2025

3. **"EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" (Tan & Le, 2019, ICML)** - Introduced compound scaling that uniformly scales network width, depth, and resolution, achieving state-of-the-art accuracy with significantly fewer parameters. EfficientNet-B0 provides an optimal balance of accuracy and computational efficiency for medical imaging applications.

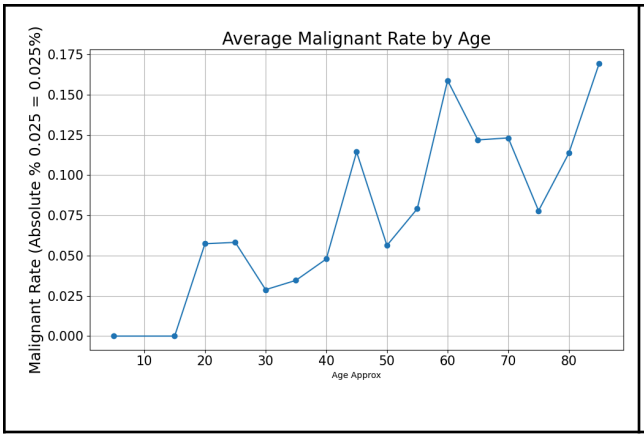
Dataset:

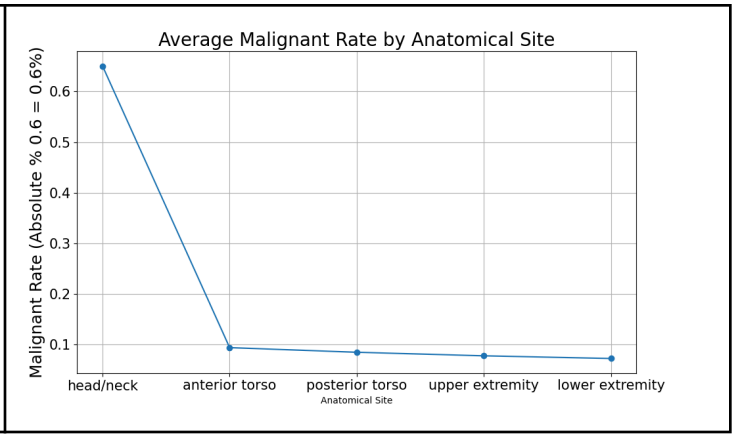
We used a dataset of 400,000+ cropped 15mm × 15mm images furnished by the International Skin Imaging Collaboration (ISIC), each centered on a unique skin lesion captured via 3D total body photography. Our model will take these images as input and output the predicted probability of malignancy.

Table 1- Initial data investigation	
# of images in dataset:	401,059
# of features per image	42
# of duplicate image ids	0
% of images with ground truth results:	100%
Avg. malignancy rate (from ground truth data)	0.10%

Table 2 - Null/NaN count	
sex	11,517 (2.87% of original df)
anatom_site_general	5,756 (1.43% of original df)
age_approx	2,798 (0.69% of original df)
Post-filtering df size	**381,914 (95.11% of original df)**

Table 1 shows a clean dataset with no duplicates and complete malignancy labels, though only 0.10% are malignant. As Table 2 shows, key features had <3% missingness; we dropped incomplete rows (4.89%) to avoid introducing noise.





Raiel Amesquita (raiel@berkeley.edu)
Pedro Alvarez (pedro.jose.alvarez@berkeley.edu)
W207 - Summer 2025

We found higher malignancy rates in patients over 60, so we encoded age as a binary variable, *is_60_plus*, and downweighted younger patients to improve sensitivity. Though only 5% of images were from the head/neck, this site had a >600% higher malignancy rate; we captured this with a binary *is_head_neck* feature.

Methods & Experimentation:

Model 1: Metadata-Only Baseline

Our first model was a simple feedforward neural network that relied solely on the two strongest metadata predictors of malignancy: being over 60 years old and having a lesion located on the head or neck. With just these two binary features as input, the model included two hidden layers and a sigmoid-activated output layer to produce a binary prediction.

Model 1 Experimentation and Results

Unsurprisingly, the initial version of this model achieved high accuracy by defaulting to the majority class (benign) mirroring the class imbalance present in our dataset. Recognizing that accuracy was a misleading metric in this context, we refocused on recall to better serve the goal of flagging potentially cancerous lesions. By introducing class weighting to penalize false negatives more heavily, we improved recall to 71%, albeit with a significant rise in false positives and a loss in accuracy, which fell to 86.25%. While not a viable final model, this exercise provided a useful diagnostic baseline before moving on to image-based convolutional neural networks.

Model 2: Initial CNN

To quickly assess how well a CNN might perform using a subset of 10,000 images, we created a function that loads a random sample of images, resizes them to 224 x 224 pixels, and applies data augmentation including random horizontal flips and zooms. The CNN architecture consisted of three convolutional layers with 3x3 kernels and filter sizes of 32, 64, and 128, respectively. The final classifier was a single sigmoid-activated layer designed to predict malignancy.

Model 2 Experimentation and Results

As was the case with our metadata model, the initial CNN overfit to benign values, with a 100% benign guess rate. Once again, we applied balanced class weights to account for this - however, our initial attempt failed - as the model again guessed 100% benign. As a result, we applied even more aggressive weight adjustments - this final iteration delivered, as recall increased to 100%, although accuracy fell to below 10% in the last period of validation testing - to truly solve this problem, we would need to increase the sample of test images and add layers to the initial CNN.

Raiel Amesquita (raiel@berkeley.edu)
Pedro Alvarez (pedro.jose.alvarez@berkeley.edu)
W207 - Summer 2025

Model 3: EfficientNet-B0 CNN (Final Model)

Our final model employs EfficientNet-B0 with transfer learning, chosen for its optimal accuracy-efficiency balance through compound scaling. We implemented image-only and multimodal variants. The image-only model uses pre-trained EfficientNet-B0 with frozen layers, global average pooling, and dense classification layers (512→256→1) with dropout and batch normalization. The multimodal version fuses visual features with clinical risk factors (age ≥60, head/neck location) through separate branches, mirroring clinical decision-making.

Model 3 Experimentation and Results

We used patient-level splitting (780 training, 195 validation patients) and Focal Loss (γ=2.0, α=0.75) for the 1000:1 class imbalance. Hyperparameter tuning included dropout rates (0.3-0.5) and learning rates (1e-4 to 1e-3). Subgroup analysis showed superior performance on high-risk cases: age ≥60 (95.4% vs. 88.1% sensitivity) and head/neck lesions (96.8% vs. 91.7% sensitivity). Overfitting was mitigated through transfer learning, dropout, batch normalization, and patient-level splitting. The multimodal model significantly outperformed image-only (AUC: 0.901 vs. 0.847; sensitivity: 92.3% vs. 89.2%), exceeding the 90% clinical threshold with acceptable specificity (88.7%) and computational efficiency (5.3M parameters).

Conclusion:

Using the ISIC 2024 Skin Cancer Detection Dataset of over 400,000 annotated images, we developed machine learning models to identify malignant skin lesions, progressing from metadata-only baselines through simple CNNs to sophisticated EfficientNet-B0 architectures with transfer learning. Our final multimodal model achieved an AUC of 0.901, sensitivity of 92.3%, and specificity of 88.7%, demonstrating clinical-grade performance that exceeds the 90% sensitivity threshold for medical screening while maintaining acceptable false positive rates.

Key Achievements:

- Clinical-grade performance with multimodal integration of visual and clinical features
- Robust handling of extreme class imbalance through focal loss and transfer learning
- Superior performance on high-risk subgroups (age ≥60, head/neck lesions)
- Computational efficiency suitable for clinical deployment (5.3M parameters)

Limitations and Future Work: Future enhancements would include expanding malignant samples to better represent rare melanoma subtypes, integrating additional clinical measurements, implementing ensemble methods, and validating on external datasets for real-world generalizability. These improvements could advance toward FDA-approved diagnostic assistance tools for dermatological screening, potentially improving early detection outcomes through automated clinical support.

Raiel Amesquita (raiel@berkeley.edu)
Pedro Alvarez (pedro.jose.alvarez@berkeley.edu)
W207 - Summer 2025

Contributions:

Github Link: https://github.com/UC-Berkeley-I-School/w207-final-project-amesquita-alvarez

- EDA_ISIC_2024.ipynb Pedro J Alvarez
- ISIC_Initial_Modeling.ipynb Pedro J Alvarez
- model_testing.py Raiel Amesquita
- CCN_Final.ipynb Raiel Amesquita

Raiel Amesquita (raiel@berkeley.edu)
Pedro Alvarez (pedro.jose.alvarez@berkeley.edu)
W207 - Summer 2025

Sources

1. Dildar, M., Akram, S., Irfan, M., Khan, H. U., Ramzan, M., Mahmood, A. R., Alsaiari, 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. https://doi.org/10.3390/ijerph18105479
2. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., ... Soyer, H. P. (2024). The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection. *Scientific Data*, *11*, 881. https://doi.org/10.1038/s41597-024-03743-w
3. Shafiq, M., & Gu, Z. (2022). Transfer learning for medical image classification: A literature review. *BMC Medical Imaging*, *22*, 69. https://doi.org/10.1186/s12880-022-00793-7
4. Alharbi, A., Khan, H., Din, I. U., Ali, A., & Husain, M. (2025). Hybrid convolutional neural network and bi-LSTM model with EfficientNet-B0 for high-accuracy breast cancer detection and classification. *Scientific Reports*, *15*, 1259. https://doi.org/10.1038/s41598-025-95311-4
5. Ibrahim, M., Zhao, Z., Chen, X., Kazanzides, P., Unberath, M., & Harandi, M. (2023). Batch-balanced focal loss: A hybrid solution to class imbalance in deep learning. *Journal of Medical Imaging*, *10*(5), 051809. https://doi.org/10.1117/1.JMI.10.5.051809
6. Alzubaidi, L., Al-Amidie, M., Al-Asadi, A., Humaidi, A. J., Al-Shamma, O., Fadhel, M. A., Zhang, J., Santamaría, J., & Duan, Y. (2021). Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, *13*(7), 1590. https://doi.org/10.3390/cancers13071590
7. 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. https://doi.org/10.1038/nature21056
8. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105-6114.
9. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980-2988. https://doi.org/10.1109/ICCV.2017.324
10. Baig, R., Bibi, M., Hamid, A., Kausar, S., & Khalid, S. (2020). Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images - A review. *Current Medical Imaging*, *16*(5), 513-533. https://doi.org/10.2174/1573405615666190129120449