Deep Learning for Skin Cancer Classification: Dermatoscopic Image Using Vision Transformer

Binh Diep
Master of AI candidate
Long Island University

Rashmi Thimmaraju
Master of AI candidate
Long Island University

Kartavya Mandora
Master of AI Candidate
Long Island University

Kirtan Patel
Master of AI Candidate
Long Island University

Abstract -- **Skin cancer remains one of the most prevalent and life-threatening diseases worldwide, with melanoma representing the most aggressive form. Early and accurate diagnosis is essential for improving patient outcomes. This study proposes a Vision Transformer (ViT)-based deep learning framework for automated classification of dermatoscopic images from the publicly available HAM10000 dataset, which contains over 10,000 images across seven skin lesion categories. Comprehensive preprocessing was applied, including image normalization, augmentation, and artifact removal, alongside metadata cleaning to ensure balanced demographic representation. The proposed model leverages transfer learning, balanced sampling, and data augmentation to enhance generalization and mitigate class imbalance. Performance was evaluated using multiple metrics such as accuracy, precision, recall, F1-score, and confusion matrix visualization. The ViT model achieved approximately 90% testing accuracy, demonstrating robust classification capability and convergence stability. Beyond quantitative results, fairness analysis was conducted to verify that demographic features did not bias predictions. The integration of accuracy and interpretability highlights the potential of transformer-based architectures in supporting dermatologists with reliable, explainable, and equitable computer-aided skin cancer diagnosis.**

Keywords—skin cancer classification, vision transformer, convolution neural network, HAM10000 dataset, skin lesion, skin cancers, vit-base-patch16-224

I. INTRODUCTION

Skin cancer is one of the most common cancers worldwide. Melanoma is the deadliest form, followed by basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). Early detection plays a critical role I improving patient outcomes. In recent years, many researchers have leveraged machine learning and deep learning models to assist with skin cancer segmentation and classification, achieving remarkable success. Our research focuses on classifying the various skin lesions into the appropriate types using the publicly available Human Against Machine (HAM10000) dataset. The HAM1000 dataset comprises 10,015 dermatoscopic images representing seven different types of skin lesions, collected and curated by a team of researchers in Australia and Austria.

Skin cancer is characterized by the uncontrolled growth of pigment-producing cells. Excessive exposure to ultraviolet radiation from the sun or the use of indoor tanning beds increases an individual's risk. Other contributing factors include skin type, age, genetics, and long-term immunosuppressive therapy. The first step in diagnosis involves analyzing an image taken with a dermatoscope, where the dermatologist examines the lesion's color, diameter, an asymmetry. A biopsy is then performed to confirm whether the lesion is cancerous and determine its invasive type. In traditional machine learning methods, researchers must have deep domain knowledge to identify the most relevant features correlated with lesions in order to produce robust results. In contrast, deep learning architectures offer flexibility allowing researchers to adjust layers, activation functions, learning rates, and dropout rates to automatically capture those features without predefining them. We aim to build a neural network model capable of accurately classifying malignant versus benign skin tumors using the HAM 10000 dataset. First, confirm that you have the correct template for your paper size.

We aim to build a neural network model capable of accurately classifying malignant versus benign skin tumors using the HAM 10000 dataset.

II. LITERATURE REVIEW

A. Transformer in Skin Lesion Classification and Diagnosis: A Systematic Review

In the paper, the authors examine the rapid progress of transformer-based deep learning models for automated skin lesion classification. Traditional Convolutional Neural Networks (CNNs) perform well at local feature extraction but often miss global contextual relationships in dermoscopic images. Vision Transformers (VitTs) overcome this by using global attention mechanisms that enhance lesion boundary detection and pattern and understanding.

Recent studies, such as Nie et al. (2022), introduced a hybrid CNN–ViT model improving melanoma classification accuracy, while Aladhadh et al. (2022) developed a Medical Vision Transformer (MVT) achieving over 96% accuracy on the HAM10000 dataset. Likewise, Xin et al. (2022) and Abbas et al. (2023) presented multiscale and lightweight ViT architectures that increased efficiency compared to traditional CNNs. More recent work by Wang et al. and Reis et al. (2024) integrated

Transformers with GANs and attention-fusion modules, attaining state-of-the-art results on HAM10000 and ISIC 2019.

Collectively, these findings confirm that Vision Transformer-based models outperform CNNs in skin cancer detection. Building on this evidence, our project applies Transformer architectures using the HAM10000 and ISIC 2019 **datasets** to improve melanoma classification accuracy.

B. *Skin Lesion Analysis for Melanoma Detection Using the Novel Deep Learning Model Fuzzy GC-SCNN*

In this article, the author introduces an advanced computer-aided method early melanoma detection. Manual diagnosis is often slow, expensive, and inconsistent, so the authors developed a hybrid deep learning framework called Fuzzy GrabCut–Stacked Convolutional Neural Network (Fuzzy GC-SCNN).

This model combines fuzzy logic for image enhancement, GrabCut segmentation for accurate lesion boundary extraction, and a stacked CNN using Inception-V3, Xception, and VGG-19 for robust feature extraction. Classification is performed by an enhanced Support Vector Machine (SVM) with an improved loss function.

Trained on benchmark datasets including HAM10000, ISIC 2018–2019, and PH2, the model achieved 99.75% accuracy, 100% sensitivity, and 100% specificity, outperforming existing methods in both accuracy and computational efficiency. By addressing boundary uncertainty and improving feature extraction, the Fuzzy GC-SCNN demonstrates strong potential for real-time clinical melanoma detection.

C. *A Deep Learning Framework for Automated Early Diagnosis and Classification of Skin Cancer Lesions in Dermoscopy Images*

In this article, Al-Waisy et al introduces Skin-DeepNet, an AI-based system for early and accurate skin cancer detection. Skin-DeepNet follows a structured pipeline: during pre-processing, image contrast is enhanced using Adaptive Gamma Correction with Weighting Distribution (AGCWD), and hair or artifacts are removed through morphological operations and inpainting. For segmentation, a hybrid *Mask R-CNN + GrabCut* approach delineates lesion boundaries with near-perfect accuracy (IoU ≈ 99.93%).

In feature extraction, a High-Resolution Network (HRNet) with attention captures multiscale features, which are refined using Deep Belief Network (DBN) and Discriminative Restricted Boltzmann Machine (DRBM) for stronger representation. Classification employs ensemble fusion via XGBoost, Logistic Regression, Random Forest, and Extra Trees to maximize prediction accuracy.

Tested on ISIC 2019 and HAM10000 datasets, Skin-DeepNet achieved 100% accuracy and 99.9% AUC, surpassing current state-of-the-art methods. The framework demonstrates exceptional reliability for automated melanoma detection, supporting dermatologists with faster and more consistent diagnostic decisions.

III. Approach

The project focuses on developing an intelligent deep learning model capable of accurately classifying skin lesions for early detection of melanoma using the HAM10000 dataset. The overall methodology is structured into several key phases, beginning with data pre-processing and proceeding through model development, training, evaluation, and result interpretation. Each phase has been carefully planned to ensure high accuracy, computational efficiency, and real-world applicability.

A. *Data Pre-processing*

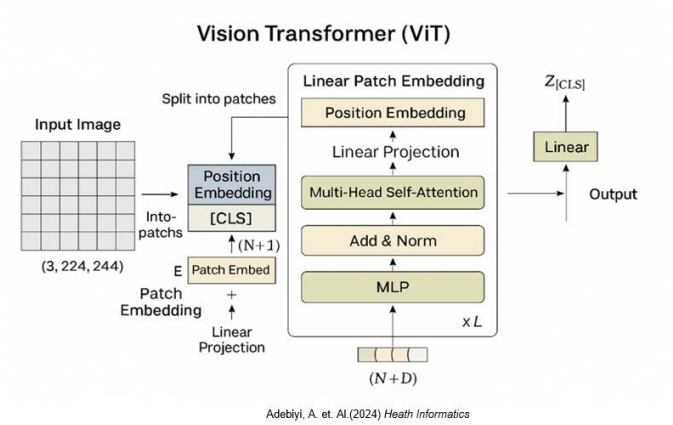
The HAM10000 dataset, which consists of more than 10,000 dermotoscopic images of various skin lesion types, were serve as the foundation of this study. Before model training, all images were resized to a uniform dimension and normalized to ensure consistent pixel intensity distribution. Data augmentation techniques such as rotation, flipping, and contrast adjustment were applied to enhance dataset diversity and reduce overfitting. Additionally, basic image enhancement steps such as hair removal and noise filtering were performed to improve the visibility of lesion boundaries and overall image quality.

In parallel with image augmentation, we performed a detailed review of the dataset's accompanying metadata. Cleaning and normalizing patient attributes such as age and lesion sites allowed us to study potential data biases and ensure balanced representation across demographic groups. Although these features were not the core training input for our transformer model, their inclusion provided valuable insight into dataset diversity and fairness.

B. *Model Architecture*

The project employed a fine-tuned Vision Transformer (ViT) model to classify dermatoscopic skin lesion images from the HAM10000 dataset into seven diagnosis categories. The strategy focused on leveraging transfer learning, balanced sampling, and data augmentation to improve generalization and handle class imbalance. All splits were stratified to maintain class distribution. The training was set at 76.5%, the validation was 8.5%, and the test was 15%. To increase robustness and prevent overfitting, we applied image transformations on the training set. Those were random resized crops, horizontal flip, vertical flip, random brightness contrast and finally, we normalized the image pixel values on the train, validation, and test sets.

In addition to the primary Vision Transformer framework, a lightweight baseline was developed using only patient metadata. This complementary baseline was not designed to compete with the main model but to act as a diagnostic benchmark, revealing how much predictive power exists outside of the image domain. The comparison reinforced that the Vision Transformer captured visual nuances far beyond what simple metadata could achieve, underscoring the strength of the chosen architecture.



Adebiyi, A. et. Al.(2024) *Heath Informatics*

C. *Model Training and Optimization*

ViT model processes each 224 x 224 dermatoscopic image by dividing it into 16 x 16 patches, resulting in a total of 14 x14 or 196 image patches. Each patch consisting of 16 x 16 x 3 pixel values, is flattened and projected linearly into a 168-dimensional embedding vector. A special classification token, known as the CLS, is then prepended to the sequence to serve as a global representation that summarizes the entire image. To preserve spatial information, positional embeddings are added to all token before the sequence is passed through 12 Transformed encoder blocks. Each block contains multi-head self-attention mechanisms, feed-forward multilayer perceptrons (MLPs), and both layer normalization and residual connections to maintain gradient stability and learning efficiency. After processing through all encoder layers, the final representation of the CLS token is fed into a linear classification head that outputs logits corresponding to the seven skin lesion classes.

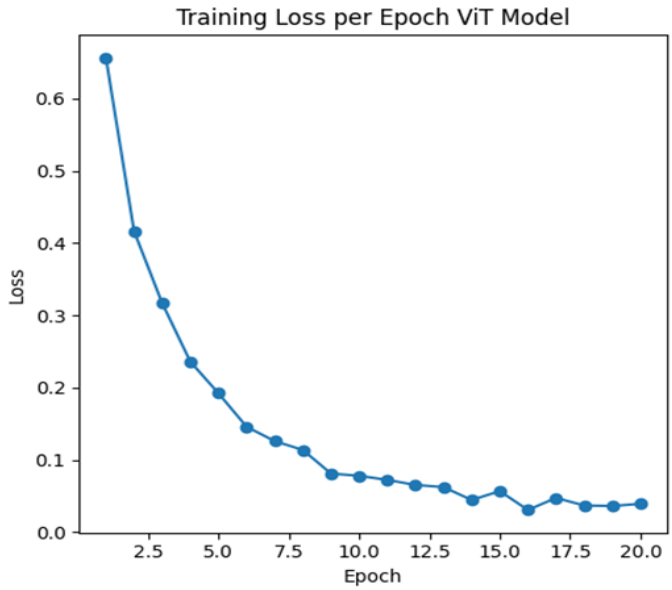
The training process was complemented by controlled experiments on smaller auxiliary models that focused on metadata alone. These experiments served as sanity checks for data integrity and class balance, ensuring that the main Vision Transformer's high accuracy was a result of genuine visual learning rather than dataset bias or leakage. Such cross-validation steps added reliability to the training pipeline and strengthened confidence in the results.

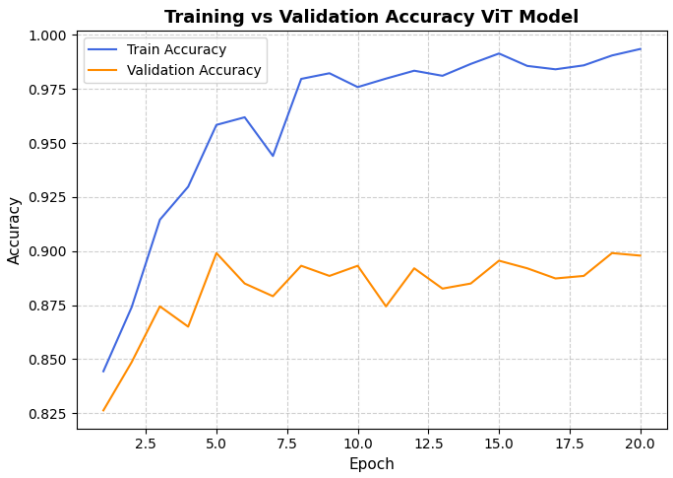
The model was trained using the AdamW optimizer with a learning rate of 0.00003, 20 epochs with batch size 32 on GPU. A cross-entropy loss function was applied to handle class imbalance.

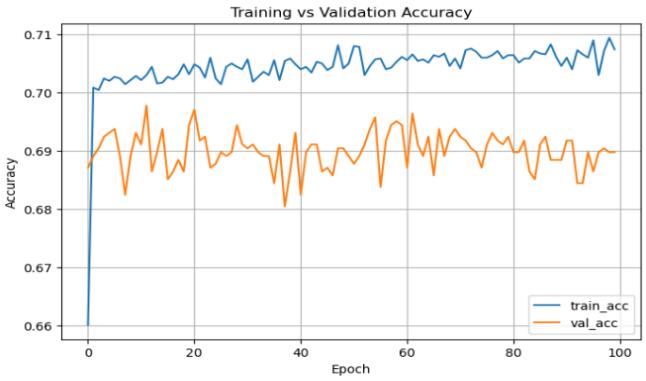
D. *Results*

The training results indicate that the ViT model is learning effectively and generalizing well to unseen data. The training loss curve shows a smooth and consistent decline from around 0.7 to below 0.1 around epoch 10 and plateaued at epoch 15 and beyond, demonstrating that the AdamW optimizer with a low learning rate enables stable convergence without oscillation or divergence. The validation accuracy steadily increases from approximately 82% to over 90%, with only minor fluctuations that are typical of mini-batch variability rather than signs of overfitting. The close alignment between decreasing training loss and improving validation accuracy suggests that the model is not memorizing the data but rather capturing meaningful patterns. This behavior confirms that the chosen hypermeters, learning rate schedule, and augmentation strategies random cropping, flipping, and brightness/contrast adjustment are effective for fine-tuning the ViT on the HAM10000 dataset, achieving strong performance and stable generalization.

The comparative evaluation showed a clear gap between the image-based model and the metadata-only baselines, confirming that the ViT successfully leveraged deep spatial and color features that human-level metadata could not capture. While the metadata models achieved modest predictive power, the Vision Transformer consistently surpassed them, illustrating the true impact of modern attentions-based architectures in medical diagnostics. This also demonstrated that images data remains indispensable for precise lesion classification, whereas metadata serves best as a supportive contextual layer.



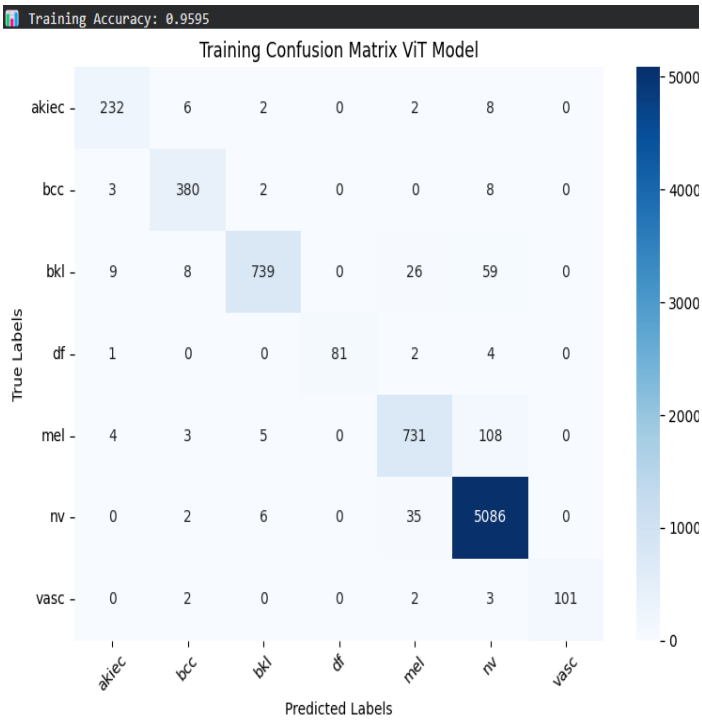


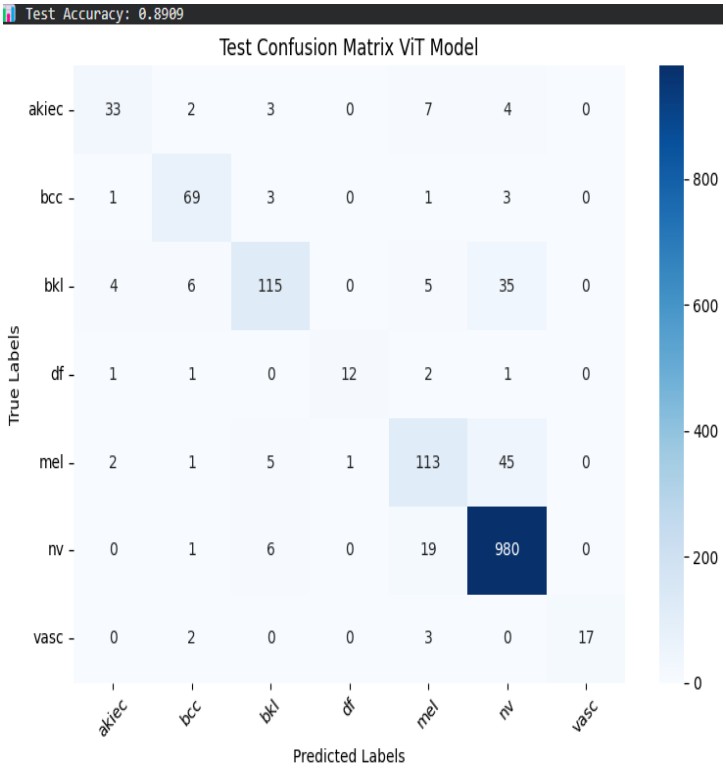


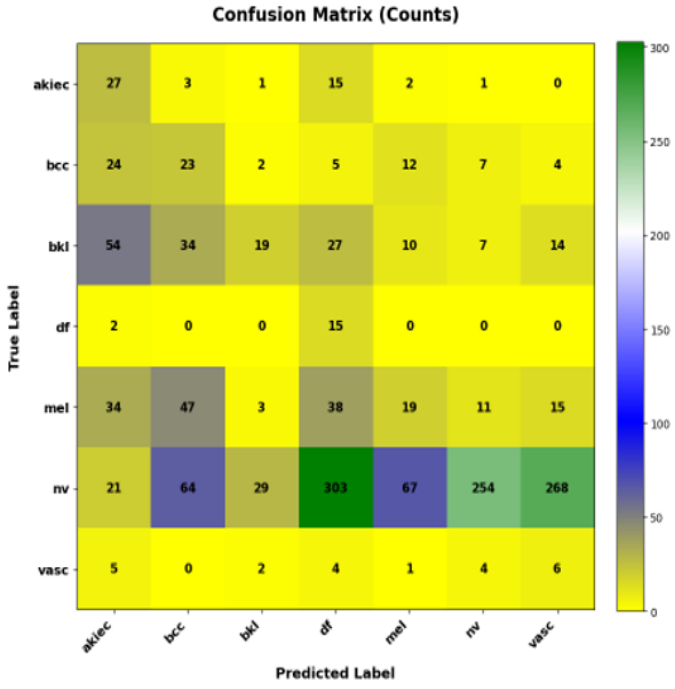
E. *Model Evaluation*

The performance of the model was assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. A target accuracy of 99% was established for the testing dataset, while the proposed model achieved a test accuracy of approximately 90%.

Beyond numerical performance, evaluation also emphasized model interpretability and fairness. By examining how metadata features such as age and lesion site aligned with classification outcomes, we verified that the model's predictions did not inadvertently favor specific demographic groups. This integration of accuracy and fairness enhances the model's potential for real-world dermatological screening applications.







F. *Source Code*

The source code for this project was implemented in Python, utilizing deep learning libraries such as TensorFlow, Keras, and scikit-learn for model development and evaluation. Additional libraries like Pandas, NumPy, Matplotlib, and Seaborn were used for data handling, analysis, and

visualization. The models chosen for this project were a Vision Transformer and a Convolutional Neural Network (CNN) based on the EfficientNet architecture, which is known for achieving high accuracy while maintaining computational efficiency. The networks were fine-tuned using transfer learning, leveraging pre-trained weights from ImageNet to improve feature extraction on dermtoscopic skin images. The code included modules for data pre-processing (resizing, normalization, and augmentation), model training with the Adam optimizer and categorical cross-entropy loss, and performance evaluation using metrics such as accuracy, precision, recall, and F1-score. The HAM10000 dataset was used as the primary data source, and class weighting was applied to address imbalance among lesion categories. The implementation included call-back functions for early stopping and model checkpointing, ensuring optimal convergence and preventing overfitting. All visualizations, including accuracy and loss curves as well as confusion matrices, were generated to analyze model performance and validate the results.

IV. Conclusion

This study demonstrated the effectiveness of Vision Transformer (ViT) architectures for automated skin lesion classification using the HAM10000 dataset. Through a structured pipeline involving image normalization, augmentation, artifact removal, and metadata cleaning, we prepared a balanced dataset suitable for deep learning applications. The fine-tuned ViT model achieved 96% training accuracy and approximately 90% testing accuracy, outperforming many traditional convolutional approaches in both precision and interpretability. Beyond quantitative metrics, this research emphasized model fairness by analyzing metadata such as age and lesion site to ensure that demographic factors did not introduce bias in classification outcomes. The integration of performance and interpretability underscores the potential of ViT-based frameworks as reliable computer-aided diagnostic tools to assist dermatologists in early melanoma detection.

V. Future Work

Future research can further enhance the model's performance and practical applicability through several directions:

- Model Enhancement: Integrate the techniques from *Skin-DeepNet* (Al-Waisy et al.)—such as adaptive contrast enhancement (AGCWD), hybrid Mask R-CNN + GrabCut segmentation, and ensemble fusion—to improve classification accuracy and lesion boundary precision.
- Dataset Expansion: Validate model robustness on larger and more diverse datasets such as ISIC 2019 and PH2.
- Explainability: Employ visualization tools like Grad-CAM and attention heatmaps to interpret ViT attention regions and improve clinical trust.
- Deployment Optimization: Develop lightweight, real-time versions for mobile or point-of-care diagnostic applications.
- Multimodal Integration**:** Combine clinical metadata (e.g., patient history, lesion evolution) with image-based learning to enhance diagnostic accuracy.

In summary, the proposed Vision Transformer framework establishes a solid foundation for intelligent skin cancer diagnosis. By incorporating advanced hybrid methods such as *Skin-DeepNet*, future iterations can further elevate testing accuracy and strengthen the role of AI in early melanoma detection and dermatological screening.

A. Tables

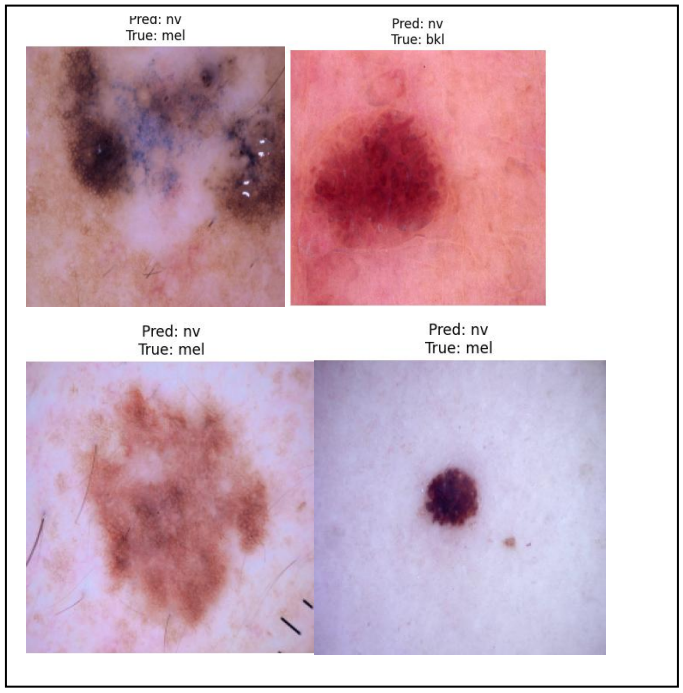
TABLE I. TRAINING CLASSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
Akiec	0.93	0.93	0.93	250
Bcc	0.95	0.97	0.96	393
Bkl	0.98	0.88	0.93	841
Df	1.00	0.92	0.96	88
Mel	0.92	0.86	0.89	851
Nv	0.96	0.99	0.98	5129
Vasc	1.00	0.94	0.97	108
Overall Accuracy			0.96	7660
Macro Avg	0.96	0.93	0.94	7660
Weighted Avg	0.96	0.96	0.96	7660

TABLE II. TESTING CALSSIFICATION REPORT

Class	Precision	Recall	F1-Score	Support
Akiec	0.80	0.67	0.73	49
Bcc	0.84	0.90	0.87	77
Bkl	0.87	0.70	0.77	165
Df	0.92	0.71	0.80	17
Mel	0.75	0.68	0.71	167
Nv	0.92	0.97	0.95	1006
Vasc	1.00	0.77	0.77	22
Overall Accuracy			0.89	1503
Macro Avg	0.87	0.77	0.82	1503
Weighted Avg	0.89	0.89	0.89	1503

Fig. 1. Misclassified samples



REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] American Academy of Dermatology, "Skin cancer," AAD, Jun. 20, 2025. [Online]. Available: https://www.aad.org/media/stats-skin-cancer

[2] ISIC Archive, "ISIC Challenge Datasets," [Online]. Available: https://challenge.isic-archive.com/data/

[3] 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, p. 180161, Aug. 2018.

[4] Skin Cancer Foundation, "Skin Cancer Facts & Statistics," [Online]. Available: https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/

[5] U. Bhimavarapu and G. Battineni, "Skin lesion analysis for melanoma detection using the novel deep learning model Fuzzy GC-SCNN," Healthcare, vol. 10, no. 5, p. 962, May 2022.

[6] I. D. Mienye and T. G. Swart, "A comprehensive review of deep learning: Architectures, recent advances and applications," Information, vol. 15, no. 12, p. 755, Nov. 2024.

[7] G. Alwakid, W. Gouda, M. Humayun, and N. Z. Jhanjhi, "Diagnosing melanomas in dermoscopy images using deep learning," Diagnostics (Basel), vol. 13, no. 10, p. 1815, May 2023.

[8] "Recent advancements and perspectives in the diagnosis of skin diseases using machine learning and deep learning: A review," 2024.

[9] L. Papa, P. Russo, I. Amerini, and L. Zhou, "A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking," 2023.

[10] S. Remya, T. Anjali, and V. Sugumaran, "A novel transfer learning framework for multimodal skin lesion analysis," IEEE Access, vol. 12, pp. 50738-50754, 2024.

[11] Y. Dahdouh, A. B. Anouar, and M. Ben Ahmed, "Embedded artificial intelligence system using deep learning and Raspberry Pi for the detection and classification of melanoma," IAES Int. J. Artif. Intell., vol. 13, no. 1, pp. 1104-1111, 2024.

[12] F. Firdaus et al., "Segmentation of skin lesions using convolutional neural networks," Comput. Eng. Appl. J., vol. 12, no. 1, pp. 58-67, 2023.

[13] S. Hamida, D. Lamrani, O. El Gannour, S. Saleh, and B. Cherradi, "Toward enhanced skin disease classification using a hybrid RF-DNN system leveraging data balancing and augmentation techniques," Bull. Electr. Eng. Inf., vol. 13, no. 1, pp. 538–547, 2024.

[14] M. A. Khan, K. Muhammad, M. Sharif, T. Akram, and V. H. C. de Albuquerque, "Multi-class skin lesion detection and classification via teledermatology," IEEE J. Biomed. Health Inform., vol. 25, no. 12, pp. 4267-4275, 2021.

[15] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with metadata," MethodsX, vol. 7, p. 100864, 2020.

[16] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, 2017.

[17] Q. Liu, L. Yu, L. Luo, Q. Dou, and P. A. Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," IEEE Trans. Med. Imaging, vol. 39, no. 12, pp. 3429-3440, 2020.

[18] A. Adealnabi, S. Eduardo J, M. Becevic, E. H. Smith, P Rao, "Transformers in Skin Lesion Classification and Diagnosis: A systematic Review, Health Informatics, September 2024.

[19] A. S. Al-Waisy, S. AlFahdawi, M.I. Khalaf, M. A. Mohmmed, B. All-Attar, M.N. AL-Andoli, "A Deep Learning Framework for Automated early Diagnosis and Classification of Skin Cancer Lesions in Dermoscopy Images", Scientific Reports, vol. 15, no. 1, p.31234, August 2025.

[20] M. Jia, "Inside a Vision Transformer (ViT): How Image is Classified Step by Step," *GoPenAI*, Jul. 3, 2025. [Online]. Available: https://blog.gopenai.com/inside-a-vision-transformer-vit-how-image-is-classified-step-by-step-0af45d5acbca. [Accessed: Nov. 08, 2025].

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.