Geometry-Aware Superpixel Graph Transformer with Metadata for Skin Lesion Classification

Muhammad Azeem[0009-0003-7713-6926], Tanveer Hussain[0000-0003-4861-8347], Amr Ahmed[0000-0002-7749-7911], and Ardhendu Behera*[0000-0003-0276-9000]

Edge Hill University, Ormskirk, Lancashire, L39 4QP, United Kingdom
{shoukatm, hussaint, ahmeda, beheraa}@edgehill.ac.uk

Abstract. Automated skin cancer classification from dermoscopic images remains challenging due to heterogeneous lesion structure, strong intra-class variability, and subtle visual differences between benign and malignant cases. Existing CNN/ViT pipelines typically rely on global or patch-level features and often combine patient metadata via late fusion, which limits spatially grounded multimodal reasoning. We present a novel **region-based graph learning framework** that explicitly models lesions as **graphs of spatially coherent superpixel regions** represented as *frozen CNN features*. To capture fine-grained lesion arrangements, we encode inter-regional geometry as edge attributes and introduce a dedicated **metadata context node** connected to all regions, providing structured integration of demographic/clinical variables within the same relational space. Node representations are updated using our *edge-aware graph transformer* followed by *attention-driven propagation*, and a final graph-level embedding for benign–malignant classification. Experiments on four public benchmarks demonstrate that explicit region-level relational modeling and graph-native multimodal fusion yield consistent gains over the state-of-the-art. Consequently, we establish a new graph-centric perspective in which CNN features are modeled as relational nodes and improved through contextual integration, yielding more expressive and robust classifications.

Keywords: Skin Lesion Classification · Superpixel Graph Transformer · Spatial Relational Modeling · Clinical Metadata Integration.

1 Introduction

Skin cancer is among the most prevalent malignancies, and melanoma remains the most lethal subtype, making an early and reliable diagnosis critical to patient outcomes [32]. Deep learning has substantially improved dermoscopic lesion classification, supported by large-scale datasets such as ISIC [20] and HAM10000 [35]. However, dermoscopic lesions are intrinsically heterogeneous; clinically meaningful cues often emerge from the co-occurrence and spatial arrangement of multiple subregions (e.g., pigment network, globules, streaks, regression structures,

* Corresponding author (beheraa@edgehill.ac.uk).

and boundary irregularity). Many CNN and ViT pipelines still represent an image as a global vector or as largely independent grid/patch tokens, which can weaken their ability to explicitly model region-to-region interactions and the geometry of subtle variations [6, 23, 21]. In addition, patient metadata (e.g., age, sex, anatomical site, clinical context) is a known source of complementary information but is frequently fused late (e.g., concatenated in the classifier), limiting its ability to condition where and how visual evidence is aggregated [3].

Recent work has explored diverse architectures for dermoscopy, including lightweight and hybrid CNNs variants [36, 28, 33, 7], attention-enhanced CNNs and sequential hybrids [14, 1, 22], and transformer-based models designed to improve global context modeling [38]. Multimodal approaches have also gained traction, ranging from contrastive and dual-stage fusion to personalized multimodal learning [13, 12, 15]. While these methods demonstrate progress, two limitations remain prominent: (i) most models operate on grid/patch tokens and do not provide an explicit, spatially grounded representation of lesion subregions and their geometric relations; and (ii) metadata are commonly treated as an auxiliary vector rather than being integrated into the core reasoning process, reducing interpretability and limiting structured multimodal interactions. Graph neural networks (GNNs) offer a natural abstraction for region-centric reasoning by encoding regions in images as nodes and edges [19]. However, only a few dermoscopy graphs employ context-aware graph reasoning [8], leaving a gap for *edge-attributed geometry-aware attention* with *metadata integrated in the graph.*

To address this gap, we propose **GeoMeta-GT**, a Geometry-Aware Superpixel Graph Transformer with Metadata for skin lesion classification. Each dermoscopic image is decomposed into superpixels to form region nodes, with node descriptors extracted from a frozen CNN. Fine-grained lesion organization is encoded through geometric edge attributes derived from inter-regional distance and orientation. Patient metadata is embedded as a dedicated context node connected to all regions, enabling message-passing-based fusion rather than late concatenation. A novel edge-aware graph transformer performs locally adaptive aggregation along geometry-informed edges, followed by attentional refinement and graph-level pooling for the final prediction.

Contributions. (1) A parameter-efficient, spatially grounded, and multimodally structured approach for empirically consistent dermoscopic lesion classification; (2) a geometry-attributed superpixel graph representation with metadata-as-node fusion for dermoscopy; (3) an edge-aware graph transformer for spatially grounded, diagnostically focused region aggregation; (4) a similarity-weighted refinement module to harmonize semantically consistent regions; and (5) consistent improvements across multiple public benchmarks using a frozen backbone.

2 Methodology

Given a dermoscopic image and associated patient metadata, the **GeoMeta-GT** pipeline (Fig. 1) constructs a *geometry-attributed superpixel graph* and performs *edge-aware attention-based message passing* to classify lesions as benign or malignant. The CNN backbone remains frozen, and all task-specific adaptation is confined to the graph module.

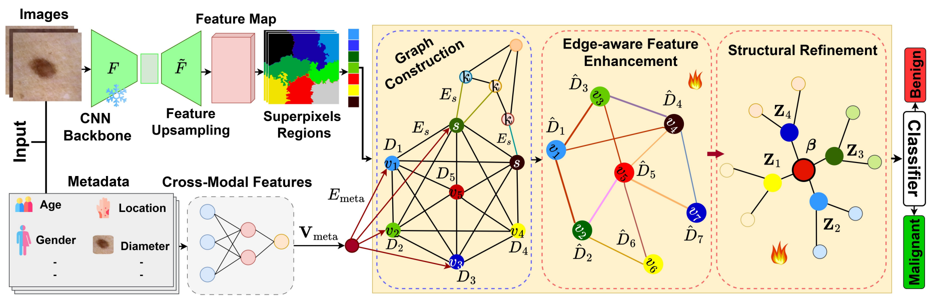

Fig. 1: **Overview of GeoMeta-GT**. The input dermoscopic image is encoded by a frozen CNN and partitioned into SLIC superpixels to form region nodes V_S with pooled deep descriptors D_S. Region-to-region edges E_S are enriched with geometric attributes (distance and orientation), while patient metadata is represented as a context node V_{meta} connected to all regions (nodes V_S). An edge-aware graph transformer and similarity-weighted structural refinement perform relational reasoning, followed by global pooling for skin lesion classification.

Superpixel decomposition and region descriptors. Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input image. We first extract deep feature representations using a pretrained frozen CNN, $F = f_{\text{CNN}}(I) \in \mathbb{R}^{C\times h\times w}$, where C is the number of channels and $h \times w$ is the spatial resolution of the feature map. Since $h < H$ and $w < W$, we upsample F to the original image size using bilinear interpolation, $\tilde{F} = \text{Interp}(F) \in \mathbb{R}^{C\times H\times W}$. In dermoscopic skin lesion imaging, the raw RGB intensities are often poorly aligned with the underlying anatomy or pathology; in particular, Intensity similarity $\not\Rightarrow$ anatomical or pathological similarity. Instead, pretrained CNN defines a task-dependent embedding space in which spatial locations that share similar anatomical or pathological characteristics are mapped to nearby feature vectors (feature similarity $\approx$ anatomical, pathological, and task-relevant semantic proximity). Consequently, performing clustering in this deep feature space produces image segments that align more closely with the true lesion structure than those obtained by clustering directly in the raw intensity domain. We therefore apply simple linear iterative clustering (SLIC) [2] directly on the features $\tilde{F}$ to decompose the image into K superpixels, $R = \{R_1, R_2, \ldots, R_K\}$, where each superpixel R_k is a spatially connected cluster whose pixels exhibit similar deep feature vectors under SLIC's joint feature–spatial distance metric. To obtain superpixel-level descriptors, we aggregate the deep features within each region. Specifically, for each superpixel R_k, we form a compact descriptor by concatenating a global context statistic (mean-pooled) and a local saliency statistic (max-pooled) features over all pixels in R_k:

$$D_k = \underbrace{\frac{1}{|R_k|} \sum_{(i,j)\in R_k} \tilde{F}(:,i,j)}_{\text{Mean pooling}} \Bigg\| \underbrace{\max_{(i,j)\in R_k} \tilde{F}(:,i,j)}_{\text{Max pooling}} \in \mathbb{R}^{2C}, \tag{1}$$

where $\|$ denotes channel-wise concatenation. The resulting descriptor D_k is used as the node feature of the k^{th} superpixel. This parameter-free aggregation provides stable, computationally efficient region representations, which are directly used as node features in the subsequent superpixel graph transformer.

Graph construction with geometric and metadata nodes. We construct a graph $G = (V, E)$ consisting of a set of superpixel nodes $V_S = \{v_1, \ldots, v_K\}$, each associated with a corresponding node feature $D_S = \{D_1, \ldots, D_K\}$, and a single metadata node V_{meta}. Each superpixel node $v_k \in V_S$ is associated with a 2D image coordinate $p_k = (x_k, y_k) \in \mathbb{R}^2$, defined as the center of the axis-aligned bounding box of the corresponding SLIC region detected in the previous step.

The spatial edges between the superpixel nodes, denoted by E_S, are constructed by n-nearest neighbor search. For each superpixel node $v_k \in V_S$ with centroid position p_k, we identify its n closest superpixel neighbors $v_s \in V_S$, $s \neq k$, based on the Euclidean distance between their centroid locations p_s and p_k. Repeating this procedure for all $v_k \in V_S$ yields the set of directed superpixel-to-superpixel edges E_S. For each edge $e_{k,s} \in E_S$ connecting node k to node s, we define a geometric edge attribute consisting of: 1) spatial distance $d_{k,s} = \|p_s - p_k\|_2$, and 2) relative orientation $\theta_{k,s} = \text{atan2}(y_s - y_k, x_s - x_k)$.

Metadata-as-node fusion. Let $\mathbf{m} \in \mathbb{R}^d$ denote the raw patient-level metadata associated with an image. We first standardize all numeric fields via z-score normalization and one-hot encode all categorical fields, producing a concatenated metadata vector $\hat{\mathbf{m}} \in \mathbb{R}^{d_m}$. We then embed these metadata into the superpixel node feature space D_k (cf. Eq. (1)) using a learnable linear projection $D_{\text{meta}} \to \phi(\hat{\mathbf{m}}) = W_m\hat{\mathbf{m}} + b_m \in \mathbb{R}^{2C}$, where W_m and b_m are trainable parameters to match the dimensionality of the superpixel node features ($2C$).

We introduce a dedicated metadata node V_{meta} with feature vector D_{meta} and connect it to all superpixel nodes V_S. Formally, we define the metadata edges as $E_{\text{meta}} = \{(V_{\text{meta}}, k), (k, V_{\text{meta}}) \mid k \in V_S\}$. The final graph G is thus given by nodes $V = V_S \cup V_{\text{meta}}$, with the respective feature $D = D_S \cup D_{\text{meta}}$ and edge set $E = E_S \cup E_{\text{meta}}$. This serves as input to the downstream graph neural network. The GNN performs relational reasoning on this graph G to jointly update all node features, thereby integrating image-derived superpixel information with patient-level metadata in a unified representation.

2.1 Edge-aware Graph Transformer for Feature Enhancement

We propose a novel *edge-aware graph transformer* to perform relational reasoning by explicitly incorporating edge attributes into the attention mechanism. Our method advances the local adaptive feature representation [31] in the graph domain, improving the i^{th} superpixel node feature descriptor $D_i \in D$ by selectively aggregating information from its adjacent superpixels.

For node i, we define a single attention head (multi-head omitted for clarity) as follows. Firstly, node features (D_i, D_j) and edge attributes $(e_{i,j})$ are linearly projected using learnable weight matrices W_1, W_2, and W_3 to compute query $\mathcal{Q}_i = W_1 D_i$, key $\mathcal{K}_j = W_2 D_j$, and the embedded edge attribute $\hat{e}_{i,j} = W_3 e_{i,j}$. The edge-aware attention score $\alpha_{i,j}$ from target node i to neighbor node j is computed by injecting the edge embedding into the key before the dot-product:

$$\alpha_{i,j} = \text{softmax}\left(\frac{\mathcal{Q}_i \bigodot (\mathcal{K}_j + \hat{e}_{i,j})}{\sqrt{d}}\right), \tag{2}$$

where d is the per-head hidden dimension and $\bigodot$ is the dot product. We then update node i by aggregating messages from its neighbors, re-injecting the edge embedding into each message to preserve edge semantics in the output:

$$\hat{D}_i = \mathcal{Q}_i + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} \big(W_4 D_j + W_5 e_{i,j}\big), \tag{3}$$

where $\mathcal{N}(i)$ denotes the set of neighbors of node i, and W_4 and W_5 are learnable weight matrices for node and edge messages, respectively.
This design offers two key advantages. (1) Since $e_{i,j}$ participates directly in the query–key interaction, the model learns to emphasize edges that are most informative for the downstream task, rather than treating all connections uniformly. (2) The edge attributes can effectively "shut down" or "boost" the influence of specific neighbors, inducing locally adaptive region features that focus on diagnostically relevant interactions (e.g., irregular borders and varied pigmentation), while suppressing contributions from noisy or less informative neighbors.

2.2 Structural Refinement and Graph-level Readout

To encourage semantic consistency among connected superpixels, we introduce a lightweight similarity-weighted propagation step applied to $\hat{D}_i$ in Eq. (3). This module refines node embeddings by aggregating information from neighboring nodes based on feature similarity, thereby "denoising" the representations produced by the edge-aware graph transformer while limiting the risk of overfitting. Concretely, we employ a learnable temperature parameter β that controls the sharpness of the similarity distribution and thus how selectively each node aggregates information from its neighborhood. Our design is inspired by the similarity-based propagation mechanism in [34], which allows visually or semantically similar superpixels to exchange information even when they are spatially distant. This is particularly beneficial in dermoscopic images, where malignant structures may appear in multiple disjoint regions of a lesion.
Given the i^{th} node embedding $\hat{D}_i$, we compute the refined embedding Z_i as

$$Z_i = \sum_{j \in \mathcal{N}(i) \cup \{i\}} \gamma_{i,j} \, \hat{D}_j, \quad \gamma_{i,j} = \text{softmax}_{j \in \mathcal{N}(i) \cup \{i\}} \big(\beta \, cosine(\hat{D}_i, \hat{D}_j)\big). \tag{4}$$

This refinement step improves the discriminative patterns that recur in the lesion, while suppressing inconsistent or spurious activations. Finally, we obtain a graph-level representation via global mean pooling over the refined node embeddings, $Z = \text{MeanPool}(\{Z_i\}_{i \in V})$, which is subsequently fed into a classifier.

3 Experiments and Results

Training Details. The model is implemented in the PyTorch framework and trained on a single NVIDIA GPU using the Adam optimizer with an initial learning rate of 1×10^{-3}, a weight decay of 1×10^{-4}, and a ReduceLROnPlateau scheduler for 50 epochs. The CNN backbone remains frozen, and only the graph transformer and classifier parameters are optimized, ensuring efficient training and fair, consistent evaluation across datasets.

Datasets. We evaluated the proposed method on four public dermoscopic benchmarks, **ISIC2024** [20], **HAM10000** [35], **PAD-UFES-20** [27], and **HIBA** [29], covering diverse imaging conditions and lesion distributions. Following state-of-the-art (SOTA) practices, all images are resized to 224×224, and the dataset is split into 80% for training and 20% for testing to enable a fair comparison.

Quantitative Evaluation and Discussion. The proposed **GeoMeta-GT** consistently outperformed SOTA methods (Table 1) on four benchmarks, demonstrating both robustness and generalization. On the ISIC2024 dataset, GeoMeta-GT achieves 98.61% accuracy, substantially exceeding the best competing multimodal contrastive model (96.69%) [12]. This gain indicates that explicitly modeling spatial relationships and integrating metadata as a dedicated graph node yields more discriminative and context-aware representations than contrastive image with metadata embedding alone. On HAM10000, the model's 98.23% surpassed the strong hybrid CNN–LSTM model (96.21%) [1] and transformer-based approaches (e.g., 94.31%) [38], highlighting that structured superpixel graphs with edge-aware attention capture fine-grained lesion structure and inter-region morphology better than sequential or global attention without explicit connectivity. On PAD-UFES-20, GeoMeta-GT's 97.17% accuracy exceeds multimodal learning baselines (95.61%) [15] and context-aware GNNs (94.29%) [8], showing stronger robustness to variability and heterogeneity in lesion appearances and imaging conditions. The consistently higher recall and F1-scores further reflect improved sensitivity to diverse lesion features. On the more challenging HIBA dataset, GeoMeta-GT's 95.41% performance markedly outperformed recent graph-based and deep learning models such as context-aware GNNs (89.19%) [8] and multimodal contrastive learning (87.51%) [12], demonstrating superior ability to handle inconsistent acquisition and imaging noise. Across all benchmarks, the empirical improvements confirm that combining geometric edge encoding, structured metadata-as-node fusion, and edge-aware relational attention produced more expressive and generalizable representations than pure CNN, transformer, sequential, or shallow GNN alternatives. These gains are especially notable on heterogeneous and challenging datasets (PAD-UFES-20 and HIBA), underpinning the practical value of geometry-aware relational reasoning in clinical skin lesion classification.

Table 1: Comparison (%) with SOTA methods across four benchmark datasets. Best are shown in **bold.**

Dataset	Method	Accuracy	Precision	Recall	F1-score
ISIC2024 [20]	Deep CNN [17]	90.10	89.81	90.10	89.81
	Lightweight CNN [36]	92.12	92.10	92.11	92.01
	Deep Framework [25]	93.48	93.24	93.41	92.72
	Ensemble Model [18]	93.51	93.31	93.51	92.81
	DSCC_Net [33]	94.17	94.17	93.76	93.93
	Hybrid CNN [28]	94.44	94.38	94.21	94.44
	Multimodal Contrastive [12]	96.69	96.14	96.41	96.69
	GeoMeta-GT	**98.61**	**98.58**	**98.64**	**98.61**
HAM10000 [35]	Context-aware GNN [8]	86.69	90.01	89.19	89.19
	MultiExCam [30]	87.54	87.45	87.49	87.54
	Fused Architecture [5]	91.31	91.21	91.45	91.30
	Soft-Attention Model [14]	93.71	93.70	93.01	93.21
	Transformer Network [38]	94.31	94.10	95.01	93.20
	Inception-ResNet [10]	95.09	95.49	95.16	95.27
	Hybrid LSTM–CNN [1]	96.21	94.75	94.93	95.55
	GeoMeta-GT	**98.23**	**98.21**	**98.26**	**98.23**
PAD-UFES-20 [27]	MDFNet [11]	80.42	78.02	79.97	78.64
	Multimodal Dual-Stage [13]	85.11	84.31	84.61	85.11
	CNN Model [22]	86.41	90.01	85.30	85.91
	CNN-Attention Hybrid [26]	91.22	91.35	91.41	91.19
	Multimodal Contrastive [12]	93.51	93.41	93.21	93.25
	Context-aware GNN [8]	94.29	94.40	94.29	94.26
	Multimodal Learning [15]	95.61	95.54	95.12	95.61
	GeoMeta-GT	**97.17**	**97.17**	**97.21**	**97.17**
HIBA [29]	Diffusion Models [37]	70.68	78.94	62.36	62.59
	MM-Skin [4]	81.65	81.41	81.65	81.58
	CNN Models [24]	82.71	83.21	71.11	81.71
	MultiExCam [30]	85.11	86.20	84.45	85.11
	Multimodal Contrastive [12]	87.51	86.40	87.51	88.51
	Context-aware GNN [8]	89.19	90.01	89.19	89.19
	GeoMeta-GT	**95.41**	**95.38**	**95.44**	**95.41**

4 Ablation Studies

Comprehensive ablations confirm that geometric encoding, metadata fusion, backbone depth, and attention design jointly drive the model's strong, stable performance across all datasets.

Impact of Geometric Edge Encoding. Incorporating geometric edge encoding (GEE) leads to consistent and substantial improvements across all datasets, as shown in Fig. 2(b). Accuracy increases from 96.14% to 98.61% on ISIC2024 (+2.47%), with the F1 score from 96.17% to 98.61% (+2.44%). On HAM10000, accuracy rises from 95.18% to 98.23% (+3.05%), and the F1 score from 95.21% to 98.23% (+3.02%). On PAD-UFES-20, accuracy improves from 94.12% to 97.17% (+3.05%), with the F1 score increasing from 94.15% to 97.17% (+3.02%). On HIBA, accuracy advances from 91.08% to 95.41% (+4.33%), while the F1 score

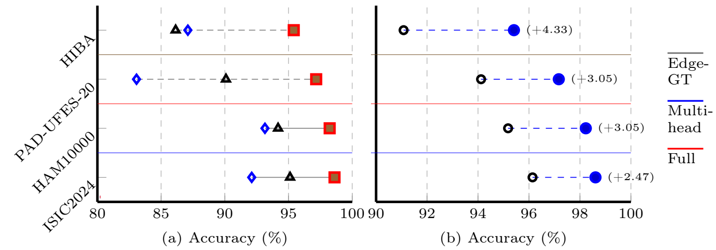

Fig. 2: Illustrates performance comparison and ablation analysis. (a) Accuracy only of model variants. (b) Impact of GEE (dumbbells) on accuracy, showing consistent and substantial gains across all datasets.

rises from 91.09% to 95.41% (+4.32%). These stable margins across datasets show that modeling spatial distance and orientation significantly increases relational expressiveness, making GEE key to strong performance and generalization.
Model Variations. Edge-aware graph transformer (Edge-GT) alone achieves 95.12%, 94.18%, 90.09%, and 86.14%, while naive multi-head concatenation further degrades performance to 92.11%, 93.15%, 83.07%, and 87.10% on ISIC2024, HAM10000, PAD-UFES-20, and HIBA, respectively. In contrast, the full model (Fig. 2(a)) achieves 98.61%, 98.23%, 97.17%, and 95.41%, showing that jointly modeling edge-aware attention, structural refinement, and metadata fusion is essential for optimal, stable performance.
Significance of Metadata. Adding patient metadata consistently improves performance from 94.12%, 95.16%, 92.08%, and 90.14% to 98.61%, 98.23%, 97.17%, and 95.41% across ISIC2024, HAM10000, PAD-UFES-20, and HIBA, respectively, demonstrating its complementary discriminative contribution.
Evaluating Backbone Sensitivity. Backbone choice significantly impacts performance. As VGG19 achieves 91.08%, 92.12%, 89.09%, and 87.06%, Inception reaches 93.12%, 94.14%, 91.10%, and 88.07%, and EfficientNet improves to 95.14%, 96.16%, 94.12%, and 92.09%. ResNet152 performs best with 98.61%, 98.23%, 97.17%, and 95.41%, on the ISIC2024, HAM10000, PAD-UFES-20, and HIBA datasets, respectively. This indicates that deeper residual representations provide stronger features for the proposed framework.
Graph Construction and Superpixel Sensitivity. We evaluated graph sensitivity to superpixel granularity ($K \in \{15, 25, 35, 50, 65, 75\}$) and neighborhood size ($n \in \{4, 6, 8, 10\}$). Performance was stable across settings, with the best accuracies at $K = 50$ and $n = 6$ (98.61%, 98.23%, 97.17%, 95.41% on ISIC2024, HAM10000, PAD-UFES-20, and HIBA, respectively). Using SEEDS [9] or Felzenszwalb [16] instead of SLIC yielded similar results, indicating the framework is largely segmentation-agnostic, though SLIC performs best overall.
Effect of Distance and Orientation Features. Incorporating orientation consistently improves performance over distance-only representation, raising best accuracy from 96.17% to 98.61%, 95.23% to 98.23%, 94.18% to 97.17%, and 91.13% to 95.41% on ISIC2024, HAM10000, PAD-UFES-20, and HIBA, respec-

tively, confirming that combining spatial distance with directional context yields more discriminative graphs and more robust classification.

5 Conclusion

We propose a geometry-aware superpixel graph framework for skin lesion classification that represents dermoscopic images as region adjacency graphs with edge-aware relational reasoning. Using frozen CNN features and encoding patient metadata as a dedicated graph node, the model captures structural and clinical context without backbone fine-tuning. Experiments and ablations demonstrate robust, superior generalization across benchmarks. Future work includes multi-class diagnosis, uncertainty modeling, and improved interpretability for clinical use. Code is available at: https://github.com/azeemchaudharyg/GeoMeta-GT

Acknowledgments. This work was supported by UK Research and Innovation (UKRI) — Economic and Social Research Council (ESRC) grant SCAnDi (ES/Y010655/1), and the EU Marie Skłodowska-Curie Actions (MSCA) grant SynAM (HORIZON-MSCA-2022-SE-01, Project: 101129996).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abohashish, S.M., Amin, H.H., Elsedimy, E.: Enhanced melanoma and non-melanoma skin cancer classification using a hybrid LSTM-CNN model. Scientific Reports **15**(1), 24994 (2025)
2. Achanta, R., Shaji, A., Smith, K., , et al.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012)
3. Ahammed, S., Cui, X., Lu, W., et al.: Skin lesion classification using dermoscopic images and clinical metadata: Insights from multimodal models. In: IEEE/CVF, CVPR Workshops. pp. 222–230 (2025)
4. Araújo, D.J., Verdelho, M.R., Bissoto, A., Nascimento, J.C., et al.: Key patches are all you need: A multiple instance learning framework for robust medical diagnosis. In: Proceedings of the IEEE/CVF, CVPR, Workshops. pp. 5231–5240 (2024)
5. Arshad, M., Khan, M.A., Almujally, N.A., et al.: Multiclass skin lesion classification and localziation from dermoscopic images using a novel network-level fused deep architecture and explainable artificial intelligence. BMC **25**(1), 215 (2025)
6. Azeem, M., Javaid, S., et al.: Neural networks for the detection of COVID-19 and other diseases: prospects and challenges. Bioengineering **10**(7), 850 (2023)
7. Azeem, M., Kiani, K., Mansouri, T., Topping, N.: SkinLesNet: Classification of skin lesions and detection of melanoma cancer using a novel multi-layer deep convolutional neural network. Cancers **16**(1), 108 (2023)
8. Azeem, M., Nazir, S., Ahmed, A., Behera, A.: Context-aware graph neural network for skin lesion classification. In: Proceedings of the 41st ACM/SIGAPP Symposium on Applied Computing. pp. 164–173 (2026)

9. Van den Bergh, M., Boix, X., Roig, G., et al.: Seeds: Superpixels extracted via energy-driven sampling. In: ECCV. pp. 13–26. Springer (2012)
10. Bozkurt, F.: Skin lesion classification on dermatoscopic images using effective data augmentation and pre-trained deep learning approach. Multimedia Tools and Applications **82**(12), 18985–19003 (2023)
11. Chen, Q., Li, M., Chen, C., et al.: MDFNet: Application of multimodal fusion method based on skin image and clinical data to skin cancer classification. Journal of Cancer Research and Clinical Oncology **149**(7), 3287–3299 (2023)
12. Christopoulos, D., Spanos, S., Baltzi, E., et al.: Skin lesion phenotyping via nested multi-modal contrastive learning. arXiv preprint arXiv:2505.23709 (2025)
13. Dai, W., Liu, R., Wu, T., et al.: Deeply supervised skin lesions diagnosis with stage and branch attention. IEEE JBHI **28**(2), 719–729 (2023)
14. Datta, S.K., Shaikh, M.A., Srihari, S.N., et al.: Soft attention improves skin cancer classification performance. In: IWIMI in MIC. pp. 13–23. Springer (2021)
15. Fan, S., Ahmed, A., Zeng, X., et al.: A personalized multimodal federated learning framework for skin cancer diagnosis. Electronics **14**(14), 2880 (2025)
16. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International journal of computer vision **59**(2), 167–181 (2004)
17. Golnoori, F., Boroujeni, F.Z., Monadjemi, A.: Metaheuristic algorithm based hyper-parameters optimization for skin lesion classification. Multimedia Tools and Applications **82**(17), 25677–25709 (2023)
18. Imran, A., Nasir, A., Bilal, M., et al.: Skin cancer detection using combined decision of deep learners. IEEE Access **10**, 118198–118212 (2022)
19. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
20. Kurtansky, N.R., D'Alessandro, B.M., Gillis, M.C., et al.: The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection. Scientific Data **11**(1), 884 (2024)
21. Leema, A.A., Balakrishnan, P., Gopichand, G., et al.: LMS-ViT: A multi-scale vision transformer approach for real-time smartphone-based skin cancer detection. Frontiers in AI **8**, 1612502 (2025)
22. Li, F., Li, M., Zuo, E., et al.: Self-contrastive feature guidance based multidimensional collaborative network of metadata and image features for skin disease classification. Pattern Recognition **156**, 110742 (2024)
23. Naseri, H., Safaei, A.A.: Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: a systematic literature review. BioMed Central (BMC) Cancer **25**(1), 75 (2025)
24. Oyedeji, M.O.: Clinical and dermoscopy image-based deep learning models for skin lesion diagnosis in clinical practice. NMAHIB **13**(1), 43 (2024)
25. Ozdemir, B., Pacal, I.: A robust deep learning framework for multiclass skin cancer classification. Scientific Reports **15**(1), 4938 (2025)
26. Pacheco, A.G., Krohling, R.A.: An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. IEEE Journal of Biomedical and Health Informatics **25**(9), 3554–3563 (2021)
27. Pacheco, A.G., Lima, G.R., Salomao, A.S., et al.: PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in Brief **32**, 106221 (2020)
28. Pradhan, J., Singh, A., Kumar, A., et al.: Skin lesion classification using modified deep and multi-directional invariant handcrafted features. Journal of Network and Computer Applications **231**, 103949 (2024)

29. Ricci Lara, M.A., Rodríguez Kowalczuk, M.V., Lisa Eliceche, M., et al.: A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population. Scientific Data **10**(1), 712 (2023)
30. Ruga, T., Caroprese, L., Vocaturo, E., et al.: MultiExCam: A multi approach and explainable artificial intelligence architecture for skin lesion classification. Computer Methods and Programs in Biomedicine p. 109081 (2025)
31. Shi, Y., Huang, Z., Feng, S., et al.: Masked label prediction: Unified message passing model for semi-supervised classification. arXiv preprint arXiv:2009.03509 (2020)
32. Siegel, R.L., Kratzer, T.B., Wagle, N.S., Sung, H., Jemal, A.: Cancer statistics, 2026. Cancer Journal for Clinicians **76**(1), e70043 (2026)
33. Tahir, M., Naeem, et al.: DSCC_Net: Multi-classification deep learning models for diagnosing of skin cancer using dermoscopic images. Cancers **15**(7), 2179 (2023)
34. Thekumparampil, K.K., Wang, C., Oh, S., et al.: Attention-based graph neural network for semi-supervised learning. arXiv preprint arXiv:1803.03735 (2018)
35. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 1–9 (2018)
36. Tuncer, T., Barua, P.D., et al.: A lightweight deep convolutional neural network model for skin cancer image classification. ASC **162**, 111794 (2024)
37. Uliana, J.J., Krohling, R.A.: Diffusion models applied to skin and oral cancer classification. arXiv preprint arXiv:2504.00026 (2025)
38. Xin, C., Liu, Z., Zhao, K., et al.: An improved transformer network for skin cancer classification. Computers in Biology and Medicine **149**, 105939 (2022)