TOWARD INTERPRETABLE SKIN CANCER DIAGNOSIS: A HYBRID EFFICIENTNET-GRAPH ATTENTION NETWORK WITH GRAD-CAM EXPLAINABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Skin lesion classification remains a challenging task because benign and malignant lesions vary substantially in morphology and appearance. This study aims to develop an accurate and interpretable deep learning framework for automated binary classification of skin lesions, explicitly modeling contextual dependencies among spatially distributed visual features. We propose a hybrid architecture combining EfficientNet-B0 with a Graph Attention Network (GAT), in which a pretrained EfficientNet-B0 backbone extracts discriminative representations from dermoscopic images, while the GAT models interactions among spatial features derived from the convolutional feature map. The refined representations are globally aggregated and subsequently classified into benign and malignant categories. To further enhance interpretability, Grad-CAM is employed to visualize the regions contributing most strongly to the model's predictions. Experiments on the publicly available "Skin Cancer: Malignant vs. Benign" dataset show that the proposed model achieves 91.21% accuracy, 91.12% precision, 91.17% recall, and a 91.14% F1-score on the independent test set, surpassing the baseline EfficientNet-B0 without GAT, alternative attention mechanisms, evaluated convolutional backbones, and results reported in related literature. Grad-CAM visualizations further confirm that the model focuses predominantly on discriminative lesion regions. These findings demonstrate that integrating graph-based attention with convolutional feature extraction enhances discriminative capability and generalization while providing clinically meaningful visual explanations for skin lesion classification.

1 INTRODUCTION

Skin cancer represents a major public health concern, with melanoma being its most aggressive form and the leading cause of mortality associated with cutaneous malignancies (Wang et al., 2025). In this context, the early and reliable differentiation between benign and malignant skin lesions is essential to facilitate appropriate clinical management. Although dermoscopy is routinely employed for the examination of skin lesions, their characterization remains challenging because of substantial morphological variability and visual similarities that can complicate diagnosis, even in expert clinical settings (Nugroho et al., 2024). Convolutional neural network (CNN)-based approaches have achieved significant advances in automated medical image analysis owing to their ability to learn hierarchical feature representations directly from data (Yu et al., 2021). However, their feature extraction mechanisms rely predominantly on local operations, which may limit the explicit modeling of interactions between spatially distant regions. This limitation can be particularly relevant in skin lesion analysis, where multiple discriminative cues distributed across different regions of a lesion may need to be interpreted jointly. Moreover, despite their strong predictive performance, deep learning models often remain difficult to interpret, which constitutes a major barrier to their acceptance and deployment in clinical settings where transparency of algorithmic decisions is essential (Hauser et al., 2022). GATs provide a promising means of addressing this limitation by explicitly modeling interactions among different feature representations through attention mechanisms (Veličković et al., 2018). Such a formulation can move beyond strictly local analysis by capturing contextual dependencies among elements within a visual representation. In parallel, Ex-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

plainable Artificial Intelligence (XAI) techniques, particularly Gradient-weighted Class Activation Mapping (Grad-CAM), can generate activation maps highlighting the image regions that contribute most strongly to a model's prediction (Selvaraju et al., 2017). These visual explanations can facilitate the interpretation of algorithmic decisions and promote greater transparency in automated dermatological classification systems.

Motivated by these considerations, we propose a hybrid and explainable framework that combines a convolutional network for learning discriminative feature representations with a Graph Attention Network for modeling contextual relationships among features. The proposed architecture is designed to jointly exploit the ability of CNNs to extract rich hierarchical visual representations and that of GATs to capture interactions among feature representations. Furthermore, Grad-CAM is incorporated into the proposed framework to provide visual explanations of the model predictions and to identify the skin regions that contribute most significantly to its decision-making process.

The main contributions of this work are as follows:

- A hybrid CNN-GAT framework for the automated classification of benign and malignant skin lesions, combining hierarchical feature extraction with contextual dependency modeling.
- An explainability mechanism based on Grad-CAM that visualizes the regions most relevant to the model prediction, thereby improving the transparency of the decision-making process.

The remainder of this paper is organized as follows. Section 2 presents related studies and existing approaches. Section 3 describes the proposed methodology and its main components in detail. Section 4 reports the experimental results, ablation studies, and comparisons with state-of-the-art methods. Finally, Section 5 concludes the paper, discusses its main limitations, and outlines directions for future research.

2 Related Work

Benign and malignant skin lesions exhibit substantial morphological heterogeneity, making their automated discrimination from dermoscopic images a challenging task. Recent studies have therefore investigated deep learning–based solutions that integrate image preprocessing, data augmentation, and either convolutional or Transformer-based architectures, particularly on the Skin Cancer: Malignant vs. Benign dataset. (Hiswati, 2021) developed a CNN model trained from scratch, preceded by an extensive preprocessing pipeline involving image vectorization, color adjustment, removal of black borders and hair artifacts, lesion segmentation, normalization, and feature selection. Class imbalance was addressed through a combination of undersampling and oversampling strategies. Despite this comprehensive preprocessing framework, the proposed CNN exhibited limited classification performance, achieving an Accuracy of 54.0%, Precision of 54.5%, Recall of 54.0%, and F1-Score of 53.0%. (Hussein et al., 2023) investigated a transfer learning framework based on AlexNet, ResNet-18, SqueezeNet, and ShuffleNet. To increase the diversity of the training samples, the authors employed several augmentation operations, including horizontal flipping, random translations of up to 30 pixels, horizontal and vertical scaling of up to 10%, and rotations of up to 30°. The original classification layers were replaced with task-specific binary classifiers to adapt the pretrained networks to the target problem. Among the evaluated models, ResNet-18 yielded the strongest results, with an Accuracy of 89.39%, Precision of 87.09%, Recall of 88.88%, and F1-Score of 88.52%. (Dagnaw et al., 2024) conducted a comparative assessment of vision Transformers and pretrained CNNs, considering ViT-Base, ViT-Large, Swin-Tiny, Swin-Small, and Swin-Base alongside VGG19, ResNet18, ResNet50, MobileNetV2, and DenseNet201. To mitigate class imbalance, SMOTE (Synthetic Minority Oversampling Technique) was adopted to synthetically generate additional samples for the minority class. ResNet50 demonstrated the most competitive overall performance, obtaining an Accuracy of 88.80%, Precision of 86.90%, Recall of 88.60%, and F1-Score of 87.80%. Similarly, (Alrabai et al., 2026) examined VGG16, DenseNet169, and DenseNet201 in conjunction with augmentation procedures based on rotations of up to 20° and horizontal and vertical flipping. DenseNet201 achieved the highest performance among the investigated architectures, with an Accuracy of 87.88%, Precision of 87.76%, Recall of 87.81%, and F1-Score of 87.78%.

Collectively, these studies demonstrate the effectiveness of conventional CNNs and transfer learning for distinguishing benign from malignant lesions, while highlighting the importance of preprocessing, data augmentation, and class-balancing strategies. Nevertheless, the existing literature largely relies on the direct application of CNN architectures to preprocessed and augmented images, with limited investigation into mechanisms capable of explicitly modeling relationships among learned visual features. In particular, the potential contribution of GATs integrated with CNN-based representations remains insufficiently explored for this binary skin-lesion classification task. This observation motivates the present study, which investigates the effect of incorporating a GAT mechanism into a convolutional backbone and systematically evaluates whether attention-based relational modeling can enhance discriminative representations and improve classification performance beyond a conventional CNN pipeline.

3 Proposed Framework

This section presents an overview of the proposed framework for binary skin lesion classification from dermoscopic images, as illustrated in Fig. 1. The overall framework is organized into three main phases: data preprocessing, model training and best-model selection, and final evaluation. The first phase prepares the input data for subsequent processing, while the second phase focuses on model learning, validation, and selection of the optimal model. In the final phase, the selected model is assessed on the independent test set using standard classification metrics, complemented by confusion matrix analysis and Grad-CAM visualization.

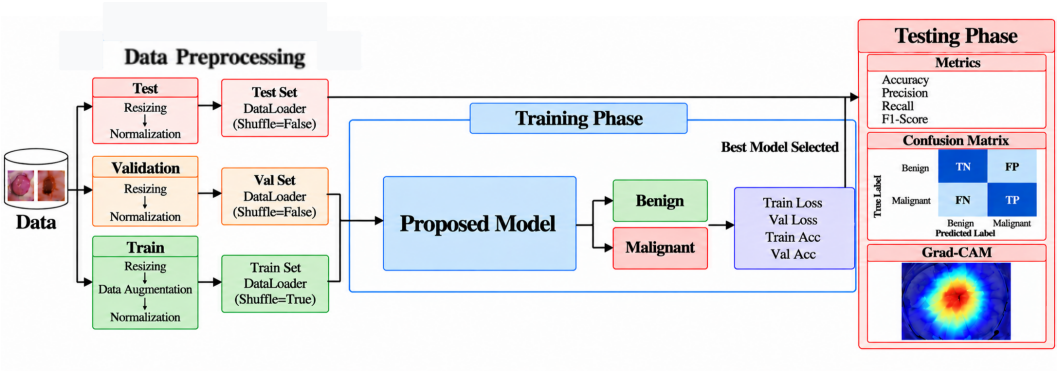


Figure 1: Overview of the proposed framework for binary skin lesion classification from dermoscopic images.

3.1 Data Used

We used the publicly available "Skin Cancer: Malignant vs. Benign" datasetFanconic (2020), previously described in the Related Work section and available on Kaggle. The dataset initially comprised 2,637 training images and 660 test images. To establish an independent validation set, 330 images were stratified from the original training set while preserving the class distribution. The resulting partition consists of 2,307 training images, 330 validation images, and 660 test images. The final class distribution is reported in Table 1.

Table 1: Distribution of the "Skin Cancer: Malignant vs. Benign" dataset after stratified partitioning.

Set	Benign	Malignant	Total
Train	1254	1053	2307
Validation	186	144	330
Test	360	300	660
Total	1800	1497	3297

3.2 Data Preprocessing

Prior to feature extraction, all images undergo a standardized preprocessing pipeline to ensure input consistency, enhance the robustness of the learned representations, and mitigate the risk of overfitting. First, all images are resized to 224×224 pixels to provide a uniform spatial resolution and satisfy the input requirements of the proposed model. The preprocessing strategy is then applied differently to the training, validation, and test sets to prevent information leakage and ensure a reliable assessment of the model's generalization capability.

For the training set , data augmentation is applied after image resizing and before normalization. As presented in Fig. 2, the augmentation strategy comprises four complementary techniques. Horizontal flipping mirrors the image along the vertical axis, whereas vertical flipping reverses it along the horizontal axis. Random rotation applies an angle uniformly sampled from the range of $-180°$ to $+180°$, thereby improving the model's robustness to variations in lesion orientation. Finally, color jittering introduces controlled variations in brightness, contrast, and saturation, each with a magnitude of 0.3, enabling the model to better accommodate variations in illumination and image acquisition conditions(Patil et al., 2025).

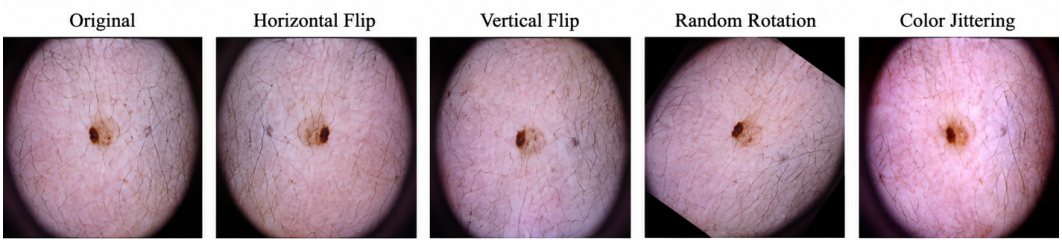


Figure 2: Data augmentation strategy applied to the training images.

For the validation and test sets, no data augmentation is applied. After resizing, both sets are normalized using the same normalization procedure as the training set. This preserves the original characteristics of the images while ensuring consistency with the input distribution expected by the proposed model. To efficiently deliver the preprocessed images and their corresponding labels to the deep learning model, each dataset subset is organized into a DataLoader for mini-batch processing. The training DataLoader uses shuffle=True to randomize the sample order at each epoch, thereby supporting stochastic optimization and reducing potential order-related bias. In contrast, shuffle=False is used for the validation and test sets to ensure consistent and reproducible evaluation while preserving the correspondence between predictions and ground-truth labels.

3.3 Proposed Model

The proposed model, illustrated in Fig. 3, is a hybrid EfficientNet-B0–GAT architecture that combines hierarchical convolutional feature extraction with explicit modeling of dependencies among spatial regions. Unlike conventional direct feature aggregation through pooling, the extracted deep features are first transformed into a graph-based representation, where the spatially distributed feature nodes are subsequently refined through an attention mechanism before being fed into the final classification stage.

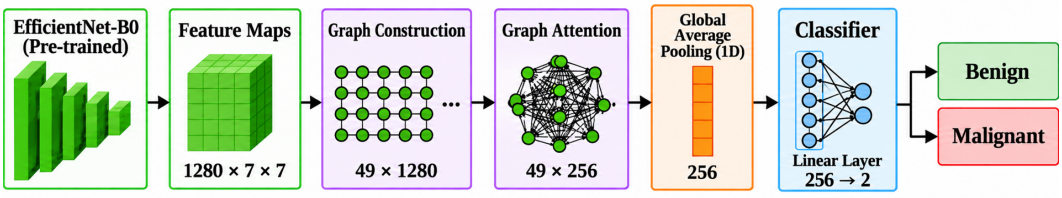


Figure 3: Proposed model composition.

3.3.1 FEATURE EXTRACTION

A pre-trained EfficientNet-B0, stripped of its classification head, is employed as the convolutional backbone. Given an input image $\mathbf{X} \in \mathbb{R}^{3\times224\times224}$, the resulting feature map is expressed as

$$\mathbf{F} = \mathcal{E}\mathrm{B0}(\mathbf{X}) \in \mathbb{R}^{1280\times7\times7}, \tag{1}$$

where $\mathcal{E}\mathrm{B0}$ denotes the EfficientNet-B0 feature extractor. The 7×7 spatial grid corresponds to 49 positions, subsequently treated as graph nodes.

3.3.2 GRAPH CONSTRUCTION

The feature map $\mathbf{F}$ is permuted and reshaped into a node-feature matrix

$$\mathbf{H}^{(0)} = \mathrm{Reshape}(\mathbf{F}) \in \mathbb{R}^{49\times1280}, \qquad \mathbf{h}_i^{(0)} \in \mathbb{R}^{1280},\ i = 1, \ldots, 49, \tag{2}$$

in which each spatial position defines a graph node, forming the input to the graph attention module. The 49 nodes are assumed fully connected, so that attention is computed over all node pairs.

3.3.3 GRAPH ATTENTION

Node features are first mapped into a 256-dimensional latent space via a learnable linear projection, $\mathbf{z}i = \mathbf{W}\mathbf{h}i^{(0)} \in \mathbb{R}^{256}$, with $\mathbf{W} \in \mathbb{R}^{256\times1280}$. For each node pair i and j, a raw attention score is computed and subsequently normalized via softmax across all 49 nodes, following the attention coefficient computation and normalization mechanism introduced by (Veličković et al., 2018), as

$$eij = \mathrm{LeakyReLU!}\left(\mathbf{a}^\top[\mathbf{z}i\|\mathbf{z}j]\right), \qquad \alpha ij = \frac{\exp(eij)}{\sum k = 1^{49}\exp(e_{ik})}, \tag{3}$$

where $\mathbf{a} \in \mathbb{R}^{512}$ is a learnable attention vector, $\|$ denotes the concatenation operator, e_{ij} represents the unnormalized attention score between nodes i and j, and α_{ij} denotes the corresponding normalized attention coefficient. Since the graph is fully connected, each node attends to all 49 nodes. The resulting attention coefficients are then used to update each node through the attention-weighted neighborhood aggregation mechanism of the GAT (Veličković et al., 2018), yielding

$$\mathbf{h}i^{(1)} = \sum j = 1^{49}\alpha_{ij}\mathbf{z}_j, \qquad \mathbf{H}^{(1)} \in \mathbb{R}^{49\times256}, \tag{4}$$

where $\mathbf{h}_i^{(1)} \in \mathbb{R}^{256}$ denotes the updated embedding of node i, while $\mathbf{H}^{(1)}$ stacks the embeddings of all 49 nodes into the refined embedding matrix, with each node representation enriched by contextual information aggregated from all graph nodes, including itself.

3.3.4 GLOBAL AGGREGATION

The 49 refined node embeddings are aggregated through Global Average Pooling (1D)(Raghavendra et al., 2023), applied independently of the classifier:

$$g_d = \frac{1}{49}\sum_{i=1}^{49}\left[\mathbf{h}_i^{(1)}\right]_d, \qquad d = 1, \ldots, 256, \tag{5}$$

yielding a compact global representation $\mathbf{g} \in \mathbb{R}^{256}$ that serves as input to the classifier.

3.3.5 CLASSIFICATION

The global representation $\mathbf{g}$ is fed into a linear classifier performing a $256 \to 2$ projection:

$$\mathbf{o} = \mathbf{W}c\mathbf{g} + \mathbf{b}c, \qquad \mathbf{W}_c \in \mathbb{R}^{2\times256},\ \mathbf{o} \in \mathbb{R}^2, \tag{6}$$

whose two output logits correspond respectively to the classes $\mathrm{Benign}, \mathrm{Malignant}$.

4 RESULTS AND DISCUSSION

4.1 TRAINING PHASE

In this experimental phase, three primary models were trained to examine the contribution of the proposed graph-based component. The first model, EfficientNet-B0+GAT, corresponds to the proposed architecture. The second model, EfficientNet-B0 without GAT, serves as the baseline, in which the extracted convolutional features are directly aggregated without any graph-based refinement. The third model, frozen EfficientNet-B0+GAT, was designed to assess the impact of freezing the convolutional backbone while preserving the GAT module. All three models were trained with a batch size of 32 for a maximum of 20 epochs, using the AdamW optimizer with a learning rate of 1×10^{-4} (0.0001) and the Cross-Entropy Loss function. An early stopping strategy with a patience of 5 epochs was applied based on the validation loss. The best-performing configuration was selected according to its highest validation accuracy, and the corresponding checkpoint was retained for subsequent evaluation on the independent test set.

Table 2: Training and validation performance of the three selected models and their number of parameters.

Model	Train Loss	Train Acc (%)	Val Loss	Val Acc (%)	Parameters (M)
EfficientNet-B0 + GAT	0.1665	94.36	0.3043	90.91	4.34
EfficientNet-B0 without GAT	0.1456	94.45	0.3102	90.30	4.01
EfficientNet-B0 + GAT with frozen backbone	0.3646	83.92	0.3911	82.12	0.33

As reported in Table 2, the proposed EfficientNet-B0 + GAT model reached a training accuracy of 94.36% and a validation accuracy of 90.91%, with corresponding losses of 0.1665 and 0.3043. This architecture comprises 4.34 million (M) trainable parameters, reflecting the joint optimization of the EfficientNet-B0 backbone together with the GAT module.

The EfficientNet-B0 without GAT baseline obtained a marginally higher training accuracy of 94.45%, yet a lower validation accuracy of 90.30%. Its training and validation losses stood at 0.1456 and 0.3102, respectively, with 4.01 M trainable parameters. Although this configuration displayed a slightly reduced training loss and a comparable validation loss, its weaker validation accuracy suggests that the proposed GAT-based design offers stronger generalization capability. This gain further supports the notion that integrating the GAT module strengthens the model's capacity to capture dependencies among spatial regions within the extracted feature representations.

By contrast, the EfficientNet-B0 + GAT with frozen backbone configuration showed markedly weaker performance, reaching training and validation accuracies of only 83.92% and 82.12%, respectively. The corresponding losses reached 0.3646 and 0.3911, while the number of trainable parameters dropped to 0.33 M, since the EfficientNet-B0 backbone's weights remained fixed throughout training. These higher losses, combined with the notably reduced accuracies, indicate that freezing the convolutional backbone constrains the network's ability to adapt its deep representations to the target classification task. Taken together, these results underline the importance of jointly fine-tuning the EfficientNet-B0 backbone and the GAT module within the proposed architecture.

4.2 TESTING PHASE

Table 3 presents the performance of the three selected models on the independent test set, evaluated in terms of Accuracy, Precision, Recall, and F1-Score.

Table 3: Performance of the three selected models evaluated on the independent test set.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
EfficientNet-B0 + GAT	91.21	91.12	91.17	91.14
EfficientNet-B0 without GAT	89.85	89.74	90.03	89.81
EfficientNet-B0 + GAT with frozen backbone	85.00	85.15	85.42	84.98

The proposed EfficientNet-B0 + GAT model achieved the strongest overall performance on the independent test set, with an Accuracy of 91.21%, a Precision of 91.12%, a Recall of 91.17%, and an F1-Score of 91.14%. These results indicate that incorporating the GAT provides an effective mechanism for refining convolutional representations and capturing dependencies among spatially distributed lesion features.

By comparison, the EfficientNet-B0 without GAT baseline reached an Accuracy of 89.85%, a Precision of 89.74%, a Recall of 90.03%, and an F1-Score of 89.81%. Although the improvement brought by the GAT component remains moderate, the proposed model consistently outperforms the baseline across all four evaluation metrics. This finding suggests that graph-based attention yields more discriminative feature representations and enhances the model's ability to generalize to previously unseen test samples.

By contrast, the EfficientNet-B0 + GAT configuration with a frozen backbone exhibited substantially lower performance, with an Accuracy of 85.00%, a Precision of 85.15%, a Recall of 85.42%, and an F1-Score of 84.98%. This performance drop shows that freezing the EfficientNet-B0 backbone restricts the adaptation of the learned convolutional representations to the specific characteristics of the skin lesion classification task, further emphasizing the importance of jointly fine-tuning the convolutional backbone and the GAT module.

Fig. 4 depicts the confusion matrices obtained on the independent test set for the three investigated configurations, offering a more granular perspective than the aggregate metrics reported in Table 3.

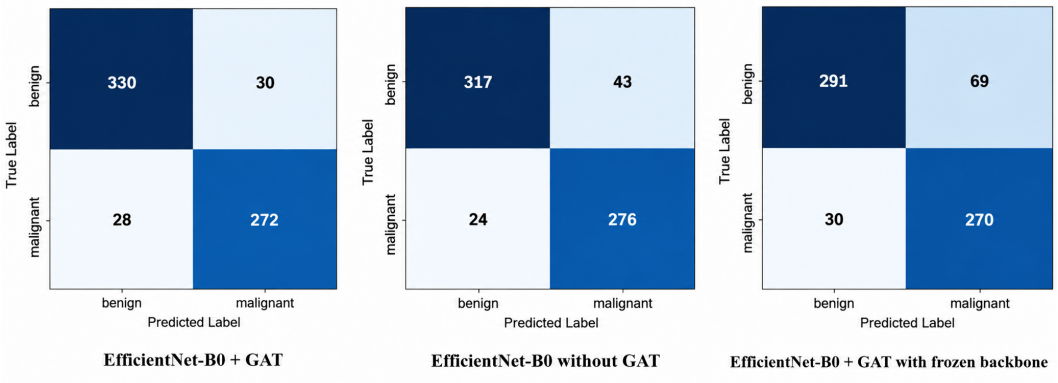


Figure 4: Confusion matrices of the three selected models on the independent test set.

For the proposed EfficientNet-B0 + GAT model, 330 True Negative (TN) benign samples and 272 True Positive (TP) malignant samples were correctly classified, together with a balanced error profile of 30 False Positives (FP) and 28 False Negatives (FN). This near-symmetric distribution of misclassifications, corresponding to error rates of 8.33% for the benign class and 9.33% for the malignant class, indicates that the graph-attention mechanism does not disproportionately favor one diagnostic category over the other. The baseline model, deprived of the GAT module, exhibits a markedly different error pattern. While it attains a slightly lower malignant-to-benign confusion rate (24 FN, i.e., 8.00%), the number of benign lesions erroneously flagged as malignant reaches 43 (i.e., 11.94%). This model therefore trades a marginal gain in sensitivity toward malignant cases for a substantially higher FP rate, an imbalance that accounts for the lower overall accuracy and F1-score reported previously. The frozen-backbone variant displays the most pronounced degradation among the three configurations. Its confusion matrix reveals 69 benign lesions misclassified as malignant and 30 malignant lesions misclassified as benign, corresponding to error rates of 19.17% and 10.00%, respectively. This deterioration, particularly marked for the benign class, is consistent with the restricted representational capacity discussed previously and confirms that freezing the convolutional backbone prevents the network from adapting its feature space to the fine-grained texture and color cues that distinguish the two lesion categories.

In aggregate, these matrices show that the improvement brought by the fully fine-tuned GAT module does not stem from a bias toward a single class, but rather from a more balanced reduction of both error types. This behavior is particularly relevant in a screening context, where an excessive FN

rate would delay the diagnosis of malignant lesions, whereas an excessive FP rate would generate unwarranted clinical burden; the proposed configuration mitigates both risks simultaneously, thereby reinforcing the value of jointly optimizing the convolutional and graph-attention components rather than relying on either mechanism in isolation.

An ablation study was conducted to separately assess the effect of the attention mechanism and the convolutional backbone. The results obtained on the independent test set are presented in Table 4.

Table 4: Ablation study on the attention mechanism and the convolutional backbone, evaluated on the independent test set.

Configuration	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
EfficientNet-B0 + GAT (Proposed)	91.21	91.12	91.17	91.14
Attention module ablation				
EfficientNet-B0 + CBAM	90.61	90.48	90.67	90.55
EfficientNet-B0 + Transformer	88.94	89.17	88.56	88.77
Backbone ablation				
MobileNetV2 + GAT	88.94	89.08	88.61	88.79
VGG16 + GAT	83.79	84.61	83.00	83.35

The analysis of the attention mechanisms demonstrates the consistent superiority of the GAT across all four evaluation metrics. Replacing the GAT with the Convolutional Block Attention Module (CBAM) results in a decrease in Accuracy from 91.21% to 90.61%, Precision from 91.12% to 90.48%, Recall from 91.17% to 90.67%, and F1-score from 91.14% to 90.55%. Replacing the GAT with a Transformer leads to a more pronounced degradation, yielding an Accuracy of 88.94%, a Precision of 89.17%, a Recall of 88.56%, and an F1-Score of 88.77%. The consistent decline across all evaluation metrics suggests that the GAT provides a more effective mechanism for modeling interactions among the extracted feature representations. The backbone analysis further confirms the importance of an appropriate convolutional representation. With MobileNetV2 + GAT, the Accuracy, Precision, Recall, and F1-Score reach 88.94%, 89.08%, 88.61%, and 88.79%, respectively, compared with 91.21%, 91.12%, 91.17%, and 91.14% for the proposed configuration. Replacing the backbone with VGG16 + GAT results in a substantially larger performance decline, with an Accuracy of 83.79%, Precision of 84.61%, Recall of 83.00%, and F1-Score of 83.35%. This systematic degradation across all four metrics highlights the impact of the quality of the convolutional representations on the performance of the graph-based refinement module.

Overall, these results indicate that the superiority of the proposed model does not arise from the GAT module in isolation, but rather from its synergy with a sufficiently expressive backbone, thereby confirming the relevance of the joint EfficientNet-B0+GAT design.

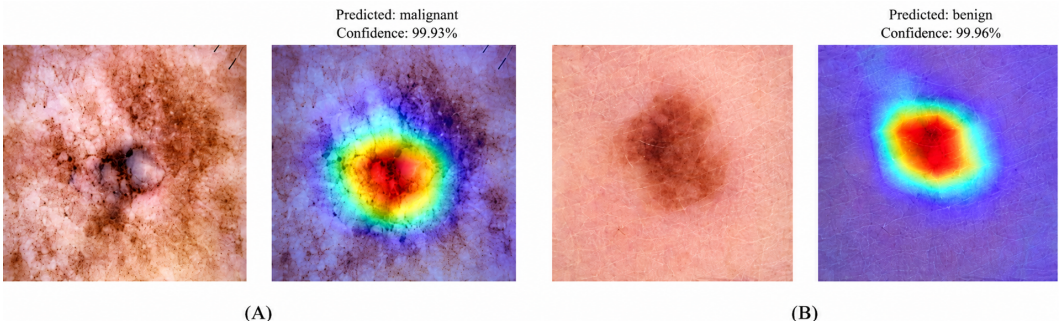


Figure 5: Grad-CAM visualizations of EfficientNet-B0+GAT on test images.

To assess the interpretability of the best-performing model, EfficientNet-B0+GAT was analyzed using Grad-CAM. As illustrated in Fig. 5, the model correctly classified a malignant lesion (A) with a confidence score of 99.93% and a benign lesion (B) with a confidence score of 99.96%. The activation maps show that the model focuses primarily on the lesion regions, particularly on the

most discriminative pigmented structures, while paying limited attention to the surrounding skin. In these visualizations, red regions correspond to areas with the highest contribution to the model's prediction, whereas blue regions indicate areas with low or negligible influence on the decision. This behavior suggests that the model relies on clinically relevant features, thereby reinforcing the reliability and interpretability of its predictions.

4.3 Comparison with State-of-the-Art Methods

Table 5 compares the proposed EfficientNet-B0+GAT model with recent approaches for binary skin lesion classification evaluated on the same dataset and, importantly, on the identical test subset, in terms of Accuracy, Precision, Recall, and F1-Score.

Table 5: Comparative study with state-of-the-art methods for skin lesion classification.

Study	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
(Hiswati, 2021)	CNN	54.00	54.50	54.00	53.00
(Hussein et al., 2023)	ResNet-18	89.39	87.09	88.88	88.52
(Dagnaw et al., 2024)	ResNet50	88.80	86.90	88.60	87.80
(Alrabai et al., 2026)	DenseNet201	87.88	87.76	87.81	87.78
Proposed Method	**EfficientNet-B0 + GAT**	**91.21**	**91.12**	**91.17**	**91.14**

The proposed method achieves the highest performance across all four evaluation metrics, with an Accuracy of 91.21%, Precision of 91.12%, Recall of 91.17%, and F1-Score of 91.14%. The closest competing model is ResNet-18 reported by Hussein et al. (Hussein et al., 2023), which obtains an Accuracy of 89.39%, Precision of 87.09%, Recall of 88.88%, and F1-Score of 88.52%. Accordingly, EfficientNet-B0+GAT improves upon ResNet-18 by 1.82% in Accuracy, 4.03% in Precision, 2.29% in Recall, and 2.62% in F1-Score. Compared with ResNet50 (Dagnaw et al., 2024), the proposed model achieves higher Accuracy, Precision, Recall, and F1-Score by 2.41%, 4.22%, 2.57%, and 3.34%, respectively. Relative to DenseNet201 (Alrabai et al., 2026), the corresponding improvements are 3.33%, 3.36%, 3.36%, and 3.36%. The performance gap is even more pronounced when compared with the CNN-based approach of Hiswati. (Hiswati, 2021), with gains of 37.21% in Accuracy, 36.62% in Precision, 37.17% in Recall, and 38.14% in F1-Score. On the whole, the comparative results consistently favor the proposed EfficientNet-B0+GAT architecture, indicating that the integration of graph-based attention with EfficientNet-B0 yields more discriminative and robust representations for distinguishing benign from malignant skin lesions than the reference methods.

5 Conclusion

This work introduced a hybrid architecture coupling an EfficientNet-B0 convolutional backbone with a GAT for the binary discrimination of skin lesions from dermoscopic imagery. By reformulating the convolutional feature map as a fully connected graph, the framework enables spatial regions to exchange contextual information through learned attention coefficients, rather than relying solely on local receptive fields. Comprehensive experiments on the "Skin Cancer: Malignant vs. Benign" benchmark confirmed that this relational refinement consistently improves discriminative capacity, with the fine-tuned GAT variant outperforming the plain convolutional baseline, alternative attention modules such as CBAM and Transformer blocks, and alternative backbones such as the lightweight MobileNetV2 and the considerably heavier VGG16. Confusion-matrix analysis further revealed that gains stem from a balanced reduction of both FP and FN rather than a bias toward a single diagnostic category, which is particularly desirable in clinical screening contexts. Grad-CAM visualizations corroborated that predictions are grounded in clinically plausible pigmented structures, reinforcing trust in the model's decision process. Nonetheless, this study is limited to a single, moderately sized public dataset and a binary formulation, leaving open questions regarding multi-class melanoma subtyping, cross-dataset generalization, and robustness to acquisition variability. Future research will therefore explore validation on larger and more heterogeneous dermoscopic cohorts, extension toward multi-class lesion taxonomies, lightweight graph-attention variants suitable for clinical deployment, and quantitative evaluation of explainability outputs against expert dermatologist annotations.

AI USE STATEMENT

In this work, generative AI tools were used solely for grammatical correction. Generative AI tools were not used to develop the research methodology, design the experiments, conduct the scientific analysis, interpret the experimental results, formulate the conclusions, or generate the scientific content of the manuscript. All core scientific content, methodological decisions, experimental procedures, analyses, and conclusions were independently developed and performed by the authors. All AI-assisted grammatical corrections were carefully reviewed and verified by the authors, who take full responsibility for the accuracy, integrity, and final content of this work.

REFERENCES

Abdelkader Alrabai, Amira Echtioui, and Fathi Kallel. Explainable transfer learning for skin cancer diagnosis. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 15(1):30, 2026. doi: 10.1007/s13721-025-00709-9.

Getamesay Haile Dagnaw, Meryam El Mouhtadi, and Musa Mustapha. Skin cancer classification using vision transformers and explainable artificial intelligence. *Journal of Medical Artificial Intelligence*, 7:14, 2024. doi: 10.21037/jmai-24-6.

Fanconic. Skin cancer: Malignant vs. benign. `https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign`, 2020.

Katja Hauser, Alexander Kurz, Sarah Haggenmueller, Roman C Maron, Christof von Kalle, Jochen S Utikal, Friedegund Meier, Sarah Hobelsberger, Frank F Gellrich, Mildred Sergon, et al. Explainable artificial intelligence in skin cancer recognition: A systematic review. *European Journal of Cancer*, 167:54–69, 2022. doi: 10.1016/j.ejca.2022.02.025.

Marselina Endah Hiswati. Deepskin: Robust skin cancer classification using convolutional neural network algorithm. *International Journal of Information and Computer*, 3(2), 2021. doi: 10.35842/ijicom.v3i2.40.

Hadeer Hussein, Ahmed Magdy, Rehab F. Abdel-Kader, and Khaled Abd Elsalam Ali. Binary classification of skin cancer using pretrained deep neural networks. *Suez Canal Engineering, Energy and Environmental Science*, 1(2):10–14, 2023. doi: 10.21608/sceee.2024.263585.1015.

Arief Kelik Nugroho, Retantyo Wardoyo, Moh Edi Wibowo, and Hardyanto Soebono. Image dermoscopy skin lesion classification using deep learning method: systematic literature review. *Bulletin of Electrical Engineering and Informatics*, 13(2):1042–1049, 2024. doi: 10.11591/eei.v13i2.6077.

Aniket Patil, Anjula Mehto, and Saif Nalband. Enhancing skin lesion diagnosis with data augmentation techniques: a review of the state-of-the-art. *Multimedia Tools and Applications*, 84(22): 25325–25364, 2025. doi: 10.1007/s11042-024-20145-7.

Paravatham VSP Raghavendra, C Charitha, K Ghousiya Begum, and VB Surya Prasath. Deep learning–based skin lesion multi-class classification with global average pooling improvement. *Journal of Digital Imaging*, 36(5):2227–2248, 2023. doi: 10.1007/s10278-023-00862-5.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, 2017. doi: 10.1109/ICCV.2017.74. URL `https://mlanthology.org/iccv/2017/selvaraju2017iccv-gradcam/`.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL `https://mlanthology.org/iclr/2018/velickovic2018iclr-graph/`.

Xingyue M Wang, Kim Borsky, Dominic W Proctor, Richard Goodall, Dominic C Marshall, William Dobell, Justin D Salciccioli, Rubeta N Matin, Joseph Shalhoub, and Naguib El-Muttardi. Trends in cutaneous melanoma mortality and incidence in european union 15+ countries between 1990

and 2019. *Journal of the European Academy of Dermatology and Venereology*, 39(7):1285–1296, 2025. doi: 10.1111/jdv.20524.

Hang Yu, Laurence T Yang, Qingchen Zhang, David Armstrong, and M. Jamal Deen. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*, 444:92–110, 2021. doi: 10.1016/j.neucom.2020.04.157.