2024 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET), Faculty of Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh, 07-08 December, 2024

2024 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET) | 979-8-3315-0947-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRPSET64863.2024.10955919

EF-SwinNet: A Hybrid EfficientNet-Swin Transformer Model for Skin Cancer Classification

Liton Sarker
Cash Transfer Modernization (CTM) Project,
Department of Social Services, Ministry of Social Welfare.
E-mail: sarker.liton@gmail.com

Ashfak Yeafi
Department of Electrical and Electronic Engineering,
Khulna University of Engineering & Technology.
E-mail: yeafiashfak@gmail.com

***Abstract*—Skin cancer ranks as the most common type of cancer worldwide. In this study, we present EF-SwinNet, a hybrid architecture combining EfficientNet and swin transformer models, designed to classify skin lesions with high accuracy. The dataset used for training and testing the model is HAM10000, which is known for its imbalance. In order to tackle this problem, we implemented several data augmentation methods to successfully alleviate the problem of classification imbalance. Our empirical findings illustrate the exceptional efficacy of the suggested hybrid model, attaining an average accuracy of 98% and an F1 score of 96%. Furthermore, we employed the Grad-CAM technique to offer a deeper understanding of the learning process of the model by graphically representing the significance of various features. Finally, we analyze the performance of our model by comparing it with current cutting-edge methods, emphasizing its improvements in the classification of skin cancer.**

***Index Terms*—Skin Cancer Classification, CNN, Swin Transformer, Deep Learning**

I. Introduction

Skin cancer, both melanoma and nonmelanoma, is one of the most prevalent cancers globally, with rising cases, especially in regions like the United States, where skin cancer cases exceed those of all other cancers combined, highlighting the need for early and accurate detection to improve outcomes [1]. Traditional diagnostic methods, including visual inspection and dermoscopy, are highly dependent on clinician expertise, which can result in variable accuracy [2]. Even with improved dermoscopic techniques, diagnostic precision remains around 75-84%, often below the level needed for consistent early detection [3]. Deep learning models, especially convolutional neural networks (CNNs), have shown promise in automating skin lesion analysis but face limitations. CNNs excel in extracting localized features but struggle with global context, which is critical for complex lesion classification [4]. Transformers, with their ability to model long-range dependencies, address this limitation. Vision Transformers and Swin Transformers, for instance, capture global and hierarchical features, outperforming CNNs in some tasks [5]. However, effectively integrating transformers and CNNs into an efficient, interpretable model for medical imaging remains challenging. This work introduces EF-SwinNet, a hybrid model designed to balance feature extraction, computational efficiency, and interpretability, crucial for clinical use. Class imbalance in skin cancer datasets like HAM10000, where common classes dominate, poses an additional challenge. EF-SwinNet combines EfficientNet and Swin Transformer to address these issues, with contributions detailed as follows:

1) We developed a novel hybrid model combining EfficientNet with the swin transformer to enhance skin cancer classification accuracy.
2) We incorporated Grad-CAM to provide visual explanations of the model's predictions, which improves transparency and clinical relevance.

The paper is structured as indicated below: Latest developments are reviewed in Section Section II. In Section Section III, the materials and procedures employed are detailed. In Section Section IV, the experiments are described and the findings are shown. Section Section V finalizes the findings of the work.

II. Related work

The field of skin cancer detection has seen significant advancements with the adoption of machine learning (ML) and deep learning models. Early approaches used handcrafted features based on asymmetry, border, color, and diameter attributes to identify lesions, with promising but limited results [6]. In another example, Saez et al. [7] achieved moderate accuracy in melanoma classification using logistic regression and artificial neural networks (ANN) to assess lesion thickness. However, these traditional ML techniques rely heavily on manual feature extraction, which can introduce subjectivity and limit diagnostic accuracy. With the advent of deep learning, CNNs have increasingly automated feature extraction for skin lesion analysis. Shete et al. [8] uses CNNs to address the challenges of feature localization, achieving 88.0% accuracy. Despite these advances, CNNs often struggle to capture the global context required for classifying complex medical images. To address this, recent studies have integrated transformers, which excel at modeling long-range dependencies. For instance, Xie et al. [9] proposed Swin-SimAM, a combination of Swin Transformer with SimAM attention, specifically for melanoma detection, while Eskandari et al. [10] applied a hierarchical transformer model for skin lesion segmentation. These transformer-based approaches have shown potential, but challenges remain in effectively combining local and global feature extraction without excessive

computational costs. Our proposed model, EF-SwinNet, improves upon previous approaches by combining the strengths of EfficientNet, known for its parameter efficiency, with Swin Transformer's hierarchical feature extraction capabilities. This hybrid architecture addresses several limitations identified in prior work. First, while CNN-based models excel in extracting localized features, they lack the ability to capture broader context across an image, a limitation addressed by Swin Transformer's shifted window attention mechanism.

III. Material and Methods

A. Dataset description and data processing

The dataset utilized for training and evaluating our model is the HAM10000 dataset, compiled by the ISIC in 2018 [11]. The dataset is an openly accessible collection of 10,015 dermatoscopic images that depict seven different types of skin diseases: basal cell carcinoma (BCC), actinic keratoses and intraepithelial carcinoma (AKIEC), dermatofibroma (DF), benign keratosis-like lesions (BKL), melanocytic nevi (NV), vascular lesions (VASC), and melanoma (MEL). The dataset distribution is as follows: 327 AKIEC images, 514 BCC images, 1,029 BKL images, 115 DF images, 1,113 MEL images, 6,705 NV images, and 142 VASC images. A major challenge associated with this dataset is the class imbalance, as illustrated in Fig. 1. Specifically, the NV class dominates, accounting for 66.9% of the total images. The dataset was

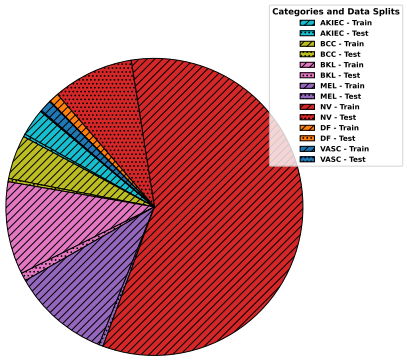


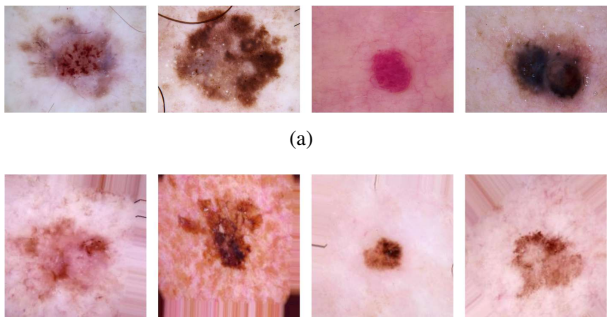
Fig. 1. Distribution of train and test data across skin cancer categories, visualized with distinct patterns for train (//) and test (..) sets. Categories include AKIEC, BCC, BKL, MEL, NV, DF, and VASC.

preprocessed by resizing all images to 384×384 pixels. Next, we divided the dataset into training and testing sets using an 80-20 ratio partition. The training set consists of 297 AKIEC, 479 BCC, 1,011 BKL, 1,067 MEL, 5,822 NV, 107 DF, and 129 VASC images, while the testing set includes 30 AKIEC, 35 BCC, 88 BKL, 46 MEL, 883 NV, 8 DF, and 13 VASC images. To mitigate the class imbalance, various data augmentation techniques were applied to the training set, artificially increasing the diversity of the underrepresented classes. These augmentations include random rotations, flips, zooms, and shifts, ensuring the model can generalize across

TABLE I
Data Augmentation Techniques and Parameters

Augmentation Technique	Description	Parameter Value
Rotation	Rotates images randomly.	180°
Width Shift	Horizontally shifts images.	10%
Height Shift	Vertically shifts images.	10%
Zoom	Zooms in/out on images.	10%
Horizontal Flip	Horizontally flips images.	True
Vertical Flip	Vertically flips images.	True
Fill Mode	Fills pixels after shifts/rotations.	fill_mode='nearest'

different categories. The augmentation parameters are provided in Table I. Fig.2 showcases examples of data before and after augmentation, demonstrating the efficacy of these techniques in balancing the dataset.



(a)

(b)

Fig. 2. (a) Images before applying augmentation.(b) Images after applying augmentation.

B. Skin Cancer Classification using EF-SwinNet

In this study, we propose a hybrid architecture that leverages the complementary strengths of the swin transformer and EfficientNet models. The swin transformer, a core component of our architecture, is built on the principles of hierarchical feature maps (HFM) and shifted window attention (SWA). HFM enables the model to manage image features across multiple scales, facilitating efficient processing of high-resolution medical images without relying on standard convolutional operations. The patch merging technique within the swin transformer reduces feature map resolution by merging adjacent patches, preserving critical image information while optimizing computational efficiency. SWA, by employing windowed self-attention, ensures both local and global contexts are effectively captured, a key requirement for identifying subtle patterns in medical imagery, such as skin lesions. EfficientNet complements the swin transformer by providing a

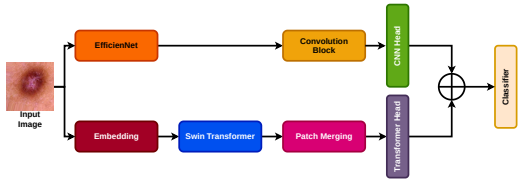


Fig. 3. Proposed architecture for the EF-SwinNet model.

TABLE II
RESULTS DERIVED FROM OUR PROPOSED EF-SWINNET MODEL.

Class	Precision	Recall	F1 score
AKIEC	0.90	0.90	0.90
BCC	0.97	0.97	0.97
BKL	0.96	0.93	0.95
DF	1.0	1.0	1.0
MEL	0.95	0.89	0.92
NV	0.99	1.0	0.99
VASC	1.0	1.0	1.0

parameter-efficient, high-performance convolutional network, which excels in handling large-scale image datasets. The model's lightweight design, coupled with fast training times, makes it particularly suitable for deployment on resource-limited devices, such as mobile platforms or embedded systems, without sacrificing classification accuracy. The hybrid model processes input images through two distinct pathways. The first pathway directs the input through the EfficientNet network, followed by a convolutional block and a CNN head that integrates a GlobalAveragePooling2D layer. The second pathway processes the input via an embedding layer, swin transformer layers, and a patch merging layer, which ultimately feeds into a transformer head composed of a GlobalAveragePooling2D layer and a BatchNormalization layer. The results from both paths are combined and then fed into a final classifier that includes a fully connected layer specifically designed to generate probabilities for different skin cancer classes. In the swin transformer configuration, we employ an 8×8 patch size, a 64-dimensional embedding space, and a multi-layer perceptron (MLP) with 128 units. This setup, when combined with EfficientNet, maximizes the model's classification potential by harnessing the swin transformer's hierarchical feature extraction alongside EfficientNet's computationally efficient structure. The synergy between these architectures yields a robust model capable of accurately predicting skin cancer classes, offering a significant tool for advancing diagnostic precision in medical imaging.

IV. EXPERIMENT AND RESULT

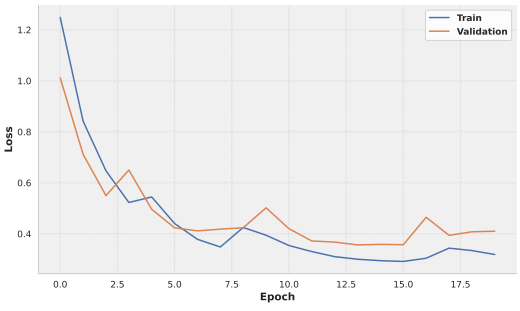


Fig. 4. EF-SwinNet training versus validation loss curve.

The EF-SwinNet model was implemented using the TensorFlow library and trained on a workstation with an NVIDIA RTX 3090 GPU, 64GB RAM, and Intel i7 CPU. The Adam optimizer was employed with a learning rate of 0.001, batch size of 32, and weight decay of 1e-4. This setup was carefully selected to ensure efficient processing and to support the high computational demands of both EfficientNet and Swin Transformer components, which are critical for handling large image datasets like HAM10000. The model was trained for 20 epochs, with performance monitored through loss curves, shown in Fig. 4. These curves indicate a consistent decrease in both training and validation loss, suggesting effective learning without overfitting. To assess EF-SwinNet's classification performance, we used standard evaluation metrics—accuracy, precision, recall, and F1 score—on a test dataset reserved specifically for this purpose [12]. Table II summarizes the model's performance across each class, and Fig. 5 provides a normalized confusion matrix for a detailed view of correct and incorrect predictions. The analysis reveals that EF-SwinNet

Fig. 5. Normalized confusion matrix showing the model's performance across skin lesion categories, with values representing the proportion of correct and incorrect predictions, adjusted for class imbalance.

performs well across most skin cancer classes. The AKIEC class attained a balanced performance with accuracy, recall, and F1 score of 90%, reflecting its ability to detect cases effectively while minimizing false positives. The BCC class achieved even higher scores, with 97% across all metrics, showcasing EF-SwinNet's strength in identifying this common type of skin cancer. For the BKL class, the model also performed well, achieving precision, recall, and F1 scores of 96%, 93%, and 95%, respectively. Both the DF and VASC classes achieved perfect scores of 100% across all metrics, highlighting the model's effectiveness in detecting these less common conditions. For the MEL (melanoma) class, however, performance was slightly lower, with a precision of 95%, recall of 89%, and F1 score of 92%. Although these metrics remain high, the reduced recall suggests that some melanoma cases were missed—a critical consideration given the aggressive nature of melanoma. For the NV class, the model achieved 99% precision, 100% recall, and 99% F1 score, underscoring EF-SwinNet's robustness in differentiating between benign and malignant lesions.

EF-SwinNet's design combines EfficientNet and Swin Trans-

former to leverage the strengths of both architectures. EfficientNet provides a lightweight, parameter-efficient convolutional backbone, which enables efficient processing of high-resolution images with fewer computational resources. This is particularly important in medical applications where computational efficiency and speed are essential. Swin Transformer, on the other hand, enhances the model's ability to capture both local and global context within images using hierarchical feature extraction and shifted window attention. This dual-pathway approach effectively captures nuanced details and global dependencies critical for distinguishing similar skin cancer classes. When comparing EF-SwinNet's results to other models (e.g., Inception-ResNet [13] with 83% accuracy and MobileNet [14] with 92% accuracy), it is evident that the hybrid structure of EF-SwinNet offers superior performance, achieving 98% accuracy and a 96% F1 score. While differences in experimental setups exist, such as variations in dataset preprocessing and augmentation techniques, EF-SwinNet's hybrid approach clearly contributes to its higher classification accuracy, emphasizing the model's adaptability in handling both common and rare skin cancer classes.

EF-SwinNet's integration of Grad-CAM visualization adds

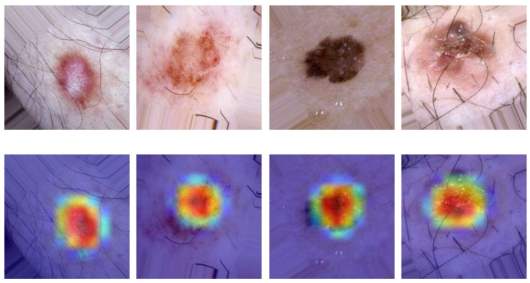


Fig. 6. Grad-CAM visualization highlighting the regions of the skin lesions the model focuses on for classification, emphasizing the key factors in its decision-making process to ensure interpretability.

interpretability, making it more clinically applicable. Grad-CAM allows clinicians to see the regions that influenced the model's decisions, enhancing trust and facilitating validation of the model's classifications. Fig. 6 shows that EF-SwinNet consistently focuses on clinically relevant regions of skin lesions, confirming its reliability as a diagnostic tool. This transparency is essential in medical imaging, where understanding the rationale behind predictions can impact diagnostic decisions.

TABLE III
COMPARISON ANALYSIS WITH OTHER PAPERS

Methods	Accuracy	Precision	F1 score
Inception-ResNet [13]	0.83	0.72	0.69
MobileNet [14]	0.92	0.87	0.84
Ensemble Net [15]	0.93	0.88	0.84
Proposed EF-SwinNet	**0.98**	**0.97**	**0.96**

V. CONCLUSION

This study introduced a hybrid model combining EfficientNet with the Swin Transformer to improve skin cancer classification. The model achieved high precision, recall, and F1 scores, particularly for BCC and NV, with Grad-CAM visualizations confirming its focus on clinically relevant regions. However, lower performance in AKIEC and MEL suggests the need for addressing data imbalance. Future work will explore oversampling methods like SMOTE and advanced augmentation techniques to improve these underrepresented classes. We also plan to expand the dataset, optimize the model for real-time clinical use, and explore advanced architectures.

REFERENCES

[1] C. Xia, X. Dong, H. Li, M. Cao, D. Sun, S. He, F. Yang, X. Yan, S. Zhang, N. Li *et al.*, "Cancer statistics in china and united states, 2022: profiles, trends, and determinants," *Chinese medical journal*, vol. 135, no. 05, pp. 584–590, 2022.

[2] N. H. Khan, M. Mir, L. Qian, M. Baloch, M. F. A. Khan, E. E. Ngowi, D.-D. Wu, X.-Y. Ji *et al.*, "Skin cancer biology and barriers to treatment: Recent applications of polymeric micro/nanostructures," *Journal of Advanced Research*, vol. 36, pp. 223–247, 2022.

[3] X. Fei, J. Wang, S. Ying, Z. Hu, and J. Shi, "Projective parameter transfer based sparse multiple empirical kernel learning machine for diagnosis of brain disease," *Neurocomputing*, vol. 413, pp. 271–283, 2020.

[4] L. Wang, L. Ding, Z. Liu, L. Sun, L. Chen, R. Jia, X. Dai, J. Cao, and J. Ye, "Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning," *British Journal of Ophthalmology*, vol. 104, no. 3, pp. 318–323, 2020.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[6] F. E. S. Alencar, D. C. Lopes, and F. M. M. Neto, "Development of a system classification of images dermoscopic for mobile devices," *IEEE Latin America Transactions*, vol. 14, no. 1, pp. 325–330, 2016.

[7] A. Sáez, J. Sánchez-Monedero, P. A. Gutiérrez, and C. Hervás-Martínez, "Machine learning methods for binary and multiclass classification of melanoma thickness from dermoscopic images," *IEEE transactions on medical imaging*, vol. 35, no. 4, pp. 1036–1045, 2015.

[8] A. S. Shete, A. S. Rane, P. S. Gaikwad, and M. H. Patil, "Detection of skin cancer using cnn algorithm," *International Journal*, vol. 6, no. 5, 2021.

[9] Z. Wang, H. Lu, J. Jin, and K. Hu, "Human action recognition based on improved two-stream convolution network," *Applied Sciences*, vol. 12, no. 12, p. 5784, 2022.

[10] S. Eskandari, J. Lumpp, and L. Sanchez Giraldo, "Skin lesion segmentation improved by transformer-based networks with inter-scale dependency modeling," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2023, pp. 351–360.

[11] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[12] A. Yeafi, M. Islam, S. K. Mondal, K. I. H. Nashad, and M. S. U. Yusuf, "A semi-supervised approach for brain tumor classification using wasserstein generative adversarial network with gradient penalty," in *2023 6th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE, 2023, pp. 1–6.

[13] H.-B. Mureşan, "Skin lesion diagnosis using deep learning," in *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2019, pp. 499–506.

[14] E. H. Mohamed and W. H. El-Behaidy, "Enhanced skin lesions classification using deep convolutional networks," in *2019 Ninth international conference on intelligent computing and information systems (ICICIS)*. IEEE, 2019, pp. 180–188.

[15] D. N. Le, H. X. Le, L. T. Ngo, and H. T. Ngo, "Transfer learning with class-weighted and focal loss function for automatic skin cancer classification," *arXiv preprint arXiv:2009.05977*, 2020.