Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Seeing the Skin Deeper: Interpretable Multi-Task Framework for Skin Lesion Diagnosis using Superpixel Graphs

J.G JERLSHIN[1], ABHISHEK RUDRA PAL[2], RITUPARNA DATTA[3], AFNAN SHAIK JAVEED[1], and VITTAL J.S[1]

[1]School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, India
[2]School of Mechanical Engineering (SMEC), Vellore Institute of Technology, Chennai, India (e-mail: abhishek.rudrapal@vit.ac.in)
[3]Accenture Technology Centers in India (ATCI), Bangalore, India

Corresponding author: Abhishek Rudra Pal (e-mail: abhishek.rudrapal@vit.ac.in).

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT Accurate and early identification of malignant skin lesions is critical for successful intervention and beneficial prognoses. Nevertheless, most current deep learning models concentrate solely on lesion classification, omitting simultaneous prediction of practically useful morphological and colorimetric attributes critical for interpretability. We propose a novel multi-task graph-based learning framework that integrates disease classification, structure attribute identification, and chromatic analysis into a unified architecture. Dermatoscopic images are converted into superpixel-level graphs of nodes representing high-resolution Swin Transformer embeddings, and edges representing spatial continuity, thereby preserving biologically meaningful skin structure. A shared Graph Attention Network (GAT) backbone enables contextual information propagation across multiple tasks and is supplemented by task-specific decoders for comprehensive inference. To compensate for class imbalance and encourage cross-task cooperation, a focal loss functional and a balanced multi-objective optimization strategy are incorporated by the framework. Moreover, to enhance clinical interpretability, GNNExplainer and attention-based visualizations of diagnostically informative regions are aligned according to dermatological heuristics. On PH2, our model achieves an F1-score of 1.000 for lesion classification, 0.881 for morph attribute prediction, and 0.895 for chromatic attribute identification, thus setting new state-of-the-art performance across all evaluation metrics. As far as we know, this is the first interpretable graph-based model for simultaneously learning diagnostic and morph-colorimetric attributes, hence pushing explainable AI for dermatological images forward for increased precision, interpretability, and clinical reliability.

INDEX TERMS Clinical Feature Prediction, Explainable Artificial Intelligence (XAI), Graph Attention Networks (GAT), Multi-Task Learning (MTL), Skin Lesion Classification, Superpixel-based Graph Representation.

I. INTRODUCTION

SKIN cancer is among the most prevalent types of malignancy across the globe, and melanoma carries the greatest risk of mortality based on its aggressive growth tendency and metastatic potential. Prompt and precise diagnosis aids successful therapeutic intervention and improved prognostic outcomes. Typically, dermatologists perform dermoscopic analysis to analyze lesions based on their morphologic attributes and colorimetric attributes, such as asymmetry, pigmentary networks, streaks, dots, globules, and blue-whitish veils. Yet, interpretation by hand is inherently subjective, prone to observer variability, and highly reliant on clinical experience. Such limitations justify a critical need for automatic systems of diagnosis employing artificial intelligence (AI) and deep learning methodologies for reduction of subjectivity and enhancement of diagnostic accuracy. Deep learning, and Convolutional Neural Networks (CNNs) in particular, has seen tremendous growth in processing of medical images. CNN-based methodologies, however, have some inherent disadvantages, such as their inability to model spatial relationships between different subregions of an image explicitly, which is of immense importance for appropriate analysis of medical images. Such a shortcoming holds particularly

true for medical images, where contextual and structural information between constituents of lesions are of immense importance for appropriate diagnosis. As compared to natural images, skin lesions consist of complex and nonsmooth patterns, for which a model should be capable of modeling local interactions between features other than pixel-wise classification. To address these challenges, Graph Neural Networks (GNNs) have emerged as a potential alternative, enabling the clear modeling of geographical interdependencies between different lesion areas. By representing dermoscopic images as graphs, where nodes are indicative of important areas (superpixels) and edges express geographical associations, GNNs are adept at perceiving topological and structural information that are usually overlooked by Convolutional Neural Networks (CNNs). Such systematic organization facilitates better understanding and enhances diagnostic precision. Nonwithstanding GNN's benefits, state-of-the-art automatic dermatological diagnosis systems are based on CNN-based end-to-end classification architectures processing dermoscopic images in a holistic manner. Nonetheless, such approaches are plagued by several shortcomings. CNNs are based on local receptive fields and pyramid feature extraction, which are not suitable for extracting long-distance dependencies. Yet, skin lesion identification requires examining long-distance relationships between constituents of lesions, e.g., between pigment networks, streaks, and regression areas. Moreover, CNNs are based on huge amounts of labeled train data for generalization purposes. Nonetheless, labeled data in medical images are limited and costly to acquire, thereby limiting CNN-based model scalability. Another limitation of CNN-based systems is their intrinsic inability to be interpretable, acting as black-box systems that provide little insight into their decision-making processes. For dermatological diagnosis, it is important to have interpretability so as to establish clinical confidence and also make AI-supported decisions align with experts' assessments. CNNs also process images as grids of dense pixels, leading to unnecessary feature extraction of local neighbors. Important dermatological features, like asymmetry, streaks, or globules, are better described as high-level structures, which are not explicitly learned by CNNs. To tackle these challenges, we propose a novel framework relying on Graph Neural Networks to transform dermoscopic images into graph representations. Here, a node represents a semantically meaningful area (superpixel), whereas edges capture the inter-location relationships between regions. This structured representation provides several advantages regarding traditional CNN-based methods. Instead of regarding complete images as dense pixel grids, our approach breaks down the lesion into superpixels by employing SEEDS (Superpixel Extraction via Energy-Driven Sampling), a compact computational segmentation method. As a result, it becomes possible for the model to focus on diagnostically relevant areas by eliminating redundancy of neighbouring pixels. Each of the graph's vertices is correlated with feature embeddings learned by a SWIN Transformer, which efficiently reveals local and global contextual dependences. Edges between nodes define geographical relationships, hence enabling the model to learn topological patterns of lesions critical for precise diagnosis. Compared to single-task classification-centered models, our framework utilizes multi-task prediction to simultaneously predict disease classification, clinical feature existence, and colorimetric attributes. The multi-dimensional approach yields a richer diagnostic framework beneficial for dermatologic practice. The framework's integration of Graph Attention Networks enhances model explanability by assigning attention weights to diagnostically critical regions, whereas GNNExplainer facilitates clear-sighted decision-making by explaining critical graph structures affecting model predictions.

This study presents a novel superpixel-based graph learning framework that significantly enhances diagnostic accuracy, interpretability, and data efficiency in automated skin lesion diagnosis. The key contributions of our research are as follows:

1) Graph-Based Representation of Skin Lesions: We introduce a superpixel-driven graph construction pipeline that encodes regional features and spatial dependencies, demonstrating superior diagnostic efficacy compared to traditional CNN-based methods.
2) Hybrid Feature Extraction via SWIN Transformer: We employ SWIN Transformers for node feature initialization, enabling the model to capture both local and global lesion characteristics while preserving fine-grained spatial details.
3) Multi-Task Learning for Comprehensive Diagnosis: Our model concurrently predicts disease classification, clinical features, and colorimetric attributes, offering a holistic diagnostic framework beneficial for dermatological practice.
4) Explainability with GNNExplainer and Attention Visualization: To enhance clinical trust, we incorporate GNNExplainer and attention-based visualization techniques, allowing dermatologists to interpret model predictions and validate their reliability.

By leveraging graph-based modeling, hybrid feature extraction, and multi-task learning, our proposed framework advances the state of AI-driven dermatological diagnostics, paving the way for enhanced accuracy, transparency, and clinical applicability.

II. RELATED WORK

A. DEEP LEARNING IN SKIN LESION DIAGNOSIS

Deep learning has significantly advanced computer-aided diagnosis (CAD) systems for dermatological applications. Convolutional Neural Networks (CNNs), particularly ResNet-50 and ResNet-101, have been widely adopted for multi-class classification tasks, achieving up to 91.7% accuracy on datasets such as HAM10000 with the help of data augmentation and transfer learning techniques [1]. Multi-task learning (MTL) frameworks that jointly model lesion classification and contextual cues like body location have demon-

strated improved robustness, reaching a mean average precision of 0.80 [2]. Despite these advances, some architectures such as Inception-v3 coupled with shallow classifiers yielded suboptimal performance—achieving only 65.8% accuracy—suggesting limitations in feature representation [3]. Ensemble learning approaches have addressed this gap by aggregating diverse models including ResNeXt, DenseNet, and Xception, attaining a recall of 94% [4]. Hybrid techniques that integrate CNN-extracted features with hand-crafted descriptors like Local Binary Patterns (LBP) have shown further performance improvements [5].

B. TRANSFORMER-BASED ARCHITECTURES

Transformer-based models have emerged as powerful alternatives to traditional CNNs. Vision Transformers (ViTs), Swin Transformers, and DinoV2 have been deployed in dermatological image analysis, with DinoV2 achieving state-of-the-art results—96.48% accuracy and an F1-score of 0.9727—alongside enhanced explainability using Grad-CAM and SHAP [6]. Fine-tuned versions of VGGNet on the ISIC Archive dataset have also demonstrated promising results, achieving 78.66% sensitivity with optimized regularization [7]. Customized CNNs leveraging global descriptors on balanced datasets have further pushed accuracy to 95.18% [8].

C. CLASSICAL AND HYBRID MACHINE LEARNING METHODS

Although deep learning dominates recent literature, classical machine learning approaches such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) remain relevant. Comparative analyses on the PH2 dataset revealed that ANNs outperformed other classifiers, achieving 92.50% accuracy [9]. Hybrid methods integrating traditional image processing (e.g., contrast enhancement and segmentation) with deep learning yielded up to 96.7% accuracy on ISIC 2017 and HAM10000 datasets [10].

D. SKIN LESION SEGMENTATION TECHNIQUES

Accurate segmentation remains foundational for downstream classification. Ensemble models combining DeepLabV3+ and Mask R-CNN have achieved high segmentation sensitivity (89.93%) and specificity (97.94%) [11]. A full-resolution convolutional network (FrCN) paired with a dynamic graph cut algorithm (DGCA) further improved classification accuracy to 97.986% by addressing over-segmentation issues [12]. Multimodal fusion of dermoscopic images, macroscopic photographs, and patient metadata has also been explored, achieving an AUC of 0.866 in melanoma detection [13].

E. GRAPH-BASED AND SUPERPIXEL-ORIENTED APPROACHES

Graph-based learning has recently garnered attention in skin lesion analysis due to its ability to encode spatial relationships and semantic structure. Superpixel Attention Networks (SANet) effectively detected lesion attributes in the ISIC 2018 Task 2 challenge [14], while GNNs have been used to model complex biological interactions in medical diagnosis tasks [15]. A hybrid of GNNs and Capsule Networks achieved 95.52% accuracy on MNIST-HAM10000, demonstrating the synergy of local and spatial reasoning [16]. Other graph-driven models using dynamic graph cuts combined with Naïve Bayes classifiers reported 94.3% accuracy in classifying benign lesions [17]. Superpixel-based feature extraction, as seen in iSLIC-enhanced frameworks, has improved segmentation precision and classification performance when paired with machine learning models [18]. More sophisticated integrations, such as the Confidence Partitioning Sampling Filtering (CPSF) framework, HDAETUNet3+, and DGCRIN, have outperformed state-of-the-art models on ISIC 2019 [19].

F. MULTI-TASK LEARNING AND GRAPH NEURAL NETWORKS

Multi-task learning has been widely applied across supervised and semi-supervised domains, particularly in health informatics, due to its efficiency in learning from limited data [20]. Within graph-based methodologies, Graph Convolutional Networks (GCNs) have been categorized into spectral and spatial domains, though challenges remain in scalability and handling non-Euclidean data structures [21].

G. REVIEWS, DATASETS, AND STANDARDIZATION

Systematic reviews consistently affirm the dominance of deep learning over classical models in CAD systems for skin lesion analysis, while emphasizing concerns around data bias, segmentation consistency, and lack of clinical validation [22], [23]. The role of explainability and model interpretability has also been increasingly highlighted in dermatological AI studies [24]. GCNs, though less common in dermatology, have shown utility across radiology and histopathology, underscoring the need for optimized graph transformations [25]. Benchmark datasets such as PH2 [26], ISIC 2016 [27], and ISIC 2018 [28] have standardized evaluation metrics for classification, segmentation, and attribute detection, promoting reproducibility and fair comparison across models.

H. POSITIONING OUR WORK

Despite significant advancements in dermatological image analysis using convolutional neural networks (CNNs), transformers, and ensemble-based architectures, existing methodologies largely operate as monolithic classification systems. These approaches typically focus on disease categorization without accounting for the rich set of clinical attributes that dermatologists rely on for nuanced diagnosis and decision-making. Furthermore, while recent efforts have begun to explore graph-based representations, they often fall short in terms of interpretability and are seldom integrated within a multi-task learning framework. In contrast to these limitations, our work addresses a critical gap by proposing an interpretable, multi-task diagnostic framework that synergis-

FIGURE 1. Representative examples of Common Nevus (left), Atypical Nevus, and Melanoma.

tically combines superpixel-based graph construction with transformer-derived embeddings. Specifically, we leverage Graph Attention Networks (GATs) to capture spatial and relational dependencies across lesion sub-regions, using superpixels as graph nodes. Each node is enriched with high-dimensional semantic features extracted from the Swin Transformer V2 backbone, enabling a more expressive and localized understanding of skin morphology. Crucially, our framework is designed to concurrently perform disease classification, morphological attribute detection, and color feature recognition—tasks that are traditionally handled in isolation. This multi-task configuration not only promotes shared learning across related objectives but also enhances model generalization and clinical relevance. By incorporating attention-based interpretability at both node and structural levels, our method facilitates transparent decision-making, a cornerstone requirement for clinical deployment. Therefore, our approach uniquely positions itself at the intersection of graph learning, transformer-based feature extraction, and explainable multi-task dermatological analysis.

III. PROPOSED METHODOLOGY

This section presents the design and implementation of our interpretable multi-task diagnostic framework, which leverages superpixel-driven graph construction, transformer-based embeddings, and attention-enhanced Graph Neural Networks (GNNs) for skin lesion analysis. The pipeline is built to address both classification and clinical attribute prediction tasks, promoting diagnostic robustness and interpretability.

A. DATASET

To evaluate the efficacy and generalizability of our proposed framework, we utilized the PH2 dataset, a publicly available and well-established benchmark in the domain of dermoscopic image analysis [26]. The dataset comprises 200 high-resolution dermoscopic images of melanocytic lesions, each meticulously annotated with both diagnostic labels and clinically relevant dermatological attributes. These images are categorized into three diagnostic groups: common nevi (80 images), atypical nevi (80 images), and melanomas (40 images), representing benign, potentially malignant, and malignant lesion types, respectively. Representative examples of these diagnostic categories from the PH2 dataset are shown in Fig. 1.

The distribution of specific categorical features across these lesion types is further analyzed and visualized in Fig. 2.

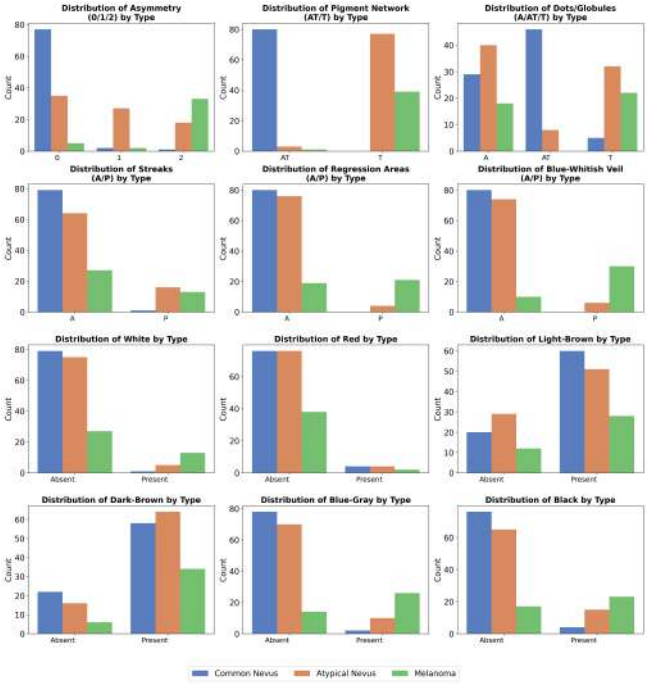

FIGURE 2. Count plots illustrating the distribution of various categorical features across the three diagnostic types: Common Nevus, Atypical Nevus, and Melanoma.

Beyond diagnostic labels, each image in the dataset is annotated with six morphological and colorimetric features that are integral to clinical decision-making. These include Asymmetry (A), which quantifies the deviation of the lesion from geometric symmetry and serves as a visual marker of irregular growth patterns often associated with malignancy; Pigment Network (P), which captures the distribution and structural organization of pigmentary patterns within the lesion and provides cues about lesion type and progression; Dots and Globules (D), referring to small pigmented structures that signal cellular proliferation and are commonly seen in both benign and malignant lesions; Streaks (S), characterized by radial projections at the lesion periphery, which are frequently observed in melanoma and indicate aggressive morphological evolution; Regression Areas (R), defined as depigmented zones within the lesion that suggest prior tissue destruction or lesion evolution, often associated with malignancy; and Color Presence (C), encompassing six clinically relevant color indicators—white, red, light brown, dark brown, blue-gray, and black. These color features are recognized by dermatologists as essential diagnostic markers, each conveying specific biological and pathological information about the lesion.

These multi-attribute annotations are incorporated into our model's learning process via a multi-task learning strategy. This design enables the framework to jointly predict the primary diagnostic class while also learning to identify the presence or absence of each clinical attribute. As such, the model not only performs classification but also mimics the diagnostic reasoning process employed by dermatologists,

thereby enhancing both accuracy and interpretability in automated skin lesion diagnosis.

B. GRAPH-BASED REPRESENTATION FOR MEDICAL IMAGING

Traditional deep learning methods for medical image analysis rely predominantly on convolutional neural networks (CNNs), which process images in a Euclidean space. While CNNs excel at capturing spatial hierarchies and local feature dependencies, they impose grid-based structural constraints that limit their ability to model non-Euclidean, relational, and global contextual dependencies. Given the complex nature of skin disease diagnosis, where lesions exhibit diverse morphological patterns, irregular structures, and interimage similarities, a graph-based approach provides a more flexible and semantically meaningful representation.

1) CNN Limitations and Motivation for Graph-Based Modeling

Convolutional Neural Networks (CNNs) have long been the dominant architecture in medical image analysis due to their ability to hierarchically extract local spatial features from grid-structured data. Formally, given an input image $I \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels respectively, CNNs apply a series of convolutional filters W_l to generate feature maps F_l through the transformation $F_l = \sigma(W_l * F_{l-1})$, where $*$ denotes the convolution operation and σ is a nonlinear activation function. Despite their local efficiency, CNNs exhibit several critical limitations when applied to dermoscopic image analysis. First, CNNs operate on fixed-size receptive fields, which restrict their ability to model global contextual relationships unless an impractically deep architecture is employed. This lack of global contextual awareness is a significant drawback for analyzing lesions, which often require understanding spatial patterns across the entire image. Second, the assumption of a regular pixel grid makes CNNs ill-suited for capturing complex morphological patterns and irregular structures commonly found in medical images. Third, CNNs are sensitive to dataset-specific biases, especially when dealing with high inter-class variability and low inter-class distinguishability—as is the case in skin lesion datasets. Finally, CNN-based methods lack the ability to capture inter-image relationships or explicitly model dependencies between distinct regions of an image, which limits their effectiveness in representing the topological and semantic complexity of dermatological features. To address these shortcomings, we adopt a graph-based representation of images, where local regions are modeled as nodes and their spatial or semantic relationships are encoded as edges. This framework enables flexible, non-Euclidean modeling of image data, facilitating both local and global reasoning in a structure-aware manner that is better aligned with the anatomical and pathological characteristics of skin lesions.

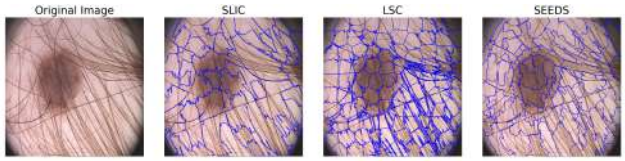

FIGURE 3. Comparison of superpixel segmentation algorithms applied to dermoscopic images: (a) SLIC, (b) LSC, and (c) SEEDS. SEEDS demonstrates superior region boundary adherence and visual coherence, aligning better with clinically relevant lesion structures.

2) Superpixel Segmentation for Region Structuring

To transform dermoscopic images into graph-compatible representations, we begin by segmenting each image into semantically meaningful regions using superpixel-based techniques. Superpixels serve as perceptually coherent clusters of pixels that share similar visual characteristics, thus enabling efficient and structured encoding of the image content. These regions subsequently form the nodes of our graph, facilitating localized, interpretable feature aggregation.

In this work, we evaluated three widely used superpixel algorithms—Simple Linear Iterative Clustering (SLIC), Linear Spectral Clustering (LSC), and Superpixel Extraction via Energy-Driven Sampling (SEEDS)—to determine the most suitable method for modeling dermoscopic lesion morphology. Each algorithm was applied to a subset of representative dermoscopic images, and the segmentation quality was assessed based on visual coherence, boundary adherence, and clinical interpretability. Fig. 3 presents an illustrative comparison of the segmentation outputs generated by these three methods.

Among the evaluated methods, SEEDS was selected as the preferred segmentation strategy. Unlike SLIC and LSC—which, while computationally efficient, often produce irregular or fragmented superpixels in areas of subtle textural transition—SEEDS exhibited a more refined segmentation that maintained alignment with dermatologically significant lesion boundaries. Its iterative energy minimization framework allows it to produce more compact, smooth, and anatomically relevant superpixels, which are essential for reliable downstream graph construction and feature learning.

The SEEDS algorithm segments an input image $I \in \mathbb{R}^{H\times W\times C}$ into N superpixels $\{S_i\}_{i=1}^{N}$, where each S_i denotes a spatially contiguous group of pixels sharing similar color and texture properties. The segmentation process is formulated as an energy minimization problem:

$$E = \lambda_1 \sum_{i=1}^{N} D_c(S_i) + \lambda_2 \sum_{i=1}^{N} D_s(S_i) + \lambda_3 \sum_{i=1}^{N} R_b(S_i), \quad (1)$$

where $D_c(S_i)$ quantifies intra-superpixel color variance—typically measured in the perceptually uniform CIELAB color space; $D_s(S_i)$ enforces spatial regularity by penalizing irregularly shaped or elongated superpixels; and $R_b(S_i)$ promotes boundary alignment by penalizing superpixel edges that cut across high image gradients. The weighting parameters $\lambda_1, \lambda_2, \lambda_3 \in$

$\mathbb{R}^+$ are empirically tuned to balance the competing objectives of color homogeneity, spatial coherence, and boundary preservation.

SEEDS operates through a three-stage hierarchical refinement process. Initially, the image is divided into a coarse grid, with each block initialized using its average color statistics. In the block update stage, neighboring blocks are iteratively exchanged to minimize the energy function, effectively merging regions with similar visual patterns. Finally, a pixel-level refinement phase adjusts boundaries at the pixel granularity, enabling the capture of subtle diagnostic features such as pigment networks, streaks, and asymmetrical extensions.

The resulting superpixel map not only reduces computational redundancy by grouping similar pixels but also preserves critical lesion structures, which are otherwise diluted in dense pixel grids. Each superpixel is later treated as a graph node, encapsulating localized information in a semantically structured format conducive to graph neural processing. By adopting SEEDS as our segmentation method, we ensure that each node in the graph corresponds to a clinically meaningful subregion of the lesion, enhancing both interpretability and learning efficacy in the subsequent graph construction and classification pipeline.

3) Graph Construction: Node and Edge Modeling

Upon segmenting the dermoscopic image into coherent superpixel regions using the SEEDS algorithm, we construct a graph-based representation to facilitate structured, context-aware analysis of lesion morphology. Unlike traditional CNNs that process images as fixed Euclidean grids with limited receptive fields, graph-based models enable the explicit encoding of both local and non-local dependencies. This is particularly advantageous in medical imaging, where lesions often exhibit irregular spatial configurations, heterogeneous textures, and context-dependent features. By representing images as graphs, we model inter-regional relationships more flexibly and accurately, thereby enhancing the interpretability and diagnostic robustness of the framework.

Formally, each dermoscopic image $I \in \mathbb{R}^{H\times W\times 3}$ is transformed into an undirected graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes and $E \subseteq V \times V$ is the set of edges. Each node v_i corresponds to a superpixel S_i, derived from the SEEDS segmentation, and represents a semantically coherent region of the image. These nodes encapsulate localized visual descriptors such as texture, color, and structural cues, which are encoded through transformer-based feature extraction. Node feature representation is derived from a pretrained SwinV2 Transformer model with window size 8×8. The input image is passed through the transformer, yielding a latent representation of shape $(1, 64, 768)$, corresponding to 64 non-overlapping patches (each of size 4×4) embedded in a 768-dimensional feature space. These patch embeddings are rearranged into a spatial grid of size 8×8 and subsequently upsampled via bilinear interpolation to match the original image resolution, resulting in a dense feature map $F \in \mathbb{R}^{768\times H\times W}$. This interpolation ensures that each pixel location (x, y) is associated with a contextualized feature vector $f(x, y) \in \mathbb{R}^{768}$ that captures local semantics and hierarchical contextual dependencies from the transformer backbone.

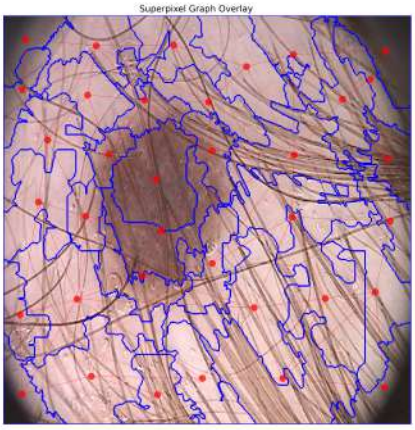

FIGURE 4. Graph construction from superpixel segmentation. Superpixels are shown with their boundaries, nodes are represented with red dots and edges represent connections between adjacent regions.

To derive node-level features, each superpixel S_i is treated as a region composed of a set of pixel locations $\{(x_j, y_j)\}_{j=1}^{|S_i|}$, and its corresponding node feature vector $f_i \in \mathbb{R}^{768}$ is computed by averaging the per-pixel transformer features over the region:

$$f_i = \frac{1}{|S_i|} \sum_{(x,y)\in S_i} f(x, y), \tag{2}$$

where $f(x, y)$ denotes the interpolated SwinV2 feature vector at pixel location (x, y). This aggregation process captures the dominant semantic signature of each superpixel, thereby enabling the encoding of rich visual representations while preserving the anatomical coherence of local regions. The resulting set of node features $\{f_i\}_{i=1}^{n}$ forms the input signal for subsequent graph neural network processing.

To illustrate this process, Figure 4 shows the superpixel map overlayed with edge connections, providing an intuitive visualization of the initial graph formation derived from the segmented lesion regions.

Graph connectivity is established based on two complementary criteria: spatial adjacency and semantic similarity. The spatial structure of the lesion is preserved by connecting nodes corresponding to directly adjacent superpixels. This results in a spatial adjacency matrix $A_{\text{spatial}} \in \{0, 1\}^{n\times n}$, defined as:

$$A_{\text{spatial}}(i, j) = \begin{cases} 1, & \text{if } S_i \text{ and } S_j \text{ share a boundary,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This spatial graph topology ensures that each node maintains direct interaction with its neighboring lesion regions, enabling effective modeling of local feature propagation.

To account for contextual relationships beyond immediate neighbors, we augment the graph with similarity-based edges that connect morphologically similar superpixels, even if they

are spatially distant. These connections are determined by computing the cosine similarity between the node feature vectors f_i and f_j. A similarity adjacency matrix $A_{\text{similarity}} \in \{0,1\}^{n\times n}$ is constructed as follows:

$$A_{\text{similarity}}(i,j) = \begin{cases} 1, & \text{if } \cos(f_i,f_j) \geq \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where τ is a threshold hyperparameter that regulates the sparsity of the similarity-based edges. This allows the model to preserve long-range dependencies and aggregate features from semantically aligned regions, which may be critical for distinguishing ambiguous or subtle lesion patterns.

The final adjacency matrix $A \in \mathbb{R}^{n\times n}$ combines both spatial and similarity-based topologies using a weighted sum:

$$A = A_{\text{spatial}} + \alpha A_{\text{similarity}}, \tag{5}$$

where $\alpha \in \mathbb{R}^+$ is a tunable parameter that modulates the influence of non-local connections. This hybrid connectivity schema enables the model to simultaneously exploit anatomical structure and semantic coherence across lesion regions.

To further enhance the discriminative capacity of the graph, each edge $(i,j) \in E$ is assigned a continuous weight w_{ij} based on the feature similarity between the connected nodes. Specifically, the edge weights are computed using a Gaussian kernel:

$$w_{ij} = \exp\left(-\frac{\|f_i - f_j\|^2}{2\sigma^2}\right), \tag{6}$$

where σ is a scaling hyperparameter that modulates the sensitivity of the similarity function, thereby influencing the distribution of edge weights. This formulation ensures that connections between semantically similar regions are emphasized, while weaker or less informative associations are naturally down-weighted. As a result, the constructed graph promotes meaningful feature propagation and mitigates the influence of noisy or redundant interactions.

The final constructed graph Figure 5 consisting of feature-enriched nodes and adaptively weighted edges—serves as the structural substrate for downstream graph neural processing.

Overall, this graph construction paradigm—integrating spatial adjacency with semantic affinity and guided by adaptive edge weighting—produces a topologically consistent and diagnostically relevant representation of dermoscopic images. By capturing both local anatomical structure and global contextual relationships, the resulting graph serves as a robust substrate for downstream graph neural network processing. This foundational design enhances the model's capacity for accurate, interpretable, and clinically actionable lesion classification and feature prediction.

4) Deep Feature Extraction

Accurate and semantically rich node feature representations are fundamental to the effectiveness of graph-based models in medical image analysis. Traditional convolutional neural networks (CNNs), though successful in capturing local texture and spatial hierarchies, are inherently constrained by their limited receptive fields and strong spatial inductive biases. These limitations hinder their ability to model global dependencies and non-local interactions—properties essential for understanding the complex and irregular morphology of skin lesions. To overcome these challenges, we adopt the Swin Transformer V2, a hierarchical vision transformer architecture that leverages shifted window self-attention to model both local and global spatial relationships efficiently.

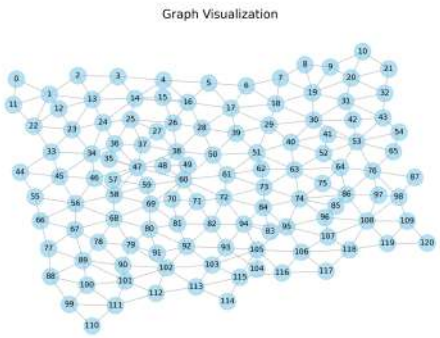

FIGURE 5. Final abstract graph structure. Nodes represent superpixel-based lesion regions, while edges denote both spatial and semantic connections.

The feature extraction process begins by partitioning the input image into non-overlapping patches of size $P \times P$. Each patch is then linearly projected into a high-dimensional embedding space, yielding an initial sequence of patch tokens:

$$X = f_{\text{patch}}(I) \in \mathbb{R}^{N\times d}, \tag{7}$$

where $I \in \mathbb{R}^{H\times W\times C}$ is the input image of height H, width W, and channels C; $N = \frac{H}{P} \cdot \frac{W}{P}$ is the total number of patches; and d is the feature embedding dimension. Unlike CNNs that operate on the entire image using kernels, the Swin Transformer performs hierarchical feature extraction through multiple transformer blocks, with each block refining its input features. The transformation at each layer l is given by:

$$X^l = f_{\text{Swin}}(X^{l-1}), \tag{8}$$

where X^l is the output of the l^{th} Swin Transformer block and f_{Swin} denotes the learned transformation function.

Standard transformers incur quadratic computational complexity $\mathcal{O}(N^2)$ due to global self-attention, which becomes infeasible for high-resolution medical images. Swin Transformer V2 addresses this bottleneck by applying Shifted Window Multi-Head Self-Attention (SW-MSA), wherein self-attention is confined to non-overlapping windows of size $M \times M$. The attention mechanism within each window is computed as:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{9}$$

where $Q,K,V \in \mathbb{R}^{M^2\times d}$ represent the query, key, and value matrices, and B is a relative position bias matrix. To enable

cross-window information flow and capture long-range dependencies, alternating layers apply a shifted window strategy, which shifts the partitioning by a fixed offset, ensuring global context aggregation without significantly increasing computational cost.

After feature extraction, the resulting token embeddings $F_{\text{trans}} \in \mathbb{R}^{N\times d}$ are reshaped into a 2D grid format suitable for spatial alignment:

$$F_{\text{grid}} = \text{reshape}(F_{\text{trans}}) \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times d}. \tag{10}$$

For instance, an input image of size 573×573 with a patch size of 4×4 yields a feature grid of 143×143, where each grid location is associated with a 768-dimensional embedding. However, to ensure compatibility with the superpixel map obtained during segmentation—which resides in the original image resolution—these features are upsampled via bilinear interpolation:

$$F_{\text{upsampled}} = \text{BilinearInterpolate}(F_{\text{grid}}, H, W) \in \mathbb{R}^{H\times W\times d}. \tag{11}$$

This interpolation ensures that each pixel in the original image domain has an associated transformer-derived feature vector, preserving spatial consistency and aligning the learned representations with the lesion's anatomical boundaries.

To derive graph node features from these dense pixel-wise embeddings, we aggregate the feature vectors within each superpixel. Let $\{S_1, S_2, ..., S_K\}$ denote the set of K superpixels. The feature representation for node k is obtained by averaging the upsampled features of all pixels contained in superpixel S_k:

$$F_k = \frac{1}{|S_k|}\sum_{(i,j)\in S_k} F_{\text{upsampled}}(i,j), \tag{12}$$

where $|S_k|$ is the number of pixels in superpixel S_k, and $F_{\text{upsampled}}(i,j) \in \mathbb{R}^d$ is the interpolated feature vector at pixel (i,j). This aggregation produces a robust, region-level descriptor that captures both the local structure and global context inherent in each lesion subregion.

The final node feature matrix $X \in \mathbb{R}^{K\times d}$ for the graph is assembled by stacking the superpixel-level feature vectors:

$$X = \{F_1, F_2, ..., F_K\}. \tag{13}$$

These transformer-enriched embeddings serve as the initial node features for downstream graph-based learning. By leveraging Swin Transformer V2, this approach ensures that each node encodes a rich representation of the lesion's visual and structural patterns, effectively bridging local detail with global semantics. This fusion of hierarchical attention, spatial alignment, and region-wise aggregation forms a robust foundation for subsequent GNN-based diagnostic inference.

C. MULTI-TASK LEARNING FRAMEWORK

To achieve comprehensive and interpretable dermatological analysis, we propose a multi-task learning (MTL) framework that concurrently performs three critical diagnostic tasks: disease classification, clinical attribute prediction, and color feature detection. This design enables the model to jointly learn from diverse supervision signals, thereby enhancing generalization and robustness. The architecture comprises a shared Graph Attention Network (GAT) backbone that extracts relational and spatially aware features from superpixel graphs, followed by three task-specific heads, each optimized for a distinct prediction objective.

1) Shared Backbone: Graph Attention Network (GAT)

The shared backbone of the proposed framework employs a multi-layer Graph Attention Network (GAT) to effectively capture both local interactions and global contextual dependencies across lesion regions in dermoscopic images. Each image is modeled as an attributed undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denotes the set of N nodes corresponding to superpixel-based regions of interest (ROIs); $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$ represents the edge set, capturing both spatial proximity and semantic similarity between ROIs; and $X \in \mathbb{R}^{N\times F}$ is the node feature matrix, with each node characterized by an F-dimensional feature vector derived from Swin Transformer embeddings. The adjacency matrix $\mathbf{A} \in \{0,1\}^{N\times N}$ encodes the underlying graph structure, where $\mathbf{A}_{ij} = 1$ indicates a connection between nodes v_i and v_j.

Within each GAT layer, the model applies a multi-head self-attention mechanism to compute task-adaptive node representations. For a given node pair (i,j), the unnormalized attention coefficient e_{ij} is computed as:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^T\left[\mathbf{W}h_i \| \mathbf{W}h_j\right]\right) \tag{14}$$

where $\mathbf{W} \in \mathbb{R}^{F'\times F}$ is a learnable weight matrix, $\mathbf{a} \in \mathbb{R}^{2F'}$ is the attention vector, h_i and h_j are the feature vectors of nodes i and j respectively, $\|$ denotes concatenation, and LeakyReLU is the leaky rectified linear unit activation function. These raw attention scores are then normalized across neighbors using the softmax function:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}(i)}\exp(e_{ik})} \tag{15}$$

where $\mathcal{N}(i)$ represents the neighborhood of node i. The final node embedding update is given by:

$$h_i' = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}\mathbf{W}h_j\right) \tag{16}$$

where σ is the ELU (Exponential Linear Unit) activation function. The backbone consists of multiple stacked GAT layers:

$$H^{(l+1)} = \sigma\left(\text{GAT}(H^{(l)}, \mathbf{A})\right) \tag{17}$$

with $H^{(0)} = X$ representing the input features. This shared backbone provides a rich representation that is then utilized by the task-specific heads.

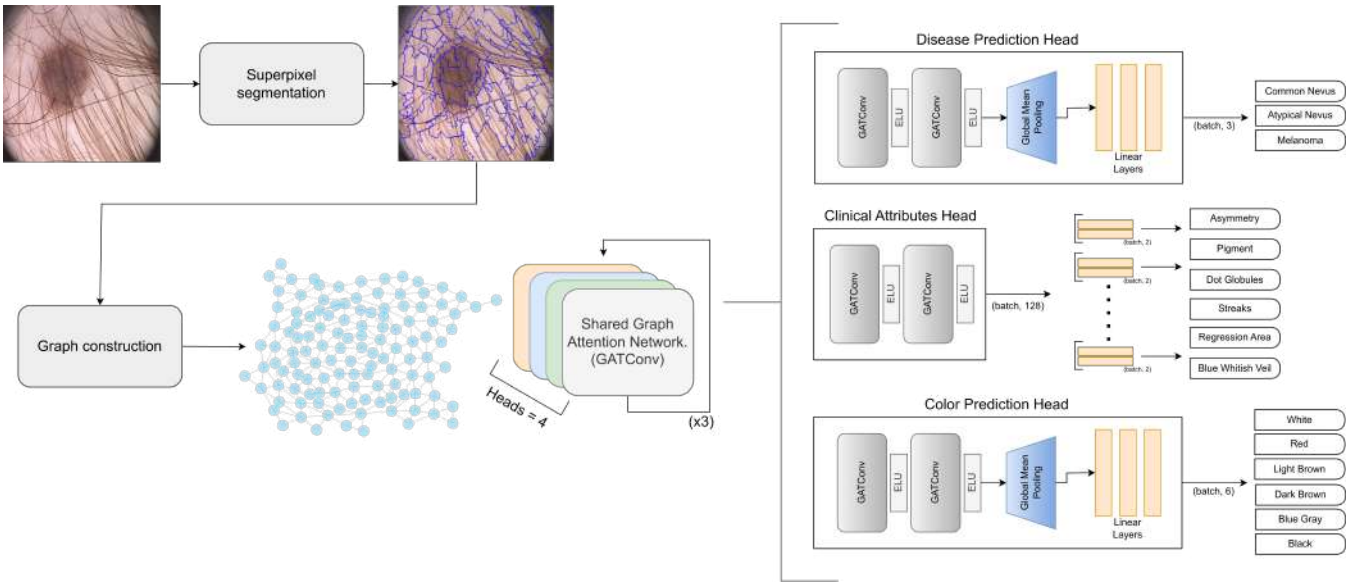

FIGURE 6. **Overall architecture of the proposed multi-task dermatological analysis framework.**

2) Task-Specific Prediction Heads

Following the shared Graph Attention Network (GAT) backbone, the model branches into three task-specific heads, each designed to optimize performance for a distinct dermatological analysis task: disease classification, clinical feature prediction, and color feature analysis. These heads operate on the high-dimensional graph representations learned by the backbone, refining them through additional processing steps tailored to each specific task. The general structure of each prediction head follows a systematic three-step approach: task-specific feature refinement, global pooling, and fully connected classification.

To enhance task-specific feature representations, additional GAT layers are employed, further refining the node embeddings obtained from the backbone. Given the adjacency matrix and the backbone output H_{backbone}, the refined node features are computed as:

$$Z = \sigma\left(\text{GAT}_2(\text{GAT}_1(H_{\text{backbone}}, \mathbf{A}), \mathbf{A})\right) \tag{18}$$

where Z represents the task-specific node embeddings, and σ is a non-linear activation function. Following this, global mean pooling is applied to aggregate node-level features into a single graph-level representation, ensuring invariance to the number of nodes in the input graph. The pooled feature vector is obtained as:

$$z = \frac{1}{N}\sum_{i=1}^{N} Z_i \tag{19}$$

where N is the number of nodes in the graph, and Z_i is the embedding of the i^{th} node. This graph-level representation serves as input to the final classification stage, where task-specific predictions are generated. The fully connected classification layer is formulated as:

$$y = f(z) = \sigma(W^3\sigma(W^2\sigma(W^1 z + b^1) + b^2) + b^3) \tag{20}$$

where W^1, W^2, W^3 and b^1, b^2, b^3 denote the trainable parameters of the fully connected layers. The choice of the activation function σ varies depending on the task: softmax activation is applied for mutually exclusive classifications such as disease categories and clinical features, while sigmoid activation is employed for multi-label classification in color prediction tasks.

a: Disease prediction

The disease classification head is designed to categorize the input dermoscopic image into one of C_d distinct skin disease classes. After extracting a task-specific feature representation, the fully connected layers transform the pooled graph embedding into a probability distribution over disease classes using the softmax activation function:

$$y_{\text{disease}} = \text{softmax}(f_{\text{disease}}(z)) \tag{21}$$

where $y_{\text{disease}} \in \mathbb{R}^{C_d}$ represents the predicted probability distribution over C_d disease categories. The classification head is trained using the categorical cross-entropy loss, ensuring that the model assigns a high probability to the correct disease class while minimizing the likelihood of incorrect predictions. The loss function is formulated as:

$$\mathcal{L}_{\text{disease}} = -\sum_{c=1}^{C_d} y_c \log(\hat{y}_c) \tag{22}$$

where $\hat{y}_c$ is the predicted probability of class c, and y_c is the corresponding ground truth label.

b: Morphological attribute prediction

Unlike disease classification, which is a single-task problem, the clinical feature prediction head simultaneously predicts

multiple lesion characteristics, each treated as an independent multi-class classification task. The clinical attributes include, for example, asymmetry ($y_{\text{asymmetry}} \in \mathbb{R}^3$), pigment pattern ($y_{\text{pigment}} \in \mathbb{R}^2$), dots and globules ($y_{\text{dots\_globules}} \in \mathbb{R}^3$), streaks ($y_{\text{streaks}} \in \mathbb{R}^2$), regression area ($y_{\text{regression\_area}} \in \mathbb{R}^2$), and blue-whitish veil ($y_{\text{blue\_whitish\_veil}} \in \mathbb{R}^2$), provide additional diagnostic insights. Given that each clinical attribute f has a distinct number of possible categories C_f, separate fully connected layers are used for each feature. The probability distribution for feature f is obtained as:

$$y_f = \text{softmax}(f_f(z)) \tag{23}$$

where $y_f \in \mathbb{R}^{C_f}$ represents the predicted class probabilities for feature f. Since clinical features are inherently independent, a multi-task optimization framework is employed, wherein each feature's prediction is optimized using an independent categorical cross-entropy loss. The overall loss for clinical feature prediction is computed as:

$$\mathcal{L}_{\text{clinical}} = \sum_f \lambda_f \mathcal{L}_{\text{feature}_f} \tag{24}$$

where λ_f is a weighting factor that adjusts the relative importance of each clinical feature prediction task. By jointly optimizing predictions for multiple lesion attributes, the model learns feature representations that generalize across various diagnostic aspects, improving robustness and interpretability.

c: Chromatic feature detection

The color feature prediction head is responsible for detecting the presence of six predefined colors in the dermoscopic image. Unlike the disease and clinical feature heads, which predict a single category per sample, the color classification task is formulated as a multi-label problem, where multiple colors can co-occur in a single lesion. To accommodate this, the final classification layer employs sigmoid activation rather than softmax, ensuring that each color label is assigned an independent probability:

$$y_{\text{color}} = \sigma(f_{\text{color}}(z)) \tag{25}$$

where $y_{\text{color}} \in \mathbb{R}^6$ represents the probability of the presence of each of the six color labels. Given the multi-label nature of this task, the binary cross-entropy loss is used for optimization, defined as:

$$\mathcal{L}_{\text{color}} = -\sum_{c=1}^{6} (y_c \log(\hat{y}_c) + (1 - y_c) \log(1 - \hat{y}_c)) \tag{26}$$

where y_c represents the true binary label indicating the presence (1) or absence (0) of color c, and $\hat{y}_c$ is the predicted probability. The use of binary cross-entropy allows the model to learn independent probability distributions for each color label, making it well-suited for multi-label classification tasks.

D. LOSS FUNCTION

To facilitate balanced optimization across heterogeneous tasks, we propose a composite loss function grounded in focal loss, tailored individually for disease classification, clinical feature prediction, and color analysis. This approach is particularly well-suited for imbalanced datasets, as it emphasizes hard-to-classify instances while mitigating the dominance of majority classes.

The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_d \mathcal{L}_{\text{disease}} + \lambda_c \mathcal{L}_{\text{clinical}} + \lambda_{col} \mathcal{L}_{\text{color}}, \tag{27}$$

where λ_d, λ_c, and λ_{col} are task-specific weighting coefficients that can either be fixed or dynamically learned during training to balance contributions from each task.

For both disease classification and clinical feature prediction, we utilize the class-weighted focal loss, defined as:

$$\mathcal{L}_{\text{focal}} = -\sum_{c=1}^{C} \alpha_c (1 - \hat{y}_c)^{\gamma} y_c \log(\hat{y}_c), \tag{28}$$

where y_c is the ground truth label, $\hat{y}_c$ denotes the predicted probability for class c, α_c is the class-specific weighting factor derived from inverse class frequency normalization, and γ is the focusing parameter that controls the degree of down-weighting for well-classified instances.

In the clinical feature prediction subtask, individual focal losses are computed separately for each feature dimension (e.g., asymmetry, pigment, dots/globules, etc.), each with its own tailored α_c vector. These are then aggregated as:

$$\mathcal{L}_{\text{clinical}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\text{focal}}^{(k)}, \tag{29}$$

where K denotes the number of clinical attributes.

For color pattern analysis, a focal binary cross-entropy loss is employed for each label in the multi-label classification setup:

$$R_{output} = A \cdot (B + C)^2 + \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right) \cdot \ln(P_i) - \int_0^{\infty} e^{-t^2} dt \cdot \cos(\omega t) + K \tag{30}$$

where M is the number of color categories, and α_m is the class-specific weight. By integrating class-aware focal loss functions into the multi-task framework, our model not only handles inter-task trade-offs effectively but also exhibits improved generalization in the presence of severe class imbalance, particularly in underrepresented clinical and color categories.

IV. RESULTS

A. EXPERIMENTAL SETUP AND GRAPH-BASED OPTIMIZATION FRAMEWORK

All experiments were implemented using the PyTorch framework and executed on an NVIDIA Tesla P100 GPU. The

dataset was initially partitioned into 80% training and 20% testing subsets. The model was trained using the Adam optimizer with an initial learning rate of 1×10^{-3}, which decayed exponentially across folds as 0.9^{fold}, in combination with a cosine annealing scheduler to promote smooth convergence and minimize the likelihood of local minima entrapment. Training proceeded for a maximum of 40 epochs per fold with a batch size of 32. Early stopping was employed with a patience threshold of 5 epochs and a minimum performance improvement criterion of 1×10^{-4}. To combat class imbalance—prevalent in multi-label dermatological datasets—focal loss was adopted with a focusing parameter $\gamma = 2$, thereby directing learning attention towards harder-to-classify and underrepresented classes. The architectural foundation of our model is a multi-task learning (MTL) framework designed to simultaneously perform disease classification, morphological attribute prediction, and chromatic feature detection. Each task was assigned an equal weight to ensure a balanced optimization landscape and uniform gradient flow during backpropagation. This approach encourages convergence across both global diagnostic tasks and localized dermatological attributes, ensuring the model's holistic proficiency.

To model spatial and topological dependencies effectively, we employed graph-based representations of dermoscopic images. Specifically, superpixel graphs were constructed using the SEEDS algorithm, which adaptively partitions each image into a variable number of perceptually homogeneous segments—ranging approximately from 40 to a maximum of 120 superpixels depending on image complexity. Each superpixel served as a graph node, and edges were defined based on spatial adjacency. Node features were extracted using a SWINv2 backbone, ensuring high-resolution, semantically rich embeddings that preserve fine-grained textural and structural cues.

To optimize the architectural and training hyperparameters, we employed Optuna—a Bayesian optimization framework known for its efficiency in high-dimensional search spaces. The objective function was defined to maximize the weighted F1-score on the validation fold, thereby aligning the tuning process with the evaluation metric most appropriate for our imbalanced classification scenario. The search space comprised both continuous and discrete hyperparameters: the learning rate was sampled from a log-uniform distribution; dropout rates were drawn from a uniform distribution over [0.1, 0.5]; and discrete architectural choices included the number of attention heads (2, 4, or 8) and hidden dimensionalities (32, 64, 128, or 256). After extensive tuning, the optimal configuration was determined to consist of three stacked graph attention layers, each comprising 32 hidden units and 4 attention heads. Task-specific heads were designed with 16 hidden units, and a dropout rate of 0.2 was identified as optimal. This configuration was consistently applied across all experiments to ensure fairness, comparability, and interpretability of the results.

This unified setup—encompassing stratified validation, graph-based spatial modeling, and automated hyperparameter tuning—underpins the reliability and generalizability of the proposed framework across diverse dermatological diagnostic tasks.

TABLE 1. Performance Metrics for Multi-Task Learning on PH2 Dataset

Task	F1-Score	Precision	Recall
Disease	1.000	1.000	1.000
Asymmetry	0.778	0.789	0.794
Pigment Network	0.987	0.988	0.988
Dots & Globules	0.799	0.799	0.800
Streaks	0.826	0.840	0.863
Regression Area	0.950	0.950	0.950
Blue-Whitish Veil	0.929	0.929	0.931
Color Features	0.895	0.897	0.894

B. EVALUATION METRICS AND QUANTITATIVE RESULTS

To comprehensively evaluate the model's performance, several metrics were employed: Accuracy, Precision, Recall, weighted F1-score, Confusion Matrix, ROC-AUC, and Average Precision. Among these, the weighted F1-score was designated as the primary metric due to its sensitivity to class imbalance and its balanced reflection of both precision and recall.

Table 1 summarizes the performance across the principal diagnostic tasks. The model attained a perfect score on disease classification ($F_1 = 1.000$), and near-perfect performance on pigment network, regression area, and blue-whitish veil, reflecting the robustness of the proposed framework across both global and fine-grained dermatological labels.

The aggregate F1-score for color feature detection was 0.895, indicating the model's proficiency in recognizing chromatic traits such as red, dark brown, and light brown—features of clinical significance in melanoma diagnosis. Similarly, morphological patterns like streaks and dots/globules demonstrated high discriminability, underscoring the efficacy of spatially structured graph modeling.

Collectively, the proposed multi-task graph-based architecture exhibited outstanding performance across all dermatological subtasks, validating its potential for real-world clinical deployment and intelligent dermatological decision support.

C. QUALITATIVE ANALYSIS AND INTERPRETABILITY

A salient advantage of Graph Neural Networks (GNNs) lies in their inherent structural interpretability—where information propagation occurs over semantically meaningful graph topologies composed of nodes (e.g., superpixels) and edges (e.g., spatial adjacency). In contrast to conventional convolutional architectures, GNNs offer an intuitive and spatially grounded framework for understanding model behavior, particularly critical in sensitive domains such as medical diagnosis. To bolster the explainability of our multi-task model and foster clinical trust, we employed two complementary interpretability techniques: GNNExplainer and graph attention visualization. GNNExplainer provides a principled approach to identify a compact subgraph and feature subset that most

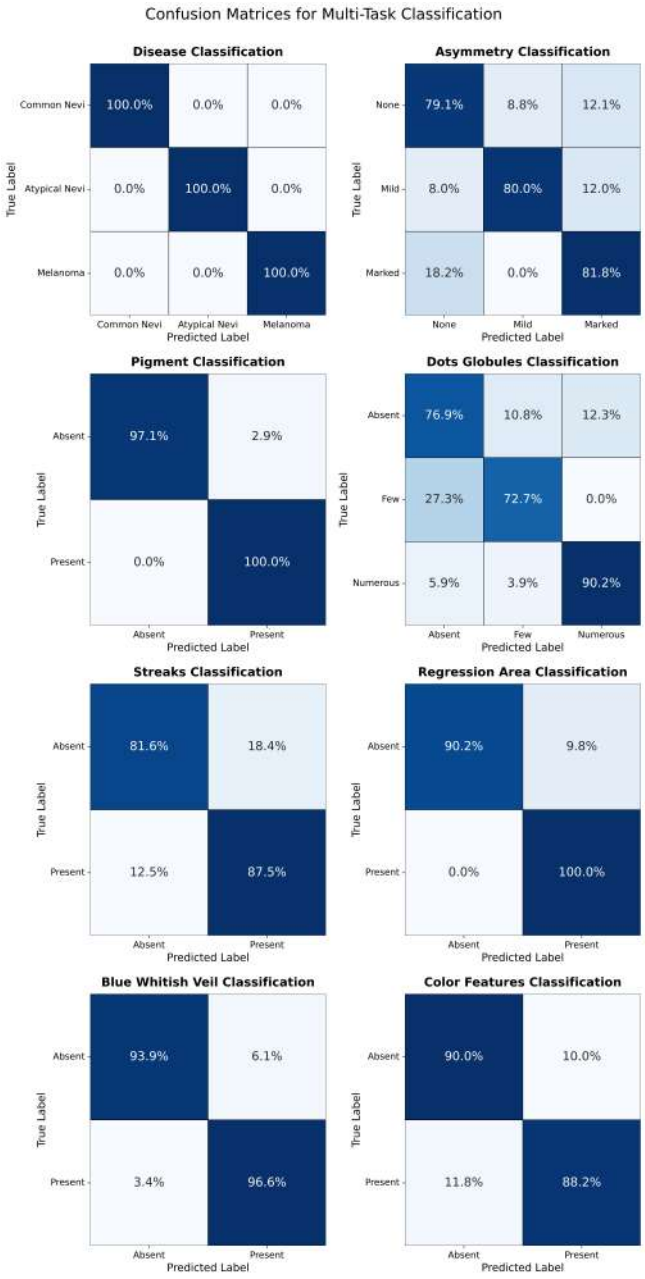

FIGURE 7. Combined confusion matrices for all diagnostic tasks, including Disease Classification, Asymmetry, Pigment Network, Dots & Globules, Streaks, Regression Area, Blue-Whitish Veil, and Color Feature Detection.

strongly influence a given prediction. Concurrently, attention score visualization—derived from the graph attention mechanism—highlights salient regions by attributing higher weights to influential nodes during message passing.

Figure 8 illustrates qualitative examples of node-level attention overlays for representative cases from each disease class: Common Nevus, Atypical Nevus, and Melanoma. The attention heatmaps reveal that the model consistently allocates higher attention to superpixels corresponding to clinically relevant features such as asymmetrical borders, irregular pigmentation, and local color variegation. This alignment with dermatological heuristics substantiates the clinical validity of the learned representations.

These qualitative insights not only corroborate the model's quantitative performance but also enhance its transparency—enabling end-users, including dermatologists and clinical researchers, to intuitively verify and trust the system's diagnostic reasoning. By grounding predictions in visibly interpretable substructures, the proposed graph-based paradigm aligns with the ethical imperative of explainability in AI-driven healthcare.

V. CONCLUSION

Our contribution is a novel multi-task graph-based learning framework for end-to-end dermatological image analysis, such as disease classification, morphological feature detection, and chromatic feature recognition in one framework. By representing dermoscopic images as graphs of superpixels and employing transformer-based node embeddings, our approach can effectively preserve strong spatial structure and semantic richness—substantial demands on precise medical image evaluation. Empirical results on the PH2 dataset confirm the efficacy of the new architecture, achieving state-of-the-art performance on all diagnostic subtasks. The model had a perfect F1-score for disease classification and showed high accuracy in detection of clinically significant morphological and chromatic features consistently. The use of focal loss with stratified validation protocols ensured label imbalance robustness, and Bayesian hyperparameter tuning with Optuna ensured better generalizability and reproducibility of experiments. Most importantly, the framework also focuses on interpretability—a highest priority clinical AI need. Through synergistic combination of the application of both GNNExplainer and attention-based visualization, we made the model's reasoning process lucid at a granular, subgraph-level. Attention heatmaps produced were shown to possess high concordance with dermatological heuristics and, as such, enhance clinical trust and enable human-AI collaboration in diagnostic decision-making. Future research on this work involves incorporating multi-modal inputs (e.g., patient metadata, clinical history) to further enhance context-aware inference, as well as testing the framework on varied datasets and real-world deployment settings. Most importantly, this work demonstrates the disruption potential of interpretable graph-based learning to augment intelligent, explainable, and clinically meaningful skin disease diagnosis.

ETHICS STATEMENT

This study did not involve any experiments on human participants or animals conducted by the authors. All data used were obtained from the publicly available PH^2 dataset, which comprises fully anonymized dermoscopic images intended for academic research. Therefore, ethical approval and informed consent were not required. The study was conducted in accordance with all relevant institutional guidelines and ethical standards.

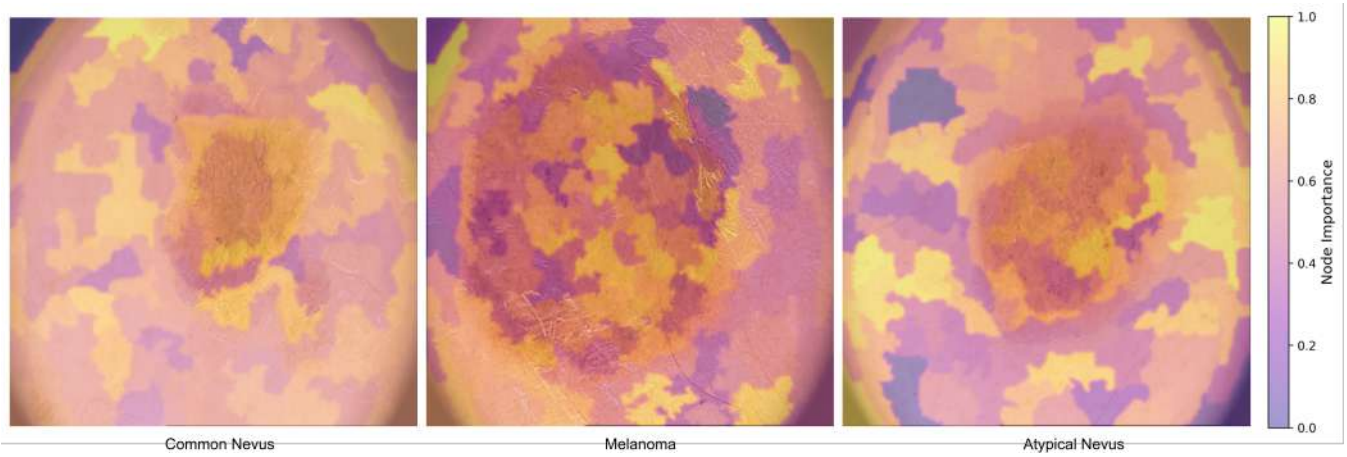

FIGURE 8. **Node-level attention visualization across disease categories. High-attention regions (brighter nodes) correspond to diagnostically significant areas such as pigment asymmetry, streaks, or regression structures. Categories: (a) Common Nevus, (b) Melanoma, (c) Atypical Nevus.**

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] M. Arshad, M. A. Khan, U. Tariq, A. Armghan, F. Alenezi, M. Younus Javed, S. M. Aslam, and S. Kadry, "A computer-aided diagnosis system using deep learning for multiclass skin lesion classification," *Computational intelligence and neuroscience*, vol. 2021, no. 1, p. 9619079, 2021.

[2] L. Haofu and J. Luo, "A deep multi-task learning approach to skin lesion classification," in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[3] P. Mirunalini, A. Chandrabose, V. Gokul, and S. Jaisakthi, "Deep learning for skin lesion classification," *arXiv preprint arXiv:1703.04364*, 2017.

[4] Z. Rahman, M. S. Hossain, M. R. Islam, M. M. Hasan, and R. A. Hridhee, "An approach for multiclass skin lesion classification based on ensemble learning," *Informatics in Medicine Unlocked*, vol. 25, p. 100659, 2021.

[5] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, "Combining deep learning and hand-crafted features for skin lesion classification," in *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2016, pp. 1–6.

[6] J. Mohan, A. Sivasubramanian, V. Ravi *et al.*, "Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable ai," *Computers in Biology and Medicine*, vol. 190, p. 110007, 2025.

[7] A. R. Lopez, X. Giro-i Nieto, J. Burdick, and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," in *2017 13th IASTED international conference on biomedical engineering (BioMed)*. IEEE, 2017, pp. 49–54.

[8] B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhattacharya, and K. Lakshmanna, "Skin lesion classification of dermoscopic images using machine learning and convolutional neural network," *Scientific Reports*, vol. 12, no. 1, p. 18134, 2022.

[9] I. A. Ozkan and M. Koklu, "Skin lesion classification using machine learning algorithms," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 5, no. 4, pp. 285–289, 2017.

[10] A. Bibi, M. A. Khan, M. Y. Javed, U. Tariq, B.-G. Kang, Y. Nam, R. R. Mostafa, and R. H. Sakr, "Skin lesion segmentation and classification using conventional and deep learning based framework," *Comput. Mater. Contin*, vol. 71, no. 2, pp. 2477–2495, 2022.

[11] M. Goyal, A. Oakley, P. Bansal, D. Dancey, and M. H. Yap, "Skin lesion segmentation in dermoscopic images with ensemble deep learning methods," *Ieee Access*, vol. 8, pp. 4171–4181, 2019.

[12] D. Adla, G. V. R. Reddy, P. Nayak, and G. Karuna, "A full-resolution convolutional network with a dynamic graph cut algorithm for skin cancer classification and detection," *Healthcare Analytics*, vol. 3, no. 100,154, 2023.

[13] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Experimental dermatology*, vol. 27, no. 11, pp. 1261–1267, 2018.

[14] X. He, B. Lei, and T. Wang, "Sanet: Superpixel attention network for skin lesion attributes detection. arxiv 2019," *arXiv preprint arXiv:1910.08995*.

[15] D. Ahmedt-Aristizabal, M. A. Armin, S. Denman, C. Fookes, and L. Petersson, "Graph-based deep learning for medical diagnosis and analysis: past, present and future," *Sensors*, vol. 21, no. 14, p. 4758, 2021.

[16] K. P. Santoso, R. V. H. Ginardi, R. A. Sastrowardoyo, and F. A. Madany, "Leveraging spatial and semantic feature extraction for skin cancer diagnosis with capsule networks and graph neural networks," *arXiv preprint arXiv:2403.12009*, 2024.

[17] V. Balaji, S. Suganthi, R. Rajadevi, V. K. Kumar, B. S. Balaji, and S. Pandiyan, "Skin disease detection and segmentation using dynamic graph cut algorithm and classification through naive bayes classifier," *Measurement*, vol. 163, p. 107922, 2020.

[18] S. Moldovanu, M. Miron, C.-G. Rusu, K. C. Biswas, and L. Moraru, "Refining skin lesions classification performance using geometric features of superpixels," *Scientific Reports*, vol. 13, no. 1, p. 11463, 2023.

[19] J. Deepa and P. Madhavan, "Optimized dynamic graph-based framework for skin lesion classification in dermoscopic images." *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 2, 2025.

[20] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.

[21] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: a comprehensive review," *Computational Social Networks*, vol. 6, no. 1, pp. 1–23, 2019.

[22] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy, "Machine learning and deep learning methods for skin lesion classification and diagnosis: a systematic review," *Diagnostics*, vol. 11, no. 8, p. 1390, 2021.

[23] J. Zhang, F. Zhong, K. He, M. Ji, S. Li, and C. Li, "Recent advancements and perspectives in the diagnosis of skin diseases using machine learning and deep learning: A review," *Diagnostics*, vol. 13, no. 23, p. 3506, 2023.

[24] H. Li, Y. Pan, J. Zhao, and L. Zhang, "Skin disease diagnosis with deep learning: A review," *Neurocomputing*, vol. 464, pp. 364–393, 2021.

[25] K. Ding, M. Zhou, Z. Wang, Q. Liu, C. W. Arnold, S. Zhang, and D. N. Metaxas, "Graph convolutional networks for multi-modality medical imaging: Methods, architectures, and clinical applications," *arXiv preprint arXiv:2202.08916*, 2022.

[26] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking,"

in *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2013, pp. 5437–5440.
[27] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016.
[28] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

FIRST A. AUTHOR received the B.Tech. degree in Computer Science and Engineering with a specialization in Artificial Intelligence from Vellore Institute of Technology (VIT), India. His research interests lie in Artificial Intelligence, Machine Learning, Deep Learning, and Large Language Models (LLMs), with a particular focus on intelligent systems and generative architectures. He is currently a prospective Master's student, aspiring to pursue advanced research in Artificial Intelligence and Machine Learning. His academic and research pursuits are driven by a deep commitment to developing transparent, efficient, and ethically aligned AI technologies.

SECOND A. AUTHOR received the B.E. degree in mechanical engineering from IIEST, Shibpur, in 2008, the M.E. degree in mechanical engineering from Jadavpur University, in 2011, and the Ph.D. degree in rehabilitation robotics from IIT Kharagpur, India, in 2021. He is currently an Assistant Professor with the School of Mechanical Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India, for the past two years. He has one year of industry experience after completing of B.E. degree at DCPL. He has two and a half years of teaching experience and one and a half years of postdoctoral research experience.

THIRD C. AUTHOR is a distinguished data scientist and researcher with extensive experience in Gen AI, Large Language Models, Optimization, Artificial Intelligence, and Machine Learning. Currently, he serves as the AI/ML Engineering Manager at Accenture, India. He has also held the position of Computer Research Scientist and Adjunct Faculty in the Department of Computer Science at the University of South Alabama, USA, and has worked as an Operations Research Scientist with Boeing Research & Technology. His research spans engineering optimization, evolutionary computation, machine learning, neural networks, and robotics. Dr. Datta has published over 80 papers in international SCI journals, book chapters, and conferences. He has edited two books with Springer, including one in the Infosys Science Foundation Series. His work has garnered over 1250 citations, reflecting his significant impact on the field. He has been invited to deliver lectures in several institutes and universities across the globe, including the University of British Columbia, Canada, Ryerson University, Canada, Trinity College Dublin (TCD), Ireland, Delft University of Technology (TUDELFT), the Netherlands, University of Western Australia (UWA), University of Minho, Portugal, University of Nova de Lisboa, Portugal, University of Coimbra, Portugal, Korea Railroad Research Institute (KRRI), Korea, Korea University and IIT Kanpur, Dr. Datta has been an invited speaker at numerous prestigious conferences and workshops globally. He holds a Ph.D. in Mechanical Engineering from the Indian Institute of Technology (IIT) Kanpur, where he specialized in evolutionary-penalty approaches for constrained optimization. In his professional career, Dr. Datta has secured several research grants, including projects funded by the US NSF IUCRC and the Korea Railroad Research Institute. He is actively involved in various professional organizations, including ACM SIGEVO and the IEEE Computational Intelligence Society, and has served as a reviewer for multiple high-impact journals.

FOURTH D. AUTHOR is a graduate of B.Tech. in Computer Science Engineering from Vellore Institute of Technology, Chennai. She has hands-on experience in various projects, including BharatVoice, a novel AI-driven Lok Sabha sessions querying system. Currently, she works as a Junior Associate at Synechron Technologies, focusing on enterprise application development, and has interned at the Centre for Development of Advanced Computing where she collaborated with a senior scientist to enhance a medical X-ray analysis tool aimed at improving diagnostic accuracy of respiratory illnesses. Her technical skills encompass programming languages like Java and Python, AI/ML full-stack development, data science, cloud technologies, and DevOps tools.

FIFTH E. AUTHOR received the B.Tech. degree in Computer science and Engineering with specialization in AI and Robotics from Vellore Institute of Technology, Chennai, in 2025. He has worked on various projects, including an obstacle detection robot using ROS and the application of quantum computing in autonomous vehicles, where he used the QML technique. His research interests include autonomous vehicles, robotics, and quantum machine learning applications. Currently working on Infosys private limited where I'm indulging in Generative AI and data science projects.