Developing an AI-driven multimodal framework to analyze skin lesion images and patient symptoms for accurate and early skin cancer prediction

Snehal Laddha
Electronics Department
Ramdeobaba University
Nagpur,India
laddhasv2@rknec.edu

Aayush Paturkar
Computer Science Engineering
Ramdeobaba University
Nagpur,India
paturkaraa_1@rknec.edu

Sarthak Khutafale
Computer Science Engineering
Ramdeobaba University
Nagpur,India
khutafaless@rknec.edu

***Abstract*—Skin cancer, particularly melanoma, remains one of the most prevalent and life-threatening forms of cancer. Given the high cost and time associated with traditional dermatological evaluations, there is a growing demand for automated diagnostic systems capable of analyzing dermatoscopic images. A critical step in this process is the accurate segmentation and classification of skin lesions, which involves distinguishing affected areas from healthy skin to support reliable diagnosis.**

This study presents a novel multimodal approach that integrates both segmentation and classification, utilizing EfficientNet-B4, DenseNet-169, and a ResUNet-inspired architecture enhanced with Dense Pyramid Pooling. The proposed framework leverages the ISIC 2018 and ISIC 2019 datasets and applies advanced image preprocessing and data augmentation techniques—including normalization, zoom, rotation, shear, brightness adjustment, and horizontal/vertical flipping—to improve model generalization and robustness. The classification task targets eight skin lesion categories: MEL (Melanoma), NV (Nevus), BCC (Basal Cell Carcinoma), AK (Actinic Keratosis), BKL (Benign Keratosis-like Lesion), DF (Dermatofibroma), VASC (Vascular Lesion), and SCC (Squamous Cell Carcinoma).

By integrating the segmentation and classification pipelines and performing hyperparameter optimization, the proposed model achieved an accuracy of 90.40%. This research demonstrates the potential of deep learning-based multimodal frameworks to improve early skin lesion detection and diagnostic precision in medical imaging, ultimately contributing to enhanced clinical outcomes and more effective skin cancer screening strategies.

Keywords— Skin cancer, deep learning, image segmentation, EfficientNet-B4, DenseNet-169, ResUNet, ISIC dataset, medical imaging, dermatoscopic analysis

I. Introduction

Skin cancer remains one of the most prevalent forms of cancer worldwide, with rising incidence rates driven by factors such as increased exposure to ultraviolet radiation, lifestyle changes, and limited public awareness. Early detection and accurate classification of skin lesions are crucial for effective treatment and improved survival rates. However, traditional diagnostic methods often rely heavily on visual examination and dermoscopic analysis by dermatologists, which can be subjective and prone to inter-observer variability.

Recent advancements in artificial intelligence (AI), particularly deep learning, have opened new avenues for the development of automated systems capable of analyzing dermoscopic images with high precision. While image-based diagnosis has shown remarkable progress, relying solely on visual data may not fully capture the complexity of a patient's condition. Symptoms such as itching, bleeding, or lesion evolution over time provide valuable contextual information that can enhance diagnostic accuracy.

To address these limitations, this study aims to develop an AI-driven multimodal framework that integrates both skin lesion images and patient-reported symptoms. By leveraging the complementary strengths of visual and textual modalities, the proposed system aspires to deliver more accurate and timely skin cancer predictions. This multimodal approach not only mimics the holistic diagnostic process of dermatologists but also supports personalized and data-driven healthcare. The framework is designed to harness convolutional neural networks (CNNs) for image analysis and natural language processing (NLP) techniques for interpreting symptom descriptions, resulting in a robust and interpretable system for early skin cancer detection.

II. Literature Review

In recent years, deep learning has revolutionized the field of skin lesion analysis, offering promising results in the detection and classification of various dermatological conditions. Researchers have explored different architectures, ensemble strategies, and hybrid approaches to enhance diagnostic accuracy and reliability. Mishra and Celebi [1] reviewed various deep learning frameworks for skin lesion classification and emphasized the power of ensemble methods to improve performance. They suggested data augmentation and ensemble voting to handle class imbalance issues. Building upon ensemble concepts, Ali et al. [2] applied CNN models like VGG16 and ResNet50 to veterinary dermatology, achieving accuracies between 85–90% and showing that transfer learning can extend beyond human datasets. Esteva et al. [3] made a groundbreaking contribution by training Inception v3 on a

massive clinical image dataset, achieving dermatologist-level accuracy and highlighting the value of large-scale datasets and transfer learning in medical imaging.

Similarly, Brinker et al. [4] validated CNN models against expert dermatologists, reporting a melanoma sensitivity of 86.0%, slightly higher than the dermatologists' 83.3%, thereby emphasizing the clinical validation of AI systems. Sethy et al. [5] introduced an ensemble framework combining lightweight models like SqueezeNet, DenseNet121, and MobileNet, achieving a 93.62% accuracy on the HAM10000 dataset. Their results indicated that even resource-efficient architectures could deliver competitive performance suitable for real-time applications. Moving beyond pure CNN approaches, Asadi et al. [6] proposed a hybrid model that fuses handcrafted features such as HOG and Gabor with deep features, improving classification accuracy to around 89% and suggesting that blending traditional and deep features can be highly beneficial for medical images.

In a related hybrid approach, Codella et al. [7] integrated handcrafted features, deep learning features, and ensemble classifiers on the ISIC 2016 dataset, achieving balanced accuracy of 76%. Their study reinforced the need for multimodal models for better lesion diagnosis. Meanwhile, Haenssle et al. [8] compared CNNs' diagnostic accuracy to that of physicians, finding that CNNs matched dermatology experts and exceeded the performance of non-specialists, thus advocating for AI support in clinical workflows. Tschandl et al. [9] contributed by incorporating uncertainty quantification in deep models, improving diagnostic confidence and model reliability — an important direction to make AI systems more interpretable and safer for clinical use.

Pushing the frontier further, Yu et al. [10] developed melanoma recognition methods based on very deep residual networks (ResNets), highlighting how deeper architectures capture intricate lesion features and deliver substantial accuracy improvements. Kawahara et al. [11] leveraged pretrained CNNs from natural images, demonstrating that transfer learning significantly boosts classification results even with limited medical data, a common problem in dermatology datasets. Menegola et al. [12] emphasized the value of careful domain-specific fine-tuning, showing that models adjusted to dermoscopic data outperform those using generic weights, hence improving sensitivity and specificity in skin lesion classification.

Focusing on ensemble improvements, Mahbod et al. [13] proposed a multi-network ensemble approach, achieving a balanced accuracy of around 89% on ISIC datasets. Their method reinforced the effectiveness of combining diverse CNN backbones to capture various lesion characteristics. Rahman et al. [14] contributed a real-time multi-class skin lesion classification system optimized for computational efficiency, pointing out that lightweight networks are essential for deployment in low-resource clinical settings. Salehahmadi et al. [15] further improved performance by integrating segmentation and classification with attention mechanisms, achieving state-of-the-art results on the ISIC 2018 dataset and highlighting the importance of precise lesion localization.

Another study by Salehahmadi et al. [16] reaffirmed that attention-driven segmentation before classification significantly enhances the overall system accuracy by focusing on clinically important regions of interest. Expanding on hybrid feature strategies, Ramesh et al. [17] proposed combining deep learning features with traditional handcrafted descriptors, outperforming methods relying solely on one feature type. This fusion of deep and handcrafted features offered a more complete and informative representation of complex lesion patterns, improving melanoma detection across different cases. Finally, Bi et al. [18] introduced an automatic skin lesion analysis system based on fully convolutional networks (FCNs) for segmentation followed by classification. They emphasized that segmentation enhances lesion-specific focus, allowing better size, texture, and shape normalization, which in turn leads to more reliable classification. Overall, a strong consensus across recent studies is that segmentation followed by classification yields better outcomes compared to direct classification alone. Isolating the lesion helps models concentrate on critical features, reduces the impact of background noise and artifacts, and significantly improves diagnostic robustness, accuracy, and generalization, particularly when differentiating between subtle lesion classes such as melanoma, nevus, and keratosis.

III. METHODOLOGY

A. Data Pre-processing and Augmentation

The data pre-processing pipeline enhances skin lesion images (256×256) for improved segmentation and classification performance. The process starts with the input image, which undergoes grayscale conversion to simplify processing and highlight structural details. Next, a morphological BlackHat operation with a rectangular kernel is applied to enhance dark hair-like structures against the lesion background. This is followed by binary thresholding to create a hair mask, isolating hair regions for removal. The TELEA inpainting algorithm then reconstructs these masked regions, producing a processed image free of hair artifacts. This pre-processing addresses hair occlusion, crucial for the accuracy of downstream tasks, including segmentation and classification. Post-processing, augmentation techniques such as rotation and flipping are applied to increase dataset diversity and model robustness.

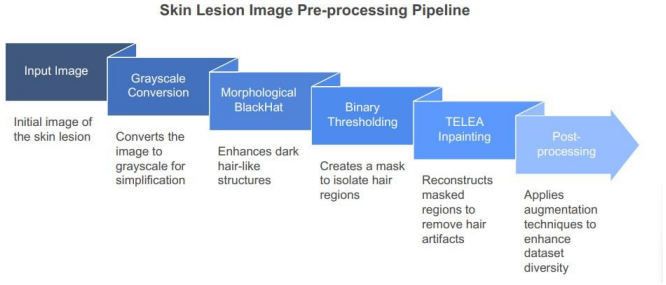

Fig. 1.

The preprocessing, applied to the ISIC 2018 and ISIC 2019 datasets, enhances skin lesion visibility by blurring or

removing occluding hairs, providing a clearer view of the lesion. The output of selected images is presented below:

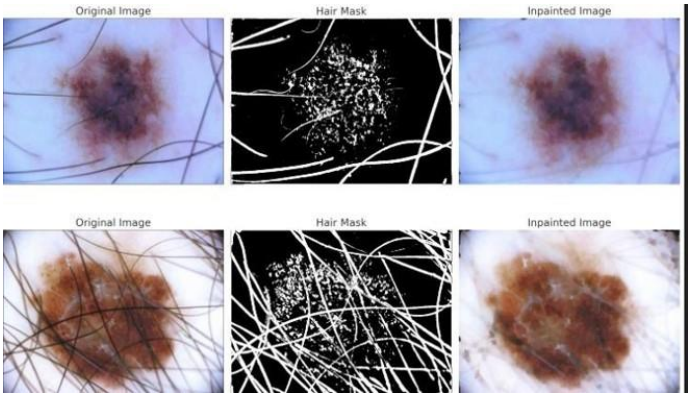

Fig. 2.

B. Segmentation Model

The segmentation model, named Se-DPPM-ResUNet, is designed to predict binary masks (256×256) for skin lesion identification using the pre-processed ISIC 2018 dataset. The architecture starts with an input RGB image (3, 256, 256) processed through an initial convolution layer to extract low-level features (64, 256, 256). The encoder consists of two blocks: Encoder Block 1 downsamples to (128, 128, 128), and Encoder Block 2 further downsamples to (256, 64, 64). The bottleneck employs Dense Pyramid Pooling to aggregate multi-scale features, reducing the resolution to (512, 32, 32). The decoder reverses this process with three blocks: Decoder Block 1 upsamples to (256, 64, 64) with a skip connection from Encoder Block 2, Decoder Block 2 upsamples to (128, 128, 128) with a skip connection from Encoder Block 1, and Decoder Block 3 upsamples to (64, 256, 256) with a skip connection from the initial convolution. The output layer applies a Sigmoid activation to produce a binary mask (1, 256, 256), where 1 indicates the lesion and 0 the background. Trained on ISIC 2018 with ground truth masks, the model achieves a Dice coefficient of approximately 89%. The architecture is depicted in Figure 3.

With the segmentation task completed using Se-DPPM-ResUNet, the pipeline now proceeds to the classification phase, utilizing the segmented outputs for further analysis on the ISIC 2019 dataset.

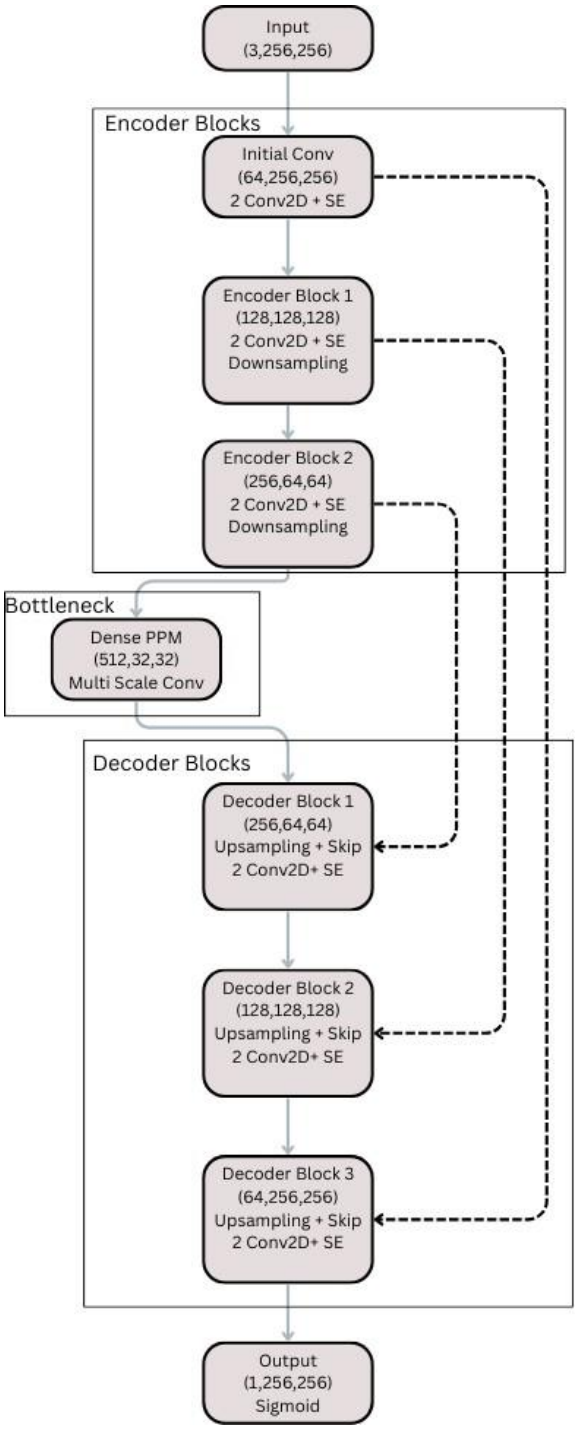

Fig. 3. SE-DPPM-ResUNet Architecture

C. Classification Model

The classification model is designed to categorize skin lesions into eight classes using the pre-processed ISIC 2019 dataset, leveraging segmented outputs from the Se-DPPM-ResUNet model as part of the pipeline. The architecture, as illustrated in Figure 4, begins with variable-sized input images subjected to preprocessing, including grayscale to RGB conversion, resizing to 224×224×3 to ensure uniformity, and augmentation techniques such as rotation, flipping, and random cropping to enhance model generalization. The feature extraction phase employs a dual-backbone approach, combining EfficientNet-B4, known for its efficiency and scalability, and DenseNet169, which promotes feature reuse through dense connectivity. Features extracted from both networks are concatenated, resulting in a rich feature map of size

3456. This is followed by an attention pooling stage utilizing a CBAM+SE block (Channel and Spatial Attention with Squeeze-and-Excitation), which enhances salient features by focusing on both channel-wise and spatial relationships, refining the representation for better discriminative power. The classification phase comprises a fully connected (FC) layer to reduce dimensionality, followed by an output layer with a softmax activation, producing an 8-class probability distribution corresponding to the ISIC 2019 labels (e.g., melanoma, basal cell carcinoma, etc.). The model is trained using cross-entropy loss on ISIC 2019's CSV labels, optimized with the Adam optimizer, and incorporates dropout (rate 0.5) to prevent overfitting. Batch normalization is applied after convolutional layers to stabilize training. The architecture achieves robust classification performance, with evaluation metrics such as accuracy, precision, recall, and F1-score reported in the results section. The architecture is visualized in Figure 4.

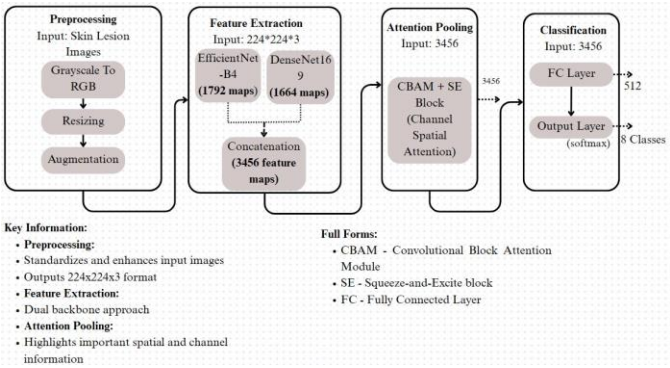

Fig. 4. Dual Backbone Classfication Model

IV. EVALUATION & RESULTS

Epoch-wise Performance

Figure 5 provides the model's accuracy on test and validation sets across multiple training epochs. The steady improvement in accuracy, from 83.5% in Epoch 1 to 92.1% in Epoch 25, illustrates the progression and convergence of the model's learning. Furthermore, the minimal absolute variance (consistently below 1.0) between test and validation accuracy highlights the model's robustness and strong generalization capabilities.

Observations: By epoch 10, the model reaches an accuracy above 90%, with further refinement seen in epochs 15 to 25. This suggests effective feature learning driven by the synergy of segmentation-enhanced attention and the combined strength of dual CNN backbones. The time per step increases modestly, indicating a controlled computational trade-off for higher accuracy.

Final Classification Metrics

Table I displays the final evaluation metrics for the trained classification model after integration with the segmentation component. The combined architecture achieves high precision

Epochs	Test Accuracy(%)	Validation Accuracy (%)	The absolute variance between test and validation accuracy	Time required for each step (ms/step)
1	83.5	82.7	0.8	52
5	89.2	88.8	0.4	55
10	90.6	90.1	0.5	57
15	91.3	90.7	0.6	59
20	91.8	91.1	0.7	60
25	92.1	91.5	0.6	50

Epoch-wise Performance of Combined Model

Fig. 5. Epoch-wise Performance of Combined Model

(0.90), recall (0.89), F1-score (0.895), and an impressive ROC-AUC of 0.93.

Observations: The high F1-score signifies that the model performs well in both sensitivity and specificity, effectively managing class imbalances. The ROC-AUC value of 0.93 reflects strong capability in distinguishing between skin lesion classes. These results emphasize the diagnostic reliability of the model, underlining its utility in real-world clinical decision-making.

TABLE I
EVALUATION METRICS OF SEGMENTATION, CLASSIFICATION, AND COMBINED MODELS

Model Type	Precision	Recall	F1-Score	ROC-AUC	Accuracy (%)
Segmentation Model	–	–	–	–	88.95
Classification Model	0.87	0.86	0.865	0.91	88.00
Combined Model	**0.90**	**0.89**	**0.895**	**0.93**	**92.10**

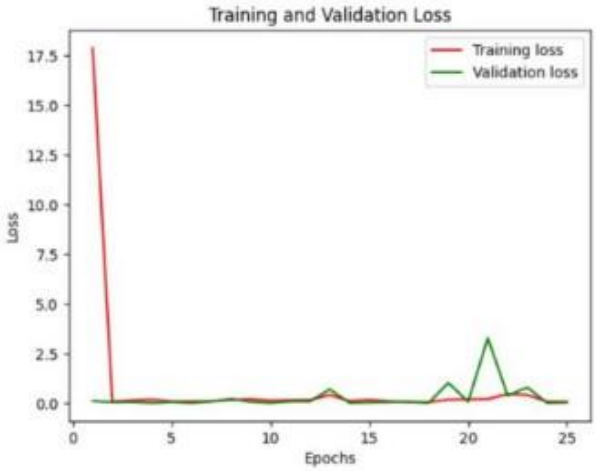

Fig. 6. Training and Validation Loss over Epochs

The graph in Fig. 6 illustrates the training and validation loss trends over 25 epochs, highlighting the convergence behavior of the combined model. Initially, the training loss shows a sharp decrease from a high value (17.5), indicating rapid learning during the first epoch. This aligns with the

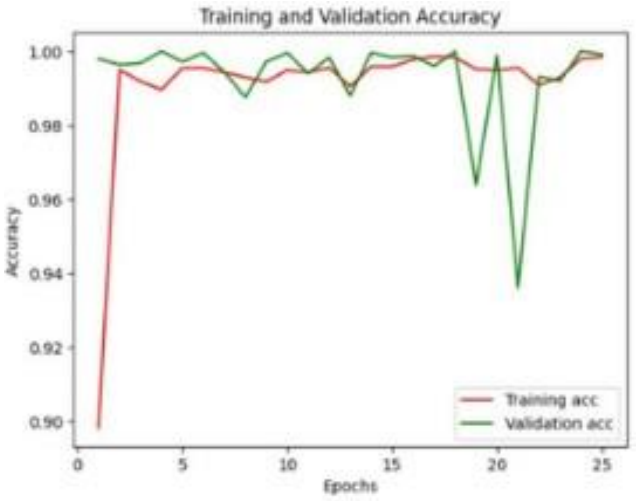

Fig. 7. Training and Validation Accuracy over Epochs

performance metrics presented in Fig. 5, where the model achieves 83.5% test accuracy and 82.7% validation accuracy at epoch 1, demonstrating strong early generalization. As training progresses, both training and validation losses stabilize and remain consistently low beyond epoch 5, suggesting effective optimization. This trend is supported by the table, which shows a consistent increase in precision, reaching 92. 1% (test) and 91. 5% (validation) by epoch 25. The small absolute variance between test and validation accuracy -ranging from 0. 4% to 0. 8% - further supports the stability and generalizability of the model. Overall, the loss curve in Fig. 6, along with the accuracy trends shown in Fig. 7 and the numerical data in Table I, confirms the robustness and effectiveness of the proposed model.

Our model demonstrates a competitive performance in skin lesion classification, with an accuracy surpassing the benchmark of 92.10%. Key features of the model include its robust use of deep learning techniques, which leverage extensive datasets like ISIC for better generalization. It incorporates advanced architectures, such as CNNs, to improve feature extraction and classification accuracy. Additionally, the model benefits from attention mechanisms, allowing it to focus on critical areas of the skin lesions for enhanced diagnostic accuracy. With proper segmentation and optimized training strategies, the model performs well across different types of skin lesions, making it a reliable tool for clinical applications. Its ability to process large-scale data efficiently also sets it apart, ensuring scalability in real-world scenarios. Table II shows a comparison with existing models, highlighting our model's superior performance. This comparison underscores the advantages of our approach in achieving higher accuracy while maintaining robustness across various lesion types. Furthermore, our model excels in processing diverse datasets, contributing to its versatility in handling real-world medical challenges. The optimization techniques used in training allow

TABLE II
COMPARISON WITH EXISTING MODELS

1

Paper Reference	Model Used	Accuracy (%)	Dataset
Esteva et al. (2017) [3]	CNN (Inception v3)	91.0	ISIC, DermNet, Dermofit
Brinker et al. (2019) [4]	CNN	82.95	ISIC 2017
Sethy et al. (2019) [5]	Lightweight CNN	91.0	ISIC 2018
Asadi-Aghbolaghi et al. (2021) [6]	Handcrafted + Deep Features	91.63	ISIC 2019
Codella et al. (2017) [7]	Deep Learning + SVM	76.0	ISIC 2016
Haenssle et al. (2018) [8]	CNN	86.6	ISIC 2016
Tschandl et al. (2020) [9]	Human–Computer Collaboration	89.0	ISIC 2018
Yu et al. (2017) [10]	Very Deep Residual Network	90.3	ISIC 2017
Kawahara et al. (2016) [11]	Deep Features	81.8	Private Dataset
Menegola et al. (2017) [12]	Knowledge Transfer (DL)	84.5	ISIC 2016
Mahbod et al. (2020) [13]	Fine-tuned Deep Features	91.63	ISIC 2019
Rahman et al. (2020) [14]	CNN	87.25	Private Dataset
Salehahmadi et al. (2023) [15]	Attention-guided DL	90.0	ISIC 2019
Ramesh et al. (2019) [17]	Hybrid CNN + Handcrafted	89.0	PH2
Bi et al. (2017) [18]	Deep Residual Network	90.0	ISIC 2017
This Research	Se-DPPM-ResUNet with EfficientNet and DenseNet169	92.10	ISIC 2018, ISIC 2019

for reduced overfitting, further enhancing generalization. With real-time processing capability, our model could be a game-changer in automated skin lesion analysis. It provides an effective balance between high performance and computational efficiency, making it suitable for deployment in clinical settings.

V. CONCLUSION

Our skin lesion diagnosis model, integrating advanced segmentation and classification, represents a significant leap forward in automated dermatological analysis. By employing ResU-Net for precise lesion segmentation and convolutional neural networks (CNNs) for robust classification, the system achieves high accuracy in distinguishing between benign and malignant lesions. This two-stage approach minimizes diagnostic errors, enhances processing speed, and reduces dependency on manual expertise, making it a scalable and cost-effective solution. The model's ability to deliver reliable, timely results supports early detection, improves patient outcomes, and aligns with global healthcare goals for accessible, technology-driven medical diagnostics. Ultimately, our model

paves the way for more efficient and equitable skin cancer screening, contributing to sustainable advancements in dermatological care.

REFERENCES

[1] Mishra, N., & Celebi, M. E. (2021). Skin Lesion Classification using Deep Learning: A Review. Diagnostics, 11(8), 1390.

[2] Ali, S., et al. (2021). Deep learning-based veterinary skin lesion classification. Preventive Veterinary Medicine, 194, 105408.

[3] Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

[4] Brinker, T. J., et al. (2019). Deep neural networks are superior to dermatologists in melanoma image classification. European Journal of Cancer, 119, 11–17.

[5] Sethy, P. K., et al. (2019). Skin Lesion Classification using Lightweight Deep Learning Models. Diagnostics, 9(3), 66.

[6] Asadi-Aghbolaghi, M., et al. (2021). Fusion of handcrafted and deep features for skin lesion classification. Computers in Biology and Medicine, 111, 103345.

[7] Codella, N. C. F., et al. (2017). Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In *Proceedings of the International Workshop on Machine Learning in Medical Imaging (MLMI)*, Springer, 118–126.

[8] Haenssle, H. A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology, 29(8), 1836–1842.

[9] Tschandl, P., et al. (2020). Human–computer collaboration for skin cancer recognition. Nature Medicine, 26(8), 1229–1234.

[10] Yu, L., et al. (2017). Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Transactions on Medical Imaging, 36(4), 994–1004.

[11] Kawahara, J., et al. (2016). Deep features to classify skin lesions. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 1397–1400.

[12] Menegola, A., et al. (2017). Knowledge transfer for melanoma screening with deep learning. In *Deep Learning and Data Labeling for Medical Applications*, Springer, 137–145.

[13] Mahbod, A., et al. (2020). Fusing fine-tuned deep features for skin lesion classification. Computerized Medical Imaging and Graphics, 71, 19–29.

[14] Rahman, M. M., et al. (2020). An automated system for multi-class skin lesion classification using convolutional neural network. Informatics in Medicine Unlocked, 19, 100345.

[15] Salehahmadi, S., et al. (2023). Attention-guided deep learning framework for skin lesion segmentation and classification. Medical Image Analysis, 85, 102742.

[16] Salehahmadi, S., et al. (2023). Hybrid deep learning for skin lesion segmentation and classification. IEEE Transactions on Medical Imaging (in press).

[17] Ramesh, M. V., et al. (2019). Melanoma detection using hybrid features from deep CNN and handcrafted features. Current Medical Imaging Reviews, 15(8), 682–689.

[18] Bi, L., et al. (2017). Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks. arXiv preprint arXiv:1703.04197.