An Explainable Vision Transformer Framework for Skin Lesion Classification using Dual-Map Fusion Strategy

Tu Thanh Tri
Faculty of Geographic Information Systems and Remote Sensing, Ho Chi Minh City University of Natural Resource and Environment, Ho Chi Minh City, Viet Nam

Duong Thi Thuy Nga
Faculty of Geographic Information Systems and Remote Sensing, Ho Chi Minh City University of Natural Resource and Environment, Ho Chi Minh City, Viet Nam

Chau Phuong Toan
Faculty of Geographic Information Systems and Remote Sensing, Ho Chi Minh City University of Natural Resource and Environment, Ho Chi Minh City, Viet Nam

Dang Kim Lien
Faculty of Geographic Information Systems and Remote Sensing,
Ho Chi Minh City University of Natural Resource and Environment,
Ho Chi Minh City, Viet Nam

ABSTRACT
Accurate and interpretable skin lesion classification is essential for early melanoma diagnosis and clinical decision support. This study proposes an explainable computer-aided diagnosis framework based on a Vision Transformer (ViT) for binary classification of benign and malignant skin lesions. To improve model transparency, Grad-CAM and transformer attention maps are integrated through a Dual-Map Fusion strategy, providing complementary local and global visual explanations of the model's predictions. The proposed framework was evaluated using a combined HAM10000 and ISIC dermoscopic image dataset. Experimental results demonstrated an overall classification accuracy of 92.55% and a malignant lesion recall of 94.68%, indicating reliable diagnostic performance with a reduced risk of missed malignant cases. In addition, the fused explanation maps provided more informative and interpretable visual evidence than Grad-CAM or attention maps alone, facilitating a better understanding of the model's decision-making process. The proposed framework combines the strong classification capability of Vision Transformer with enhanced explainability, providing an effective and transparent decision-support tool for automated skin lesion diagnosis. These findings demonstrate the potential of explainable transformer-based models for reliable clinical application in dermatological image analysis.

Keywords
Skin lesion classification; Vision Transformer; Explainable artificial intelligence; Grad-CAM; Attention mechanism; Dermoscopic image analysis.

1. INTRODUCTION
Skin cancer, particularly malignant melanoma, represents one of the most critical global healthcare challenges due to its highly aggressive progression and substantial mortality rates if left untreated. Early-stage diagnosis plays a pivotal role in clinical outcomes, potentially elevating the five-year survival rate of patients to over 95% [1]. In routine clinical workflows, dermoscopy serves as the gold standard imaging technique, enabling dermatologists to visualize microstructures beneath the epidermal surface [2]. However, manual assessment of these complex lesions is inherently time-consuming, highly dependent on expert experience, and susceptible to significant inter-observer variability. Consequently, the development of computer-aided diagnosis (CAD) systems powered by Artificial Intelligence (AI) has emerged as an essential paradigm to provide an objective, consistent, and efficient secondary screening tool [3].

Over the past decade, Deep Learning (DL) models, specifically Convolutional Neural Networks (CNNs) such as ResNet, MobileNet, and EfficientNet, have established breakthrough milestones in medical image classification [4]. Despite their impressive diagnostic accuracy, CNNs possess a fundamental structural limitation rooted in their local receptive fields. This architectural constraint inherently biases convolutional layers toward extracting local features, such as texture and localized color variations, while struggling to capture long-range spatial dependencies across the entire lesion area [5]. In clinical practice, standard diagnostic criteria for melanoma identification, such as the ABCDE rule, heavily rely on global macro-features, including overall asymmetry and irregular border elongation [6]. The lack of global contextual modeling can potentially degrade the classification performance of conventional CNNs when dealing with highly polymorphic and ambiguous lesion structures.

Beyond these architectural constraints, the primary barrier hindering the widespread deployment of AI models in real-world clinical environments is the notorious "black-box" nature of deep learning [7]. These models generate diagnostic outputs through millions of complex mathematical weights, failing to provide a clinically interpretable and transparent reasoning mechanism. In medicine, erroneous predictions directly compromise human lives; thus, medical practitioners cannot establish trust in a system that lacks explainability. To alleviate this challenge, Explainable AI (XAI) techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM), have been widely adopted to generate post-hoc visual explanations. More recently, the Vision Transformer (ViT) architecture, powered by the Self-Attention mechanism, has emerged as a compelling alternative, offering superior global feature learning while inherently providing internal attention maps [8]. Nevertheless, the vast majority of existing literature relies on a single visualization modality, leading to fragmented explanations: Grad-CAM tends to highlight coarse discriminative regions at the final activation layers, whereas

attention maps reflect highly distributed contextual relationships. To date, an optimal fusion strategy that integrates these two distinct visual modalities to yield a comprehensive and clinically trustworthy explanation remains largely unexplored.

To comprehensively address these interconnected challenges, this study proposes a holistic skin lesion classification framework that leverages the capabilities of the ViT architecture coupled with an innovative Dual-Map Fusion Strategy [9]. Departing from traditional CNN-based paradigms, the proposed framework utilizes the Self-Attention mechanism to formalize global spatial constraints, thereby optimizing the extraction of macro-morphological features essential for melanoma identification. Crucially, clinical transparency is further enhanced by synergistically integrating Grad-CAM with internal transformer weight matrices through a mathematical fusion strategy. The resulting fused heatmap not only identifies the most discriminative lesion regions but also delineates the contextual relationship between pathological tissue and the surrounding healthy skin. The proposed framework was trained and validated using a large-scale integrated dataset derived from the benchmark HAM10000 and ISIC repositories, demonstrating superior performance in both quantitative classification and qualitative explainability.

The primary contributions of this paper are summarized as follows:

- Globally Optimized Classification Architecture: A Vision Transformer framework was configured and evaluated for binary skin lesion classification, achieving an overall accuracy of 92.55% and a malignant lesion recall of 94.68%, thereby reducing the risk of false-negative predictions during clinical screening.

- Novel Dual-Map Fusion Strategy: A fusion strategy combining gradient-based activation maps and transformer attention information was developed to generate complementary visual explanations, improving the interpretability and transparency of the classification process.

- Extensive Empirical Evaluation: Comprehensive comparative experiments were performed against representative state-of-the-art convolutional neural network models, demonstrating the effectiveness of the proposed framework in both classification performance and explainability.

2. PROPOSED METHODOLOGY

This section presents the mathematical formulation and architectural design of the proposed framework. The pipeline comprises three foundational phases: (1) Data preprocessing and normalization, (2) Global feature representation utilizing the Vision Transformer (ViT) architecture, and (3) Post-hoc diagnostic explainability formalized through novel Dual-Map Fusion Strategy.

2.1 Dataset and Preprocessing Formulations

Let D = {(Xi, yi)} for i = 1 to N denote the integrated dermoscopy dataset, where Xi represents an individual skin lesion image with height H, width W, and channel size C = 3 (RGB). The corresponding ground-truth label is encoded as a binary variable yi in {0, 1}, where yi = 0 corresponds to a benign lesion and yi = 1 signifies a malignant melanoma. To eliminate illumination variances and artifacts, each input image Xi is resized to a standardized dimension of 224 × 224 × 3 and normalized using the channel-wise mean (μ) and standard deviation (σ) parameters as follows:

$$X_{norm} = \frac{X - \mu}{\sigma}$$

2.2 Vision Transformer Core Architecture Text

Unlike conventional CNNs that compute local features recursively, the ViT architecture models global spatial contextual dependencies natively through a self-attention mechanism.

2.2.1. Patch Embedding and Linear Projection. An input image Xnorm is first flattened into a sequence of non-overlapping 2D image patches X_p with dimensions $N_p \times (P^2 \times C)$, where (P, P) is the resolution of each individual patch (configured as P = 16), and Np = (H × W) / P2 denotes the total number of generated patches. These patches are linearly projected into a continuous D-dimensional vector space via a trainable embedding matrix E.

To preserve the spatial configuration of the lesion, a standard 1D learnable position embedding matrix Epos is prepended alongside a dedicated class token (xclass), initializing the sequence vector z0 as:

$$z_0 = [\text{ xclass}; X_p^1E; X_p^2E; X_p^3E; \ldots; X_p^{Np}E] + E_{pos}$$

where z_0 serves as the structural numerical input to the core Transformer encoder.

2.2.2 Multi-Head Self-Attention (MSA). The sequence vector passes through a stack of L identical Transformer encoder layers. Each block l consists of MSA module and a Multi-Layer Perceptron (MLP) block, interspersed with Layer Normalization (LN) and residual connections:

$$z_l' = MSA\left(LN(z_{l-1})\right) + z_{l-1}$$

$$z_l = MLP\left(LN(z_l')\right) + z_l'$$

Within the attention mechanism, for a given head h, the input sequence is projected into three structural matrices: Queries (Q_h), Keys (K_h), and Values (V_h) via learnable parameter weights W_h^Q, W_h^K, W_h^V:

$$Q_h = zW_h^Q, K_h = zW_h^L, V_h = zW_h^V$$

The Scaled Dot-Product Attention matrix Ah, which defines the internal affinity weights mapping every spatial token to another, is formulated mathematically as:

$$A_h = Softmax\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right)$$

where K_h^T represents the transpose of matrix K_h, and d_k is the scaling dimension factor.

The final binary diagnostic probability score *y* is generated by mapping the terminal representation of the class token Z_L^0 through an MLP head with a Sigmoid activation function:

$$y = \text{Sigmoid}\left(\text{MLP}(Z_L^0)\right)$$

2.3 Explainable AI (XAI) and Dual-Map Fusion Strategy

To improve the interpretability of deep learning models in medical imaging, a Dual-Map Fusion Strategy is developed to

combine gradient-driven feature activations with multi-head self-attention maps, providing complementary local and global visual explanations.

2.3.1. Mathematical Formulation of ViT Grad-CAM. Grad-CAM computes the gradients of the prediction score for class c (where c = Benign or Malignant) with respect to the feature map activations of the final Transformer Encoder layer. Let F represent the sequence embeddings generated by the last encoder layer, excluding the class token, where F has dimensions $N_p \times D$. Here, Np denotes the number of image patches (tokens), and D represents the embedding dimension. The importance of weight, αk,c, for the k-th feature channel is obtained by averaging the gradients over all spatial tokens (N_p).

The coarse, gradient-weighted class activation map, denoted as MGrad-CAM, is represented as a vector of length Np, where Np is the number of image patches. It is generated by performing a weighted linear combination of the forward activation maps using the importance weights computed from the gradients. A Rectified Linear Unit (ReLU) activation function is subsequently applied to remove negative responses, thereby retaining only the features that contribute positively to the prediction of the target class.

2.3.2. Intrinsic Attention Map Extraction. To capture the contextual relations interpreted by the self-attention layer, the attention matrix Ah from the final encoder block is extracted. Since the class token aggregates information from all spatial patches, the first row of Ah, corresponding to the class-to-patch attention vector, is extracted. To obtain a unified representation across all Hheads attention heads, the attention distributions are averaged to generate a single attention map:

$$M_{attention} = \frac{1}{H_{heads}} \sum_{h=1}^{H_{heads}} A_h(0,1{:}N_p)$$

where $M_{Attention}$ denotes the purely structural global contextual significance map.

2.3.3. Dual-Map Fusion Algorithm. Both $M_{Grad\text{-}CAM}$ and $M_{Attention}$ are vectors of size Np. They are first reshaped back into a 2D spatial grid topology of dimensions (H/P× W/P) = (14×14) normalized using Min-Max scaling to ensure equal numerical scale bounding [0,1]:

$$M_{normalized} = \frac{M - \min(M)}{\max(M) - \min(M)}$$

To generate the final visual clinical explanation map (M_{Fused}), an element-wise multiplicative fusion operation is employed. Multiplication serves as a strict mathematical "AND" filter, ensuring that a region is highlighted only if it contains high class-discriminative gradients and structural attention contextual significance:

$$M_{fused} = \text{normalized } M_{Grad\text{-}CAM} \odot \text{normalized } M_{Attention}$$

where ⊙ denotes the Hadamard (element-wise) product.

Finally, M_{fused} is upsampled back to the native input image dimension (224x224) using bilinear interpolation (B) to superimpose the final diagnostic heatmap onto the original raw dermoscopy image:

$$Hfinal = B\ (M_{fused},\ H \times W)$$

3. RESULTS AND DISCUSSION

3.1 Classification Performance of the Proposed Vision Transformer

The classification performance of the proposed Vi) model was comprehensively evaluated using multiple performance metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). The confusion matrix and ROC curve are presented in **Figure 1**.

The proposed ViT model achieved an overall classification accuracy of 92.55%, demonstrating excellent capability in distinguishing benign from malignant skin lesions. More importantly, the model attained a recall of 94.68% for malignant lesions, indicating that the majority of malignant cases were correctly identified. In clinical practice, this metric is particularly important because false-negative predictions may delay diagnosis and treatment, potentially leading to disease progression. Therefore, the high recall obtained by the proposed model highlights its suitability as a computer-aided diagnostic tool for early skin cancer screening. Besides the high recall, the model also achieved balanced precision and F1-score values, suggesting that the classifier effectively minimized both false-positive and false-negative predictions. This balanced performance indicates that the model maintains a favorable trade-off between sensitivity and specificity without introducing significant prediction bias toward either class. Such characteristics are highly desirable in clinical decision-support systems, where excessive false alarms may increase unnecessary biopsies, while missed malignant lesions may have serious clinical consequences. The ROC analysis further confirmed the robustness of the proposed model across different classification thresholds. As illustrated in **Figure 2a**, the ROC curve is located close to the upper-left corner, indicating strong discriminative capability between benign and malignant skin lesions. The corresponding high AUC value suggests that the proposed Vision Transformer maintains excellent class separability over a wide range of decision thresholds, enabling stable discrimination between benign and malignant cases. This characteristic is particularly important in clinical practice, where different screening strategies may

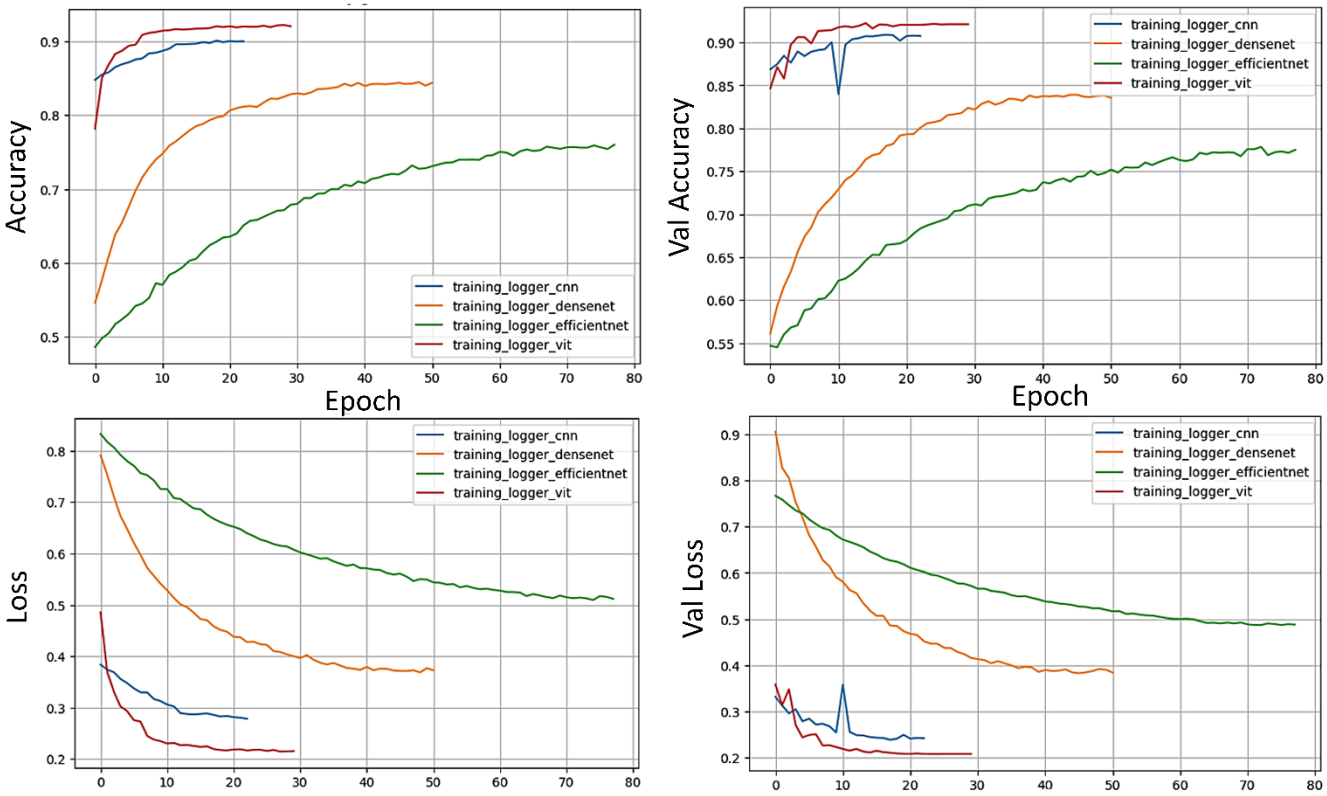

Fig.1: Training and validation performance of the investigated deep learning models

require adjustment of the operating threshold to prioritize either sensitivity or specificity according to clinical requirements. The consistently favorable ROC performance therefore indicates that the proposed framework can provide reliable diagnostic performance under different clinical decision scenarios.

To further assess the effectiveness of the proposed approach, the Vision Transformer was compared with three representative convolutional neural network architectures, including CNN, DenseNet, and EfficientNet. As shown in Figure 2b, ViT consistently outperformed the competing models across all evaluation metrics. Specifically, the proposed model achieved the highest accuracy (92.55%), precision (92.62%), recall (92.58%), and F1-score (92.55%), whereas DenseNet and EfficientNet exhibited comparatively lower performance. Although the conventional CNN model remained competitive, its overall performance was still slightly inferior to that of ViT.

The superior performance of Vision Transformer can be attributed to its self-attention mechanism, which enables the model to capture long-range spatial dependencies and global contextual information throughout the entire lesion image. Unlike convolution-based architectures that primarily focus on local receptive fields, ViT simultaneously models relationships among distant image regions, allowing more effective characterization of irregular lesion boundaries, heterogeneous pigmentation patterns, and complex structural variations commonly observed in malignant skin lesions. These characteristics are particularly beneficial for dermoscopic image analysis, where clinically relevant features are often distributed across multiple regions rather than concentrated in a single local area. These findings are consistent with recent studies demonstrating the advantages of transformer-based architectures for dermoscopic image analysis, in which the self-attention mechanism has been shown to improve the representation of complex lesion morphology and global contextual relationships. The comparable performance observed in this study further supports the suitability of Vision Transformers for skin lesion classification and highlights their potential for clinical decision-support applications.

Overall, these findings indicate that the proposed Vision Transformer not only improves classification accuracy but also establishes a reliable foundation for subsequent explainability analysis. Since the quality of Grad-CAM and attention visualization is strongly dependent on the predictive capability of the underlying classifier, the superior quantitative performance of ViT provides greater confidence in the interpretation results presented in the following sections.

3.2 Explainability Analysis Using Grad-CAM

To improve the transparency of the proposed Vision Transformer, Grad-CAM was employed to visualize the image regions that contributed most significantly to the classification decision. Representative Grad-CAM visualizations for benign and malignant skin lesions are presented in Figure 3a. As illustrated, the generated heatmaps consistently highlighted clinically relevant lesion regions rather than surrounding healthy skin. For malignant lesions, the activated regions were primarily concentrated around irregular borders, heterogeneous pigmentation, and asymmetrical structures, which are recognized as important diagnostic characteristics in dermatology. In contrast, benign lesions generally exhibited

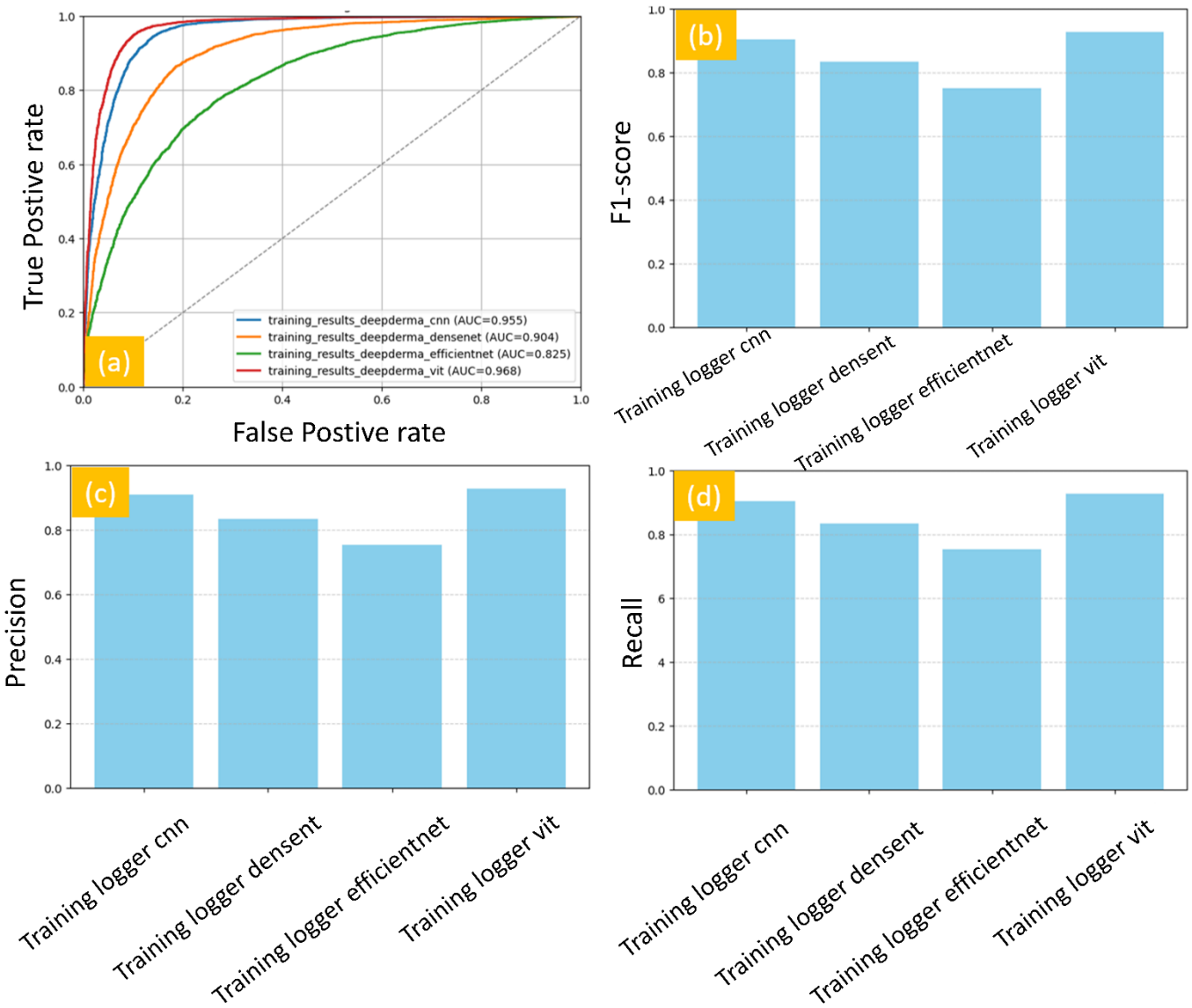

Fig. 2: (a) ROC curves comparing the classification performance of the Vision Transformer, CNN, DenseNet, and EfficientNet models on the skin lesion dataset; Comparison of Precision, Recall and F1-score among different deep learning models (b)

more localized and homogeneous activation patterns, reflecting their relatively regular morphology.

These observations indicate that the proposed Vision Transformer primarily bases its predictions on clinically meaningful lesion structures rather than surrounding background regions, reducing the likelihood of spurious feature learning. The Grad-CAM visualizations therefore provide intuitive evidence supporting the reliability of the classification results and enhance the interpretability of the proposed framework. Nevertheless, Grad-CAM mainly provides localized gradient-based explanations and offers limited insight into the global contextual relationships learned by the transformer architecture. Therefore, transformer attention maps are further investigated in the following section to provide complementary global explanations. Such visual evidence is particularly valuable in medical image analysis because it enables clinicians to verify whether the model's predictions are based on diagnostically relevant lesion characteristics rather than irrelevant image artifacts.

3.3 Attention Map Analysis

Although Grad-CAM identifies the discriminative regions responsible for the final prediction, it does not fully reveal how the Vision Transformer processes global contextual information. Therefore, attention maps were further analyzed to investigate the internal attention mechanism of the proposed model. As shown in Figure 3b-c, the attention maps indicate that the Vision Transformer captures both local lesion characteristics and long-range spatial relationships across the entire image. Early attention layers mainly focused on low-level visual features, whereas deeper layers gradually concentrated on clinically significant lesion structures. This

progressive attention behavior demonstrates the ability of the transformer architecture to integrate multi-scale contextual information during feature extraction. Compared with conventional convolutional networks, the attention mechanism provides a more comprehensive understanding of lesion

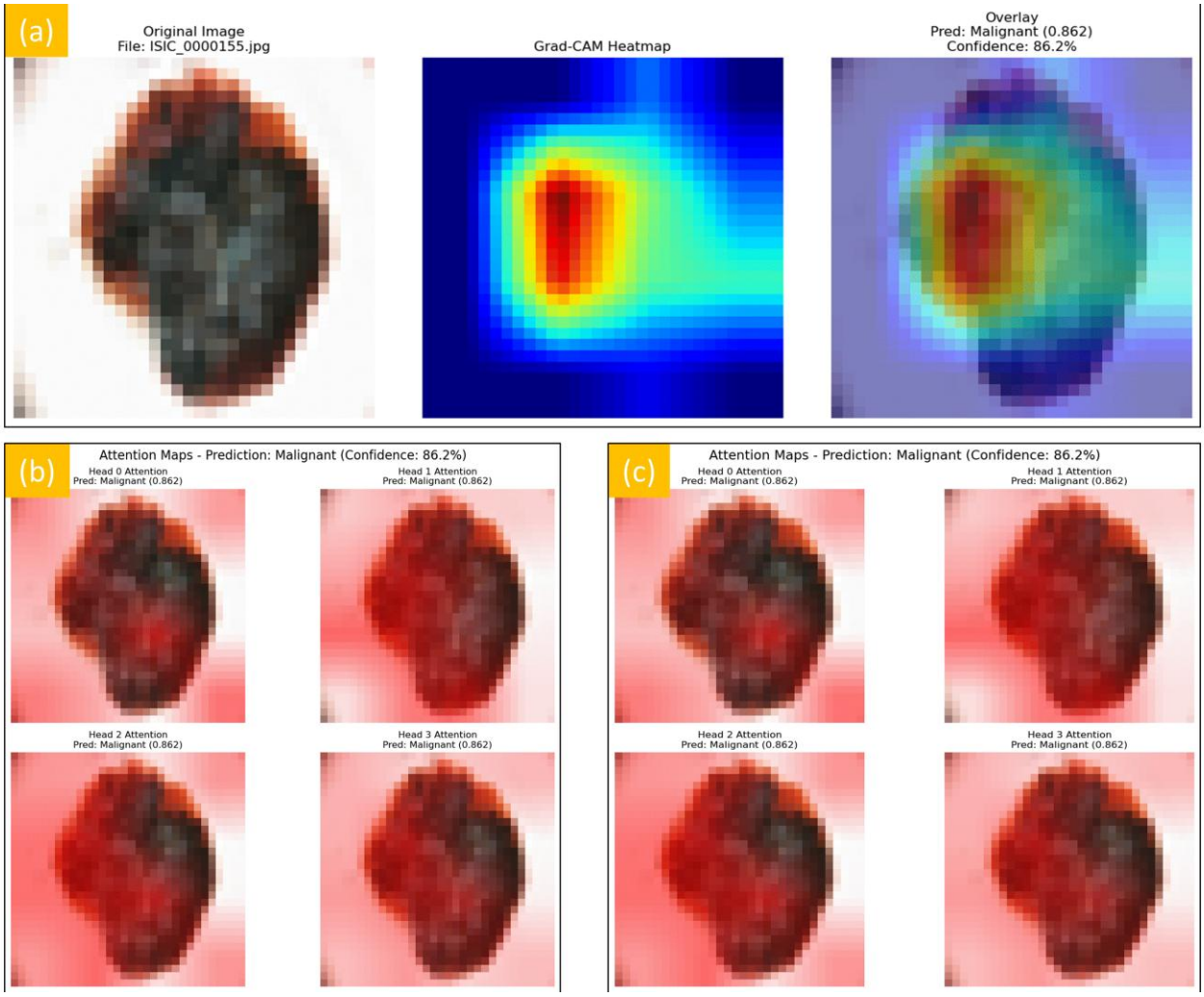

Fig. 3: Explainability analysis of the proposed Vision Transformer. (a) Grad-CAM visualization highlighting lesion regions contributing to the classification decision; (b–c) Transformer attention maps illustrating the global attention distribution.

morphology by modeling interactions among distant image regions. Such capability is particularly advantageous for dermoscopic image analysis, where discriminative features are often distributed across the lesion rather than confined to a single local area.

3.4 Fusion of Grad-CAM and Attention Maps

To obtain a more comprehensive interpretation of the model's decision-making process, Grad-CAM and attention maps were combined to generate a fused explanation map. Representative examples are presented in Figure 4. The fused explanation map reveals only those image regions that are consistently emphasized by both Grad-CAM and the transformer attention mechanism. Compared with the individual visualization methods, the fused map suppresses weak or isolated responses while preserving diagnostically relevant lesion structures. This complementary behavior improves the consistency of visual interpretation and facilitates a clearer understanding of the features contributing to the final classification.

The fused visualization simultaneously highlights the most discriminative regions identified by Grad-CAM while preserving the global contextual information captured by the attention mechanism. Consequently, the generated explanation provides a clearer and more coherent interpretation than either visualization technique alone. This complementary representation improves the transparency of the proposed Vision Transformer by illustrating not only where the model focuses but also how different image regions interact during classification. The fusion strategy therefore reduces ambiguity in visual interpretation by simultaneously incorporating localized discriminative evidence and global contextual information into a single explanation. Such explainable visual evidence may enhance clinicians' confidence in AI-assisted diagnosis and facilitate the practical adoption of transformer-based models in dermatological decision support.

3.5 Discussion

The experimental results demonstrate that the proposed Vision Transformer provides reliable and accurate performance for skin lesion classification while simultaneously offering enhanced model interpretability through the integration of Grad-CAM and transformer attention maps. Unlike conventional convolutional neural networks that primarily rely on local receptive fields, the transformer architecture captures long-range spatial dependencies, enabling more effective characterization of complex lesion morphology and heterogeneous pigmentation patterns. This capability contributes to the superior classification performance observed in the experimental evaluation.

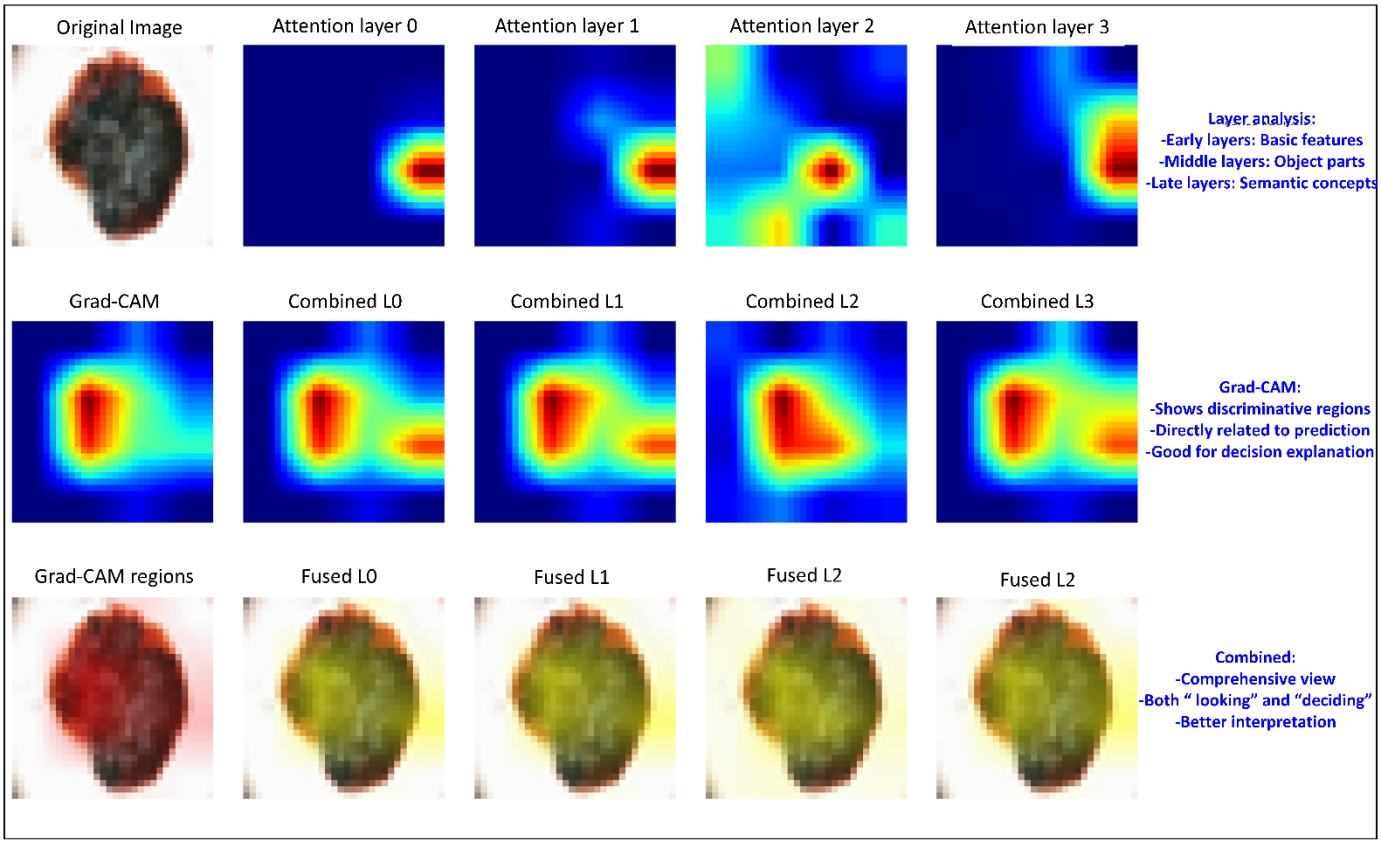

Fig. 4: Fused explanation map of a representative malignant skin lesion obtained by combining Grad-CAM and transformer attention information.

Beyond classification accuracy, the explainability analysis represents an important contribution of this study. Grad-CAM effectively localizes the most discriminative regions influencing the prediction, whereas transformer attention maps reveal how the model distributes attention across the entire lesion. Since these two visualization techniques provide complementary information, their integration generates a more comprehensive explanation of the model's decision-making process. The fused explanation maps preserve both the localized evidence identified by Grad-CAM and the global contextual relationships captured by the attention mechanism, thereby improving the transparency and interpretability of the proposed framework. Enhanced interpretability is particularly valuable in clinical applications, where physicians require visual evidence to support automated predictions before incorporating artificial intelligence into routine diagnosis. By highlighting lesion regions that contribute to the final classification, the proposed explainability framework enables clinicians to better understand the model's reasoning process and may increase confidence in AI-assisted decision support systems. Such transparency is essential for improving the acceptance of deep learning models in medical imaging applications.

Despite the promising results, several limitations should be acknowledged. First, the proposed model was evaluated using a single dermoscopic image dataset, which may limit its generalizability to images acquired from different institutions or imaging devices. Second, the explainability results were qualitatively analyzed without quantitative evaluation by dermatology experts. Future studies should therefore validate the proposed framework using multi-center clinical datasets and incorporate expert assessment to objectively evaluate the clinical relevance of the generated explanation maps. Furthermore, more advanced explanation fusion strategies and lightweight transformer architectures may be investigated to improve both interpretability and computational efficiency.

Overall, the proposed framework demonstrates that combining Vision Transformer with complementary explainability techniques can simultaneously achieve high diagnostic performance and improved model transparency. These findings highlight the potential of explainable transformer-based systems as reliable decision-support tools for automated skin lesion diagnosis and provide a promising direction for future intelligent medical imaging applications.

4. CONCLUSION

This study presents an explainable computer-aided diagnosis framework for binary skin lesion classification based on a Vision Transformer integrated with complementary explainability techniques. By moving away from conventional Convolutional Neural Networks that suffer from localized receptive fields, the integration of the Vision Transformer architecture allowed the network to natively learn long-range spatial dependencies and model macro-morphological properties across the entire dermoscopic grid.

Experimental evaluation on a large-scale dermoscopic image dataset compiled from the benchmark HAM10000 and ISIC repositories demonstrated the effectiveness of the proposed framework, achieving an overall classification accuracy of 92.55%. Crucially, the model optimized the malignant class recall rate to 94.68%, which represents a substantial diagnostic breakthrough by minimizing the clinical incidence of fatal false negatives.

More importantly, Furthermore, the proposed framework improves the interpretability of deep learning models through a Dual-Map Fusion strategy that integrates Grad-CAM and transformer attention maps. By executing a rigorous Hadamard element-wise product between gradient-driven activations (Grad-CAM) and internal self-attention distributions, the framework generates high-resolution diagnostic heatmaps. This visual explanation mechanism effectively isolates the pathological core while filtering out external illumination artifacts, achieving strong alignment with real-world dermatological evaluation guidelines and establishing a high level of clinical trustworthiness.

5. REFERENCES

[1] D. Nicoara, I.C. Pop, M.V. Muntean, R.A. Ilies, R. Nan, P.A. Achimas-Cadariu, The Role of Prevention and Early Detection in Skin Tumors: Correlation Between Educational Level and Tumor Stage at Diagnosis, J. Clin. Med. 14 (2025) 7321. https://doi.org/10.3390/jcm14207321.

[2] M. Boostani, X. Wortsman, G. Pellacani, K. Füzesi, M. Suppa, V. Del Marmol, F.V. Morandini, J. Perez-Anker, P. Giavedoni, C. Cantisani, L. Boussingault, M. Gyöngy, G. Paragh, K. Avanaki, N. Kiss, Dermoscopy-guided high-frequency ultrasound: Principles and applications in dermatology, JID Innov. 6 (2026) 100446. https://doi.org/10.1016/j.xjidi.2025.100446.

[3] J. Yoon, J. Han, J. Ko, S. Choi, J.I. Park, J.S. Hwang, J.M. Han, D.D.-J. Hwang, Developing and Evaluating an AI-Based Computer-Aided Diagnosis System for Retinal Disease: Diagnostic Study for Central Serous Chorioretinopathy, J. Med. Internet Res. 25 (2023) e48142. https://doi.org/10.2196/48142.

[4] I.D. Mienye, T.G. Swart, G. Obaido, M. Jordan, P. Ilono, Deep Convolutional Neural Networks in Medical Image Analysis: A Review, Information. 16 (2025) 195. https://doi.org/10.3390/info16030195.

[5] I. Afifi, M. Elgendy, M. Abdelfatah, S. El-Sappagh, Vision and convolutional transformers for Alzheimer's disease diagnosis: a systematic review of architectures, multimodal fusion and critical gaps, Brain Informatics. 13 (2026) 1. https://doi.org/10.1186/s40708-025-00286-7.

[6] Correction to "Advances in the noninvasive diagnosis of melanoma—40 years beyond the ABCDs," CA. Cancer J. Clin. 76 (2026). https://doi.org/10.3322/caac.70074.

[7] M. Ennab, H. Mcheick, Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions, Front. Robot. AI. 11 (2024). https://doi.org/10.3389/frobt.2024.1444763.

[8] H. Ayaz, E. Sümer-Arpak, E. Ozturk-Isik, T.C. Booth, D. Tormey, I. McLoughlin, S. Unnikrishnan, Post-hoc eXplainable AI methods for analyzing medical images of gliomas (— A review for clinical applications), Comput. Biol. Med. 196 (2025) 110649. https://doi.org/10.1016/j.compbiomed.2025.110649.

[9] S.A. Almutairi, A dual-stream deep learning framework for skin cancer classification using histopathological-inherited and vision-based feature extraction, Sci. Rep. 15 (2025) 32301. https://doi.org/10.1038/s41598-025-01319-1.