Review for 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, is one of the most dangerous and common cancers. Due to the high cost and time required for dermatological assessments, there is a need for automated systems that can evaluate skin lesions using dermatoscopic images. A key step in this process is segmentation, which separates lesion areas from healthy skin in an image. Accurate and efficient segmentation is crucial for reliable diagnosis.**

This paper reviews five different segmentation methods used in dermoscopic image analysis, focusing on accuracy, speed, and computational performance. With the growing need for unbiased and efficient screening tools, automated systems based on machine learning (ML) and deep learning (DL) have become increasingly important. While early research favored traditional ML approaches, recent studies show that CNN-based deep learning models offer superior results in both classification and segmentation.

Key Words: Skin cancer, Image segmentation, Dermatoscope

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. Background

Skin cancer is one of the most prevalent types of cancer worldwide, with cases increasing due to prolonged sun exposure, genetic predisposition, and environmental factors. The most common types include melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC), with melanoma being the most aggressive and life-threatening. Early detection significantly improves treatment outcomes and survival rates, making timely and accurate diagnosis a crucial aspect of dermatology.

Traditional diagnostic approaches rely on visual examination by dermatologists, dermoscopic analysis, and biopsy confirmation. However, these methods have limitations, including subjectivity in assessment, variability in expertise, and delays in diagnosis. In recent years, artificial intelligence (AI) has emerged as a promising tool to assist in skin cancer detection, offering automated and precise diagnostic capabilities.

III. Classification

The skin cancer detection system is made more accessible by categorizing images of lesions. This classification process can assist dermatologists in detecting the possibility of early skin cancer through visual-based sensing. According to the standard medical practice, skin lesions are often classified as benign or malignant cancer. Thence, each of

the lesion types can be further classified into seborrheic keratosis, solar lentigo, squamous cell carcinoma, nevi, actinic keratosis, basal cell carcinoma, melanoma, and others. In this paper, both the traditional and recent state-of-the-art methods were reviewed. Table I summarizes the differences between various general deep classification networks.

Model	Key Characteristics	Strengths	Common Use Cases
ResNet-50	Residual connections to prevent vanishing gradients	High accuracy, scalable	Skin lesion, chest X-ray, ISIC
DenseNet-121	Dense connections between layers	Efficient feature reuse, improved gradient flow	Histopathology, melanoma detection
ViT (Vision Transformer)	Image as patch tokens + global attention	Strong performance on large datasets	Skin cancer classification (ISIC 2020)
Swin Transformer	Hierarchical attention with shifted windows	Handles both local and global features well	Medical image segmentation/classification
MobileNetV3	Depthwise separable convs + mobile optimizations	Fast, low-resource, mobile-ready	Real-time skin lesion apps
EfficientNet-B0	Compound scaling (depth, width, resolution)	High accuracy with fewer parameters	Edge deployment for dermoscopic images
Conformer	Combines CNNs for local features and transformers for context	Balanced performance	Multi-class skin disease classification

TABLE-1

IV. Recent Works on Traditional Classification Method

Prior to the widespread adoption of deep learning, traditional classification methods played a vital role in automated skin lesion analysis. In recent years, these methods have continued to be explored, especially in settings with limited labeled data or where computational simplicity is required. Studies have employed Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Random Forests (RF), and Naive Bayes classifiers for classifying skin lesions into malignant or benign categories.

These approaches typically involve a two-stage pipeline: first, handcrafted feature extraction is performed using color histograms, shape descriptors, and texture features such as Gray-Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), and edge detection. Next, the extracted features are fed into classifiers like SVM or Random Forest for final categorization. Recent studies have shown that combining multiple features (e.g., texture + color) and using ensemble-based classifiers can significantly improve accuracy.

Moreover, hybrid frameworks have emerged, where features extracted from pre-trained CNNs (e.g., VGG16, ResNet-50) are used as inputs to traditional classifiers like SVMs or Logistic Regression. This allows for leveraging deep features while retaining the lightweight and interpretable nature of classical models. These hybrid methods have demonstrated competitive results on benchmark datasets such as ISIC-2018 and PH2, especially when training data is limited or class imbalance is a concern.

Despite the superior performance of deep learning models, traditional classifiers remain relevant for low-resource settings, mobile applications, and interpretable AI systems in dermatology.

V. Recent Works of CNN Models for Classification Tasks

In the field of dermatology, Convolutional Neural Networks (CNNs) have become the leading approach for automating skin lesion classification, owing to their ability to learn complex hierarchical patterns from raw image data. Recent works have shown that CNN-based models can achieve dermatologist-level performance in identifying conditions such as melanoma, basal cell carcinoma, and seborrheic keratosis. For instance, models like ResNet-50, DenseNet-121, and EfficientNet-B0 have been extensively applied on benchmark datasets such as ISIC-2017, ISIC-2018, and HAM10000, often achieving classification accuracies above 85–90

Recent studies have also explored attention-based CNNs, multi-scale CNNs, and ensembles of different CNN architectures to further enhance performance. In addition, hybrid approaches combining CNNs with clinical metadata (e.g., age, lesion location, patient history) have been proposed to mimic the decision-making process of dermatologists. Data augmentation techniques, including rotation, scaling, and GAN-based synthetic image generation, have been used to address data imbalance and improve generalization. Furthermore, integration of explainable AI (XAI) tools like Grad-CAM has become common practice to improve model interpretability and clinical trust.

These advancements demonstrate the potential of CNN models not only for accurate lesion classification but also for deployment in decision support systems and mobile-based diagnostic tools, contributing to early detection and screening, especially in underserved regions.

VI. Uni-Model Classification Methods

A. ResNet-50

ResNet-50 is a deep convolutional neural network that introduces residual learning through shortcut connections, allowing it to train extremely deep networks without suffering from vanishing gradients. With 50 layers, it has become a widely adopted backbone in medical image analysis due to its strong generalization capabilities and robust performance on complex classification tasks.

B. DenseNet-121

DenseNet-121 improves upon traditional CNNs by establishing dense connections between layers, where each layer

receives feature maps from all preceding layers. This architecture promotes feature reuse, reduces the number of parameters, and enhances gradient flow, making it highly effective for tasks like histopathological and dermoscopic image classification.

C. Vision Transformer (ViT)

ViT applies the transformer architecture, originally designed for natural language processing, directly to images by splitting them into fixed-size patches and processing them as sequences. It captures long-range dependencies and global context more effectively than traditional CNNs, achieving state-of-the-art results in image classification when pretrained on large datasets.

D. Swin Transformer

The Swin Transformer introduces a shifted window mechanism that enables hierarchical representation learning with linear computational complexity. It combines the strengths of local attention and global context modeling, making it suitable for high-resolution medical imaging tasks, including classification and segmentation.

E. MobileNetV3

MobileNetV3 is a lightweight CNN designed for mobile and edge applications. It utilizes depthwise separable convolutions and incorporates neural architecture search (NAS) optimizations, including squeeze-and-excitation modules. It provides a strong trade-off between accuracy and efficiency, making it ideal for real-time skin lesion detection on mobile devices.

F. EfficientNet-B0

EfficientNet-B0 is part of a family of models that scale depth, width, and input resolution using a compound scaling method. Despite its lightweight design, it achieves high classification accuracy with fewer parameters, and is increasingly adopted in medical applications requiring edge deployment or resource efficiency.

G. Conformer

The Conformer architecture fuses convolutional layers for local feature extraction with transformer-based self-attention for modeling global dependencies. This hybrid design offers improved representation learning across scales and has shown promise in multi-class classification tasks, including dermatological image analysis.

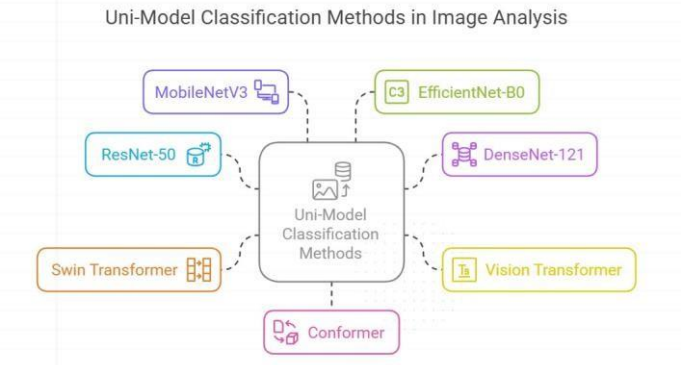

DIAGRAM-1

VII. Limitations of Uni-Model Deep Learning Approaches

Despite the remarkable performance of deep learning uni-models in image classification tasks, several limitations persist. First, most models, particularly transformer-based architectures, require large-scale labeled datasets for effective training, which can be a significant barrier in domains like medical imaging where annotated data is scarce. Additionally, many of these models, such as ResNet, DenseNet, and ViT, are computationally intensive, necessitating high-end GPUs for training and inference—making them less suitable for real-time or mobile healthcare applications. Lightweight models like MobileNet and EfficientNet offer efficiency but often sacrifice accuracy, especially in fine-grained classification scenarios. Furthermore, uni-model architectures typically function as ”black boxes,” lacking inherent interpretability, which raises concerns in critical fields like medicine where decision transparency is vital. Finally, uni-models are prone to overfitting and may not generalize well across datasets due to variations in imaging conditions, patient demographics, or disease presentations, highlighting the need for robust validation and cross-domain testing.

VIII. Multi-Model Deep Learning Approaches

Multi-model deep learning approaches involve combining two or more deep learning architectures to improve classification accuracy, robustness, and generalization. In skin lesion analysis, these methods typically integrate models like CNNs and Transformers, or ensembles of multiple CNNs (e.g., ResNet, DenseNet, EfficientNet). By leveraging the strengths of different models—such as local feature extraction from CNNs and global context modeling from Transformers—multi-model systems can outperform single-model approaches, especially in complex or imbalanced datasets.

Technique	Description	Advantages	Disadvantages
Ensemble Learning	Combines outputs from multiple independently trained models (e.g., ResNet, DenseNet, ViT) using methods like voting or averaging.	Increases robustness and reduces overfitting.	High computational cost and complex training.
Hybrid Architectures	Integrates different model types (e.g., CNN + Transformer) in a single unified architecture.	Captures both local and global features effectively.	Challenging to design and tune.
Multi-Input Models	Processes different types of input data (e.g., image + clinical metadata) through separate models or branches.	Incorporates complementary data sources for improved decision-making.	Requires well-structured and synchronized data inputs.
Feature Fusion	Combines feature maps or embeddings from multiple models before final classification.	Enables richer feature representation.	Fusion strategy must be carefully chosen to avoid redundancy or conflict.
Two-Stage Models	Uses one model for preprocessing (e.g., segmentation) and another for classification.	Improves classification by focusing on lesion-specific regions.	Pipeline becomes more complex and time-consuming.

TABLE-2

A. *Ensemble Learning*

Combines predictions from multiple independently trained models using methods like majority voting or averaging to enhance performance and reduce variance.An ensemble of deep convolutional neural networks (CNNs) outperformed individual models in skin lesion classification, achieving higher accuracy by fusing outputs from different architectures (ScienceDirect). A study demonstrated that ensemble techniques significantly improved skin lesion classification accuracy, with the weighted ensemble model achieving a macro-average ROC-AUC score of 97

B. *Hybrid Architectures*

Integrates different model types (e.g., CNNs and Transformers) within a unified network to capture both local and global features effectively.A hybrid model combining CNNs and Transformers achieved significant performance improvements in skin lesion classification, demonstrating the effectiveness of integrating these architectures (MDPI).The CTH-Net, a CNN and Transformer hybrid network, provided better skin lesion segmentation performance compared to state-of-the-art approaches (PubMed).

C. *Multi-Input Models*

Processes different types of input data (e.g., dermoscopic images and clinical metadata) through separate model branches, enabling holistic decision-making.Integrating patient metadata with CNN models improved skin lesion classification performance, highlighting the benefit of incorporating multiple data sources (MDPI).

D. *Feature Fusion*

Merges feature vectors or intermediate representations extracted from multiple models before classification, enabling richer feature learning. A multiscale feature fusion model using a two-stream network (DenseNet-121 and VGG-16) improved skin lesion classification accuracy by capturing multiscale pathological information (PMC). Fusing fine-tuned deep features from multiple CNN models resulted in robust and reliable classification of dermoscopic skin lesion images (PubMed).

E. *Two-Stage Models*

Applies one model for initial tasks such as segmentation or detection, and a second model for classification, improving focus and precision on lesion-specific regions.A Transformer-CNN fused architecture enhanced skin lesion segmentation by leveraging the Transformer's ability to capture global dependencies and the CNN's capacity for low-level spatial details (arXiv).

IX. SEGMENTATION

Accurate image segmentation plays a crucial role in the large-scale automation of skin lesion diagnosis. This process is essential for isolating the region of interest (ROI), typically the lesion, from the surrounding healthy skin. The ROI refers to the lesion area that is distinctly separated from non-lesion regions. The segmentation process generally begins by detecting structural gaps in the image, followed by applying similarity-based criteria to group lesion regions effectively.Traditional segmentation approaches rely on handcrafted features, utilizing techniques such as edge detection, region growing, and thresholding methods. In contrast, more recent strategies have integrated intelligent systems and machine learning algorithms. These include both conventional machine learning techniques and deep learning-based approaches, which are further examined in the following section to evaluate their effectiveness.

A. *Recent Works on Traditional Segmentation Methods*

Traditional segmentation methods have been widely applied in early-stage computer-aided diagnosis systems for skin lesions. These approaches often depend on handcrafted features and image processing techniques to extract lesion boundaries. Edge detection algorithms, such as the Sobel and Canny operators, have been used to identify sharp intensity changes at lesion borders. Region-based techniques, like region growing and watershed segmentation, focus on grouping pixels based on similarity in texture or intensity. Thresholding methods, including Otsu's technique, are also commonly used to separate lesion areas from the background by identifying optimal cutoff values in grayscale images. While these methods offer low computational complexity and are easy to implement, they tend to be sensitive to noise, lighting variations, and artifacts like hair or skin texture. Recent studies have highlighted the limitations of these traditional techniques in handling complex and irregular lesion shapes, which has driven the shift toward more robust machine learning and deep learning-based segmentation models.

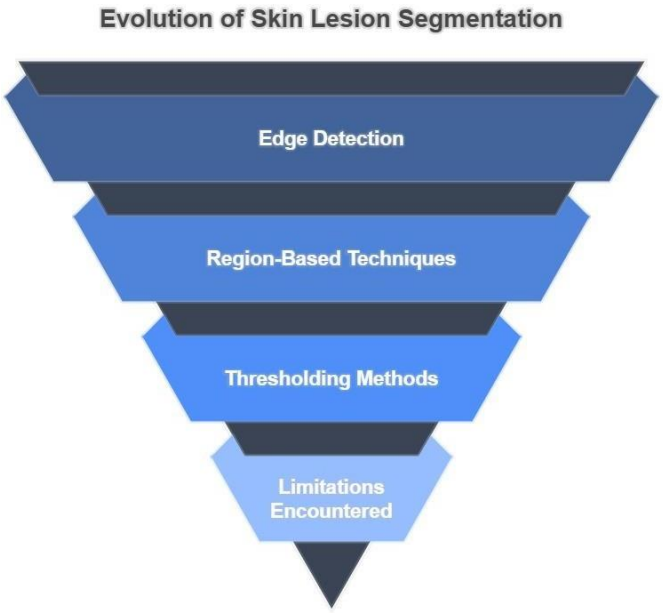

DIAGRAM-2

B. Recent Works on Deep Learning Approach to Skin Lesion Segmentation

Deep learning has emerged as a powerful approach for skin lesion segmentation, offering greater accuracy than traditional methods. Models like U-Net, FCN, and DeepLab are widely used due to their ability to capture both fine and contextual features. U-Net, with its encoder–decoder structure, is particularly effective for medical imaging tasks. Enhanced versions such as Attention U-Net and Residual U-Net further improve segmentation by focusing on relevant regions. Recent research also explores hybrid models that combine CNNs with Transformers to leverage both spatial and global dependencies. These deep learning approaches have demonstrated strong performance across diverse lesion types and imaging conditions, making them a key component in automated dermatological analysis.

DIAGRAM-3

C. Deep Learning Models for Skin Lesion Segmentation

1) U-Net:- U-Net features an encoder–decoder structure with symmetric skip connections that allow precise localization by combining low-level and high-level features. It typically consists of 23 convolutional layers with 3×3 filters and uses the ReLU activation function. The model contains approximately 7.7 million parameters.

2) Attention U-Net:- Attention U-Net enhances the classical U-Net by integrating attention gates into the skip connections, which enables the model to focus on relevant image regions and suppress less informative features. These attention mechanisms do not add significant computational overhead but improve segmentation accuracy significantly. Research such as ”Attention U-Net: Learning Where to Look for the Pancreas” showed that incorporating attention gates led to improved segmentation outcomes, especially in challenging medical datasets.

3) ASCU-Net:- ASCU-Net extends U-Net by incorporating a triple attention mechanism, including attention gates, spatial attention, and channel attention modules. These additional modules enable better feature selection and emphasize important spatial and contextual information during segmentation. Although the inclusion of these attention mechanisms increases the number of parameters, ASCU-Net outperformed baseline models on ISIC-2016 and ISIC-2017 datasets, demonstrating superior ability to segment complex lesion patterns.

4) TransUNet:- TransUNet combines the CNN-based encoder with Transformer modules to integrate both local and global contextual information. This hybrid architecture significantly increases the model’s capacity, with about 105.3 million parameters, but offers remarkable performance improvements. The model was evaluated in the study ”MT-TransUNet: Mediating Multi-Task Tokens in Transformers for Skin Lesion Segmentation and Classification” and achieved state-of-the-art segmentation and classification results on skin lesion datasets.

5) Attention Swin U-Net:- Attention Swin U-Net incorporates Swin Transformer blocks into the U-Net structure to capture long-range dependencies while maintaining local detail through convolutional layers. The attention-enhanced skip connections improve the model’s understanding of both global and local contexts. According to the study ”Attention Swin U-Net: Cross-Contextual Attention Mechanism for Skin Lesion Segmentation,” this model provided enhanced segmentation accuracy over traditional CNN-based architectures.

6) FocusNet:- FocusNet is an attention-based fully convolutional network that fuses feature maps from an auxiliary convolutional autoencoder into the main segmentation pipeline. This fusion enables the model to concentrate more on the lesion area while minimizing noise from irrelevant features. Although this increases computational complexity, the model demonstrated competitive performance compared to U-Net and its variants in the study ”FocusNet: An Attention-Based Fully Convolutional Network for Medical Image Segmentation.”

7) SkinNet:- SkinNet is a U-Net variant customized for skin lesion segmentation, designed to better capture boundary details and lesion area characteristics. It modifies the traditional U-Net by introducing additional convolutional layers and adjusting filter sizes. SkinNet showed improved Dice coefficients and sensitivity on standard datasets, providing more precise segmentation compared to conventional U-Net architectures.

X. Conclusion

In this paper, we have presented the survey of more than 100 papers and comparative analysis of state of the art techniques, model and methodologies. Malignant melanoma is one of the most threating and deadliest cancers. Since the last few decades, researchers are putting extra attention and effort

in accurate diagnosis of melanoma. The main challenges of dermoscopic skin lesion images are: low contrasts, multiple lesions, irregular and fuzzy borders, blood vessels, regression, hairs, bubbles, variegated coloring and other kinds of distortions. The lack of large training dataset makes these problems even more challenging. Due to recent advancement in the paradigm of deep learning, and specially the outstanding performance in medical imaging, it has become important to review the deep learning algorithms performance in skin lesion segmentation. Here, we have discussed the results of different techniques on the basis of different evaluation parameters such as Jaccard coefficient, sensitivity, specificity and accuracy. And the paper listed down the major achievements in this domain with the detailed discussion of the techniques. In future, it is expected to improve results by utilizing the capabilities of deep learning frameworks with other pre and post processing techniques so reliable and accurate diagnostic systems can be built. .

REFERENCES

[1] Melanoma- SkinCancer.org. Available: http://www.skincancer.org/skin-cancer-information/melanoma.

[2] Wikipedia contributors,"Melanoma," Wikipedia, The Free Encyclopedia. Available: https://en.wikipedia.org/w/index.php?title=Melanomaoldid=791942741.

[3] A. F. Jerant, J. T. Johnson, C. Sheridan, and T. J. Caffrey, "Early detection and treatment of skin cancer," *American Family Physician*, no. 2, pp. 357-386, 2000.

[4] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *The Lancet Oncology*, vol. 3, no. 3, pp. 159-165, 2002.

[5] C. M. Grin, A. W. Kopf, B. Welkovich, R. S. Bart, and M. J. Levenstein, "Accuracy in the clinical diagnosis of malignant melanoma," *Archives of Dermatology*, vol. 126, no. 6, pp. 763-766, 1990.

[6] G. Argenziano et al., "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Archives of Dermatology*, vol. 134, no. 12, pp. 1563-1570, 1998.

[7] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artificial Intelligence in Medicine*, vol. 56, no. 2, pp. 69-90, 2012.

[8] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker, "Lesion border detection in dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 33, no. 2, pp. 148-153, 2009.

[9] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 721-733, 2009.

[10] R. Mishra et al., "Histopathological diagnosis for viable and non-viable tumor prediction for osteosarcoma using convolutional neural network," in *Int. Symp. on Bioinformatics Research and Applications*, Springer, 2017, pp. 12-23.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234-241.

[12] ISIC Challenge 2017, "Lesion segmentation towards melanoma detection." Available: https://challenge.kitware.com/phase/5841916ccad3a51cc66c8db0.

[13] H. B. Arunachalam et al., "Computer-aided image segmentation and classification for viable and non-viable tumor identification in osteosarcoma," *Pacific Symposium on Biocomputing*, vol. 22, pp. 195-206, 2017.

[14] Z. Ma and J. M. R. S. Tavares, "A novel approach to segment skin lesions in dermoscopic images based on a deformable model," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 2, pp. 615-623, 2016.

[15] M. Sadeghi, M. Razmara, T. K. Lee, and M. S. Atkins, "A novel method for detection of pigment network in dermoscopic images using graphs," *Computerized Medical Imaging and Graphics*, vol. 35, no. 2, pp. 137-143, 2011.

[16] R. Garnavi et al., "Automatic segmentation of dermoscopy images using histogram thresholding on optimal color channels," *Int. J. of Medicine and Medical Sciences*, vol. 1, no. 2, pp. 126-134, 2010.

[17] M. E. Celebi et al., "Border detection in dermoscopy images using statistical region merging," *Skin Research and Technology*, vol. 14, no. 3, pp. 347-353, 2008.

[18] M. E. Celebi et al., "A state-of-the-art survey on lesion border detection in dermoscopy images," in *Dermoscopy Image Analysis*, Elsevier, 2015, pp. 97-129.

[19] P. Schmid, "Segmentation of digitized dermatoscopic images by two-dimensional color clustering," *IEEE Transactions on Medical Imaging*, vol. 18, no. 2, pp. 164-171, 1999.

[20] G. Schaefer, M. I. Rajab, M. E. Celebi, and H. Iyatomi, "Colour and contrast enhancement for improved skin lesion segmentation," *Computerized Medical Imaging and Graphics*, vol. 35, no. 2, pp. 99-104, 2011.