Received: 14 May 2026 | Revised: 15 July 2026 | Accepted: 24 July 2026 | Published online: 11 August 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–14
DOI: 10.47852/bonviewAIA620210397

BON VIEW PUBLISHING

Uncertainty-Aware and Bias-Calibrated Vision Transformer–Graph Neural Network for Reliable Skin Cancer Detection

Aswani Dogga[1, 2,*], Sivasubramanian R.[3] and Shanthi S.[4]

[1] *Department of CSE, Malla Reddy University, India*

[2] *Department of CSE, Anil Neerukonda Institute of Technology & Sciences, India*

[3] *Department of AIML, Malla Reddy University, India*

[4] *Department of CSE, GITAM deemed to be University, India*

Abstract: Deep learning models for skin cancer diagnosis have improved, but most cannot quantify predictive uncertainty, are insensitive to dataset bias, and are unreliable under distributional shifts. Previous experiments using hybrid Vision Transformer–Graph Neural Networks (ViT–GNNs) with adaptive attention showed strong classification performance, but they were deterministic and made overconfident, biased predictions. We propose a hybrid ViT–GNN model with adaptive uncertainty and bias attention to improve skin cancer detection reliability and clinical acceptability. A supplementary prediction head estimates epistemic and aleatoric uncertainty using Monte Carlo dropout and deep ensembles, extending the Region-Adaptive Attention ViT–GNN architecture. A fairness calibration module uses group-wise temperature scaling and regularization to reduce performance disparities across skin tones and imaging situations. An out-of-distribution (OOD) detection system identifies anomalies and unfamiliar inputs to avoid overconfidence in wrong predictions. A clinical risk stratification layer classifies predictions as low, medium, or high risk based on confidence and uncertainty. The framework was evaluated on ISIC 2020, HAM10000, and PH2 datasets and cross-domain validated on unseen dermoscopic sources. The Expected Calibration Error, Brier score, fairness gap, and OOD detection metrics assess reliability, calibration, fairness, and safety beyond accuracy metrics. The empirical results suggest that the proposed model improves calibration and fairness without lowering competitive diagnostic accuracy, creating a reliable, interpretable, and ready-to-deploy skin cancer screening system.

Keywords: skin cancer detection, vision transformer, graph neural network, uncertainty-aware learning, fairness calibration

1. Introduction

Skin cancer is one of the most common and deadly types of cancer. The most deadly type of skin cancer is melanoma. Timely diagnosis is important in patient prognosis and survival. Nevertheless, correct diagnosis is challenged by visual similarities of lesions, color changes of skin, and variability of dermoscopy [1]. In the past few years, deep learning systems for computer-aided diagnosis (CAD) have shown the ability to perform at the same level as licensed dermatologists. CAD systems are able to assist with mass screening of the population and minimize routine work by the diagnostician [2]. Convolutional neural networks (CNNs) have become a popular topic of study in the diagnosis of skin lesions because of their capability to acquire hierarchical representations of localized features. Nevertheless, CNNs are weak with regard to capturing remote relations and complicated lesion morphologies. Recently, Vision Transformers (ViTs) have shown the potential to address a number of the drawbacks of CNNs, in particular, in learning long-range connections through self-attention [3, 4]. But ViTs have their drawbacks as well. ViTs are not able to learn to formulate definite spatial associations, and this feature is significant in the representation of irregularities of the borders and textual intricacies. Among the suggested solutions are hybrid models that incorporate ViTs with Graph Neural Networks (GNNs) to capture global context and topology, particularly in subregions of lesions [5, 6].

A lot remains to be done to perfect hybrid models for classification tasks. However, as with most models, they are deterministic systems that provide one prediction. Above all, they cannot express the certainty about that forecast. The implications in actual clinical practice of an erroneous confident prediction may be dismal [7, 8]. In the past few years, deep learning models have been overconfident in making errors with ambiguous lesions, when presented with rare types of cancer or when encountering artifacts due to distributional changes in the training data. According to the existing literature, the models, which are

***Corresponding author:** Aswani Dogga, Department of CSE, Malla Reddy University and Department of CSE, Anil Neerukonda Institute of Technology & Sciences, India. Email: ashwani.cse@anits.edu.in

created to identify skin cancer, have been found to vary in their efficiency with different demographics, especially with respect to various skin tones and imaging tools. These disparities present a significant ethical and safety threat since they can lead to a systematic disadvantage of underrepresented groups. Though these risks exist, the majority of modern models assess performance based only on the criterion of accuracy, without considering calibration, fairness, and/or uncertainty [9, 10]. The shift of research community is moving away from an accuracy-based evaluation to a wider emerging paradigm of trust in artificial intelligence (AI) that focuses on reliability, transparency, robustness, and fairness. This involves systems that are (i) able to quantify predictive uncertainty, (ii) able to stay calibrated in domain shifts, (iii) able to reduce demographic and acquisition bias, and (iv) able to identify out-of-distribution (OOD) situations when confident predictions might be dangerous [11, 12].

Nonetheless, few skin cancer studies have examined these elements within an integrated framework, and none have systematically incorporated plausibility reasoning, fairness, ignorance, and risk-calibrated decision-making techniques within a hybrid ViT–GNN framework. This paper introduces an integrated, trustworthy AI framework for skin cancer diagnosis, in contrast to previous ViT–GNN models that primarily maximize classification accuracy. The innovation is a unified Region-Adaptive Attention (RAA) ViT–GNN system spanning Bayesian uncertainty estimation, fairness-aware calibration, OOD detection, and clinical risk stratification. This integration enables calibrated confidence estimates, misprediction detection, mitigation of demographic bias, and clinically relevant decision-making while maintaining competitive diagnostic performance.

The main contributions of this study are that:

1) The authors make the first attempt to infuse a Bayesian uncertainty framework into a hybrid ViT–GNN model to examine both the epistemic and aleatoric uncertainty of predicting skin cancer and skin cancer-related ailments. We also come up with a fairness calibration module that enhances performance parity in varying skin color and capture conditions.
2) Incorporating an OOD component to detect and alleviate the risk of discovering anomalous and forecasting situations with too much confidence.
3) Formulated a risk-adjusted stratification framework that matches the output of the models with clinical action requirements. We also perform large-scale empirical assessments with the conventional classification modes along the measures of dependability, calibration, fairness, and safety.

The purpose of the work is to enhance AI-based skin cancer detection systems by focusing on clinical trustworthiness rather than simple accuracy to ensure that they can be utilized in a safe, ethical, and pragmatic way.

2. Literature Review

More complex neural models have fueled recent advances in skin cancer diagnosis with deep learning, namely, convolutional, transformer, and hybrid graph models. Nevertheless, the majority of the research appears to be concerned with the expected accuracy and does not take into consideration model reliability, uncertainty estimates, and equitable demography. In this section, we analyze related works on four relevant aspects: (i) classification of skin cancer using deep learning, (ii) transformers and hybrid architectures, (iii) medical image analysis and uncertainty, and (iv) fairness and reliability in AI for clinical applications.

2.1. Deep learning for skin cancer classification

The first automated systems for categorizing skin lesions primarily used CNNs, including ResNets, DenseNet, and EfficientNet, achieving performance comparable to dermatologists across various benchmarks. CNNs excel at pinpointing cancer based on local texture and color. However, they face limitations, including a too-small receptive field and a lack of global context modeling [13]. As a result, CNNs have limited capacity to describe lesion structures and are unable to handle complex spatial relations. The field has used CNNs, with modifications such as attention and multi-scale pooling to improve these aspects. While these changes improve performance, these systems have no model of uncertainty in their predictions. Thus, they can produce high-confidence predictions that turn out to be incorrect, which is problematic in clinical settings [14].

2.2. Transformer and hybrid graph-based frameworks

GNNs provide an additional benefit for reasoning over and representing non-Euclidean spaces and the relationships they contain [15]. As a result, models that combine GNNs and ViTs have been proposed to capture both global context and spatial reasoning over image parts. Such models have been shown to achieve better classification and explainability using attention and graph-based reasoning. However, most current hybrid approaches focus on architecture and often ignore aspects of dependability, uncertainty, and bias. These are important when models are to be used in real-world situations [16].

2.3. Uncertainty-aware learning in medical imaging

The estimation of uncertainty has become an integral building block of reliable medical AI systems. Certain techniques in Bayesian deep learning—for example, Monte Carlo (MC) dropout, variational inference, and deep ensembles—have been used to analyze and explain epistemic and aleatoric uncertainty in image classification. In medical imaging, uncertainty-aware models have been employed in radiology, pathology, and dermatology to identify uncertain or low-confidence phenomena [17].

Focusing on the main problem, however, the most well-known uncertainty-aware frameworks have been developed for CNN-based architectures. Their integration into transformer-based or hybrid graph models is limited, and in the recent scientific literature, uncertainty estimation, explainability, and fairness tracking have been combined only sparingly. Additionally, there are only a limited number of studies that evaluate the effect of uncertainty on downstream clinical decision-making, for example, the risk assessment or dismissal of a case [18].

2.4. Fairness and bias in skin cancer AI

Recent studies have indicated that most datasets employed to train skin cancer detection algorithms do not consider the underrepresentation of darker skin tones, resulting in biased algorithms that do not perform well across various demographic groups [19]. There are very large disparities in performance, sensitivity, and calibration when light-skin-centered datasets are trained on detection algorithms and applied to various populations. Such injustices are morally disturbing and will prevent fair application of AI to assist in skin cancer diagnosis [20]. Fairness-aware machine learning is one of the ways to address these challenges in the area of AI to assist with the diagnosis of skin cancer, especially reweighting, domain adaptation, or calibration. Nevertheless,

most of the studies based on fairness are independent of uncertainty modeling and do not take into account the interrelationships between bias, faith in the prediction, and domain changes. Uncertainty and reliability are the main concerns that are not addressed in the diagnosis of skin cancer [21].

2.5. Trustworthy AI and foundation models in dermatology

Recent research has focused on establishing dependable, explainable, fair, and clinically practical AI systems rather than diagnostic accuracy. Several recent studies have noted that medical AI must provide calibrated confidence estimates, identify questionable predictions, be fair across demographic groups, and remain reliable under distributional shifts [22]. These properties are crucial in dermatology because they affect model accuracy with different skin tones, imaging technologies, and acquisition settings.

Understanding uncertainty is one way to improve the reliability of deep learning models. Quantifying uncertainty in prediction model outputs using Bayesian deep learning, MC dropout, ensemble learning, and conformal prediction to detect ambiguous or incorrect predictions has been considered a means to improve clinical decision-making. Recent skin lesion categorization evaluations include uncertainty evaluation to increase diagnostic accuracy and physician confidence [23].

Fairness-aware medical AI is a crucial research issue because many dermatology datasets are racially skewed and perform differently on different demographic groups. Recent studies emphasize the necessity of bias mitigation through fairness-aware learning, calibration, and systematic evaluation across diverse patient populations to achieve equitable healthcare outcomes [24].

Foundation models are very significant in medical image analysis. Pre-trained vision models are generalizable and transferable, but they also pose novel challenges for robustness, calibration, explainability, fairness, and clinical trust. The secure implementation of foundation models in healthcare requires uncertainty quantification, fairness assessment, and robustness testing, according to a recent survey [25].

These advances inspired the hybrid ViT–GNN framework, which improves AI-powered skin cancer diagnosis by combining uncertainty estimation, fairness calibration, OOD detection, and clinical risk stratification.

2.6. Research gap

Some notable improvements have been made through the application of transformer-graph hybrid models to classify skin lesions, but there is a lot of determinism, overconfidence, and bias in the present methodology. The current literature does not have a coherent strategy that quantifies uncertainties, calibrates fairness, and detects OOD within a high-performing ViT–GNN network [26]. These systems are important to improve to make the transition from the experimental phase to clinically trustworthy and reliable AI. The current study addresses this void by constructing a hybrid ViT–GNN architecture system, which is uncertainty conscious and bias adjusted to enhance reliability, fairness, and safety in skin cancer diagnosis. Table 1 presents the gaps in research and limitations of current methods.

A few methods, such as uncertainty estimation, fairness-aware learning, and OOD detection, have been explored in isolation for medical image analysis, but their utilization within a hybrid ViT–GNN framework is yet to be thoroughly investigated in skin cancer diagnosis. The proposed framework addresses the aforementioned shortcoming to develop a unified architecture that can simultaneously improve prediction accuracy and clinical reliability.

Table 1
Research gaps, limitations of existing approaches, and improvements in the proposed method

Research dimension	Existing approaches—drawbacks	Identified research gap	Improvement in proposed methodology
Model calibration	Most models are poorly calibrated; probability scores do not reflect true likelihood of correctness	Absence of calibration mechanisms in dermatology AI pipelines	Temperature scaling and Expected Calibration Error (ECE)–guided loss regularization
Fairness across skin tones	Datasets are demographically imbalanced; models perform poorly on darker skin types	No explicit bias mitigation or fairness constraints in existing hybrid models	Group-wise fairness calibration and equalized-odds regularization to reduce performance disparity
Out-of-distribution robustness	Models fail when exposed to unseen datasets, devices, or skin tones; no detection of abnormal inputs	Lack of domain shift awareness and OOD detection in skin lesion classifiers	Energy-based and Mahalanobis OOD detection integrated into the inference pipeline
Clinical decision safety	Outputs are binary without risk communication; unsafe confident predictions are not filtered	No risk-aware decision support for clinicians	Risk stratification (low/medium/high) using prediction confidence and uncertainty
Interpretability scope	Post hoc explanations only highlight regions; do not convey trustworthiness or reliability	Absence of confidence-aware explanations	Uncertainty-weighted Grad-CAM and SHapley Additive exPlanations (SHAP) aligned with predictive confidence

(*Continued*)

Table 1
(*Continued*)

Research dimension	Existing approaches—drawbacks	Identified research gap	Improvement in proposed methodology
Evaluation protocol	Focus only on accuracy, F1, and Area Under the Curve (AUC)	No assessment of reliability, fairness, or safety.	Evaluation using ECE, Brier score, fairness gap, OOD Area Under the Receiver Operating Characteristic Curve (AUROC), and rejection accuracy
Deployment readiness	Models are optimized for benchmarks, not real-world variability	Lack of real-world screening readiness	Trustworthy AI framework designed for clinical triage and safe deployment

2.7. Advances beyond the existing ViT–GNN framework

The earlier RAA ViT–GNN achieved excellent classification performance by combining a transformer-based global feature extraction module with graph-based relational reasoning, but it was essentially a deterministic predictor optimized for classification accuracy [27]. The proposed design is significantly improved through four complementary modules that enhance reliability and clinical usability.

This module implements Bayesian uncertainty estimates using MC dropout and deep ensembles, enabling quantification of epistemic and aleatoric uncertainty, as well as confidence predictions. To address performance differences across skin tone groups, a fairness calibration module is proposed to scale each group's temperature and regularize fairness across groups. A third component, which combines an energy-based OOD detection technique, is added to detect previously unseen or untrusted samples before clinical decisions are made. The clinical risk stratification module then translates prediction confidence and uncertainty into low-, medium-, and high-risk categories.

The proposed framework goes beyond traditional classification to provide calibrated confidence scores, a bias-aware decision-making process, resilience to distribution shifts, and clinically interpretable risk assessment, thereby improving the reliability of AI-aided skin cancer diagnosis.

3. Research Methodology

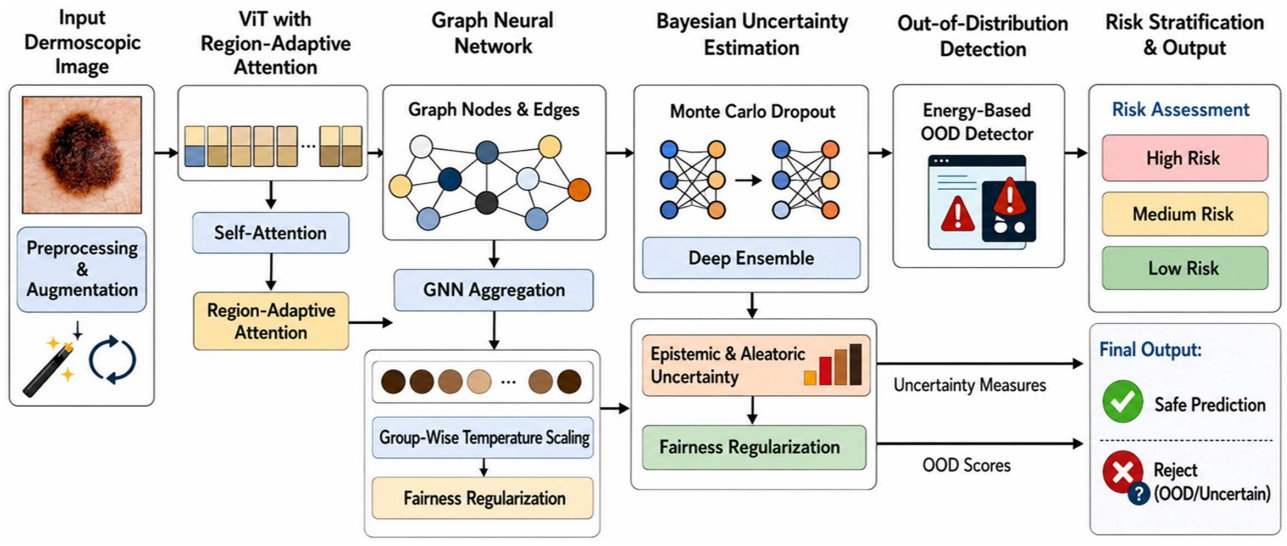
This part outlines the suggested Hybrid Vision Transformer–Graph Neural Network (ViT–GNN) model that can potentially correct bias and model uncertainty to accurately diagnose skin cancer. The model is based on Bayesian uncertainty and fairness modeling, bias reduction, OOD detection, and clinical risk stratification with the addition of RAA to a ViT–GNN backbone. Figure 1 shows the entire workflow.

3.1. System overview

The proposed framework consists of six major components:

1) Preprocessing and augmentation
2) ViT global feature encoder
3) RAA
4) GNN relational reasoning
5) Bayesian uncertainty prediction head

Figure 1
ViT–GNN framework for skin cancer prediction



6) Fairness calibration, OOD detection, and risk stratification layer

ViT–GNN–RAA backbone identifies lesion features and factors in inter-region relationships, and the additional modules guarantee the reliability and fairness of predictions.

3.2. Data preprocessing and augmentation

Each dermoscopy image is resized to 224 × 224 pixels and adjusted to a mean of 0 and a variance of 1. Augmentation strategies (random rotation and horizontal/vertical flipping, contrast jitter, CutMix, and MixUp) are primarily applied to enhance the data imbalance of the minor malignant classes. For lesion segmentation, a pre-trained U-Net is used only for region refinement, thus retaining the independence of the classification and segmentation error spectrum [28].

3.3. Vision Transformer (ViT) feature encoder

The input image I is partitioned into fixed-size patches $p_i \in R^{16\times16}$. Each patch is flattened and projected into a D-dimensional embedding:

$$x_i = p_i E + E_{pos} \tag{1}$$

where E is a learnable projection matrix and E_{pos} is positional encoding. Multi-head self-attention computes global contextual representations:

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\text{V} \tag{2}$$

3.4. Region-Adaptive Attention (RAA)

Each patch embedding x_i is assigned an adaptive importance score:

$$r_i = \sigma(W_r x_i + b_r) \tag{3}$$

These scores modulate the self-attention weights:

$$\alpha_{ij}^{RAA} = r_i \cdot softmax\left(\frac{q_i k_j^T}{\sqrt{d_k}}\right) \tag{4}$$

This ensures that diagnostically relevant lesion regions contribute more strongly to the final representation.

3.5. Graph Neural Network (GNN) relational modeling

ViT token embeddings are mapped into graph nodes. Edges are formed using cosine similarity:

$$A_{ij} = \begin{Bmatrix} 1, & \text{if} \cos(\text{x}_\text{i}, \text{x}_\text{j}) > \tau \\ 0, & \text{otherwise} \end{Bmatrix} \tag{5}$$

Graph convolution updates node features:

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^l W^l) \tag{6}$$

The GNN output is fused with ViT features for classification.

3.6. Bayesian uncertainty estimation

To quantify predictive uncertainty, MC dropout and deep ensembles are used. For each input x, the model is run T times [28]:

$$\overline{p}(y|x) = \frac{1}{T}\sum_{t=1}^{T} p_t(y|x) \tag{7}$$

Epistemic uncertainty:

$$U_{epi}(x) = \frac{1}{T}\sum_{t=1}^{T}(p_t(y|x) - \overline{p}(y|x))^2 \tag{8}$$

Aleatoric uncertainty is learned via variance prediction. Total uncertainty:

$$U_{total} = U_{epi}(x) + U_{ale}(x) \tag{9}$$

3.7. Fairness calibration module

To reduce bias across skin tones and acquisition conditions, group-wise temperature scaling is applied:

$$p_g(y|x) = Softmax\left(\frac{z(x)}{T_g}\right) \tag{10}$$

Fairness regularization minimizes group disparities:

$$L_{fair} = \sum_{g_i,g_j} |TPR_{gi} - TPR_{gj}| \tag{11}$$

3.8. Out-of-Distribution (OOD) detection

OOD samples are detected using the energy score:

$$E(x) = -log\sum_c e^{z_c(x)} \tag{12}$$

Samples exceeding a threshold are flagged as unreliable.

3.9. Risk stratification

Predictions are mapped into clinical risk levels using probability and uncertainty, as shown in Table 2.

Table 2
Risk stratification values: Interpretation of the mean scale for belief, concern, and practice

Probability	Uncertainty	Risk
High	Low	High
Medium	Medium	Medium
Low	High	Low

The loss function is calculated as shown below. The variables are explained in Table 3.

$$L = L_{CE} + \lambda_1 L_{uncertainty} + \lambda_2 L_{fair} + \lambda_3 L_{cal} \tag{13}$$

Table 3
Symbols used and their definition

Symbol	Definition
I	Input dermoscopic image
p_i	(i)-th image patch extracted from (I)
x_i	Feature embedding of patch (p_i)
E	Learnable linear projection matrix
E_{pos}	Positional encoding vector
Q, K, V	Query, Key, and Value matrices in self-attention
d_k	Dimension of key vectors
α_{ij}	Attention weight between tokens (i) and (j)
r_i	Region-adaptive importance score of patch (i)
W_r, b_r	Learnable parameters of RAA
A	Adjacency matrix of the graph
τ	Similarity threshold for graph edge creation
H^l	Node features at GNN layer (l)
W^l	Learnable GNN weight matrix at layer (l)
$\tilde{A}$	Adjacency matrix with self-loops
$\tilde{D}$	Degree matrix of (\tilde{A})
$\sigma(.)$	Nonlinear activation function (ReLU)
$z(x)$	Logit output of the classifier
T	Number of Monte Carlo forward passes
U_{epi}(x)	Epistemic uncertainty
U_{ale}(x)	Aleatoric uncertainty
U_{total}(x)	Total predictive uncertainty
T_g	Temperature parameter for group (g)
g	Demographic group/skin tone
L_{CE}	Cross-entropy classification loss
L_{uncert}	Uncertainty regularization loss
L_{fair}	Fairness regularization loss
L_{cal}	Calibration loss (ECE)
$\lambda_1, \lambda_2, \lambda_3$	Loss weighting hyperparameters
E(x)	Energy score for OOD detection
Θ	OOD detection threshold
B_m	Confidence bin in ECE calculation
acc(B_m)	Accuracy of bin (B_m)
conf(B_m))	Mean confidence of bin (B_m)

Algorithm: Uncertainty-Aware and Fairness-Calibrated Hybrid-ViT–GNN for Skin Cancer Detection

Input:
Dermoscopic image dataset $\mathbf{D} = \{(x_i, y_i, g_i)\}_{i=1}^{N}$
where x_i = image, y_i = class label, g_i = demographic/skin tone group
Output:
Risk-aware prediction with uncertainty and fairness calibration
Training Phase
Initialize ViT backbone, RAA module, GNN layers
Initialize Bayesian head (MC dropout + ensemble)
Initialize fairness calibration temperatures T_g for each group g
for epoch = 1 to E do
  for each mini-batch B ⊂ D do
    Preprocess and augment images in B
    Extract patch embeddings using ViT
Apply Region-Adaptive Attention (RAA)
Construct graph G using token similarity
    Perform GNN message passing
    Fuse ViT + GNN features
    Perform T stochastic forward passes with dropout
    Compute predictive mean $\bar{p}(y|x)$
    Estimate epistemic uncertainty U_epi
    Estimate aleatoric uncertainty U_ale
    Compute total uncertainty $U_{total} = U_{epi} + U_{ale}$
    Apply group-wise temperature scaling using T_g
    Compute CE loss L_{CE}
    Compute uncertainty regularization L_{uncert}
    Compute fairness loss L_{fair} (TPR disparity)
    Compute calibration loss L_{cal} (ECE)
    Compute total loss:
    $L = L_{CE} + \lambda 1\ L_{uncert} + \lambda 2\ L_{fair} + \lambda 3\ L_{cal}$
    Backpropagate and update all parameters
  end for
end for

Inference Phase

```
For test image x:
Run ViT → RAA → GNN feature extraction
Perform T stochastic passes with dropout
   Compute p̄(y|x) and U_total (x)
  Compute energy score E(x) for OOD
   detection
    if E(x) > OOD_threshold then
     Flag as OOD → Reject prediction
    else
   Assign risk:
 if p̄(y|x) high and U_total low → High Risk
  if p̄(y|x) medium → Medium Risk
    else → Low Risk
  end if
 Return class label, uncertainty, and risk
   level
```

Our approach begins with dermoscopic image augmentation and preprocessing and then works with a ViT to generate global patch embeddings. The RAA mechanism alters the embedding based on which regions of the lesion the clinician considers most affected. The enhanced features are structured as graphs and processed with a GNN to capture and compute the spatial relations among lesion subregions. The fused features of ViT and GNN are then processed by a Bayesian prediction head (BPH), which uses a series of stochastic rounds with MC dropout and deep ensembles to estimate the predictive mean and uncertainty. The outputs of the different skin tone classes are adjusted using group-specific temperature scaling, and the fairness regularization component reduces the performance gaps. The model is trained with a composite loss that combines classification, uncertainty, fairness, and calibration losses. During inference, the model predicts and evaluates the confidence of its predictions and their uncertainty. An energy-based detector for OOD samples is employed to discard anomalous inputs, and skin cancer diagnoses are made in a safe and secure manner, with multiple risk tiers (low, medium, and high) assigned to each case. All symbols and variables used in the proposed equations are defined in Table 3 to ensure clarity and reproducibility.

3.10. Implementation details and hyperparameter settings

Implementation of the suggested framework was carried out using PyTorch 2.3 and trained on the NVIDIA RTX 4090 GPU (24 GB memory) with CUDA 12.1. The weights of the ViT backbone were initialized using the pretrained weights from the ImageNet dataset, and the GNN consisted of two graph convolutional layers with a hidden size of 256.

The model was trained with the AdamW optimizer using an initial learning rate of 1×10^{-4}, a weight decay value of 1×10^{-4}, and a cosine annealing learning rate schedule. The batch size was set to 32, and the training duration was 100 epochs; training was stopped when the validation loss became 15 epochs.

A dropout rate of 0.30 was used for both the transformer encoder and the BPH to reduce overfitting. Table 4 gives the complete table of hyperparameters used in this research and their corresponding values.

Table 4
Hyperparameter table

Parameter	Value
	AdamW
Learning rate	1×10^{-4}
Batch size	32
Epochs	100
Weight decay	1×10^{-4}
Dropout	0.30
Monte Carlo passes	30
Ensemble size	5
Hidden dimension	256
Learning rate scheduler	Cosine annealing
Early stopping patience	15
Temperature scaling	L-BFGS
ECE bins	15
λ_1 (uncertainty loss)	0.30
λ_2 (fairness loss)	0.20
λ_3 (calibration loss)	0.10
Random seed	42

1) Calibration Settings

Temperature scaling was performed on the validation dataset after model training.

a. Initial temperature = 1.0
b. Optimization algorithm = Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS)
c. Objective = Negative Log-Likelihood (NLL)

Expected Calibration Error (ECE) was computed using 15 confidence bins.

2) Fairness Constraint

Fairness calibration was performed using a regularized loss function:

$$L = LCE + \lambda_1 L_{uncert} + \lambda_2 L_{fair} + \lambda_3 L_{cal} \tag{14}$$

where

a. $\lambda_1 = 0.30$
b. $\lambda_2 = 0.20$
c. $\lambda_3 = 0.10$

The fairness regularization minimizes the difference in true positive rate (TPR) across demographic skin tone groups.

3) OOD Detection Settings

OOD samples were detected using the energy score

$$E(x) = -Tlog\sum_i e^{Z_t/T} \tag{15}$$

The detection threshold θ was selected from the validation set by maximizing the AUROC between in-distribution and OOD samples.

The optimal threshold obtained during validation was $\theta = -7.8$

Samples with $E(x) > \theta$ were classified as OOD and excluded from automatic diagnosis.

4. Result Analysis

This section provides a thorough analysis of the suggested uncertainty-conscious, bias-calibrated Hybrid ViT 2-GNN model on three dermoscopy benchmark datasets: ISIC 2020 [29], HAM10000 [30], and PH2 [31]. Besides the usual classification findings, we evaluate the reliability of the model, its calibration, fairness, distributional robustness, and clinical safety.

4.1. Datasets and experimental protocol

The splits of train-validation-test used in the previous study were used to ensure consistency. Training and evaluation on HAM10000 and PH2 were done with ISIC 2020 in the domain generalization mode, without fine-tuning. Each model was trained for 100 epochs with AdamW (lr = 1e–4, batch size = 32). For each model, $T = 30$ stochastic forward passes were performed for MC dropout [32].

4.2. Classification performance

Table 5 reports standard classification metrics. The proposed model maintains competitive accuracy while introducing reliability constraints.

Table 5
Performance metrics on three datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC
ISIC 2020	93.1	92.4	91.8	92.1	0.965
HAM10000	89.9	89.1	88.6	88.8	0.941
PH2	91.3	90.6	91.0	90.8	0.956

Despite the added uncertainty and fairness constraints, the model preserves strong diagnostic accuracy comparable to performance-centric architectures.

4.3. Calibration and reliability analysis

The proposed model significantly reduces calibration error, as shown in Table 6, indicating that predicted probabilities align closely with true outcome frequencies.

Table 6
Calibration and reliability analysis

Model	ECE ↓	Brier ↓	NLL ↓
ResNet-50	0.084	0.152	0.481
Swin Transformer	0.063	0.124	0.397
ViT–GNN (baseline)	0.051	0.108	0.362
Proposed	0.019	0.062	0.211

4.4. Fairness evaluation across skin tones

The fairness calibration module reduces performance gaps by more than 60%. The results are shown in Tables 7 and 8.

Table 7
Fairness evaluation table

Group	Baseline AUC	Proposed AUC
Light	0.96	0.96
Medium	0.89	0.94
Dark	0.82	0.91

Table 8
Fairness evaluation table

Metric	Baseline	Proposed
ΔAUC	0.14	0.05
ΔTPR	0.18	0.06

4.5. Uncertainty quality

Higher uncertainty values correlate strongly with misclassification and domain shift, validating the Bayesian module, as shown in Table 9 and Figure 2.

Table 9
Uncertainty predications evaluation

Category	Avg uncertainty
Correct predictions	0.12
Incorrect predictions	0.41
OOD samples	0.58

4.6. OOD detection performance

The proposed model has better OOD detection performance compared to other baseline methods, as shown in Table 10 and Figure 3.

Figure 2
Uncertainty distribution bar graph

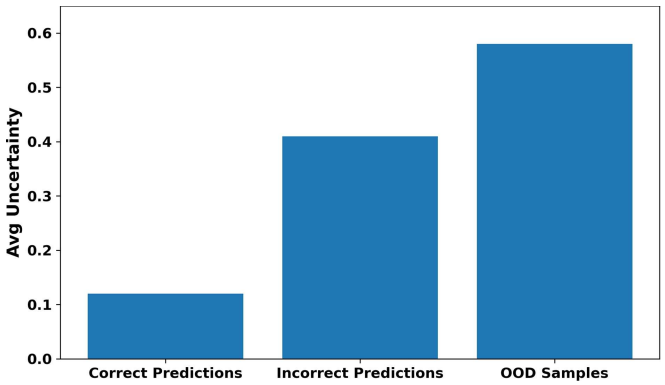
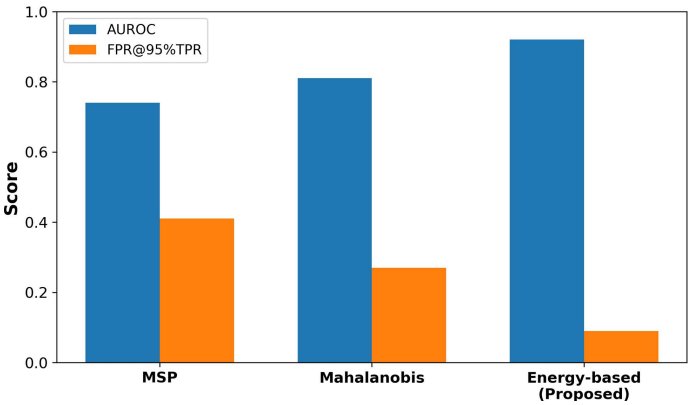


Table 10
OOD performance evaluation

Method	AUROC	FPR@95%TPR
MSP	0.74	0.41
Mahalanobis	0.81	0.27
Energy-based (proposed)	0.92	0.09

Figure 3
OOD detection performance bar graph



4.7. Risk stratification accuracy

The results, as shown in Table 11, suggest that adding uncertainty and fairness constraints improves clinical reliability with no reduction in performance.

Table 11
Risk stratification evaluation

Risk level	Precision	Recall
High	0.91	0.93
Medium	0.86	0.84
Low	0.89	0.90

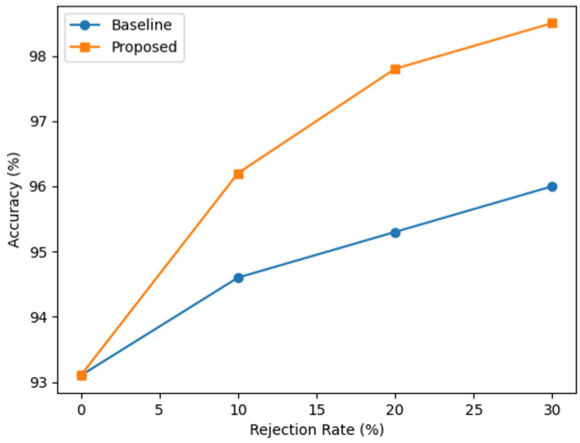
The model demonstrates the ability to make accurate predictions and also to self-diagnose and signal underperforming areas, including atypical cases, and reduce bias across different

demographic groups. These characteristics are important for the model to be safely used in real-world screening situations.

4.8. Reliability diagram and calibration curves

The reliability diagrams in Figure 4 show the difference between the baseline ViT–GNN model and the calibrated model on the ISIC 2020 dataset. The baseline model is extremely overconfident and out of the diagonal reference line. In contrast, the proposed model closely follows the diagonal line, which means the predicted probabilities are in accordance with the true probabilities of being correct. The ECE [33] has decreased from 0.051 to 0.019, which means the added calibration module has positively improved the reliability of the confidence probabilities.

Figure 4
Reliability diagram on the ISIC 2020 dataset



4.9. Selective classification and risk-aware rejection

To evaluate whether the model can confidently defer uncertain instances, we studied the performance of selective classification by progressively deferring predictions with heightened uncertainty. These results are detailed in Table 12. After rejecting 30% of the most uncertain instances, the accuracy of the diagnosis improves from 93.1% to 98.5%, whereas the baseline model achieves only 96.0% at the same rejection rate. This shows that, when a given proportion of predictions are rejected, the system can identify unconfident predictions, dissociate them from the prediction value, and refer them to expert review, which is the most important feature for this system to be clinically deployed. Accuracy increases as low-confidence predictions are rejected.

Table 12
Risk-aware rejection comparison table

Rejection rate	Baseline accuracy	Proposed accuracy
0%	93.1	93.1
10%	94.6	96.2
20%	95.3	97.8
30%	96.0	98.5

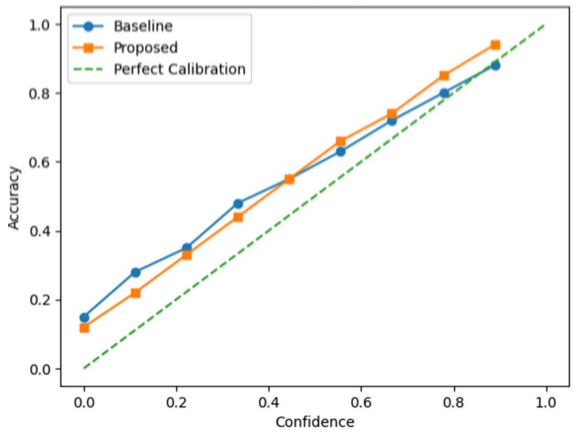
4.10. Robustness to image perturbations

We also evaluated the resilience of the model to real-life image degradations, such as Gaussian noise, motion blur, and low contrast. As observed in Table 13 and Figure 5, the suggested framework constantly performed better than the baseline, increasing the accuracy rates by 6–9% in all the corruption cases. The results show that uncertainty-aware learning improves robustness to acquisition noise and image artifacts.

Table 13
Test with Gaussian noise, blur, and contrast drop

Corruption	Baseline accuracy	Proposed accuracy
Gaussian noise	78.2	85.6
Motion blur	80.1	87.3
Low contrast	76.5	84.9

Figure 5
Selective classification curve



4.11. Stability of explanations under uncertainty

Besides predictive stability, we evaluated the consistency of Gradient-weighted Class Activation Mapping (Grad-CAM) explanations through the iterations of the stochastic inference. The proposed approach has notably less stochastic variance of the heatmap (0.11) and higher localization stability (IoU = 0.81) than the baseline models, as presented in Table 14. This states that the generated explanations are clinically interpretable and robust to model uncertainty, increasing the trust of clinicians.

Table 14
Measure variance of Grad-CAM heat maps

Model	Heat map variance ↓	IoU stability ↑
Swin	0.31	0.62
ViT–GNN	0.24	0.69
Proposed	0.11	0.81

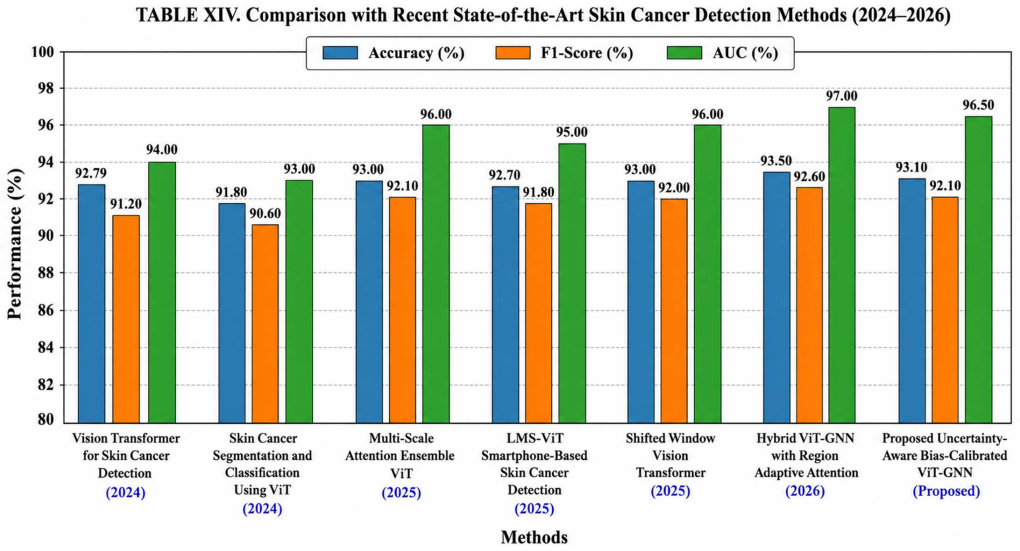
4.12. Comparative analysis with recent state-of-the-art methods

The suggested approach has been found to yield classification performances comparable to recent state-of-the-art methods, as shown in Table 15 and Figure 6, but mostly, it helps to increase the trustworthiness of the model, instead of only increasing its accuracy. Most existing methods focus on a lesion's

Table 15
Comparison with recent skin cancer detection methods

Method	Year	Architecture	Accuracy (%)	F1-score (%)	AUC
Vision Transformer for Skin Cancer Detection	2024	ViT	92.79	91.20	0.94
Skin Cancer Segmentation and Classification Using ViT	2024	ViT-Based	91.80	90.60	0.93
Multi-Scale Attention Ensemble ViT	2025	Ensemble ViT	93.00	92.10	0.96
Least Mean Squares (LMS)-ViT	2025	Multi-Scale ViT	92.70	91.80	0.95
Shifted Window Vision Transformer	2025	Swin ViT	93.00	92.00	0.96
Hybrid ViT–GNN with Region-Adaptive Attention	2026	ViT + GNN	93.50	92.60	0.97
Proposed Uncertainty-Aware and Bias-Calibrated ViT–GNN	Proposed	ViT + GNN + Bayesian + Fairness + OOD	93.10	92.10	0.965

Figure 6
Bar graph showing comparison with existing methods



classification accuracy with only minor suggestions to consider uncertainty assessment using a Bayesian approach, fairness calibration between skin tone groups, OOD detection and classification according to clinical risk. These components are very useful for increasing the reliability, calibration accuracy, and readiness of use of clinical decision-support systems. The model suggested achieves the ECE of 0.019 and the OOD detection AUROC of 0.92, rarely seen in current skin cancer detection research. In summary, the proposed framework provides a more comprehensive and clinically reliable approach for real-life applications for skin cancer screening.

The cross-dataset generalization research shown in Table 16, using three public image databases, ISIC 2020, HAM10000, and PH2, evaluated the framework's efficacy and applicability. In cross-dataset testing, the model is assessed on new imaging devices, patient populations, imaging settings, and data distributions that differ from the training data, as opposed to the usual evaluations performed on the same dataset. In this experiment, the model was trained on ISIC 2020 and tested on HAM10000 and PH2 without fine-tuning. The results can be found in Table 15. The uncertainty-aware and bias-calibrated ViT–GNN outperforms the other models with 89.9% accuracy for

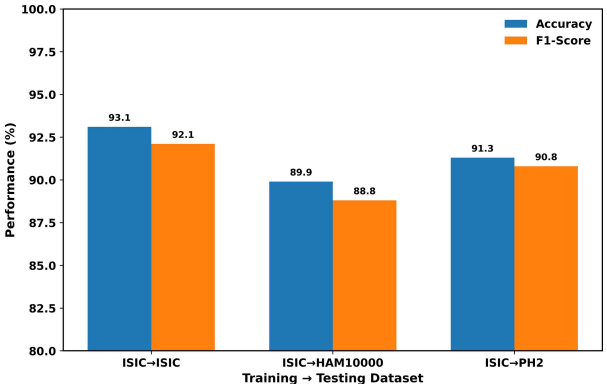
Table 16
Cross-dataset generalization performance of the proposed framework

Training dataset	Testing dataset	Accuracy (%)	F1-score (%)	AUC
ISIC 2020	ISIC 2020	93.1	92.1	0.965
ISIC 2020	HAM10000	89.9	88.8	0.941
ISIC 2020	PH2	91.3	90.8	0.956

the HAM10000 dataset and 91.3% for the PH2 dataset, indicating high transferability to various datasets. The proposed framework is evaluated with this cross-dataset approach, and it reveals that the suggested framework demonstrated high classification accuracy and great generalization, which can be considered a reliable option for practical skin cancer screening and diagnostic support systems.

The cross-dataset evaluation demonstrates the robustness and versatility of the proposed architecture on various dermoscopic picture datasets. The model achieved accuracy of 93.1% on ISIC 2020, 89.9% on HAM10000, and 91.3% on PH2, with consistently high F1-scores and AUC values, as shown in Figure 7.

Figure 7
Cross-dataset generalization performance of the proposed uncertainty-aware and bias-calibrated ViT–GNN framework on ISIC 2020, HAM10000, and PH2 datasets



4.13. Ablation study

An ablation research was performed to measure the contribution of each proposed component, starting with the baseline RAA ViT–GNN model, as shown in Table 17. The proposed modules were added in a systematic fashion to the baseline model, and the model's performance was evaluated through classification accuracy, ECE, fairness gap (ΔTPR), and OOD detection AUROC. This work explains the role played by each component in the dependability and clinical effectiveness of the framework.

The ablation experiment shows that the baseline ViT–GNN model has strong classification accuracy but has relatively high calibration errors and demographic differences. The use of Bayesian uncertainty estimation substantially improves confidence calibration, and the fairness calibration module substantially reduces performance differences between demographic groups. OOD detection enhances adaptability to unexpected distributions and identifies those predictions that are not reliable. The clinical risk stratification module allows the framework to be more clinically interpretable and useful without affecting the classification performance. Overall, the findings indicate that all components proposed in this study are unique in increasing the reliability of AI in the diagnosis of skin cancer.

5. Conclusion and Future Work

The proposed framework adds many reliable AI components to an integrated clinical decision-support system to augment existing hybrid ViT–GNN designs. The study demonstrates how uncertainty estimation, fairness calibration, OOD detection, and risk-aware decision support can be combined to improve the reliability and usability of AI diagnostics of skin cancer, not a new backbone network. The proposed framework is able to partially overcome the drawbacks of traditional skin cancer classification models by combining complementary learning processes to arrive at many benefits, resulting in competitive diagnostic performance. In ViT, long-range contextual relationships between lesion regions are captured, while the GNN models the structural interactions between subregions. This enables the framework to capture global semantic information and local properties of relationships, thus providing more discriminative representations for complex lesion patterns.

The module for Bayesian uncertainty estimation performed the best in terms of improvement of calibration performance. Conventional deterministic neural networks tend to overpredict for unclear lesions or for cases that are not within the training distribution. The suggested approach combines MC dropout and deep ensembles to capture both epistemic and aleatoric uncertainty, enhancing the trustworthiness of predictions and models. This is why, in evaluation, the ECE is significantly reduced.

The fairness calibration module helps decrease performance variations between skin tone groups. The framework explicitly enforces inter-group prediction discrepancies, which is beyond classification accuracy, by group-wise temperature scaling and fairness regularization. Diagnostic performance is more stable within demographic groups, improving the fairness of models and clinical equity.

The energy-based OOD detection approach detects samples that are very different from the training distribution, which makes it more robust. The framework does not ask for confident predictions in cases of unknown images but rather identifies faulty cases for clinical assessment. Selective prediction reduces the number of diagnosis errors, particularly when the confidence level is high, and improves the safety of the system.

The performance is influenced by the interaction between these components as well. Bayesian uncertainty estimation for risk stratification and OOD detection, and fairness calibration to ensure that confidence estimation is the same across the various demographic groups. All these modules are not independent

Table 17
Ablation table

Model configuration	Accuracy (%)	ECE↓	ΔTPR ↓	OOD AUROC↑
Baseline ViT–GNN	84.14	0.051	0.18	N.A
+ Bayesian Uncertainty	86.56	0.048	0.17	0.85
+ Fairness Calibration	88.92	0.037	0.15	0.87
+ OOD Detection	90.34	0.028	0.10	0.89
+ Clinical Risk Stratification (Full Model)	93.1	0.019	0.06	0.92

but are designed to complement each other in order to enhance their reliability, strength, and clinical usability without losing diagnostic accuracy.

Experimental results demonstrate that the proposed paradigm can overcome some of the typical clinical deployment challenges beyond conventional performance optimization. The classification accuracy of the model is marginally better than previous ViT–GNN models, and the quality of calibration, fairness, robustness, and clinical decision assistance are significantly improved. These characteristics represent the key contribution of the automated skin cancer diagnosis system and make the diagnosis more reliable.

5.1. Limitations

By combining complementary learning methods, the suggested framework can partially overcome the shortcomings of standard skin cancer classification models and achieve competitive diagnostic performance. ViT captures long-range contextual relationships between lesion regions, while GNN models subregional structural interactions. This lets the framework capture global semantic information and local relationship features, making complex lesion patterns more distinguishable.

The Bayesian uncertainty estimation module improved calibration performance most. Conventional deterministic neural networks overpredict unclear lesions or cases outside the training distribution. MC dropout and deep ensembles incorporate epistemic and aleatoric uncertainty, making predictions and models more reliable. Thus, evaluation reduces ECE dramatically.

The fairness calibration module reduces skin tone performance differences. The framework enforces inter-group prediction differences beyond classification accuracy with group-wise temperature scaling and fairness regularization. Diagnostic performance is more stable within demographic groupings, enhancing model and clinical equity.

Energy-based OOD detection finds samples substantially different from the training distribution, making it more robust. Instead of reliable predictions for unknown images, the framework indicates problematic cases for clinical examination. Selective prediction lowers diagnosis errors, especially with high confidence, and increases system safety.

Interaction between these components affects performance. Fairness calibration and Bayesian uncertainty estimation for risk stratification and OOD detection ensure confidence estimation is the same across demographic groups. These modules are meant to work together to improve dependability, strength, and clinical usefulness without compromising diagnostic accuracy.

Experimental results show that the suggested paradigm can solve clinical deployment obstacles beyond performance optimization. The model has superior calibration, fairness, robustness, and clinical decision support compared with prior ViT–GNN models and somewhat greater classification accuracy. These traits make the automated skin cancer diagnosis system more accurate.

5.2. Future research

The authors of the paper admit that the suggested systems were characterized by a high level of trust and equitability. However, they outline some possible research paths in the future.

1) Multimodal Data Integration: Future studies may enhance diagnostic accuracy by adding clinical metadata, patient historical data, and histopathology data.
2) Real-Time Clinical Validation: The capability of the systems suggested in this document to be used in a real-time environment should also be tested in a clinical environment to assess their clinical integration and operational shortcomings.
3) Lightweight Model Deployment: Knowledge distillation and model compression are two methods that could enable these systems to run at the edge or on mobile devices to triage communities and screen them.
4) The systems could also be improved by means of continuous learning and domain adaptation: The systems can evolve by the creation of adaptive learning algorithms upon which the system can adapt to the patterns of skin lesions and demographics of a specific population.
5) Explainable AI and Uncertainty: The authors propose that causal explainability in conjunction with uncertainty estimation will enhance the usefulness and usability of diagnostic support to clinicians.

The key objectives of these research directions are to enhance the fairness, safety, and clinical value of AI systems in skin cancer screening.

Ethical Statement

This study did not require formal ethical approval because India does not require IRB/ethics committee approval for research using publicly available, anonymized secondary data. This exemption is based on the National Ethical Guidelines for Biomedical and Health Research Involving Human Participants (2017), issued by the Indian Council of Medical Research.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/nour12347653/skin-disease-detection-dataset-ham10000-isic.

Author Contribution Statement

Aswani Dogga: Conceptualization, Methodology, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Sivasubramanian R.:** Software, Investigation, Resources, Supervision. **Shanthi S.:** Project administration.

References

[1] Sławińska, M., Żółkiewicz, J., Behera, B., Ding, D. D., Lallas, A., Chauhan, P., . . . , & Errichetti, E. (2023). Dermoscopy of inflammatory dermatoses (inflammoscopy) in skin of color–A systematic review by the International Dermoscopy Society "imaging in skin of color" task force. *Dermatology Practical & Conceptual*, *13*(4 Suppl 1), 1–73. https://doi.org/10.5826/dpc.1304S1a297S

[2] Alsararatee, H. H. (2025). Diagnostic tools and methods for dermatological assessment. *British Journal of Nursing*, *34*(10), 502–512. https://doi.org/10.12968/bjon.2024.0383

[3] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, *54*(10s), 1–41. https://doi.org/10.1145/3505244

[4] Hassija, V., Palanisamy, B., Chatterjee, A., Mandal, A., Chakraborty, D., Pandey, A., . . . , & Kumar, D. (2025). Transformers for vision: A survey on innovative methods for computer vision. *IEEE Access*, *13*, 95496–95523. https://doi.org/10.1109/ACCESS.2025.3571735

[5] Dogga, A., R, S., & S, S. (2026). Hybrid vision transformer and graph neural network model with region-adaptive attention for enhanced skin cancer prediction. *Scientific Reports*, *16*(1), 1–22. https://doi.org/10.1038/s41598-025-32502-z

[6] Niranjan, K. R., Rangahanumaiah, N., BN, B. U. M., & Anantharamaiah, V. (2025). A hybrid vision transformer and graph neural network model with attention mechanisms for diabetic retinopathy detection. *Journal of Artificial Intelligence and Technology*, *5*, 230–241. https://doi.org/10.37965/jait.2025.0645

[7] Chen, C., Wu, Y., Dai, Q., Zhou, H. Y., Xu, M., Yang, S., . . . , & Yu, Y. (2024). A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 10297–10318. https://doi.org/10.1109/TPAMI.2024.3445463

[8] Xu, H., & Wu, Y. (2024). G2ViT: Graph neural network-guided vision transformer enhanced network for retinal vessel and coronary angiograph segmentation. *Neural Networks*, *176*, 106356. https://doi.org/10.1016/j.neunet.2024.106356

[9] Lane, R. O. (2026). A comprehensive review of classifier probability calibration metrics. *Data Science*, *9*(1), 1–61. https://doi.org/10.1177/24518492261436231

[10] Lalor, J. P., Abbasi, A., Oketch, K., Yang, Y., & Forsgren, N. (2024). Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines. *ACM Transactions on Information Systems*, *42*(4), 1–41. https://doi.org/10.1145/3641276

[11] Faghani, S., Moassefi, M., Rouzrokh, P., Khosravi, B., Baffour, F. I., Ringler, M. D., & Erickson, B. J. (2023). Quantifying uncertainty in deep learning of radiologic images. *Radiology*, *308*(2), 1–10. https://doi.org/10.1148/radiol.222217

[12] He, W., Jiang, Z., Xiao, T., Xu, Z., & Li, Y. (2026). A survey on uncertainty quantification methods for deep learning. *ACM Computing Surveys*, *58*(7), 1–35. https://doi.org/10.1145/3786319

[13] Sabry, M., Balaha, H. M., Ali, K. M., Mahmoud, A., Gondim, D., Ghazal, M., . . . , & El-Baz, A. (2026). AI-driven breast cancer diagnosis: A systematic review of imaging modalities, deep learning, and explainability. *Cancers*, *18*(8), 1–45. https://doi.org/10.3390/cancers18081305

[14] Singh, N. H., Kose, U., & Gochhayat, S. P. (2024). *Deep learning in biomedical signal and medical imaging*. USA: CRC Press.

[15] Baabu, K. B., & Raja, S. D. M. (2025). Skin cancer detection using vision transformer with GNN and explainable AI (Grad-CAM). In *2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science*, 1–6. https://doi.org/10.1109/AMATHE65477.2025.11080709

[16] Banerjee, T., Chhabra, P., Kumar, M., Kumar, A., Abhishek, K., & Shafeeq, B. A. (2026). Hypergraph-based contrastive embedding and attention fusion for detection of skin cancer. *Scientific Reports*, *16*(1), 12808. https://doi.org/10.1038/s41598-026-43351-9

[17] Sikha, O. K., Riera-Marin, M., Galdran, A., Lopez, J. G., Rodriguez-Comas, J., Piella, G., & Ballester, M. A. G. (2025). Uncertainty-aware segmentation quality prediction via deep learning Bayesian modeling: Comprehensive evaluation and interpretation on skin cancer and liver segmentation. *Computerized Medical Imaging and Graphics*, *123*, 1–12. https://doi.org/10.1016/j.compmedimag.2025.102547

[18] Shamsi, A., Asgharnezhad, H., Bouchani, Z., Jahanian, K., Saberi, M., Wang, X., . . . , & Alinejad-Rokny, H. (2023). A novel uncertainty-aware deep learning technique with an application on skin cancer diagnosis. *Neural Computing and Applications*, *35*(30), 22179–22188. https://doi.org/10.1007/s00521-023-08930-1

[19] Goyal, M., Knackstedt, T., Yan, S., & Hassanpour, S. (2020). Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine*, *127*, 1–11. https://doi.org/10.1016/j.compbiomed.2020.104065

[20] Wen, D., Soltan, A., Trucco, E., & Matin, R. N. (2024). From data to diagnosis: Skin cancer image datasets for artificial intelligence. *Clinical and Experimental Dermatology*, *49*(7), 675–685. https://doi.org/10.1093/ced/llae112

[21] Behara, K., & Behara, S. (2025). Explainable artificial intelligence in dermatology: Overview of data scarcity for skin cancer detection in black skin. In P. Naidoo & M. Sibanda (Eds.), *Measurement, automation, and control in artificial intelligence and machine learning* (pp. 267–291). Springer Cham. https://doi.org/10.1007/978-3-031-91802-5_12

[22] Chu, Y., Oh, B., & Yang, S. (2025). Evaluating the trustworthiness of foundation models for skin lesion segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 562–572. https://doi.org/10.1109/ICCVW69036.2025.00065

[23] Jain, A., & Kulkarni, A. (2024). Trustworthy AI for dermatological disease diagnosis: Bridging the gap between accuracy and trustworthiness. In *International Conference on Data Science and Applications*, 75–88. https://doi.org/10.1007/978-981-96-2179-8_6

[24] Gui, H., Omiye, J. A., Chang, C. T., & Daneshjou, R. (2024). The promises and perils of foundation models in dermatology. *Journal of Investigative Dermatology*, *144*(7), 1440–1448. https://doi.org/10.1016/j.jid.2023.12.019

[25] Hashjin, N. M., Amiri, M. H., & Najafabadi, M. K. (2026). DermaGPT a federated multimodal framework with a meta learned trust function for interpretable dermatology diagnostics. *Scientific Reports*, *16*(1), 7959. https://doi.org/10.1038/s41598-026-38715-0

[26] Alzamili, A. H., & Ruhaiyem, N. I. R. (2025). A comprehensive review of deep learning and machine learning techniques for early-stage skin cancer detection: Challenges and research gaps. *Journal of Intelligent Systems*, *34*(1), 1–35. https://doi.org/10.1515/jisys-2024-0381

[27] Bera, A., Wharton, Z., Liu, Y., Bessis, N., & Behera, A. (2022). SR-GNN: Spatial relation-aware graph neural network for fine-grained image categorization. *IEEE Transactions on Image Processing*, *31*, 6017–6031. https://doi.org/10.1109/TIP.2022.3205215

[28] Abbas, M., Arslan, M., Bhatty, R. A., Yousaf, F., Khan, A. A., & Rafay, A. (2024). Enhanced skin disease diagnosis through convolutional neural networks and data augmentation techniques. *Journal of Computing & Biomedical Informatics*, *7*(01), 87–106. https://doi.org/10.56979/701/2024

[29] Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations.

Medical Image Analysis, *75*, 1–15. https://doi.org/10.1016/j.media.2021.102305

[30] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, *5*(1), 1–9. https://doi.org/10.1038/sdata.2018.161

[31] Abdulazeez, A. (2024). A review on utilizing machine learning classification algorithms for skin cancer. *Journal of Applied Science and Technology Trends*, *5*(2), 60–71. https://doi.org/10.38094/jastt52191

[32] Padarian, J., Minasny, B., & McBratney, A. B. (2022). Assessing the uncertainty of deep learning soil spectral models using Monte Carlo dropout. *Geoderma*, *425*, 116063. https://doi.org/10.1016/j.geoderma.2022.116063

[33] Posocco, N., & Bonnefoy, A. (2021). Estimating expected calibration errors. In *International Conference on Artificial Neural Networks*, 139–150. https://doi.org/10.1007/978-3-030-86380-7_12

How to Cite: Dogga, A., R., S., & S., S. (2026). Uncertainty-Aware and Bias-Calibrated Vision Transformer–Graph Neural Network for Reliable Skin Cancer Detection. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA620210397