AI-Based Skin Lesion Classification Using EfficientNet & Transfer Learning

Juna Kim
Western University
jkim2938@uwo.ca

Diya Patel
Western University
dpate524@uwo.ca

Brian Lee
Western University
tlee463@uwo.ca

Chelsea Ye
Western University
cye68@uwo.ca

Marc Crasto
Western University
mcrasto2@uwo.ca

Raiyan Butt
Western University
rbutt23@uwo.ca

Kamal Chahal
Western University
kchaha22@uwo.ca

***Abstract*—Skin lesions can go unnoticed by non-experts, and diagnosis relies on visual assessment and biopsy, leading to delays in care. This paper develops and evaluates a deep learning approach for a three-class (malignant, benign, and non-neoplastic) skin lesion classification. The Fitzpatrick17k dataset was augmented to 21,864 labelled images, with an approximate train-test-validation split of 85-10-5. The model uses an EfficientNetV2-S pre-trained on ImageNet, swapping the final layer for three-class prediction. The model was trained using cross-entropy loss and AdamW, with separate learning rates for the backbone and classifier head. To improve robustness and reduce overfitting, data augmentation techniques, MixUp, and automatic mixed precision were applied. The best-performing model achieved 0.9137 overall accuracy, macro F1-score of 0.8753, and malignant recall of 0.8905. These results suggest that the proposed method could support early triage and lesion screening for non-experts.**

The code for this work is available at https://github.com/junakim0118/WCS_SkinCancerDiagnosticsAI.

I. Introduction

A. Motivation

Skin cancers are among the most common skin diseases globally, and their ability to metastasize while often being difficult to distinguish makes early detection important [1]. Studies show that early detection of melanoma, a severe form of skin cancer, results in a 99% five-year survival rate [2]. During the COVID-19 pandemic, delays in melanoma diagnoses resulted in disease progression, with melanomas advancing to a larger breslow thickness and increased ulceration [3]. Current diagnostic pathways rely on visual assessment and subsequent biopsying, which can result in low specificity and high rate in unnecessary biopsies [4]. Thus, there is potential benefit in an easily accessible application that can preemptively identify the severity of skin lesions. Previous application of AI-based decision-support tools in primary care settings suggest that such a tool could maintain a low risk of false negatives, while reducing unnecessary excisions when used by trained physicians [5].

B. Related Works

A systematic review and meta-analysis observed large variability in skin cancer diagnostic accuracy among physicians with varying specialties and years of experience. The study explored the idea that dermatologists best classify skin lesions when combining dermoscopy with clinical examination, in correlation to their level of experience [6]. In parallel, previous AI and machine learning-based approaches using convolutional neural networks (CNNs) for skin lesion classification have performed similarly to experienced dermatologists, highlighting their potential as decision or triage support tools in clinical settings [5], [7].

In many medical imaging applications, CNNs initialized with ImageNet-pretrained weights have shown success in image classification through transfer learning. These models leverage features learned from large natural image datasets to improve convergence and generalization when trained on limited datasets. Recent publications have highlighted the benefit of EfficientNet architectures (B0-B7) in model training. These architectures scale network depth, width, and resolution to achieve optimized accuracy, with newer variants such as EfficientNetV2 demonstrating further improvements in training efficiency and performance when fine-tuned with ImageNet pretrained weights. [8]

Dermatological imaging datasets can carry class imbalances, particularly for malignant lesions. Such Imbalances can bias models towards majority classes and significantly hinder a models ability to generalize. Prior work have often handled this issue through class-weighted loss functions, weighted random sampling, and image augmentation to improve specificity and sensitivity among minority classes [9]. Image augmentation techniques often include image resizing, horizontal and vertical flipping, rotations, and RandAugment to artificially increase data diversity [9]. Image augmentation combined with regularization methods such as MixUp allow the model to better simulate real-world variability in dermatological images [10].

The method used in this paper builds upon these strategies. The proposed approach integrates transfer learning with an EfficientNet-based CNN architecture, trained on an artificially balanced dataset using multiple augmentation strategies to develop a robust skin lesion classification model.

C. Problem Definition

This paper explores the development and evaluation of an AI-based application that can be used to assist in the early identification of skin lesions using image-based analysis. The approach applies deep learning and computer vision techniques on a dataset of 16, 577 clinical images of skin lesions, classified as benign, malignant, and non-neoplastic based on atlas metadata. The proposed method aims to provide an accessible, preliminary decision-support tool that can assist clinicians in distinguishing potentially malignant lesions from benign cases, with hope to reduce unnecessary biopsies, improve diagnostic specificity, and mitigate delays in care while maintaining a low risk of false-negative assessments.

II. Methodology

A. Dataset

Fitzpatrick17k dataset was used for this study that consists of 16, 577 skin lesion images classified into three distinct classes: malignant, benign, and non-neoplastic. Malignant lesions refer to cancerous cells or tumours capable of spreading, benign lesions are non-cancerous lesions that do not spread, and non-neoplastic lesions are abnormal tissue growths not caused by uncontrolled cell proliferation. This dataset distinguishes itself with other datasets as the images with diverse skin tones were selected for the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) [11]. Any duplicate/null images were taken out of the dataset beforehand [12]. The data was split into training, validation, and testing sets with a ratio of 85:5:10 respectively.

This dataset distinguishes itself with other datasets as the images with diverse skin tones [Fig. 1] were selected for the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) [11]. Any duplicate/null images were taken out of the dataset beforehand [12]. The data was split into training, validation, and testing sets with a ratio of 85:5:10 respectively.

B. Data Preprocessing

To enhance the dataset performance for CNN-based image recognition, PyTorch's torchvision.transforms library was utilized for implementing a preprocessing pipeline. All images were resized to a fixed size to keep consistent input dimensions, and random horizontal/vertical flips and random rotations were applied to reduce overfitting. Then, the images were converted into tensors and were normalized using ImageNet's mean and standard deviation to align with the ImageNet's pretrained weights for faster and more stable training. In the strong augmentation setting, RandAugment (2 ops, magnitude 9) was used during training to apply additional random transformations and increase image variety, while small randomly-sized patches of image area were hidden to improve robustness.

This workflow ensures that the data is well-structured, sound, efficient, and appropriate for skin lesion image classification.

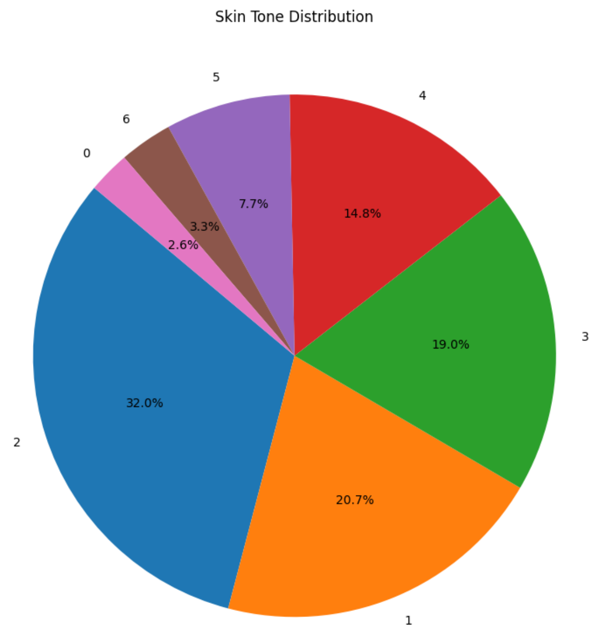

Fig. 1. Pie chart illustrating the skin tone distribution of the dataset. Each number represents the corresponding Fitzpatrick Skin Type (FST) scale as below: 0 – N/A, 1 – pale/ivory, 2 – fair, 3 – medium, 4 – olive/light brown, 5 – brown, 6 –dark brown.

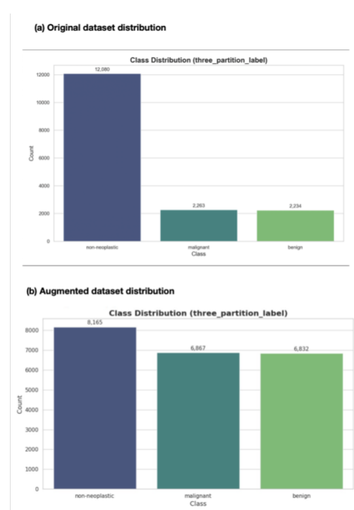

Fig. 2. Class distribution of the skin lesion dataset. (a) Original dataset showing significant class imbalance with non-neoplastic lesions dominating the dataset. (b) Augmented dataset after applying image augmentation techniques.

C. Model Architecture and Training

A CNN-based approach was utilized with a pretrained EfficientNetV2-S model, and fine tuning for speed and efficiency during model training. The pretrained EfficientNetV2-S model was employed for 20 epochs, and the default classifier layer was replaced with a custom linear classifier designed for the three-class output. A cross-entropy loss function with label smoothing of 0.00556 was introduced for the classification model. The AdamW optimizer with a head learning rate of 0.00119, a backbone learning rate of 0.00015, and a weight decay of 2.55924e-05 was selected using hyperparameter tuning. The model was trained using a batch size of 48 with a CUDA-enabled GPU, in addition to automatic mixed precision (AMP) to speed up training and lower GPU memory usage. During training, MixUp was applied as an additional data augmentation technique to reduce overfitting, where pairs of images were randomly blended together and their labels were mixed in the same proportion. The validation set was used during hyperparameter tuning to select the best model.

D. Evaluation Metrics

The evaluation of the model describes the model's ability to differentiate between true-positive and true-negative cases. To satisfy such expectation, this study focuses on evaluation metrics including precision, recall (sensitivity), and F1-score. Their respective formulas are:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1\ Score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

The confusion matrix is also introduced to visually display a detailed result of the model. It provides where the model could be improved and which part it is struggling the most to classify.

III. Results

A. Overall Model Performance

After careful hyperparameter tuning, the model successfully reached overall a test accuracy of 91.37%. For the ability to detect skin cancer (malignant), the model achieved a precision of 87.76%, a recall of 89.05%, and F1-Score of 88.40%.

TABLE I
PERFORMANCE OF THE MODEL FOR CLASSIFYING SKIN LESIONS

Class	Precision	Recall	F1
Malignant	87.76%	89.05%	88.40%
Benign	80.27%	80.00%	80.13%
Non-neoplastic	94.17%	93.94%	94.06%
		Overall Accuracy	**91.4%**

B. Classification Results (Confusion Matrix)

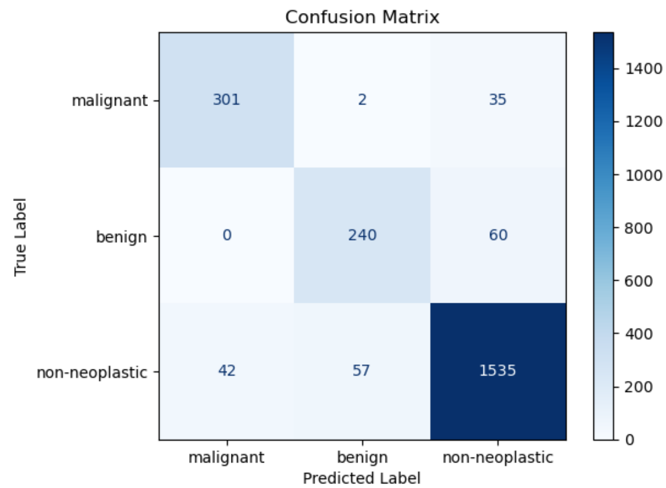

Fig. 3. Confusion matrix illustrating the performance of CNN-based skin cancer classification model. The model achieved high classification recall for non-neoplastic (94%) and malignant (89%), but benign exhibited slightly lower precision (80%).

Examining performance by class shows that the model correctly classified 301 malignant cases, while two were misclassified as benign and 35 as non-neoplastic. For benign lesions, 240 cases were correctly classified, while 60 were misclassified as non-neoplastic and none as malignant. For non-neoplastic lesions, the model correctly classified 1,535 cases, and misclassified 42 and 57 cases as malignant and benign, respectively.

The model performed well altogether[Fig. 3], although the recall for malignant is slightly lower than the overall model accuracy. It clearly struggles classifying benign cases with a recall of 80%; however, the confusion matrix shows that no benign cases are misclassified as malignant. To a large extent, these results spot the areas where the model requires to be improved. Additionally, utilizing the Fitzpatrick dataset, the model learned and performed well across different skin tones [Fig. 4].

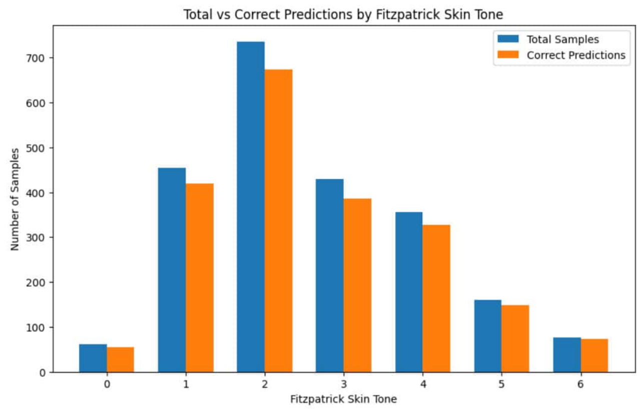

Fig. 4. Chart illustrating the number of correct predictions made by the model compared to the total number of data for each FST. The model managed to achieve high accuracy throughout the diverse skin tones.

IV. Conclusion

This study investigated the effectiveness of CNN-based machine learning models for medical image classifications. Utilizing the pretrained EfficientNetV2-S model, the study demonstrated that the model can achieve a high accuracy of 91.37% in classifying skin lesions, suggesting the potential of implementing CNN models for medical applications. Although it indicated lower recall for benign cases, most of the malignant cases—the most important classification for the ethical deployment of the model—were correctly predicted. Since misclassifying malignant cases would be critical to patients, the model should focus on minimizing false negatives.

Despite these results, the model still misclassified 37 malignant as benign and non-neoplastic; therefore, it is too early to consider the model as a final diagnosis as it still struggles with variances. Instead, it should be treated as a decision support triage tool that has the potential to be used in healthcare.

V. Future Work

Future research should prioritize improving the generalizability of the model and reducing false negatives by incorporating larger, heterogeneous datasets that capture a wide range of skin lesion forms. To further enhance classification precision, we intend to investigate advanced data augmentation techniques—such as synthetic lesion generation via Generative Adversarial Networks (GANs)—to better train the model for rare cases.

Additionally, we see significant potential in multi-modal learning where dermoscopic images are analyzed along with genetic markers, which may further reduce malignant false negatives. Our ultimate objective is to deploy and implement these models within real-world medical environments, where their values can be evaluated as a testing tool for dermatologists in clinical practice.

VI. Limitations

Despite the high performance achieved in this study, several limitations were acknowledged by the model. As the model is fundamentally rooted in pattern recognition, it struggles with classifying edge cases and underrepresented skin lesion images in the training dataset. In essence, instances such as rare subtypes of melanoma may result in lower accuracy. Furthermore, the model's reliance on high-quality dermoscopic images may limit its ability to classify images that were not precisely taken.

VII. Ethical Considerations

The deployment of AI-based diagnostic models must ensure patient privacy and the secure handling of sensitive medical imagery. To prevent bias, it is critical to verify that diagnostic accuracy remains consistent across all skin types.

Furthermore, it must be emphasized that AI should function as a decision-support tool, not a replacement for clinical diagnostic tools. It is essential that patients do not rely solely on AI advice for self-diagnosis; instead, these tools should assist early actions. Maintaining cautiousness is essential for the ethical standards and patient safety to remain the highest priorities.

VIII. Acknowledgements

The Model was created by the team consisting of Juna Kim, Diya Patel, Brian Lee, Chelsea Ye, Marc Crasto, Raiyan Saleem, Kamal Chahal, and with the support of WCS.

References

[1] M. Wang, X. Gao, and L. Zhang, "Recent global patterns in skin cancer incidence, mortality, and prevalence," *Chinese Medical Journal*, vol. 138, pp. 185–192, Dec. 2024. [Online]. Available: https://journals.lww.com/cmj/fulltext/2025/01200/recent_global_patterns_in_skin_cancer_incidence,.6.aspx

[2] J. B. Heistein, U. Acharya, and S. K. R. Mukkamalla, "Cancer, malignant melanoma," PubMed, Feb. 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK470409/

[3] A. Soni, E. Purcell, B. Lim, G. Marcaccini, I. Seth, and W. M. Rozen, "The silent spread: A systematic review of delayed melanoma diagnosis and disease progression during the covid-19 pandemic," *JEADV Clinical Practice*, vol. 4, pp. 982–1006, Aug. 2025.

[4] J. K. Rivers and D. S. Rigel, "Ruling out melanoma: A practical guide to improving performance through non-invasive gene expression testing (family practice)," Skin Therapy Letter, Feb. 2019. [Online]. Available: https://www.skintherapyletter.com/family-practice/melanoma-non-invasive-gene-expression-testing-2/

[5] P. Papachristou, M. Söderholm, J. Pallon, M. Taloyan, S. Polesie, J. Paoli, C. D. Anderson, and M. Falk, "Evaluation of an artificial intelligence-based decision support for detection of cutaneous melanoma in primary care – a prospective, real-life, clinical trial," *British Journal of Dermatology*, vol. 191, pp. 125–133, Jan. 2024.

[6] J. Y. Chen, K. Fernandez, R. P. Fadadu, R. Reddy, M.-O. Kim, J. Tan, and M. L. Wei, "Skin cancer diagnosis by lesion, physician, and examination type," *JAMA Dermatology*, vol. 161, pp. 135–146, Nov. 2024. [Online]. Available: https://jamanetwork.com/journals/jamadermatology/article-abstract/2826310

[7] H. K. Jeong, C. Park, R. Henao, and M. Kheterpal, "Deep learning in dermatology: A systematic review of current approaches, outcomes and limitations," *JID Innovations*, vol. 3, Aug. 2022.

[8] K. K., K. S., A. K. J., and C. B., "Enhancing skin cancer classification using efficient net b0–b7 through convolutional neural networks and transfer learning with patient-specific data," *Asian Pacific Journal of Cancer Prevention*, vol. 25, pp. 1795–1802, May 2024.

[9] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, "Data augmentation for skin lesion analysis," in *Lecture Notes in Computer Science*, vol. 11041, Oct. 2018, pp. 303–311.

[10] N. Lama, R. J. Stanley, B. Lama, A. Maurya, A. Nambisan, J. Hagerty, T. Phan, and W. Van Stoecker, "Lama: Lesion-aware mixup augmentation for skin lesion segmentation," *Journal of Imaging Informatics in Medicine*, vol. 37, pp. 1812–1823, Feb. 2024.

[11] M. Groh, C. Harris, L. R. Soenksen, F. Din-Houn Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1820–1828.
[12] K. Abhishek, A. Jain, and G. Hamarneh, "Investigating the quality of dermamnist and fitzpatrick17k dermatological image datasets," *Scientific Data*, vol. 12, no. 1, Feb. 2025.