Skin Cancer Detection using SLICE-3D Data (ISIC 2024)

Giovanni Budi [1] **Andrew Chisolm** [2] **Noah Parker** [2]

Abstract

This paper evaluates a skin cancer detection system aimed at overcoming accessibility challenges in underserved regions. Utilizing the ISIC SLICE-3D dataset, which consists of smartphone-quality images, the system employs EfficientNet architectures to classify skin lesions as malignant or benign. We compare models trained on high-quality 2019 ISIC data, smartphone-quality 2024 ISIC data, and a fine-tuned transfer learning model to assess their performance. Our results demonstrate the promise of transfer learning in enhancing diagnostic accuracy and specificity, while addressing challenges related to class imbalance and data variability.

1. Introduction

Skin cancer remains a global health challenge, with early detection playing a vital role in improving surviving rates. Despite advancements in dermoscopic imaging that have enhanced diagnostic accuracy, the reliance on specialized equipment limits accessibility, particularly in underserved regions. The lack of affordable and practical diagnostic tools prevents early intervention, exacerbating the burden of skin cancer in resource constrained settings. Addressing the gap requires innovative solutions that can adapt to the realities of limited medical infrastructure.

Our project aims to tackle this issue by developing an accessible and robust skin cancer detection system that leverages smartphone-quality images. Using the ISIC SLICE-3D dataset, which features such images, we employ EfficientNet architectures to classify skin lesions as malignant or benign. By comparing models trained on high-quality 2019 ISIC data with those trained on lower-quality 2024 ISIC data, we explore the potential of transfer learning to bridge the gap in diagnostic accuracy between datasets of differing quality. This approach seeks to make advanced diagnostic capabilities more widely available, even in areas lacking specialized imaging equipment.

To achieve this, we followed a systematic methodology. First, we replicated the ISIC 2019 winner's methodology using the 2019 dataset transformed into binary classifications. Next, we tested the model's performance on the 2024 dataset to evaluate the impact of image quality. Finally, we fine-tuned the pre-trained 2019 model on the 2024 dataset to enhance its performance on lower-quality images. This step-by-step approach allowed us to assess the impact of image quality and transfer learning while addressing challenges such as class imbalance and data variability, ultimately contributing to the development of a more inclusive and effective diagnostic system.

2. Data

This paper utilizes two datasets, both used in ISIC challenges. The first is from ISIC's 2019 challenge. The 2019 data is the combination of the BCN_20000 Dataset, the HAM10000 Dataset, and the MSK Dataset (Combalia et al., 2019; Tschandl et al., 2018; Codella et al., 2018). This data was collected from the Hospital Clínic in Barcelona, the Department of Dermatology at the Medical University of Vienna, Austria, the skin cancer practice of Cliff Rosendahl in Queensland, Australia, and the Memorial Sloan-Kettering Cancer Center in New York. The 25,331 total images were captured using dermatoscopes and were labeled into nine different categories, some benign and some malignant. Metadata relating to the demographics of the patient were also included.

The second dataset utilized in this paper is the 3D SLICE dataset, used in ISIC's 2024 challenge. Rather than being captured from a dermatoscope, these images were extracted from 3D total body photographs (Kurtansky et al., 2024). The 400,000 images are cropped to a 15 mm x 15 mm scale, and they closely resemble the quality of a cell phone image. The images were gathered from nine sites in Australia, Europe, and the United States.

[1] Darla Moore School of Business, University of Sourth Carolina, Columbia, South Carolina [2] Molinaroli College of Engineering and Computing, University of Sourth Carolina, Columbia, South Carolina. Correspondence to: Giovanni Budi <gbudi@email.sc.edu>, Andrew Chisolm <chisolmi@email.sc.edu>, Noah Parker <nhparker@email.sc.edu>.

Preliminary work.

3. Related Work

4. Experimentation

4.1. Experimental Setup

To validate the methodology and ensure proper setup, we replicated the original ISIC 2019 paper experiment, using the skin lesion categories provided in the dataset. This process involved carefully aligning file locations, model parameters, and software packages to the specifications detailed in the paper. Once replication was completed, we adapted the dataset to our specific goal: binary classification of malignant vs. benign skin lesions. This enabled a direct comparison of the 2019 and 2024 datasets unde r consistent conditions.

4.2. Data Preperation

The ISIC 2019 dataset initially contained multi-class labels for various skin lesion types. To meet our binary classification objective, we transformed the dataset by mapping each lesion category to either ”malignant” or ”benign.” This transformation was guided by medical domain knowledge and consistent with classifications used in prior research. After transformation, the dataset was restructured and split into trainin g and validation sets.

The ISIC 2024 SLICE-3D dataset, in contrast, already provided binary classification labels, eliminating the need for additional transformation. We directly utilized the dataset for training and evaluation, focusing on addressing its challenges, such as lower image quality and significant class imbalance (1:20 malignant-to-benign ratio).

4.3. Binary Classification Models

We trained three models for binary classification:

4.3.1. ISIC 2019 Model

- Trained exclusively on the 2019 dataset with the binary-transformed labels, leveraging its high-quality images and balanced (1:2 malignant-to-benign) class ratio.

- The model architecture is based on the pretrained EfficientNet-B0 and fine-tuning on the 2019 dataset to adapt to its specific features.

4.3.2. ISIC 2024 Model

- Trained on the 2024 dataset with no label transformation, addressing its significant class imbalance (1:20 ratio) and smartphone-quality images.

- The model architecture follows the same pretrained EfficientNet-B0 backbone and fine-tuning on the 2024 dataset to handle the challenges of lower-quality images and class imbalance.

4.3.3. Fine-Tuned 2019/2024 Model

- This model is based on the best-performing 2019 model, which has been pretrained on the ISIC 2019 dataset and fine-tuned for optimal performance.

- For the 2019/2024 Fine-Tuned Model, we froze certain layers of the 2019 model including the conv_stem, block 0, and block 1 layers, while fine-tuning the higher layers on the ISIC 2024 dataset. This allows the model to retain the general features learned from the 2019 dataset while adapting to the more challenging 2024 dataset.

4.4. Image Preprocessing for Training and Validation

4.4.1. Training

During training, the following transformations are applied to augment the dataset and improve model generalization:

- **Random Resized Crop:** Randomly resizes and crops the image to a specified size, improving the model’s robustness to different scales.

- **Random Flip:** Random horizontal and vertical flips to increase the variety of image orientations. Rotation: Random rotation (within a specified range) to improve the model’s invariance to rotation.

- **ColorJitter:** Random adjustments to the brightness, contrast, saturation, and hue of the image to simulate different lighting conditions.

- **Cutout:** Randomly masks out a portion of the image, forcing the model to focus on less prominent features and improving generalization.

- **Normalization:** The pixel values are normalized using the dataset’s mean and standard deviation to standardize the input for the model.

- **ToTensor:** Converts the image to a PyTorch tensor, which is the format required for input into the model.

4.4.2. Validation

During validation, the following transformations are applied to ensure consistent, deterministic behavior while still allowing some degree of variability:

- **CenterCrop:** Crops the image to a fixed size from the center to ensure a consistent region of interest is evaluated.

- **Fixed Crops:** When necessary, crops the image at predefined, deterministic positions for validation.
- **Flipping:** Horizontal flip may be applied for testing, but this is generally less aggressive than in training.
- **Normalization:** Normalizes the image using the same mean and standard deviation as used in training.
- **ToTensor:** Converts the image to a tensor for use in evaluation.

4.5. Training and Validation

Each model was evaluated using a 5-fold cross-validation framework to ensure robustness and generalizability across unseen data. Training parameters, data augmentation techniques, and class weights were carefully optimized to mitigate the challenges posed by the datasets. The Cross-Entropy Loss with Class Weights was used during training to address the class imbalance, particularly in the 2024 dataset, where the malignant -to- benign ratio is 1:20.

Detailed metrics for all models and folds are summarized in Tables 1, 2, and 3, and discussed in subsequent sections. The results highlight the performance trade-offs between datasets, models, and configurations, providing a clear pathway for further optimization in clinical contexts.

The following training parameters were employed to ensure the model trained efficiently while mitigating overfitting and achieving optimal performance:

- Batch Size: 20
- Initial Learning Rate: 0.000015
- Learning Rate Adjustment: After 25 epochs with no improvement, the learning rate was reduced by a factor of 5.
- Training Steps: 60 epochs
- Mean and Standard Deviation for Normalization: Set to [0.0, 0.0, 0.0] and [1.0, 1.0, 1.0] respectively, ensuring the input images were normalized accordingly

These parameters were essential to managing the training process, ensuring stability, and preventing overfitting, especially with the imbalanced data.

4.6. Evaluation Setup

We employed a diverse set of metrics, including accuracy, sensitivity, specificity, AUC, and ROC curve analysis, to comprehensively evaluate our models' performance in predicting these classifications.

- Accuracy: The percentage of correct predictions over the total dataset.
- Sensitivity: The proportion of malignant cases correctly identified.
- Specificity: The proportion of benign cases correctly identified.
- Area Under the Curve (AUC): Measures the trade-off between sensitivity and specificity, evaluating model discrimination.
- ROC Curve Analysis: Graphical representation of model performance across classification thresholds.
- Class-Wise Precision and Recall: Evaluates per-class performance, particularly important for addressing class imbalance.

5. Positive Outcomes

Our evaluation of the 2019, 2024, and fine-tuned 2019/2024 models demonstrates several key strengths in handling varying image qualities and class imbalances:

1. **Improved Performance with Fine-Tuning:** The fine-tuned 2019/2024 model outperformed both the standalone 2019 and 2024 models, achieving a 2.3% increase in accuracy and a 1.4% increase in F1 score. This highlights the value of fine-tuning and transfer learning in adapting the model to the 2024 dataset, even with lower-quality images.
2. **Enhanced Specificity:** The fine-tuned model exhibited a notable improvement in specificity, rising by 2.8compared to the 2024 model. This suggests that the model is more effective at distinguishing between benign and malignant cases, reducing the number of false positives.
3. **Effective Transfer Learning:** The success of the fine-tuned model indicates that leveraging the 2019 model's architecture and weights provided valuable knowledge transfer, enabling the model to better handle the challenges posed by the 2024 dataset with smartphone-quality images.

6. Limitations

Despite these positive outcomes, certain limitations need attention:

1. **High Sensitivity:** Both the 2024 model and the fine -tuned 2019/2024 model exhibited a drop in sensitivity, particularly the fine-tuned model, which showed a

decrease of 5.4%. This could be attributed to the extreme class imbalance in the 2024 dataset (1:20 ratio of malignant to benign cases), which likely hindered the model's ability to detect malignant cases as effectively.

2. **Class Imbalance Impact:** The sensitivity issue underscores the challenge posed by class imbalance, which impacted the model's ability to identify malignant cases, especially in the 2024 dataset. The relatively balanced 1:2 ratio in the 2019 dataset did not face this issue to the same extent, contributing to better sensitivity in that model.

These results suggest that while transfer learning and fine-tuning have significantly improved specificity and overall accuracy, further efforts are needed to address the sensitivity issue, likely through strategies for mitigating class imbalance.

7. Chameleon Cloud Experience

We started using Chameleon Cloud for this project as a way to avoid having to run the program on our own devices. There was a fairly steep learning curve, and we had quite some difficulty getting multiple users to access the same instance once we had a reservation set up. We ended up fooling the computer into thinking we were all the same user by all using the same public and private key, which most certainly is not the intended approach. We also forgot to properly back up our files before the instance ended, and from then there's no way to get the files back, so we ended up losing some progress. The sum of all these small technical difficulties motivated us to switch to doing our computation on Kaggle instead. With some more patience and foresight, we could have used Chameleon Cloud more effectively, but given our time constraints, we opted to stick with what we were familiar with.

8. Conclusions and Future Steps

lthough class weighting was implemented in the loss function to address the class imbalance, the extreme imbalance in the 2024 dataset (1:20 ratio) may still have hindered the model's ability to detect malignant cases effectively. The class weights helped by giving more importance to the minority class, but the disparity between malignant and benign cases might still pose challenges for detection. Further techniques, such as oversampling the minority class, more aggressive class balancing, or modifications to the model architecture, could be explored to enhance sensitivity for malignant cases.

Another potential contributing factor to the lower sensitivity is the nature of the 2024 dataset, which consists of smartphone-quality images. These images may contain less detail or more noise compared to the high -resolution images from the 2019 dataset, which could further complicate the model's ability to detect subtle features associated with malignant lesions. While the fine-tuned model improved specificity and accuracy, enhancing sensitivity for malignant cases remains a priority for future work.

Overall, the fine-tuned 2019/2024 model provided the best balance of performance, particularly in terms of specificity. This outcome underscores the benefits of transfer learning, where leveraging the 2019 model's learned features helped improve performance on the lower-quality 2024 images. However, the drop in sensitivity signals a potential area for further refinement. While class weighting in the loss function helped mitigate the impact of class imbalance, the extreme imbalance in the 2024 dataset remains a challenge.

To address the lower sensitivity for malignant cases, future work will explore additional techniques such as more aggressive class balancing, oversampling, or further fine-tuning the model to better capture subtle malignant features. Additionally, further experiments could evaluate whether image quality improvement techniques, such as data augmentation or super-resolution methods, could help the model better handle the lower-quality 2024 images and enhance its sensitivity.

By continuing to refine the model's ability to detect malignant cases, we hope to achieve better overall performance and greater generalization across datasets with varying image quality.

References

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Combalia, M., Codella, N. C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A. C., Puig, S., et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

Kurtansky, N. R., D'Alessandro, B. M., Gillis, M. C., Betz-Stablein, B., Cerminara, S. E., Garcia, R., Girundi, M. A., Goessinger, E. V., Gottfrois, P., Guitera, P., et al. The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection. *Scientific Data*, 11(1):884, 2024.

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic

Table 1. 2019 Model Results

FOLD	EPOCH	ACCURACY	F1	MEAN AUC	WEIGHTED ACCURACY	SPECIFICY	SENSITIVITY
0	60	0.7735	0.8114	0.8826	0.7934	0.7330	0.8538
1	60	0.7777	0.8138	0.8812	0.7971	0.7365	0.8576
2	60	0.7811	0.8241	0.8699	0.7857	0.7717	0.7998
3	60	0.7833	0.8212	0.8835	0.7993	0.7503	0.8483
4	60	0.7799	0.8164	0.8780	0.7967	0.7435	0.8500
		0.7791	0.8174	0.8790	0.7944	0.7470	0.8419

Table 2. 2024 Model Results

FOLD	EPOCH	ACCURACY	F1	MEAN AUC	WEIGHTED ACCURACY	SPECIFICY	SENSITIVITY
0	30	0.8752	0.9308	0.9045	0.8249	0.8805	0.7692
1	20	0.8940	0.9420	0.9334	0.8637	0.8966	0.8308
2	30	0.9104	0.9507	0.9411	0.8716	0.9153	0.8280
3	30	0.8715	0.9283	0.9055	0.8435	0.8745	0.8125
4	40	0.8624	0.9229	0.9316	0.8476	0.8640	0.8312
		0.7791	0.8174	0.8790	0.7944	0.7470	0.8419

images of common pigmented skin lesions. *Scientific Data*, 5(180151), 2018.

Table 3. 2019/2024 Fine Tuned Model Results

FOLD	EPOCH	ACCURACY	F1	MEAN AUC	WEIGHTED ACCURACY	SPECIFICY	SENSITIVITY
0	40	0.8873	0.9379	0.9186	0.8312	0.8932	0.7692
1	60	0.9116	0.9523	0.9319	0.8212	0.9193	0.7231
2	50	0.9122	0.9518	0.9373	0.8625	0.9185	0.8065
3	30	0.9188	0.9560	0.9016	0.8446	0.9268	0.7625
4	60	0.9024	0.9468	0.9166	0.8253	0.9104	0.7403
		0.9065	0.9490	0.9212	0.8370	0.9136	0.7603