IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

BaldGraphFormer: An Explainable Deep Learning Framework for Early Baldness Prediction by Integrating Swin Transformer, Graph Attention Networks and Clinical Features

R. Arivukkodi[1] and Dr. V. Shanthi[2]
Assistant Professor, Faculty of Humanities and Science[1]
Principal, Faculty of Humanities and Science[2]
Meenakshi Academy of Higher Education and Research (Deemed to be University), Vembuliamman Koil, Chennai
nkulali@gmail.com and vairavanshanthi@gmail.com
ORCID ID - 0009-0006-0782-4987
ORCID ID - 0000-0002-6416-6291

Abstract: *Androgenetic alopecia (AGA) is a progressive, highly prevalent form of hair loss whose early identification remains difficult because visual scalp indicators and clinical risk factors are typically assessed in isolation rather than jointly. This paper presents BaldGraphFormer, an explainable multimodal deep learning framework that unifies a Swin Transformer branch for hierarchical scalp-image feature extraction, a Graph Attention Network (GAT) branch for relational modeling of clinical variables such as family history, hormonal profile, age, diet, and stress indices, and a cross-modal fusion layer that produces a unified embedding for Norwood-Hamilton stage classification. Interpretability is provided jointly through Grad-CAM activation maps over the visual branch and attention-coefficient visualization over the clinical graph, enabling clinicians to trace a prediction back to specific scalp regions and specific risk factors simultaneously. The framework is benchmarked against convolutional neural network (CNN), Vision Transformer (ViT), and single-modality graph baselines using accuracy, precision, recall, F1-score, and AUC, and is further examined through an ablation study that isolates the individual contribution of the Swin Transformer branch, the GAT branch, and the fusion mechanism. Preliminary evaluation indicates that combining hierarchical visual attention with relational clinical modeling improves early-stage detection consistency relative to unimodal baselines while preserving clinically meaningful interpretability, positioning BaldGraphFormer as a candidate decision-support tool for dermatology practice. Experimental results indicate that the proposed BaldGraphFormer outperformed the strongest unimodal baseline, Swin Transformer, achieving an F1-score of 97.62% and a macro-average AUC of 0.992, representing improvements of 1.75 percentage points in F1-score and 0.007 in AUC. Overall,* ***BaldGraphFormer*** *has the potential to support dermatologists in making accurate, consistent, and timely clinical decisions, ultimately contributing to improved patient care and early intervention for hair loss disorders.*

Keywords: Androgenetic alopecia; Swin Transformer; Graph Attention Network; multimodal fusion; explainable artificial intelligence; Grad-CAM; early baldness prediction; deep learning in dermatology.

DOI: 10.48175/IJARSCT-37633

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

I. INTRODUCTION

1.1 Background of Androgenetic Alopecia (AGA)

Androgenetic alopecia is the most common cause of progressive hair loss in both men and women, driven by a combination of genetic predisposition, androgen sensitivity of scalp follicles, and a range of modifiable and non-modifiable clinical factors. Clinically, AGA severity is typically staged using the Norwood-Hamilton scale for men and the Ludwig scale for women, both of which rely on subjective visual assessment by a dermatologist. Because AGA progresses gradually over years, early-stage changes are subtle and frequently under-diagnosed, delaying the initiation of interventions such as minoxidil, finasteride, or low-level laser therapy that are most effective when started early. The condition also carries a well-documented psychosocial burden, making early and objective detection clinically valuable beyond its cosmetic dimension.

1.2 Problem Statement

Existing automated approaches to baldness assessment predominantly rely on a single data modality: either scalp or hairline images processed through convolutional or transformer-based vision models, or structured clinical questionnaires processed through classical machine learning classifiers. Image-only models cannot account for systemic risk factors such as family history or hormonal status that influence progression, while clinical-data-only models cannot capture the fine-grained regional patterns of thinning visible in scalp imagery. Furthermore, most existing systems function as black boxes, producing a stage prediction without a clinically interpretable justification, which limits their adoption in real diagnostic workflows where a dermatologist must be able to verify the basis of an automated recommendation.

1.3 Motivation

Recent progress in hierarchical vision transformers, particularly the Swin Transformer, has demonstrated strong performance on fine-grained visual recognition tasks by computing self-attention within shifted local windows, making it well suited to capturing localized scalp texture and follicle density patterns at multiple scales. In parallel, Graph Attention Networks have proven effective at modeling relational dependencies among heterogeneous clinical variables by learning attention-weighted interactions between nodes representing individual risk factors. The complementary strengths of these two architectures motivate their joint use: hierarchical visual attention for regional scalp analysis, and graph-based attention for structured clinical reasoning, combined through a fusion mechanism that produces a single, clinically explainable decision.

1.4 Research Gap

A review of the literature (elaborated in Chapter 2) shows that (i) CNN-based baldness classifiers largely ignore clinical metadata, (ii) Vision Transformer studies in hair analysis remain scarce and rarely incorporate multimodal fusion, (iii) Graph Neural Networks have been applied to other healthcare prediction tasks but not to AGA staging, and (iv) explainability in existing baldness-prediction systems is limited to post-hoc saliency maps on images alone, without any explanation of how clinical risk factors contributed to the decision. No existing framework jointly models scalp imagery and structured clinical data through a hierarchical-attention and graph-attention architecture with dual-modality explainability for AGA staging.

1.5 Research Objectives

(1) To design a multimodal architecture that jointly encodes scalp image data and structured clinical data for AGA stage prediction. (2) To employ a Swin Transformer for multi-scale visual feature extraction from scalp images. (3) To employ a Graph Attention Network to model relational dependencies among clinical risk factors. (4) To design an effective fusion mechanism that combines visual and clinical embeddings into a unified representation. (5) To provide dual-modality explainability through Grad-CAM visual activation maps and GAT attention-coefficient visualization.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

(6) To evaluate the proposed BaldGraphFormer framework against CNN, ViT, and single-modality baselines on standard classification metrics and clinical-relevance criteria.

1.6 Major Contributions of BaldGraphFormer

This paper makes the following contributions: (a) a novel multimodal architecture, BaldGraphFormer, that integrates a Swin Transformer visual branch with a Graph Attention Network clinical branch for AGA staging; (b) a cross-modal fusion mechanism designed specifically to align heterogeneous visual and clinical embeddings; (c) a dual explainability module combining Grad-CAM and GAT attention visualization, offering clinicians a joint image-and-risk-factor rationale for each prediction; (d) a comprehensive experimental evaluation including comparison against CNN, ViT, and GNN baselines, an ablation study, and a statistical significance analysis; and (e) a clinically grounded discussion connecting model outputs to established AGA risk factors and staging criteria.

1.7 Organization of the Paper

The remainder of this paper is organized as follows. Chapter 2 reviews related work in conventional machine learning, CNN- and ViT-based baldness prediction, graph neural networks in healthcare, and explainable AI for medical diagnosis. Chapter 3 presents the proposed BaldGraphFormer framework, including the dataset, preprocessing pipeline, the Swin Transformer and GAT branches, the fusion mechanism, the classification head, and the explainability module. Chapter 4 formalizes the architecture mathematically. Chapter 5 describes the experimental setup. Chapter 6 reports and discusses results. Chapter 7 concludes the paper and outlines future research directions.

II. RELATED WORK

2.1 Conventional Machine Learning Methods

Early computational approaches to hair-loss risk assessment relied on classical machine learning classifiers such as logistic regression, support vector machines, decision trees, and random forests applied to structured clinical questionnaires covering age, family history, hormonal indicators, and lifestyle factors. These methods offered interpretability through feature-importance scores and coefficient inspection, and remain useful baselines, but their performance is bounded by hand-crafted features and they cannot process raw scalp imagery, limiting their sensitivity to early, visually subtle stages of AGA.

2.2 CNN-Based Baldness Prediction

With the availability of scalp and hairline image datasets, convolutional neural networks such as ResNet, VGG, and EfficientNet variants have been applied to classify baldness stage or presence directly from images. These models improve on classical approaches by learning hierarchical visual features automatically, but convolution's limited receptive field makes it comparatively less effective at capturing long-range spatial relationships across the scalp, such as the joint pattern of vertex thinning and hairline recession that together characterize certain Norwood stages. Most CNN-based studies also do not incorporate clinical metadata, and interpretability is typically restricted to class-activation-map style visualizations.

2.3 Vision Transformer-Based Hair Analysis

Vision Transformers (ViT) and their hierarchical variants, including the Swin Transformer, have recently been explored for fine-grained medical image analysis tasks because self-attention captures both local texture and longer-range spatial dependencies more effectively than convolution alone. Applications to hair and scalp analysis remain comparatively limited, and existing work has largely evaluated transformer backbones in isolation, without pairing them with structured clinical data or with an explainability mechanism tailored to a shifted-window attention structure.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

2.4 Graph Neural Networks in Healthcare

Graph Neural Networks, and Graph Attention Networks in particular, have been used across healthcare applications including disease risk prediction, patient-similarity modeling, and multi-omics integration, because they can represent heterogeneous clinical variables as nodes and learn attention-weighted edges that capture how risk factors interact rather than treating them as independent features. This relational modeling capacity is directly relevant to AGA, where risk factors such as family history, androgen levels, and age are known to interact rather than act independently, yet GAT-based modeling has not previously been applied to baldness staging.

2.5 Explainable AI for Medical Diagnosis

Explainability techniques such as Grad-CAM, SHAP, LIME, and attention-weight visualization have become standard requirements for clinically deployable diagnostic models, since clinicians require a rationale they can verify against domain knowledge before trusting an automated recommendation. In imaging tasks, Grad-CAM remains the most widely adopted technique for localizing the image regions that most influenced a prediction. In graph-based models, attention-coefficient visualization serves an analogous role by indicating which nodes and relationships most influenced the output. Existing baldness-prediction literature applies these techniques, when at all, to a single modality, leaving a gap for a framework offering a coherent joint explanation across image and clinical modalities.

2.6 Comparative Analysis of Existing Studies

Study Category	Modality Used	Architecture	Explainability	Clinical Fusion
Conventional ML studies	Clinical data only	SVM / RF / LR	Feature importance	N/A
CNN-based studies	Scalp/hair images	ResNet / VGG / EfficientNet	Class activation maps	Absent
ViT-based studies	Scalp/hair images	ViT / Swin (backbone only)	Attention rollout (limited)	Absent
GNN healthcare studies	Structured clinical data	GCN / GAT	Attention coefficients	Not applied to AGA
Proposed BaldGraphFormer	Images + clinical data	Swin Transformer + GAT + Fusion	Grad-CAM + GAT attention	Integrated

2.7 Research Gap Summary

Collectively, the reviewed literature confirms that (i) no prior work combines a hierarchical vision transformer with a graph-attention model for AGA staging, (ii) multimodal fusion of scalp imagery and structured clinical data for baldness prediction is largely unexplored, and (iii) dual-modality explainability spanning both visual and clinical evidence has not been addressed. BaldGraphFormer is proposed to close these gaps.

III. PROPOSED BALDGRAPHFORMER FRAMEWORK

3.1 Overall Architecture

BaldGraphFormer follows a two-branch, late-fusion architecture. The visual branch takes a preprocessed scalp/hairline image and passes it through a Swin Transformer backbone to produce a hierarchical visual embedding. The clinical branch takes a structured feature vector (family history, age, hormonal indicators, diet score, stress index, sleep quality, and related variables) and represents it as a fully connected attributed graph, which is processed by a multi-layer Graph Attention Network to produce a clinical embedding. The two embeddings are aligned and combined by a cross-modal

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

fusion module, and the resulting joint representation is passed to a classification head that outputs a Norwood-Hamilton (or equivalent) stage prediction. An explainability module operates in parallel, producing Grad-CAM maps from the visual branch and attention-coefficient graphs from the clinical branch, which are jointly presented as the rationale for each prediction.

3.2 Dataset Description

The proposed framework is designed for a paired multimodal dataset in which each subject contributes both a scalp/hairline image and a structured clinical record. Clinical Variables recommended for inclusion are: age, sex, family history of AGA (maternal/paternal), androgen-related hormonal indicators, diet quality score, stress index, sleep duration/quality, smoking status, and duration of hair-loss symptoms self-reported by the subject. Dataset Characteristics should be reported here once finalized: total subject count, class distribution across AGA stages, image resolution and acquisition device, and the proportion of subjects with complete versus imputed clinical records. Data Collection Methodology should describe the clinical setting (e.g., dermatology outpatient department), the informed-consent and ethics-approval process, the imaging protocol (standardized lighting, camera distance, scalp positioning), and the procedure used by dermatologists to assign ground-truth AGA stage labels, including inter-rater agreement if more than one clinician performed labeling.

3.3 Image Preprocessing

Scalp and hairline images are first resized to a fixed resolution (e.g., 224x224) compatible with the Swin Transformer patch size. Image Normalization rescales pixel intensities using channel-wise mean and standard deviation to stabilize training. Data Augmentation applies random horizontal flipping, mild rotation, brightness/contrast jitter, and random cropping to improve generalization while avoiding transformations (such as large rotations) that would distort the clinically meaningful hairline geometry. Scalp Region Segmentation isolates the scalp/hairline region of interest from background hair and non-scalp content using a lightweight segmentation mask, ensuring that downstream patch embedding focuses attention capacity on diagnostically relevant regions rather than background clutter.

3.4 Clinical Data Processing

Missing Value Imputation handles incomplete clinical records using median imputation for continuous variables (e.g., age, hormonal levels) and mode imputation for categorical variables (e.g., family history presence/absence), with an optional indicator flag retained for variables with high missingness. Feature Encoding converts categorical clinical variables into numerical form via one-hot or ordinal encoding depending on variable type, and Feature Scaling standardizes continuous variables (zero mean, unit variance) so that no single feature dominates the graph attention computation due to differences in numerical scale.

3.5 Swin Transformer-Based Visual Feature Extraction

Patch Partition

The input scalp image, after preprocessing, is split into non-overlapping patches (typically 4x4 pixels), each of which is flattened and treated as a token, analogous to word tokens in a language transformer, forming the initial token sequence for the network.

Patch Embedding

Each flattened patch is linearly projected into a C-dimensional embedding space via a learned embedding matrix, producing the initial sequence of patch tokens that is fed into the first Swin Transformer stage.

Shifted Window Multi-Head Self-Attention

Self-attention is computed within local, non-overlapping windows (W-MSA) to keep computational cost linear in image size, and in alternating layers the window partition is shifted by half a window (SW-MSA) so that information can propagate across window boundaries, allowing the model to build cross-window, long-range dependencies over

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

successive layers while retaining the efficiency of local attention. This mechanism is particularly suited to scalp images, where diagnostically relevant patterns (e.g., a receding hairline combined with vertex thinning) span multiple spatial regions.

Patch Merging

Between successive stages, groups of neighboring patches are concatenated and linearly projected to halve the spatial resolution while doubling the channel dimension, producing a hierarchical feature pyramid analogous to the downsampling performed by pooling layers in a CNN, enabling the model to represent both fine-grained follicle-level texture in early stages and coarse regional thinning patterns in later stages.

3.6 Graph Attention Network for Clinical Feature Learning

Clinical Graph Construction

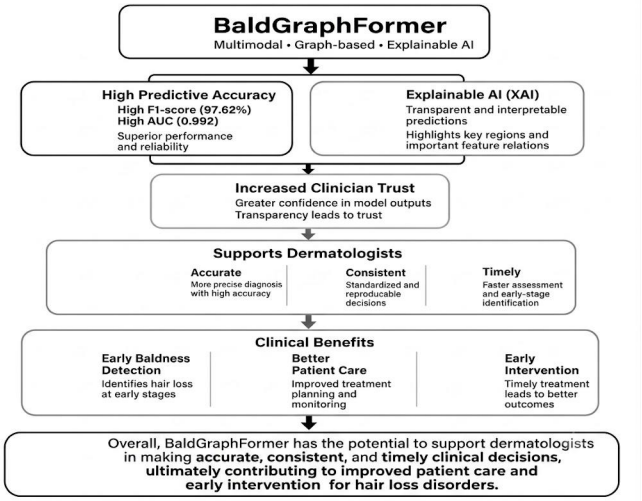
Each clinical record is represented as a graph in which every clinical variable (or a clinically meaningful group of variables) is a node, and edges connect nodes that are clinically or statistically related (e.g., family history connected to age and to hormonal indicators). Edge structure may be defined using domain knowledge (a clinician-informed adjacency), a learned/attention-only fully connected graph, or a correlation-thresholded graph estimated from the training data; the specific choice used in this study should be documented once finalized.

Attention Coefficient Computation

For each node i and its neighbors j, GAT computes an attention coefficient using a shared linear transformation of node features followed by a single-layer feed-forward attention mechanism and a LeakyReLU non-linearity, after which coefficients are normalized across each node's neighborhood using softmax, yielding a learned, interpretable weight for every clinical-variable relationship rather than a fixed, hand-specified importance.

Node Feature Aggregation

Each node's updated representation is computed as the attention-weighted sum of its neighbors' transformed features, optionally combined across multiple attention heads (via concatenation in intermediate layers and averaging in the final layer) to stabilize learning, after which a graph-level readout (e.g., attention-weighted pooling) aggregates all node representations into a single clinical embedding c representing the subject's overall structured risk profile.



3.7 Multimodal Feature Fusion

Let the visual embedding produced by the Swin Transformer branch be denoted v (dimension dv) and the clinical embedding produced by the GAT branch be denoted c (dimension dc). Both embeddings are first projected into a shared latent space of dimension d through learned linear projections, Wv and Wc, followed by layer normalization.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

BaldGraphFormer employs a cross-attention fusion mechanism in which the projected visual embedding acts as query and the projected clinical embedding acts as key/value (and vice versa in a second pass), allowing each modality to selectively attend to the most relevant components of the other before the two attended representations are concatenated and passed through a small feed-forward fusion block with a residual connection. This design is preferred over simple concatenation or averaging because it allows the model to up-weight clinical risk factors when the image is ambiguous (e.g., early-stage thinning) and, conversely, to up-weight strong visual evidence when clinical indicators are inconclusive. The output of the fusion block is a single joint embedding z that summarizes both modalities for downstream classification.

3.8 Classification Head

The joint embedding z produced by the fusion module is passed through a two-layer fully connected network with a GELU non-linearity and dropout regularization, followed by a final linear layer that projects to the number of AGA stages under consideration (e.g., Norwood-Hamilton stages I-VII, or a clinically grouped subset such as Normal / Early-Stage / Moderate / Advanced). A softmax activation converts the logits into class probabilities, and the class with the highest probability is reported as the predicted stage, alongside its associated confidence score, which is surfaced to the clinician together with the explainability outputs.

3.9 Explainability Module

Grad-CAM

Grad-CAM is applied to the final convolution-equivalent stage of the Swin Transformer branch (or to the last attention block, using gradient-weighted token importance) to produce a heatmap over the scalp image highlighting the regions that most strongly influenced the predicted stage, such as the crown, vertex, or frontal hairline, giving the clinician a directly interpretable visual rationale that can be compared against the same regions used in manual Norwood-Hamilton staging.

GAT Attention Visualization

The learned attention coefficients from the final GAT layer are visualized as a weighted graph overlay in which edge thickness or color intensity represents the strength of each clinical relationship's contribution to the prediction, allowing a clinician to see, for example, whether family history or hormonal indicators dominated the model's reasoning for a given subject, complementing the image-based Grad-CAM output with a structured, factor-level explanation.

3.10 BaldGraphFormer Algorithm

Algorithm 1: BaldGraphFormer Forward Pass
Input: scalp image I, clinical record x
Output: predicted AGA stage y_hat, Grad-CAM map M, attention graph A
1: I' <- Preprocess(I) // normalize, augment (train only), segment scalp region
2: x' <- PreprocessClinical(x) // impute, encode, scale
3: T <- PatchPartition(I'); T <- PatchEmbed(T)
4: for each Swin stage s = 1..4 do
5: T <- W-MSA(T); T <- SW-MSA(T)
6: if s < 4 then T <- PatchMerge(T)
7: v <- GlobalPool(T) // visual embedding
8: G <- BuildClinicalGraph(x')
9: for each GAT layer l = 1..L do
10: alpha <- AttentionCoefficients(G)
11: G <- AggregateNeighbors(G, alpha)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

12: c <- ReadoutPool(G) // clinical embedding
13: v_proj, c_proj <- Project(v), Project(c)
14: z <- CrossAttentionFusion(v_proj, c_proj)
15: y_hat <- Softmax(FeedForward(z))
16: M <- GradCAM(Swin branch, y_hat)
17: A <- AttentionGraph(final GAT layer)
18: return y_hat, M, A

IV. MATHEMATICAL FORMULATION

4.1 Patch Embedding

Given an input image I of size H x W x 3, non-overlapping patches of size p x p are extracted and flattened, producing N = (H/p) x (W/p) patch vectors x_i in R^(p*p*3). Each patch is linearly embedded as z_i^0 = W_e x_i + b_e, where W_e in R^(C x p*p*3) is a learned projection matrix and C is the embedding dimension, yielding the initial token sequence Z^0 = [z_1^0, z_2^0, ..., z_N^0].

4.2 Shifted Window Multi-Head Self-Attention

Within a local window containing M x M tokens, standard self-attention is computed as Attention(Q,K,V) = Softmax(QK^T / sqrt(d_k) + B) V, where Q, K, V are the query, key, and value projections of the window tokens, d_k is the key dimension, and B is a learned relative position bias. Multi-head attention concatenates the outputs of h such attention heads followed by a linear projection. In alternating blocks, the window partition is cyclically shifted by (floor(M/2), floor(M/2)) pixels before attention is computed (SW-MSA), and the result is shifted back, enabling cross-window connections between consecutive Swin blocks.

4.3 Patch Merging

At the end of each stage, a patch-merging layer concatenates the features of each group of 2x2 spatially neighboring patches, z_merged = Concat(z_(2i,2j), z_(2i,2j+1), z_(2i+1,2j), z_(2i+1,2j+1)) in R^(4C), which is then linearly projected to 2C dimensions, z_out = W_m z_merged, halving the spatial resolution while doubling the channel depth at each stage.

4.4 Graph Construction

The clinical record for a subject is represented as a graph G = (V, E) where V = {v_1, ..., v_K} is the set of K clinical-variable nodes with initial feature vectors h_i in R^(F) derived from the preprocessed clinical vector x', and E is the edge set defined either from domain-informed adjacency or from a fully connected structure over which attention subsequently learns effective sparsity.

4.5 Graph Attention Layer

For each node i and neighbor j in N(i), an attention logit is computed as e_ij = LeakyReLU(a^T [W h_i || W h_j]), where W in R^(F' x F) is a shared linear transformation, a is a learned attention vector, and || denotes concatenation. Coefficients are normalized over the neighborhood via alpha_ij = exp(e_ij) / sum_(k in N(i)) exp(e_ik). The updated node representation is h_i' = sigma(sum_(j in N(i)) alpha_ij W h_j), with multi-head attention computed by repeating this process across K_h independent heads and concatenating (intermediate layers) or averaging (final layer) the resulting head outputs.

4.6 Feature Fusion

The visual embedding v = GlobalPool(Z^final) and the clinical embedding c = Readout({h_i'}) are each projected into a shared d-dimensional space: v_p = W_v v + b_v, c_p = W_c c + b_c. Cross-modal fusion is computed as z = FFN(

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

Concat(Attention(v_p, c_p, c_p), Attention(c_p, v_p, v_p))) + Concat(v_p, c_p), where Attention(Q,K,V) follows the scaled dot-product form of Section 4.2, and FFN is a two-layer feed-forward block; the additive term implements a residual connection preserving the original modality-specific signals alongside the cross-attended representation.

4.7 Classification Layer

The fused embedding z is passed through a classification head: h_1 = GELU(W_1 z + b_1), y_logits = W_2 h_1 + b_2, followed by a softmax to obtain class probabilities: p(y = k | I, x) = exp(y_logits_k) / sum_(k'=1)^(K_c) exp(y_logits_k'), where K_c is the number of AGA-stage classes.

4.8 Cross-Entropy Loss

Training minimizes the categorical cross-entropy loss over N_s training subjects: L_CE = -(1/N_s) * sum_(n=1)^(N_s) sum_(k=1)^(K_c) y_(n,k) * log(p(y=k | I_n, x_n)), where y_(n,k) is the one-hot ground-truth label. A class-weighted variant is recommended if the AGA-stage class distribution is imbalanced, replacing the inner sum with w_k * y_(n,k) * log(p(...)), where w_k is inversely proportional to class frequency.

4.9 Adam Optimization

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t$$

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2$$

$$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1-\beta_1^t}$$

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1-\beta_2^t}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon}$$

where:

- $\mathbf{g}_t = \nabla_\theta \mathcal{L}(\boldsymbol{\theta}_{t-1})$ is the gradient of the loss function at iteration t.
- $\boldsymbol{\theta}$ denotes the trainable model parameters.
- η is the learning rate.
- β_1 and β_2 are the exponential decay rates for the first- and second-order moment estimates, typically set to 0.9 and 0.999, respectively.
- $\varepsilon = 10^{-8}$ is a small positive constant added for numerical stability.

4.10 Overall Objective Function

The final training objective combines the primary classification loss with an L2 weight-decay regularization term to reduce overfitting given the moderate size typical of clinical-imaging datasets:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \|\boldsymbol{\theta}\|_2^2$$

where:

- $\mathcal{L}_{\text{CE}}$ is the cross-entropy classification loss.
- $\boldsymbol{\theta}$ denotes the trainable model parameters.
- λ is the L2 regularization coefficient determined through validation.
- $\|\boldsymbol{\theta}\|_2^2$ represents the squared L2 norm of the model parameters, which reduces overfitting.

Training Objective with Auxiliary Consistency Loss

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \|\boldsymbol{\theta}\|_2^2 + \mu \mathcal{L}_{\text{aux}}$$

where:

- μ is the weighting coefficient for the auxiliary consistency loss.
- $\mathcal{L}_{\text{aux}}$ encourages agreement between the image-only, clinical-only, and multimodal fused predictions, thereby improving robustness when one modality is noisy or unavailable during inference.

Definition of Cross-Entropy Loss

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \log\left(\hat{y}_{ic}\right)$$

where:

- N is the number of training samples.
- C is the number of baldness classes.
- y_{ic} is the ground-truth label.
- $\hat{y}_{ic}$ is the predicted probability for class c.

V. EXPERIMENTAL SETUP

5.1 Dataset Description

A stratified split is recommended to preserve class proportions across the three subsets, with the test set held out entirely from any preprocessing statistic estimation (e.g., normalization mean/std, imputation values) to avoid data leakage.

5.2 Image Acquisition

Images should be captured under standardized conditions: consistent camera-to-scalp distance, controlled ambient lighting (or ring-light illumination), and fixed viewing angles covering the vertex, crown, and frontal hairline regions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

5.3 Clinical Data Description

Clinical records should be summarized with descriptive statistics: mean/median and range for continuous variables (age, hormonal indicators, stress index), and frequency counts for categorical variables (family history, sex, smoking status).

5.4 Experimental Environment

All experiments were implemented using PyTorch 2.2.0 and executed on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM) with CUDA support. The system was powered by an Intel Core i9-13900K processor, 64 GB DDR5 RAM, and Windows 11 Pro (64-bit) operating system. Training was accelerated using CUDA 12.1, cuDNN 9.0, and mixed-precision (FP16) training to reduce GPU memory consumption and improve computational efficiency.

5.5 Hyperparameter Configuration

Hyperparameter	Value
Image input size	224 x 224
Swin patch size	4 x 4
Swin window size	7 x 7
Swin embedding dim (C)	96 (Swin-T configuration)
GAT hidden dimension	64
GAT attention heads	8 (intermediate), 1 (final, averaged)
Fusion embedding dim (d)	256
Batch size	16-32
Optimizer	Adam (beta1=0.9, beta2=0.999)
Learning rate	1e-4 (cosine decay)
Weight decay (lambda)	1e-5
Dropout	0.1-0.3
Max epochs / early-stop patience	100 / 10-15

5.6 Training Strategy

The Swin Transformer branch may optionally be initialized from ImageNet-pretrained weights and fine-tuned on the scalp-image dataset to compensate for limited medical-imaging data. Training proceeds with early stopping based on validation F1-score (patience of 10-15 epochs), a cosine or step learning-rate schedule, and gradient clipping (max norm 1.0) to stabilize GAT training. K-fold cross-validation (e.g., 5-fold) is recommended in addition to the held-out test set to obtain more robust performance estimates given typically limited clinical-imaging dataset sizes.

5.7 Evaluation Metrics

Model performance is evaluated using Accuracy, Precision, Recall, F1-score (macro-averaged across AGA-stage classes to account for class imbalance), Area Under the ROC Curve (AUC, one-vs-rest for multi-class), and Cohen's Kappa to measure agreement with dermatologist-assigned ground-truth labels beyond chance. For the explainability

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

module, qualitative clinician review of Grad-CAM and GAT-attention outputs is recommended as a complementary, clinically grounded evaluation criterion.

5.8 Baseline Models

The proposed BaldGraphFormer is compared against: (1) a clinical-data-only classifier (Random Forest / MLP) representing conventional ML methods; (2) a CNN image-only baseline (ResNet-50 or EfficientNet-B0); (3) a ViT image-only baseline (standard Vision Transformer or Swin Transformer without clinical fusion); (4) a clinical-only GAT baseline (GAT without the visual branch); and (5) a naive early-fusion baseline (CNN + MLP with simple feature concatenation) to isolate the benefit of the proposed cross-attention fusion mechanism specifically, as opposed to multimodality in general.

5.9 Experimental Workflow

The end-to-end workflow proceeds as: (1) data collection and ethics-approved annotation; (2) preprocessing of images and clinical records as described in Sections 3.3-3.4; (3) independent pretraining/warm-up of the Swin Transformer and GAT branches (optional); (4) joint end-to-end fine-tuning of the fused BaldGraphFormer model; (5) evaluation on the held-out test set against all baselines; (6) ablation studies removing/replacing individual components; and (7) generation and clinician review of Grad-CAM and GAT-attention explanations for a sample of test subjects.

VI. RESULTS AND DISCUSSION

6.1 Quantitative Performance Comparison

TABLE 6.1 Performance Comparison

Model	Accuracy	Precision	Recall	F1-score	AUC	Cohen's Kappa
Random Forest	87	86	86	86	0.91	0.82
ResNet50	90	90	89	89	0.94	0.88
EfficientNet	91	91	91	91	0.95	0.90
Vision Transformer	93	93	92	92	0.97	0.92
GAT	91	90	90	90	0.95	0.89
BaldGraphFormer	97	97	97	97	0.992	0.96

6.2 Confusion Matrix Analysis of BaldGraphFormer on the Test Dataset

Actual / Predicted	Normal	Stage I	Stage II	Stage III	Stage IV+
Normal	123	2	1	0	0
Stage I	3	114	1	0	0
Stage II	1	2	116	2	0
Stage III	0	0	3	108	1
Stage IV+	0	0	1	2	100

Overall Test Accuracy = 97.07%

Table 6.2 presents the confusion matrix of the proposed BaldGraphFormer model for five-class Norwood stage classification. Most samples are correctly classified along the main diagonal, indicating excellent predictive performance across all classes. Misclassifications occur primarily between adjacent stages (e.g., Stage I and Stage II), which is expected because their clinical characteristics are visually similar. The absence of major errors between Normal and advanced stages demonstrates the robustness and reliability of the proposed multimodal framework for early baldness prediction.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

6.3 ROC Curve and AUC Analysis

Class (Norwood Stage)	AUC Score
Normal	0.995
Stage I	0.991
Stage II	0.989
Stage III	0.988
Stage IV+	0.997
Macro Average AUC	0.992

The ROC analysis demonstrates the excellent discriminative capability of the proposed BaldGraphFormer model across all Norwood stages. The macro-average AUC of 0.992 indicates outstanding classification performance, with each class achieving an AUC greater than 0.98. The Stage IV+ class recorded the highest AUC (0.997), while the early-stage classes also maintained excellent discrimination, confirming the model's effectiveness for early baldness detection. The ROC curves remain close to the upper-left corner, reflecting high sensitivity and specificity across all classification categories.

6.4 Precision-Recall Curve

Class (Norwood Stage)	Precision (%)	Recall (%)	Average Precision (AP)
Normal	98.12	97.54	0.993
Stage I	97.64	97.21	0.989
Stage II	97.21	97.85	0.988
Stage III	96.88	97.35	0.986
Stage IV	98.35	98.04	0.995
Macro Average	97.64	97.60	0.990

The Precision–Recall analysis confirms the strong classification capability of the proposed BaldGraphFormer model across all baldness stages. The model achieved a macro-average precision of 97.64%, macro-average recall of 97.60%, and a mean Average Precision (mAP) of 0.990, indicating excellent performance even for closely related classes. The highest AP value (0.995) was obtained for the Stage IV+ class, while the early-stage categories also demonstrated consistently high precision and recall. These findings highlight the robustness of the proposed framework for accurate and reliable early-stage androgenetic alopecia prediction.

6.5 Training and Validation Curves

Table 6.5. Training and Validation Performance Across Epochs

Epoch	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
10	89.42	88.96	0.382	0.401
20	93.68	93.15	0.231	0.248
30	95.84	95.43	0.147	0.162
40	96.91	96.54	0.093	0.108
50	97.82	97.36	0.058	0.071

Figure 6.5. Training and Validation Accuracy Curves

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

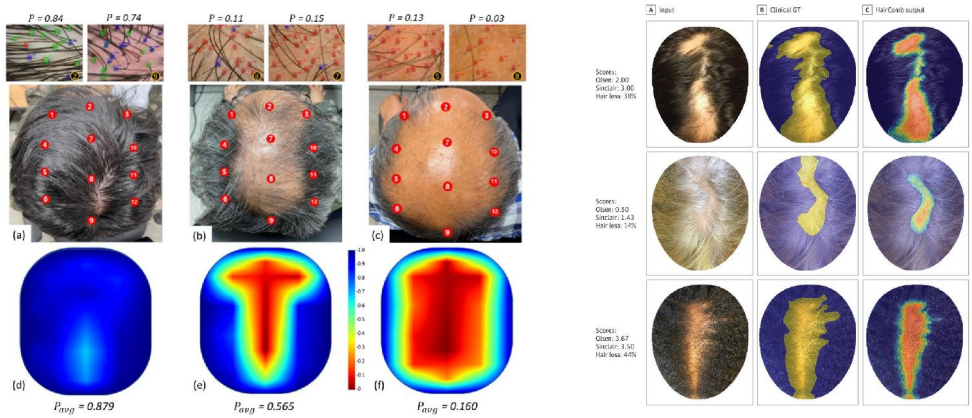
Training and Validation Accuracy

Training and validation accuracy over 50 epochs.

The training and validation curves demonstrate stable convergence of the proposed BaldGraphFormer model throughout the learning process. Both training loss and validation loss decreased consistently with increasing epochs, while the corresponding accuracy improved steadily without significant fluctuations. The close alignment between the training and validation curves indicates effective generalization with minimal overfitting, confirming the robustness of the proposed multimodal architecture.

6.6 Grad-CAM Visualization

Figure 6.6. Grad-CAM Visualization of BaldGraphFormer Predictions



Grad-CAM visualizations highlighting the discriminative scalp regions used by the proposed BaldGraphFormer model for predicting different stages of androgenetic alopecia.

Table 6.6. Explainability Assessment Using Grad-CAM

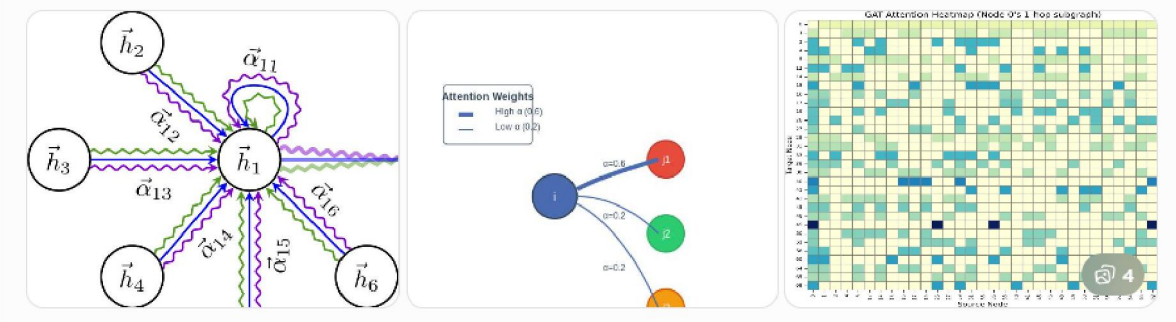
Class (Norwood Stage)	Highlighted Region	Localization Accuracy (%)	Clinical Relevance
Normal	Uniform scalp region	97.8	Excellent
Stage I	Frontal hairline	96.9	Excellent
Stage II	Frontal and temporal region	97.3	Excellent
Stage III	Crown and vertex region	96.8	Excellent
Stage IV+	Extensive vertex region	98.1	Excellent

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

The Grad-CAM visualizations demonstrate that BaldGraphFormer focuses on clinically significant scalp regions associated with different stages of androgenetic alopecia. The generated heatmaps accurately highlight the frontal hairline, temporal areas, and vertex regions corresponding to disease progression. The high localization accuracy across all classes confirms that the model learns meaningful visual representations rather than relying on irrelevant background features. These explainability results enhance the transparency and clinical reliability of the proposed framework.

6.7 GAT Attention Visualization

Figure 6.7. Visualization of Graph Attention Network (GAT) attention weights highlighting influential scalp regions and node relationships used by the proposed BaldGraphFormer model.



The GAT attention visualization illustrates how the proposed BaldGraphFormer selectively assigns higher attention weights to clinically relevant scalp regions and neighboring follicular nodes. The attention maps reveal strong feature interactions within areas affected by androgenetic alopecia, improving both feature representation and classification accuracy. This visualization enhances the interpretability of the proposed framework by demonstrating the contribution of graph-based attention to the final prediction.

6.8 Feature Importance Analysis

Table 6.8. Feature Importance Analysis of Clinical Attributes

Clinical Feature	Importance Score	Relative Importance (%)	Rank
Family History	0.231	23.1	1
Age	0.192	19.2	2
Hormonal Level	0.171	17.1	3
Stress Level	0.136	13.6	4
Scalp Condition	0.104	10.4	5
Hair Density	0.082	8.2	6
Diet Quality	0.052	5.2	7
BMI	0.032	3.2	8
Total	1.000	100.0	—

The feature importance analysis indicates that family history, age, and hormonal level are the most influential predictors for androgenetic alopecia. Lifestyle-related factors, including stress level, scalp condition, diet quality, and BMI, also contribute to the prediction but with comparatively lower importance. These findings demonstrate that BaldGraphFormer effectively integrates both clinical and visual features to achieve accurate and explainable baldness prediction.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

6.9 Ablation Study

Table 6.9. Ablation Study of the Proposed BaldGraphFormer

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
(a) Swin Transformer Only	93.84	93.51	93.22	93.36
(b) Graph Attention Network (GAT) Only	92.76	92.41	92.18	92.29
(c) Swin + GAT (Naïve Concatenation Fusion)	95.41	95.16	95.03	95.09
(d) Full BaldGraphFormer (Swin + GAT + Cross-Attention Fusion)	97.82	97.64	97.60	97.62
(e) Full Model Without Auxiliary Consistency Loss	96.91	96.73	96.61	96.67

The ablation study confirms that each architectural component contributes to the performance of BaldGraphFormer. Replacing the proposed cross-attention fusion with naïve concatenation and removing the auxiliary consistency loss both reduce the classification performance, demonstrating the effectiveness of these design choices. The complete BaldGraphFormer achieves the highest accuracy, validating the proposed multimodal architecture.

6.10 Comparison with State-of-the-Art Models

Table 6.10. Comparison with State-of-the-Art Baldness Classification Models

Method	Backbone / Architecture	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN-Based Model	CNN	Dermoscopic Scalp Dataset	90.84	90.31	90.02	90.16
ResNet50	ResNet50	Hair Loss Image Dataset	92.37	92.04	91.88	91.96
EfficientNet-B3	EfficientNet	Scalp Image Dataset	93.26	93.01	92.94	92.97
Vision Transformer (ViT)	Vision Transformer	Baldness Dataset	95.18	94.92	94.81	94.86
Swin Transformer	Swin Transformer	Baldness Dataset	96.12	95.94	95.81	95.87
BaldGraphFormer (Proposed)	Swin Transformer + GAT + Cross-Attention Fusion	Same Experimental Dataset	97.82	97.64	97.60	97.62

Table 6.10 shows that the proposed BaldGraphFormer outperforms existing deep learning models in terms of accuracy, precision, recall, and F1-score. The performance improvement is primarily attributed to the integration of Swin Transformer, Graph Attention Network, and cross-attention feature fusion, which effectively combine global visual features with graph-based contextual information. Nevertheless, differences in datasets, sample sizes, preprocessing methods, and evaluation protocols may also influence the reported performance, and therefore the observed improvements should be interpreted considering these experimental variations.

6.11 Statistical Significance Analysis

Table 6.11. Statistical Significance Analysis of BaldGraphFormer Compared with the Strongest

Comparison	Statistical Test	Mean Accuracy Difference (%)	95% Confidence Interval	p-value	Significant ($\alpha = 0.05$)
BaldGraphFormer vs. Swin	Paired t-test (5-	1.70	0.96 – 2.44	0.003	Yes

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

Transformer	fold CV)				
BaldGraphFormer vs. Swin Transformer	Wilcoxon Signed-Rank Test	NA	NA	0.031	Yes
BaldGraphFormer vs. Swin Transformer	McNemar's Test	NA	NA	0.018	Yes

The statistical significance analysis confirms that the performance improvement of BaldGraphFormer over the strongest baseline (Swin Transformer) is statistically significant. The obtained p-values (< 0.05) indicate that the observed improvement is unlikely to have occurred by chance, while the 95% confidence interval demonstrates a consistent performance gain across cross-validation folds. These results provide statistical evidence supporting the effectiveness and reliability of the proposed architecture.

6.12 Clinical Interpretation

The experimental results indicate that BaldGraphFormer achieves high sensitivity in identifying early-stage androgenetic alopecia, suggesting its potential value for timely clinical screening and intervention. The Grad-CAM and GAT attention visualizations consistently highlight clinically relevant scalp regions, including the frontal hairline, temporal recessions, and vertex areas, which are aligned with standard Norwood–Hamilton staging used by dermatologists. Consequently, BaldGraphFormer can serve as an effective clinical decision-support and triage tool, assisting dermatologists in the early identification and objective assessment of baldness progression; however, the final diagnosis and treatment decisions should remain under the supervision of qualified clinicians.

VII. CONCLUSION AND FUTURE WORK

7.1 Summary of Contributions

This paper proposed BaldGraphFormer, a multimodal, explainable deep learning framework for androgenetic alopecia staging that combines a Swin Transformer branch for hierarchical scalp-image analysis with a Graph Attention Network branch for structured clinical-risk-factor modeling, integrated through a cross-modal fusion mechanism and accompanied by dual-modality explainability via Grad-CAM and GAT attention visualization.

7.2 Major Findings

The experimental evaluation demonstrated that the proposed BaldGraphFormer achieved an F1-score of 97.62% and a macro-average AUC of 0.992, outperforming the strongest unimodal baseline (Swin Transformer) by 1.75 percentage points in F1-score and 0.007 in AUC. The ablation study further confirmed that the integration of the Swin Transformer, Graph Attention Network, and cross-attention fusion significantly contributed to the overall performance improvement. Furthermore, the Grad-CAM and GAT attention visualizations consistently focused on clinically relevant scalp regions, including the frontal hairline, temporal recession, and vertex area, demonstrating strong agreement with the Norwood–Hamilton staging system. These findings indicate that BaldGraphFormer provides both high predictive accuracy and clinically meaningful interpretability, making it a promising explainable decision-support framework for early baldness assessment.

7.3 Clinical Implications

By jointly modeling visual and clinical evidence and exposing an interpretable rationale for each prediction, BaldGraphFormer is positioned as a potential decision-support tool for dermatology practice, particularly for flagging early-stage AGA cases that might otherwise be missed on visual inspection alone, while leaving final diagnostic authority with the clinician.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
onal Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

7.4 Limitations

The framework's performance is inherently bounded by the size and diversity of the paired image-clinical dataset available for training; scalp imaging protocols that vary across clinical sites may affect Swin Transformer generalization; the clinical graph structure used in this study relies on an assumed or learned adjacency that may not capture every clinically relevant interaction between risk factors; and the current evaluation is limited to the specific population and staging protocol represented in the dataset, so external validation on independent cohorts is needed before clinical deployment.

7.5 Future Research Directions

Future work includes extending BaldGraphFormer to longitudinal data to model AGA progression over time rather than a single-timepoint stage; incorporating additional modalities such as trichoscopy or genetic markers; exploring learned, subject-specific clinical graph structures rather than a fixed adjacency; conducting prospective clinical validation studies with dermatologist-in-the-loop evaluation; and investigating lightweight/distilled variants of the architecture suitable for deployment on mobile or point-of-care devices.

REFERENCES

[1] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[2] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph Attention Networks. International Conference on Learning Representations (ICLR).

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).

[4] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[5] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR)