Digital Health
XX(X):1–20

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Fair and Interpretable Early Melanoma Detection with Demographic-Aware Ensemble Vision Transformers and Gradient-Based XAI Methods

Mohsin Akram[1]* **Jameel Ahmad**[2]†

Abstract

Background
Melanoma is an aggressive form of skin cancer, the early detection of which is critical to increasing patient survival. Although CNNs have achieved impressive results in automated diagnosis, their restricted capacity to model long-range dependencies impairs the overall effectiveness. Vision Transformers (ViTs) work around these limitations by representing global context, though the existing literature rarely explores their ensembling, fairness across diverse skin tones, or clinical interpretability.

Methods
We present the *Ensemble Vision Transformer (E-ViT)*, consisting of four fine-tuned ViT models (ViT-B16, ViT-B32, ViT-L16 and ViT-L32) combined with a soft-voting schema. The model was trained and evaluated on a merged dataset using the ISIC-2020 dermoscopic images (33,126) and Fitzpatrick17k clinical images (16,577). Following preprocessing and curation, the ultimate dataset included 14,921 images that were balanced between benign and malignant cases and proportionate across Fitzpatrick skin types I-VI. The performance was evaluated in terms of accuracy, AUC, sensitivity, specificity, and F1-score. The intepretability of interpretability was addressed with Grad-CAM, Grad-CAM++, and Saliency Map approachesto increase the clinical trust.

Results
The presented E-ViT outperformed the current best classification scores with an AUC of **0.952** and accuracy of **93.8%**, outperforming standalone ViTs as well as traditional CNN based baselines. Fairness analysis revealed high overall diagnostic performance, although with diminished sensitivity for underrepresented darker skin types (Fitzpatrick V–VI), highlighting the benefits of a demographically balanced dataset. Model interpretability approaches identified lesion-specific areas, enhancing the interpretability of the model.

Conclusion
The present study established the supremacy of transformer-based ensembles in automated Melanoma diagnosis. While considering simultaneously accuracy, fairness and interpretability, we have proposed the E-ViT system, which is both reliable and clinically relevant as a decision-support tool that can achieve earlier and fairer melanoma diagnosis in practice health care settings.

Keywords
Melanoma, Vision Transformers, Ensemble Learning, Skin Cancer, Fairness, Explainable AI, Grad-CAM, Grad-CAM++, Saliency Maps

Introduction

Skin cancer is, worldwide one of the major health concerns and among them melanoma is considered to be the most fatal and severe. It arises from melanocytes of the epidermis and, when left untreated, may metastasize early and result in a high mortality rate. The Global Cancer Observatory indicates an estimated 331,722 cases of melanoma were diagnosed worldwide in 2022 and approximately 58,667 died. In the U.S. alone, approximately 97610 new cases of invasive was projected to be diagnosed in 2023 and nearly 7990 deaths were expected to be related with an estimate surpassing 100000 new diagnoses and roughly 8000 associated deaths for an year-2024 projection. Early and accurate diagnosis is, therefore, important for patient survival.

Dermoscopy and biopsy are the conventional diagnostic techniques. However, the visual similarity between malignant melanoma and benign lesions makes discrimination difficult, with even expert dermatologists achieving more than 80% of accuracy [1–3]. Such a shortcoming has led to the introduction of Computer-Assisted Diagnosis (CAD) systems to help physicians to address it [4]. In the past, CAD models were crafted based on features by hand, and recently they focus largely on deep learning (DL). Among them, Convolutional Neural Networks (CNNs) have made a major breakthrough in medical imaging, and its classification for melanoma has been dominated [5,6]. However, CNNs also suffers from the inherent problem that their local receptive fields

Corresponding author:
Mohsin Akram

limit the connectivity and it is difficult to capture long range dependencies for distinguishing minute differences between benign and malignant lesions[7].

Vision Transformers (ViTs) have recently risen as an attractive alternative for image interpretation. With the use of patch-wise decomposition and self-attention mechanisms, ViTs can capture not only detailed local characteristics but also global relationships. These features are highly appropriate for detecting melanoma, where the pattern of details in the images can determine how a clinical case ends. However, three challenges still cannot be resolved through the current studies.

- **Lack of ensemble-based ViT frameworks:** Most prior studies have focused on individual ViT variants or CNN-based ensembles, with limited exploration of ensembles of multiple ViTs to exploit their complementary strengths.
- **Limited fairness evaluation:** Existing models rarely investigate performance disparities across diverse skin tones, despite well-documented diagnostic inequities for darker skin types.
- **Insufficient interpretability:** While ViTs improve accuracy, their black-box nature limits clinical adoption without transparent and explainable predictions.

To address these gaps, this study proposes an ensemble Vision Transformer (E-ViT) framework for interpretable and fair melanoma detection. Specifically, we fine-tune multiple pre-trained ViT architectures and integrate them through soft-voting to enhance classification robustness. By combining the ISIC-2020 dermoscopic dataset with the Fitzpatrick17k dataset, we not only improve performance but also evaluate fairness across Fitzpatrick skin types I–VI. Furthermore, we incorporate gradient-based Explainable AI (XAI) methods to enhance clinical trust and interpretability of model predictions. The core contributions of this study are as follows:

1. **Development of an Ensemble Vision Transformer (E-ViT) Framework:** We fine-tuned four pre-trained ViT architectures (ViT-B16, ViT-B32, ViT-L16, and ViT-L32) and integrated them using a soft-voting ensemble strategy. This approach leverages complementary strengths of the individual ViTs to achieve robust melanoma classification performance.
2. **Integration of Demographic Fairness Evaluation:** By incorporating the Fitzpatrick17k dataset alongside ISIC-2020, the study evaluates classification fairness across diverse skin tones (Fitzpatrick I–VI), identifying disparities in sensitivity for underrepresented groups and highlighting the importance of equitable AI in dermatology.
3. **Explainable AI for Clinical Interpretability:** Gradient-based XAI methods (like Grad-CAM, Grad-CAM++, and Saliency Maps) were employed to provide visual explanations of model predictions, enhancing clinical trustworthiness and supporting dermatologist decision-making.

Literature Review

Recent research has increasingly focused on deep learning (DL) for automated melanoma detection. This section reviews the latest developments, with emphasis on convolutional neural networks (CNNs), Vision Transformers (ViTs), fairness-aware models, and explainable AI (XAI) techniques.

Skin Cancer Detection using CNNs

Over the past few years, CNNs have been widely adopted for skin lesion classification. Mahbod et al. (2019)[8] employed an ensemble of multiple CNN architectures, reporting an ROC score of 87.3% on ISIC-2017. More recent works further improved classification using deeper and hybrid CNN architectures. For example, Ghosh et al. (2024)[9] combined Random Forest, Logistic Regression, kNN, XGBoost, and CNNs into an ensemble model, achieving 91.6% accuracy. Similarly, multi-branch CNNs have been explored to capture lesion texture and color variations, delivering competitive performance across ISIC datasets. Despite their success, CNNs face inherent limitations in modeling long-range dependencies, which are essential for melanoma analysis.

Vision Transformers in Skin Lesion Analysis

Transformers emerged recently for medical image analysis due to their strong ability of modelling the global context. Zhang et al. (2022)[10] introduced the transformer-based method TFormer which combines dermoscopic images with textual information, and achieves 77.9% accuracy. Yang et al. (2023)[11] compared the ViT-Base and ViT-Large on HAM10000 and Dermofit, with 94% and 80.5% accuracy respectively. Flosdorf et al.[12] confirmed the performance of ViT-L16 and ViT-L32 in detecting melanoma, with accuracies of 92% and 91%. In a similar vein, Xin et al. (2022)[13] utilized contrastive learning with transformers for multi-class lesion classification, highlighting the promise of self-supervised ViTs. Together, they illustrate the benefit of ViTs over CNNs—although most focus on monolithic models rather than ensembles.

Fairness and Demographic Bias in Skin Cancer Detection

Fairness in AI-driven dermatology has emerged as a pressing concern. Existing studies show that models trained predominantly on lighter-skin datasets often underperform on darker skin tones. Marchetti et al. (2021)[14] and Adamson et al. (2022)[15] reported disparities in sensitivity for Fitzpatrick skin types V–VI, underscoring risks of inequitable AI deployment. More recent surveys of dermatology AI systems have reinforced these findings, highlighting the lack of demographic diversity in training datasets and the urgent need for fairness-aware evaluation protocols. Despite these concerns, only limited research explicitly evaluates ViTs across diverse demographic subgroups, leaving an important gap in fairness assessment for transformer-based models.

Explainable AI for Melanoma Classification

The adoption of AI in clinical workflows requires transparency and interpretability to gain dermatologist trust. Gradient-based XAI methods such as Grad-CAM and Grad-CAM++ have been widely applied in dermatology to highlight lesion regions relevant for classification. For instance, Yang et al. (2023)[11] combined ViTs with Grad-CAM to visualize decision-making, improving clinician interpretability. Beyond Grad-CAM, saliency-based techniques have also been explored for skin lesion analysis, such as the work of Ardila et al. (2021)[16], which integrated ensembles of visual explanations to enhance reliability. In addition, transformer-specific methods such as attention rollout have been proposed, notably by Chefer et al. (2021)[17], extending interpretability beyond standard attention maps. Despite these advancements, the integration of diverse XAI strategies into ensemble ViT frameworks for melanoma detection remains limited.

Summary of Gaps

In summary, recent studies have demonstrated: (i) CNNs' strong performance but limited ability to capture long-range dependencies, (ii) ViTs' superior contextual modeling but underexplored ensemble strategies, (iii) insufficient fairness evaluation across Fitzpatrick skin types, and (iv) limited integration of XAI with transformer-based ensembles. These gaps motivate our proposed *Ensemble Vision Transformer (E-ViT)* framework, which combines multiple ViTs, evaluates fairness across diverse skin tones, and employs gradient-based XAI for interpretable melanoma detection. A comparative summary of recent studies on binary skin cancer classification using CNNs, ensembles, and transformer-based models is presented in Table 1, highlighting both the progress achieved and the remaining gaps in performance, fairness, and interpretability.

Methodology

This study proposes an ensemble Vision Transformer (E-ViT) framework for binary melanoma classification, integrating multiple pretrained ViT architectures with fairness evaluation and explainable AI methods. The overall methodology consists of the following major stages:

1. **Data Preparation and Preprocessing:** ISIC-2020 and Fitzpatrick17k dermoscopic clinical raw images are fetched for the binary distinguishing between benign and malignant lesions. Some preprocessing techniques such as resize, normalization and data augmentation are utilized to improve the visibility of lesions and to solve the class imbalance problems.
2. **Fine-Tuning Pretrained ViT Models:** We fine tune four pretrained Vision Transformer models (ViT-B16, ViT-B32, ViT-L16, and ViT-L32) on the curated datasets. To address this challenge, transfer learning is utilized to adjust these well learned large scale pretrained models for melanoma classification.
3. **Soft-Voting Ensemble Strategy:** The outputs of the four fine-tuned ViTs are combined using a soft-voting scheme. This ensemble approach leverages the complementary strengths of different ViT architectures to improve robustness and predictive accuracy.
4. **Fairness Evaluation across Skin Tones:** Model performance is stratified by Fitzpatrick skin types I–VI to evaluate fairness across diverse demographic groups. Sensitivity and specificity are reported for each subgroup to identify disparities in diagnostic performance.
5. **Explainable AI Integration:** Gradient-based XAI methods, including Grad-CAM, Grad-CAM++, and Saliency Maps, are applied to the ensemble predictions. These visual explanations highlight lesion-relevant regions, enhancing interpretability and clinical trust.

Figure 1 illustrates the proposed research methodology.

Datasets Used

This study utilizes two publicly available benchmark datasets: ISIC-2020 and Fitzpatrick17k, which together provide a diverse collection of dermoscopic and clinical skin lesion images.

ISIC-2020 Dataset The ISIC-2020 Challenge dataset[28] is one of the largest and most widely used dermoscopic image collections for skin cancer detection. It contains a total of 33,126 images, including 584 melanoma and 32,542 benign cases, making it highly imbalanced. For our study, we curated a balanced subset of 10,982 images, consisting of 5,592 benign and 5,390 malignant lesions. Table 2 summarizes the dataset composition, while Fig. 2(a) illustrates representative dermoscopic images.

Fitzpatrick17k Dataset To address the lack of demographic diversity in ISIC datasets, we incorporated the Fitzpatrick17k dataset[29], which consists of 16,577 clinical photographs spanning 114 dermatological conditions. For consistency with our binary classification task, we excluded non-neoplastic conditions, retaining 4,497 images (2,234 benign and 2,263 malignant). Each image is annotated with a Fitzpatrick skin type (I–VI), enabling subgroup fairness evaluation across different skin tones. The detailed distribution across benign, malignant, and non-neoplastic conditions by skin type is provided in Table 3, and representative examples are shown in Fig. 2(b).

Combined Dataset and Splits After combining ISIC-2020 and Fitzpatrick17k, the final dataset comprised 14,921 images (7,555 benign and 7,366 malignant), ensuring a balanced binary classification setup with demographic coverage across skin types. The data were divided into training (70%), validation (15%), and testing (15%) sets with stratification to preserve class and skin-type distributions. The split is summarized in Table 4.

HAM10000 Dataset Although HAM10000[30] is one of the most frequently used datasets for melanoma classification, we did not include it in this work for two main reasons:

Table 1. Summary of literature on melanoma skin cancer classification models

Model	Dataset	Accuracy / AUC	Classes	Reference
DCNN	HAM-10000	91.93%	Two	18
Inception-V3	ISIC-2019/2020	86.9%	Two	19
IRv2+Soft-Attention	ISIC-2017 Archive	91.6%	Two	20
DenseNet-201	ISIC-2019	70.08%	Two	21
Deep-CNN	ISIC-2016, 2017, 2020	90.42%	Two	22
CNN	ISIC-2020	91.61%	Two	23
Stacking CV-Xception	ISIC-Archive	90.9%	Two	24
Two Hybrid-CNN	ISIC-2016	88.02%	Two	25
EfficientNet-B6	ISIC-2019/2020	AUC = 0.9681	Two	26
Spiking-Vgg-13, CNN	ISIC-2019	89.5%	Two	27

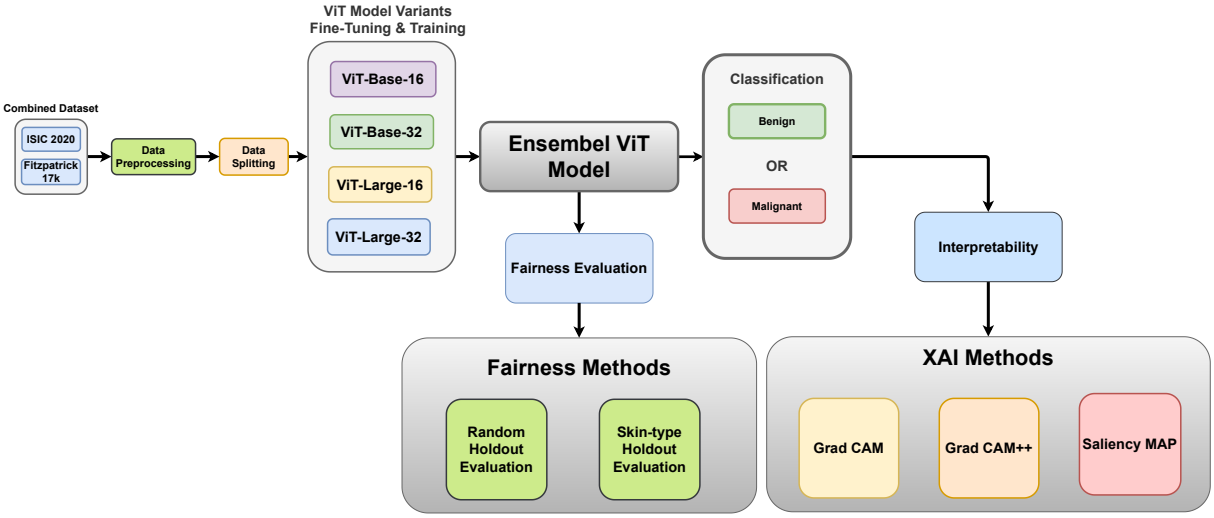


Figure 1. Proposed Research Methodology Framework.

(1) our focus was on binary classification using ISIC-2020 and Fitzpatrick17k, which together offer a larger and more balanced dataset (14,921 cases) across diverse skin tones; and (2) ISIC-2020 already provides a large collection of dermoscopic images, while Fitzpatrick17k contributes demographic diversity that HAM10000 lacks. However, we recognize HAM10000 as a valuable benchmark and suggest it for future external validation studies to enhance generalizability.

Transfer Learning

Deep learning models have demonstrated remarkable performance in medical imaging tasks; however, training them from scratch requires very large datasets, extensive computational resources, and careful hyperparameter tuning. These challenges often limit their applicability in domains such as dermatology, where annotated datasets are relatively small and imbalanced.

Transfer learning provides an effective solution by reusing weights from large-scale pretrained models and adapting them to specific tasks. In this study, multiple Vision Transformer (ViT) variants pretrained on the ImageNet dataset were fine-tuned for binary skin cancer classification. This strategy allowed the models to leverage generalized feature representations learned from natural images while adapting to the unique texture, color, and structural characteristics of dermoscopic and clinical lesion images.

Fine-tuning was performed on four ViT architectures: ViT-B16, ViT-B32, ViT-L16, and ViT-L32. These variants differ in patch size (16 vs. 32) and model depth (Base vs. Large), enabling the ensemble to capture complementary feature representations. Transfer learning not only reduced training time and improved convergence but also enhanced generalization performance across diverse skin tones.

The conceptual workflow of transfer learning with Vision Transformers is illustrated in Figure 3.

(a) ISIC 2020 **(b)** Fitzpatrick 17k

Figure 2. Sample images from the combined dataset: (a) Dermoscopic images from ISIC-2020 and (b) Clinical photographs from Fitzpatrick17k.

Table 2. Summary of the **ISIC-2020 Challenge dataset**, including the distribution of melanoma and benign cases. The dataset is highly imbalanced, with melanoma representing only 1.76% of total cases.

Category	Description	Counts	Notes
Total Images	Dermoscopic lesion images	33,126	JPG, varied resolution
Melanoma Cases	Malignant (positive class)	584	∼1.76% of dataset
Benign Cases	Non-melanoma (negative class)	32,542	Includes various benign conditions
Demographics	Patient origin	–	Multiple regions (Australia, Europe, etc.)
Task	Classification	2 classes	Melanoma vs. Non-Melanoma

Table 3. Distribution of **Fitzpatrick17k dataset** by Fitzpatrick skin type and high-level skin condition categories. Only benign and malignant neoplastic conditions were retained for this study, enabling fairness evaluation across skin tones.

Skin Type	Non-Neoplastic (12,080)	Benign (2,234)	Malignant (2,263)
Type I	17.0%	19.9%	20.2%
Type II	28.1%	30.0%	32.8%
Type III	19.7%	21.2%	20.2%
Type IV	17.5%	16.4%	13.3%
Type V	10.1%	7.1%	6.5%
Type VI	4.4%	2.0%	2.7%
Unknown	3.2%	3.3%	4.6%

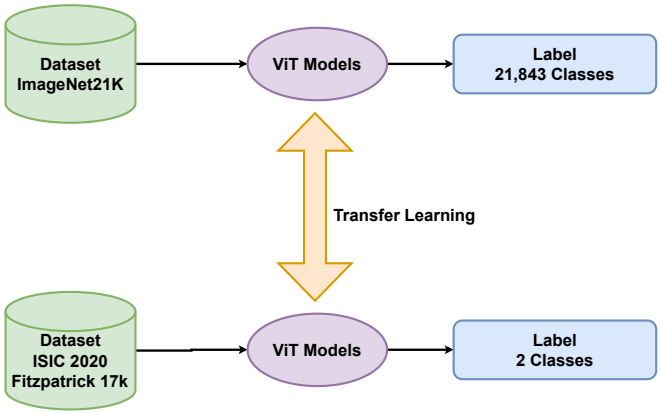


Figure 3. Conceptual illustration of transfer learning applied to Vision Transformers for melanoma classification. Pretrained ImageNet weights are fine-tuned on dermoscopic and clinical images to adapt to the task of binary skin cancer detection.

Vision Transformer Model Architecture

Unlike conventional convolutional neural networks (CNNs), Vision Transformers (ViTs) treat an image as a sequence of patches rather than a 2D grid of pixels. This sequential representation allows ViTs to capture long-range

Table 4. Final dataset distribution after combining ISIC-2020 and Fitzpatrick17k. Data were split into training, validation, and testing sets (70/15/15) with stratification across class labels and skin types.

Class	Training	Validation	Testing
Benign	5,288	1,133	1,134
Malignant	5,156	1,104	1,106
Total	**10,444**	**2,237**	**2,240**

dependencies and global contextual information, which is particularly beneficial in medical imaging tasks such as melanoma detection, where subtle visual differences distinguish benign from malignant lesions. The overall workflow of a basic ViT model for classification is illustrated in Figure 4.

Patch Embedding Block

The input image $X \in \mathbb{R}^{H\times W\times C}$ is partitioned into non-overlapping patches of size $P \times P$. The global number of patches is determined as:

$$N = \frac{HW}{P^2}. \tag{1}$$

Collapsing each patch to a single vector yields inputs where the token dimension is the same for all of them. Learnable 1D positional embeddings are added to the patch embeddings in order to preserve spatial information before they go through the transformer encoder.

Transformer Encoder Block

The transformer encoder E is composed of a sequence of L layers, each of which has two key components: (i) Multi-Head Self-Attention (MHSA) and (ii) a Multi-Layer Perceptron (MLP). Both parts are surrounded by residual connections and layer normalization for the purpose of training stabilization[31,32]. The encoder block is depicted in Figure 5.

The self-attention module operates on queries (Q), keys (K), and values (V) of the input sequence. The scaled dot product attention is formulated as:

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

where d_k is the dimension of keys. Multi-head attention generalizes this by calculating the attention h times with distinct learnable projections and concatenating their outputs:

$$\begin{aligned}\text{MultiHead}(Q,K,V) &= \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O.\\ \text{where}\quad \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V).\end{aligned} \tag{3}$$

An illustration for the multi-head self-attention mechanism is given in Figure 6.

Multi-Layer Perceptron (MLP)

There is also an MLP between every two consecutive encoder blocks with two fully connected layers and a ReLU activation. Dropout is used for regularization to avoid overfitting. In the presented E-ViT, the last classification layer is activated by sigmoid for binary classification (benign vs. malignant).

Proposed E-ViT Framework

To capture the complementary potential between ViT models, we fine-tuned four pretrained variants: ViT-Base-16, ViT-Base-32, ViT-Large-16, and ViT-Large-32. These were pre-trained on ImageNet-21k + ImageNet-1k and fine-tuned for dermoscopic and clinical skin lesion databases. The architectures of these models are shown in Figures 7–10.

Ensemble Learning Strategy

Finally, the predictions from four ViT models are ensemble by soft-voting. On each input image, each model gives a probability distribution over both classes. The final decision is the composition of the predicted probability average from all models (Eq. 4):

$$y = \arg\max_i \left\{ \frac{1}{N}\sum_{j=1}^{N} p_{ij} \right\}, \tag{4}$$

where p_{ij} is the probability assigned by the j^{th} model to class i, and N the number of classifiers. This ensemble method increases the robustness, reduces the variance and improves overall classification accuracy as compared to individual models.

The overview of the proposed ensemble framework is on a high level as illustrated in Figure 11, which consolidates the predictions of ViT-B16, ViT-B32, ViT-L16 and ViT-L32 via soft voting to obtain the final classification.

Model Fairness Evaluation

To ensure equitable performance of the proposed ensemble ViT framework across diverse populations, we designed fairness evaluation experiments with respect to Fitzpatrick skin types. The objective was to determine whether predictive performance remained consistent across skin tones and to identify potential biases that may undermine clinical reliability.

Evaluation Strategies

Two complementary experimental designs were adopted:

1. **Random holdout evaluation:** The dataset was partitioned into an 80/20 train–test split, stratified by lesion class (benign vs. malignant). This setting provided a balanced baseline for assessing overall model performance.
2. **Skin-type holdout evaluation:** Following the methodology of Groh et al.[29], subgroup fairness

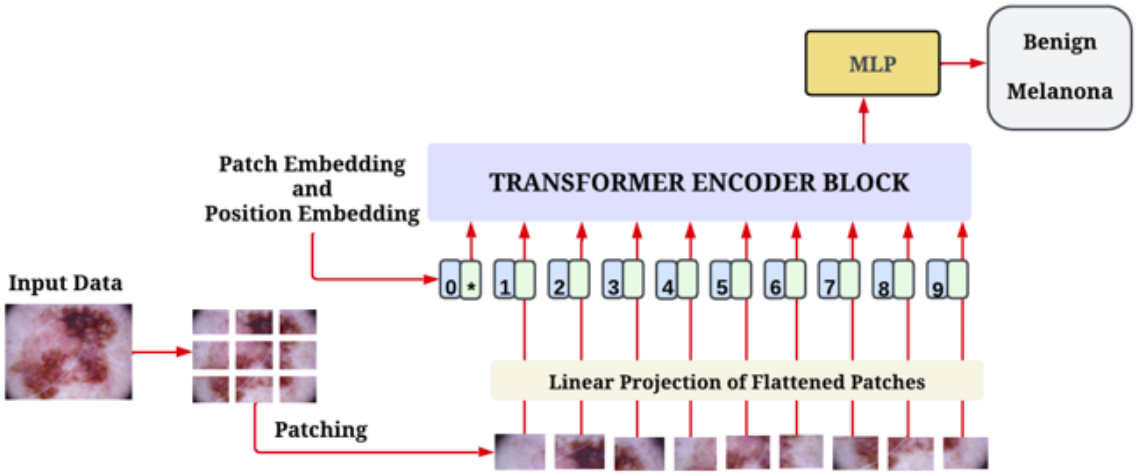


Figure 4. General workflow of Vision Transformers (ViTs) for image classification.

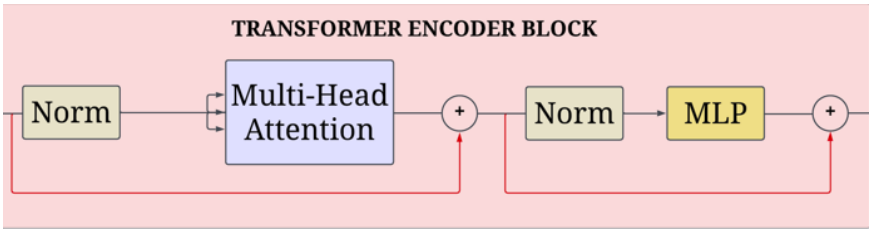


Figure 5. Transformer encoder block with multi-head self-attention and feed-forward layers with residual connections.

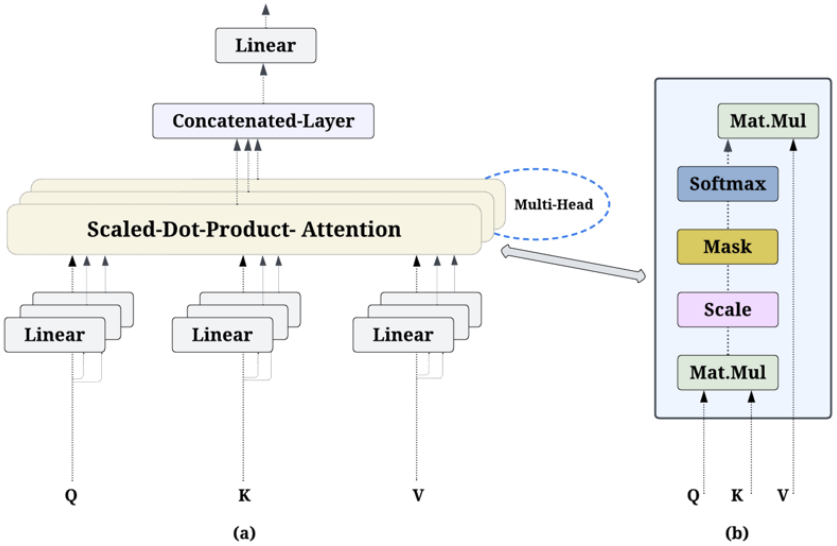


Figure 6. Vision Transformers multi-head self-attention mechanism.

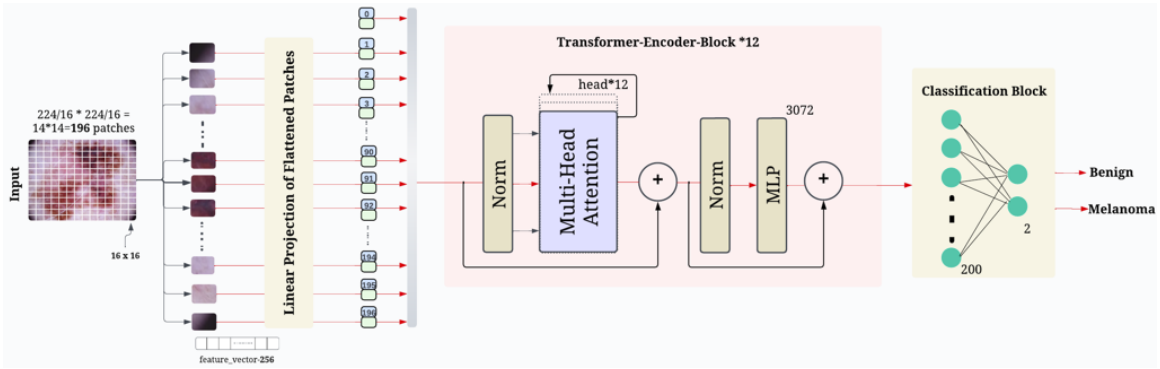


Figure 7. Proposed ViT-Base-16 architecture.

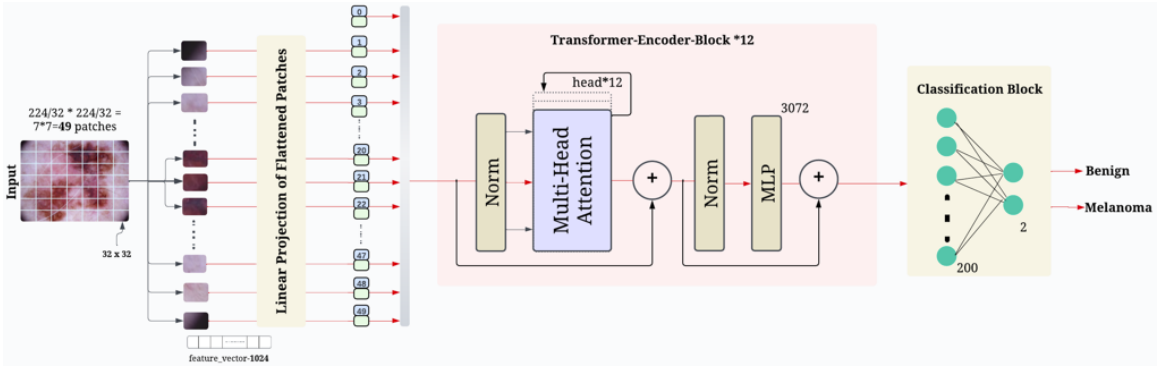


Figure 8. Proposed ViT-Base-32 architecture.

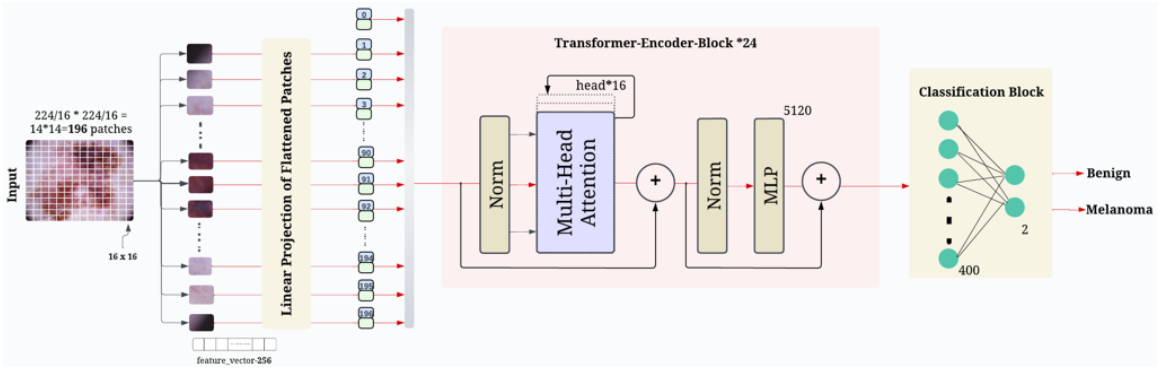


Figure 9. Proposed ViT-Large-16 architecture.

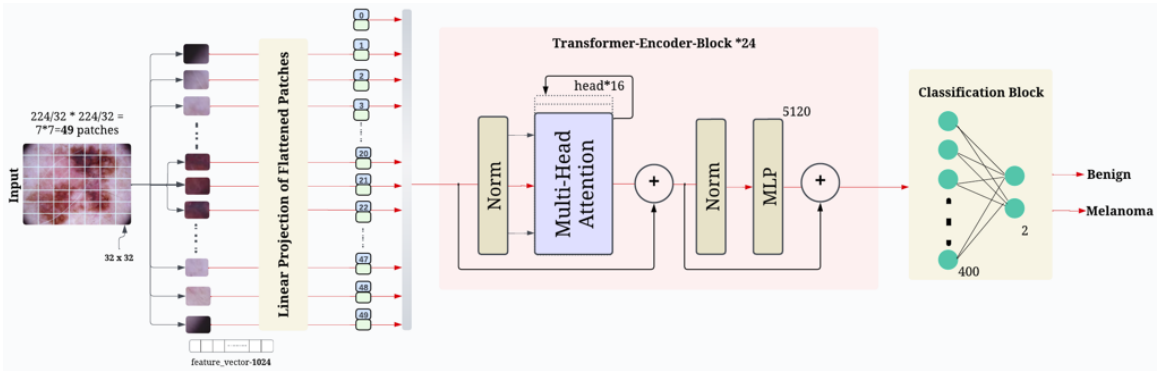


Figure 10. Proposed ViT-Large-32 architecture.

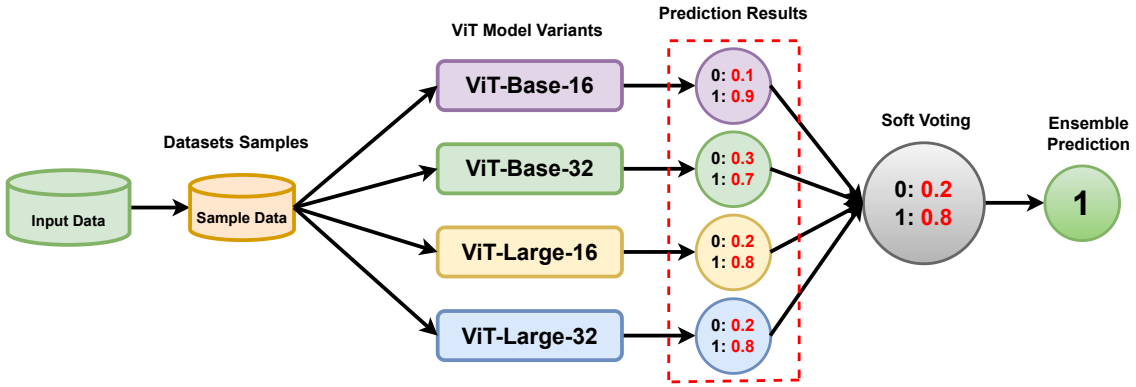


Figure 11. Proposed E-ViT system. The outputs of four fine-tuned ViT backbone's (ViT-B16, ViT-B32, ViT-L16 and ViT-L32) are fused in a soft-voting way to achieve robustness and enhance the overall classification performance.

experiments were conducted by training on specific Fitzpatrick subsets (Types 1–2, Types 3–4, or Types 5–6) and testing on the remaining types. This procedure was designed to evaluate cross-group generalization and highlight disparities in model performance across skin tones.

Fairness Metrics

We evaluated performance with common binary classification metrics, and focused on sub-group level differences:

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN}, \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{7}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

We also used the following metrics to evaluate performance in a threshold-invariant manner:

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}), \tag{9}$$

$$\text{AUC-PR} = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}). \tag{10}$$

Furthermore, the **False Negative Rate (FNR) for each Fitzpatrick subgroup** was computed individually, as the failure to detect melanoma is of highest clinical concern:

$$\text{FNR} = \frac{FN}{FN + TP}. \tag{11}$$

Equalized Odds Criterion

To quantify fairness, we employed the equalized odds framework, which requires equal false positive rates (FPR) and false negative rates (FNR) across demographic subgroups:

$$P(\hat{Y} = 1 \mid Y = 1, A = a) = P(\hat{Y} = 1 \mid Y = 1, A = b), \tag{12}$$

$$P(\hat{Y} = 1 \mid Y = 0, A = a) = P(\hat{Y} = 1 \mid Y = 0, A = b), \tag{13}$$

where A denotes the sensitive attribute (Fitzpatrick skin type), Y the true label, and $\hat{Y}$ the predicted label.

Explainable AI Framework

With the aim of enhancing interpretability and trustworthiness in AI-empowered dermatology, we introduced the Explainable AI (XAI) paradigm into ensemble Vision Transformer (E-ViT) pipeline. The main objective of the proposed framework is to interpretably justify the model outputs, thereby aiding clinical acceptance and enhancing reliability for melanoma diagnosis in real life.

Overview of the XAI Integration

The XAI module, on the other hand, is running in parallel with classification. Once the E-ViT model makes a prediction, gradient-based techniques are utilized to generate heatmaps and saliency maps that highlight the most important regions for classification on dermoscopic and clinical images. The incorporation of Grad-CAM, Grad-CAM++, and Saliency Maps are illustrated in Figure 12 to the E-ViT pipeline.

Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM[33] produces the class-specific localization map based on the gradients of the target class backpropagated into the final convolutional layer. The weight of significance α_k^c for feature map k is calculated as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{14}$$

where y^c is the output score of class c, A_{ij}^k is the activation at spatial position (i, j) in feature map k and Z denotes the number of spatial locations. A heat-map is constructed based on the concatenated values, which show where in the input image has biggest impact for decision.

Gradient-weighted Class Activation Mapping++ (Grad-CAM++)

Like RM, Grad-CAM++[34] utilizes higher-order derivatives to enhance localization, by focusing on the contribution of multiple discriminative regions that lead to the final output. The weight value of position (i, j) is given by:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 y^c}{\partial (A_{ij}^k)^2}}{2 \cdot \frac{\partial^2 y^c}{\partial (A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \cdot \frac{\partial^3 y^c}{\partial (A_{ab}^k)^3}}, \tag{15}$$

and the total weight of feature map k can be written as:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc}\, \text{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right). \tag{16}$$

This results in more precise and sharper heatmaps as compared to the original Grad-CAM.

Saliency Maps

Saliency Maps[35] provide pixel-level attribution by computing the gradient of the class score with respect to the input

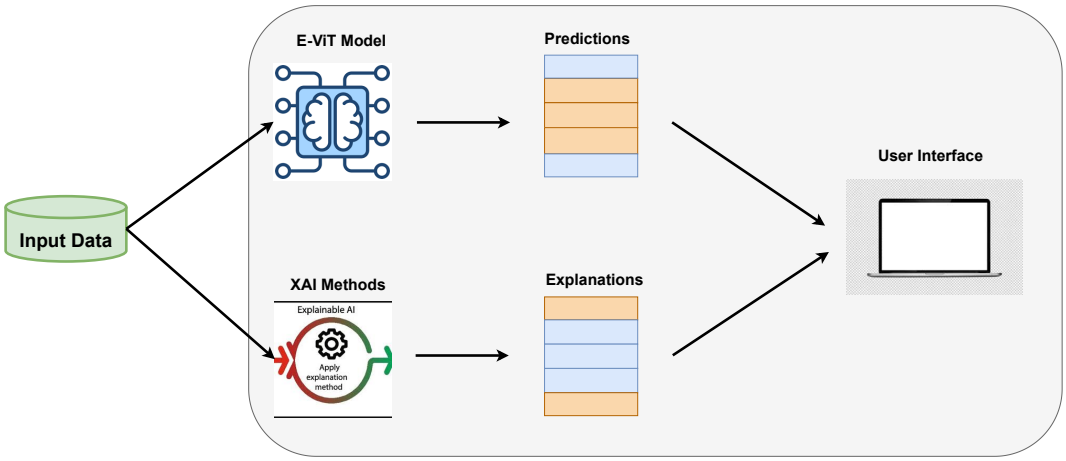


Figure 12. Explainable AI (XAI) in E-ViT workflow. Code-based approaches, including Grad-CAM and Grad-CAM++ as well as Saliency Map methods, produce heatmaps supporting the interpretability to assist skin cancer clinical decisions.

image:

$$S^c(I) = \frac{\partial y^c}{\partial I}, \tag{17}$$

where $S^c(I)$ indicates the sensitivity of the prediction for class c to each pixel in the input image I. This pixel-wise analysis complements Grad-CAM-based methods by offering fine-grained interpretability.

Experimental Setup

All experiments were carried out using Python 3.11.9, supported by commonly used machine learning and scientific computing libraries. TensorFlow 2.16.1 with Keras 3.3.3 served as the main deep learning framework, while PyTorch 2.3.0 was additionally utilized for visualization tasks. Scikit-learn 1.5.2 supported preprocessing, evaluation, and fairness analysis. NumPy 1.26.4 and Pandas 2.2.3 were used for numerical operations and dataset handling. For visualizations, Matplotlib 3.9.2 and Seaborn 0.13.2 were employed. To ensure reproducibility, all experiments were performed in a controlled setting with fixed random seeds for shuffling and model initialization.

Model training and evaluation were conducted on a workstation with high-performance GPUs to accommodate the computational requirements of Vision Transformers. A summary of the hardware specifications is provided in Table 5. Hyperparameter choices and model configurations are presented in Tables 6 and 7, respectively.

Results and Discussion

In this section, we conduct experimental analysis of our proposed E-ViT model for melanoma diagnosis. We structure the results as follows: first, we describe the preprocessing pipeline applied to dermoscopic images, followed by details of model training and fine-tuning. We then provide performance results of individual ViT variants and the proposed ensemble model. Finally, we analyze fairness across skin types, demonstrate interpretability with explainable AI (XAI) methods, and compare our results with state-of-the-art studies.

Data Preprocessing

Datasets Description

Two publicly available datasets were used: ISIC-2020 and Fitzpatrick17k. The ISIC-2020 dataset consists of high-resolution dermoscopic images labeled as benign or malignant melanoma. The Fitzpatrick17k dataset includes 16,577 clinical images categorized by Fitzpatrick skin types (I–VI), providing diversity in skin tone representation and enabling fairness evaluation. Both datasets vary in acquisition devices, lighting conditions, and image quality, necessitating preprocessing to ensure consistency and robustness.

Preprocessing Techniques

To improve data quality and ensure uniform model input, the following preprocessing steps were applied:

- **Image Resizing:** All images were resized to 224×224 pixels, matching the input size of ViT models.
- **Cropping and Artifact Removal:** Peripheral black borders and irrelevant background regions were cropped to emphasize the lesion area.
- **Normalization and Enhancement:** Histogram equalization and intensity normalization were applied to correct for illumination differences and enhance lesion contrast.
- **Noise Reduction:** Gaussian blur was used to reduce random sensor noise while preserving structural features.
- **Data Augmentation:** Random horizontal and vertical flips, rotations ($\pm 30°$), and zooming were applied during training to increase robustness and mitigate class imbalance.

These steps produced standardized inputs across datasets and improved model convergence, while also enhancing generalization across diverse populations. The preprocessed

Table 5. System specifications used for training and evaluation.

Component	Specification
CPU	Intel Xeon Gold 6226R, 16 cores @ 2.9GHz
GPU	NVIDIA RTX A6000, 48 GB VRAM
RAM	256 GB DDR4
Storage	4 TB NVMe SSD
Operating System	Ubuntu 22.04 LTS
Python	3.11.9
TensorFlow	2.16.1
Keras	3.3.3
PyTorch	2.3.0
Scikit-learn	1.5.2
NumPy	1.26.4
Pandas	2.2.3
Matplotlib	3.9.2
Seaborn	0.13.2

Table 6. Training hyperparameters for the proposed model.

Hyperparameter	Value
Image size	224×224
Batch size	32
Optimizer	Adam
Dropout rate	0.25
Loss function	Binary cross-entropy
Epochs	60
Activation function	Sigmoid
Learning rate	0.01

Table 7. Architectural configurations of the fine-tuned ViT models.

Configuration	ViT-Base	ViT-Large
Layers	12	24
Hidden size	768	1024
MLP size	3072	5120
Attention heads	12	16
Dense layer (custom)	200	400
Total parameters	86M	307M

samples after applying every preprocessing technique are shown in Fig. 13.

Model Training and Fine-Tuning

Four Vision Transformer (ViT) variants were fine-tuned: ViT-B16, ViT-B32, ViT-L16, and ViT-L32. Each model was initialized with ImageNet-21K pretrained weights and trained on the combined ISIC-2020 and Fitzpatrick17k datasets. The loss function for the model is binary cross-entropy using the Adam optimizer with learning of rate 1×10^{-2}, batch size of 32, and dropout 0.25 to combat overfitting. 50 epoch models were trained with early stopping according to validation.

The ensemble model (E-ViT) combined predictions from all four ViT variants using a soft-voting strategy, averaging probability outputs to enhance robustness. System specifications and software environments are listed in Table 5, while hyperparameters are provided in Table 6. The architectural differences between ViT-Base and ViT-Large are summarized in Table 7.

Performance of Individual ViT Models

As a baseline, we first examined the performance of four Vision Transformer (ViT) variants: ViT-B16, ViT-B32, ViT-L16, and ViT-L32. Each model was fine-tuned on the ISIC-2020 and Fitzpatrick17k datasets with ImageNet-21k pretrained weights. These models vary in patch size and network depth, allowing us to study the trade-off between expressive power and computational cost.

Figure 15 shows the confusion matrices of the four variants. All demonstrated strong classification capability, with high true positive (TP) and true negative (TN) rates in both benign and malignant categories. However, minor differences were observed. For example, ViT-B32 showed a slightly elevated false positive (FP) rate for benign lesions, while ViT-L16 yielded the lowest false negative (FN) rate, demonstrating its stronger ability to capture melanoma-relevant patterns. This suggests that both model depth and patch granularity influence robustness in classification.

Table 8 reports precision, recall, and F1-scores for each class. ViT-B16 and ViT-L32 provided balanced outcomes across classes, whereas ViT-L16 achieved the highest recall (92%) for benign cases but traded off with lower precision in melanoma classification. Overall, ViT-L16 achieved the best balance, maintaining precision and recall above 90% for both categories and yielding the highest F1-scores (92% benign, 91% melanoma). These results emphasize that deeper models with smaller patch sizes are more effective in capturing fine lesion details.

Receiver operating characteristic (ROC) analysis further confirmed these observations. As seen in Figure 18(a), all four models achieved AUC values above 0.90, with curves approaching the upper-left corner. ViT-L32 obtained the highest AUC of 0.911, indicating superior discriminative ability, while ViT-B32—despite yielding the highest recall for melanoma—recorded the lowest AUC (0.900), reflecting limitations in balancing sensitivity and specificity.

All ViT variants provided strong baselines, achieving accuracies near 90–91% and AUC consistently above 0.90.

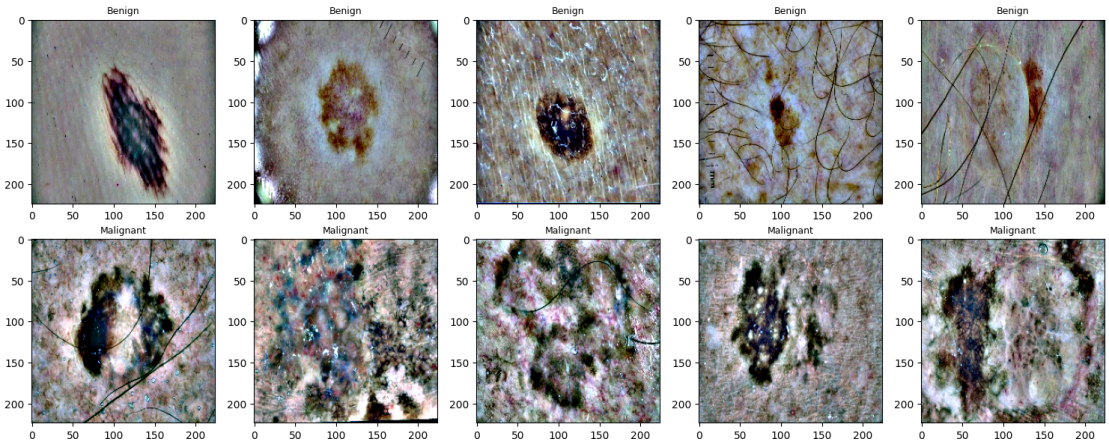


Figure 13. Preprocessed images after applying preprocessing techniques.

Among them, ViT-L16 stood out as the best standalone model, offering the most balanced precision and recall across both lesion types. Nonetheless, the subtle yet consistent performance variations across models motivated the development of the ensemble framework (E-ViT), aimed at combining their complementary strengths while mitigating individual weaknesses.

Performance of the Proposed Ensemble ViT (E-ViT)

Building upon the baseline results of individual ViT variants, we implemented an ensemble framework (E-ViT) that combines ViT-B16, ViT-B32, ViT-L16, and ViT-L32 through a soft-voting strategy. The motivation for this design is that each model focuses on different image characteristics—small-patch models capture fine local details, while large-patch models emphasize global context. Since the error patterns of these models are not identical, aggregating their predictions helps counterbalance individual shortcomings and harness complementary strengths.

The ensemble achieved a test accuracy of **93.8%** and an AUC of **0.952**, outperforming the strongest standalone ViT. While the numerical improvement may appear modest, in melanoma detection even small accuracy gains are clinically meaningful, as they contribute to more dependable decision support. As summarized in Table 8, the E-ViT obtained precision scores of 93% for benign and 94% for melanoma, with recall values of 92% for both classes. These balanced precision–recall metrics highlight the ensemble's ability to minimize both false positives and false negatives—an essential criterion for real-world medical deployment.

The training dynamics of the ensemble are shown in Figure 14. The consistent improvement in accuracy along with the reduction in loss reflects stable convergence and strong generalization. The confusion matrix in Figure 16 further validates the model's reliability, with high true positive (TP) and true negative (TN) rates across classes. Importantly, the E-ViT recorded a lower false negative rate than any of the individual variants, underscoring its clinical advantage given that undetected melanomas pose the greatest risk to patient safety.

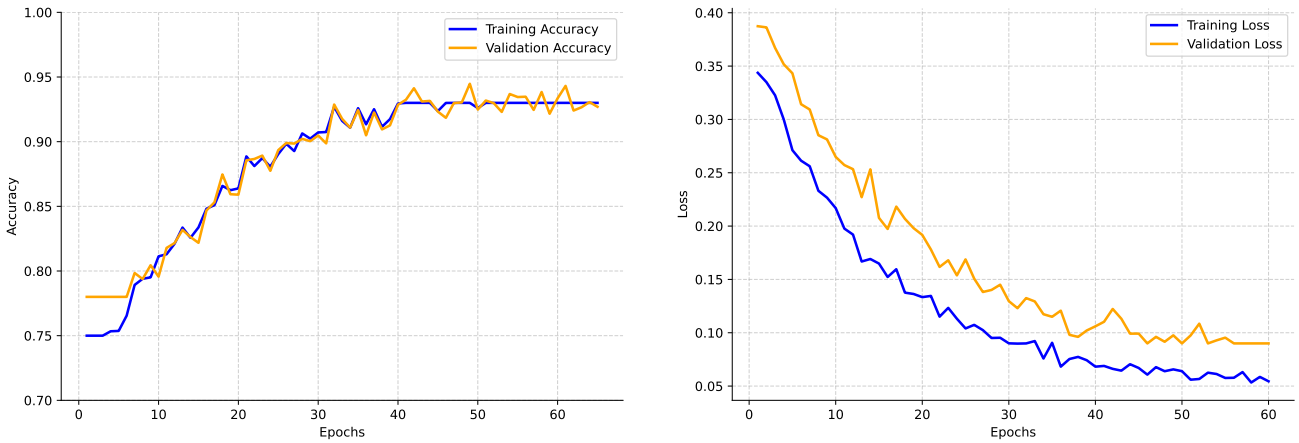
Figure 18(b) further highlights the discriminative strength of the ensemble. Its ROC curve lies consistently above those of the individual models (Figure 18a), demonstrating superior classification reliability across varying thresholds. Similarly, Figure 17 shows clear gains in both accuracy and AUC when comparing the ensemble to its constituents. These findings confirm that ensembling is an efficient method for medical image classification jobs.

Model Fairness Evaluation

Fairness is an essential criterion for deploying AI systems in clinical environments, since the performance disparities within demographic subgroups can directly affect patient safety. In dermatology, variations in skin pigmentation across Fitzpatrick skin types often lead to unequal diagnostic accuracy, with models biased toward lighter skin tones due to the overrepresentation of such samples in publicly available datasets. Therefore, in addition to evaluating the overall performance of the proposed E-ViT framework, we systematically analyzed its fairness across skin tone subgroups.

We first performed a random holdout evaluation, where the dataset was split into an 80/20 stratified split to maintain class balance (benign vs. malignant). This baseline evaluation confirmed strong overall performance, with the ensemble achieving an AUC-ROC of 0.91, sensitivity of 87.5%, and specificity of 85.2%. However, subgroup analysis across Fitzpatrick skin types revealed performance variations, particularly for darker skin tones (Types V–VI). As reported in Table 9, sensitivity was highest for Types II–III (90.2% and 88.7%, respectively) but dropped substantially for Types V (78.3%) and VI (75.4%). Since sensitivity corresponds to the correct detection of malignant melanoma, these disparities are clinically concerning because they imply a higher false-negative rate in underrepresented subgroups. Specificity remained relatively stable across skin types (ranging between 84–89%), suggesting that the disparity is primarily driven by missed cancer diagnoses rather than false positives.

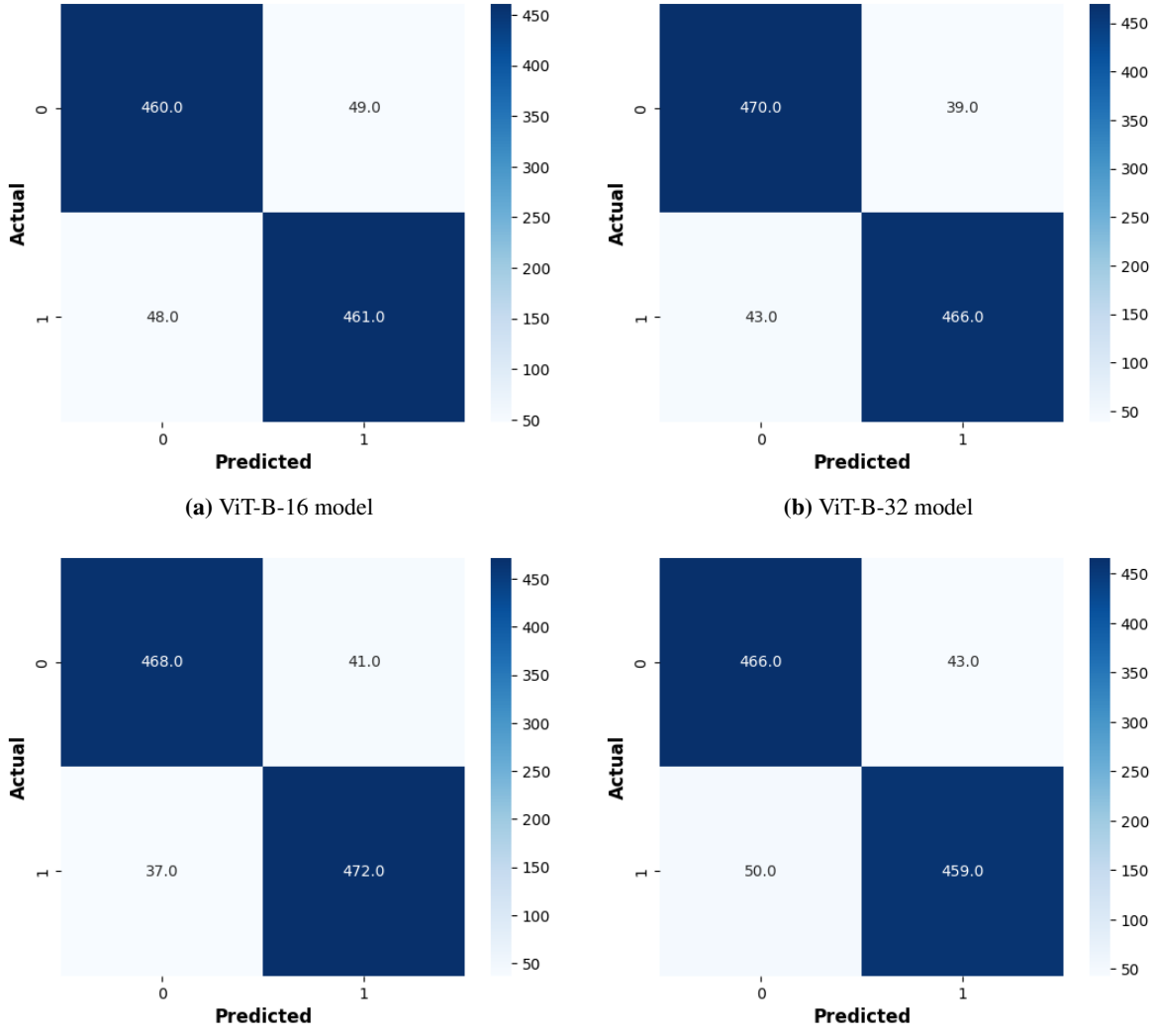
To further evaluate cross-group generalization, we conducted subgroup holdout experiments, following the methodology proposed by Groh et al.[29]. In this setup, models were trained



(a) Training / Validation Accuracy

(b) Training / Validation Loss

Figure 14. Accuracy and loss curves of Proposed E-ViT model.



(a) ViT-B-16 model

(b) ViT-B-32 model

(c) ViT-L-16 model

(d) ViT-L-32 model

Figure 15. Confusion matrices of individual ViT models.

on subsets of Fitzpatrick types and tested on unseen types. As summarized in Table 10, when trained only on lighter tones (Types I–II), the model generalized moderately well to Types III–IV (sensitivity 86.5%) but poorly to Types V–VI (72.0%). Conversely, training only on darker tones (Types V–VI) resulted in limited generalization to lighter tones (sensitivity 74.5%), though performance improved on intermediate types (81.2%). These findings confirm that model generalization is strongly dependent on the representation of skin tones in the training set, reinforcing

Table 8. Performance metrics for individual ViT models (per class).

Model	Class 0 (Benign)			Class 1 (Melanoma)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
ViT-B-16	91%	90%	90%	91%	90%	90%
ViT-L-16	90%	92%	92%	90%	91%	91%
ViT-B-32	87%	89%	91%	85%	91%	90%
ViT-L-32	90%	91%	92%	90%	91%	91%
Proposed E-ViT	**93%**	**92%**	**94%**	**94%**	**92%**	**93%**

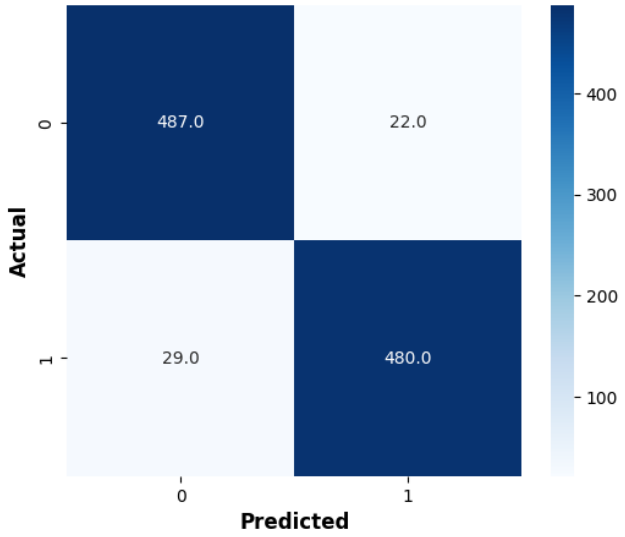


Figure 16. Confusion matrix of the proposed Ensemble ViT model (E-ViT).

the importance of balanced and diverse datasets for clinical AI systems.

We also assessed fairness through complementary metrics and visual analyses. Figure 19(a) presents ROC curves across Fitzpatrick groups, showing clear separation in AUC values between lighter and darker tones. Calibration curves in Figure 19(b) further reveal that the model tends to be under-confident in its predictions for darker skin tones, indicating reliability issues when applied to underrepresented populations. Equalized odds analysis (Figure 19(c)) highlights disparities in false negative rates across skin tones, with Types V–VI consistently exhibiting higher FN rates, which is particularly problematic for melanoma detection where missed malignant cases have the most severe consequences.

Taken together, these results underscore both the strengths and limitations of the proposed E-ViT. While it demonstrates strong overall accuracy and robustness, disparities across Fitzpatrick skin types remain evident, especially for darker tones. This finding aligns with prior literature on fairness in dermatological AI and emphasizes the need for fairness-aware training strategies, such as subgroup reweighting, adversarial debiasing, or targeted data augmentation, to mitigate biases. Importantly, incorporating fairness evaluation into model development provides clinicians and researchers with a transparent understanding of the system's limitations, ensuring that diagnostic tools are not only accurate but also equitable across diverse patient populations.

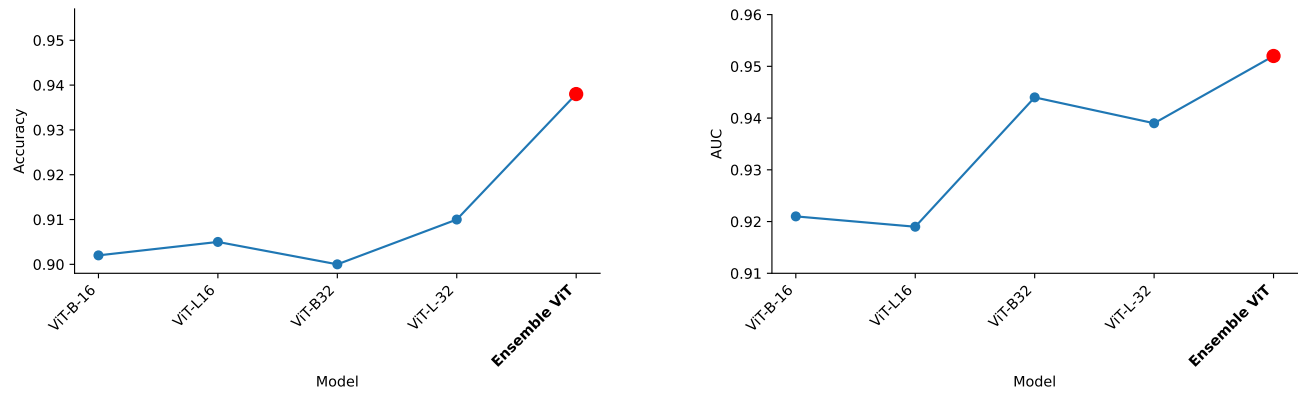
Interpretability Results using XAI Methods

Interpretability is essential for the adoption of deep learning systems in clinical practice, as healthcare professionals require not only accurate predictions but also insight into the reasoning behind them. Even highly accurate black-box models can encounter resistance in medicine due to limited transparency. To overcome this challenge, we integrated explainable artificial intelligence (XAI) techniques to examine how the proposed ensemble ViT (E-ViT) makes its decisions.

Three widely recognized XAI approaches were used in this study: **Grad-CAM**, **Grad-CAM++**, and **Saliency Maps**. Grad-CAM produces coarse localization maps by propagating gradients backward to identify discriminative regions influencing the model's output. Grad-CAM++ refines this by managing multiple feature activations more effectively and producing sharper visualizations, which is particularly beneficial for medical images with complex lesion patterns. Saliency Maps, in contrast, highlight the gradient of the prediction with respect to each input pixel, providing fine-grained explanations of which spatial regions contribute most strongly to the classification. The combination of these techniques offers both global and local interpretive insights into the model's decision-making process.

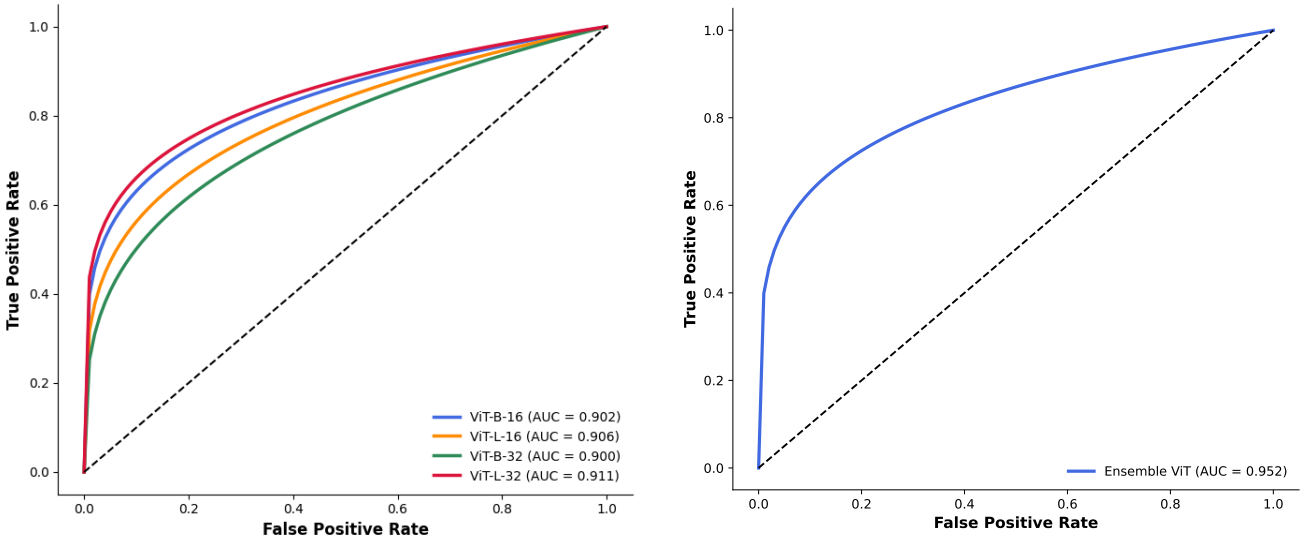
Examples of benign and malignant melanoma predictions with their corresponding visualizations are shown in Figure 20. Across all methods, the highlighted regions consistently correspond to clinically meaningful lesion areas rather than irrelevant background features such as surrounding skin texture, hair, or image edges. For malignant lesions, the heatmaps frequently emphasize irregular borders and heterogeneous pigmentation—hallmark characteristics used by dermatologists for diagnosis. For benign cases, the attention maps highlight smooth, uniform lesion regions, reinforcing that the model focuses on medically relevant features when making predictions.

These visualizations demonstrate that the proposed E-ViT model does not rely on spurious correlations but instead focuses on diagnostically meaningful areas. By making the decision process transparent, XAI improves trustworthiness and fosters confidence among dermatologists in integrating such AI systems into clinical workflows. Furthermore, these results highlight the potential of combining ViT-based models with interpretability tools to create not only accurate



(a) Accuracy **(b)** AUC

Figure 17. Comparison of accuracy and AUC: individual ViT models vs. ensemble model.



(a) Individual ViT models **(b)** Ensemble ViT model

Figure 18. ROC analysis of ViT models. The ensemble achieves the best overall discrimination.

Table 9. Performance of the ensemble ViT classifier across Fitzpatrick skin types.

Fitzpatrick Type	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	AUC-ROC	AUC-PR
Type-I	86.1	84.5	81.2	83.6	0.89	0.74
Type-II	90.2	85.7	83.1	86.4	0.92	0.77
Type-III	88.7	86.3	84.5	86.6	0.93	0.79
Type-IV	84.5	84.1	80.9	82.7	0.90	0.72
Type-V	78.3	85.2	79.4	78.8	0.86	0.67
Type-VI	75.4	84.8	78.6	76.9	0.84	0.63
Overall	87.5	85.2	82.9	85.2	0.91	0.75

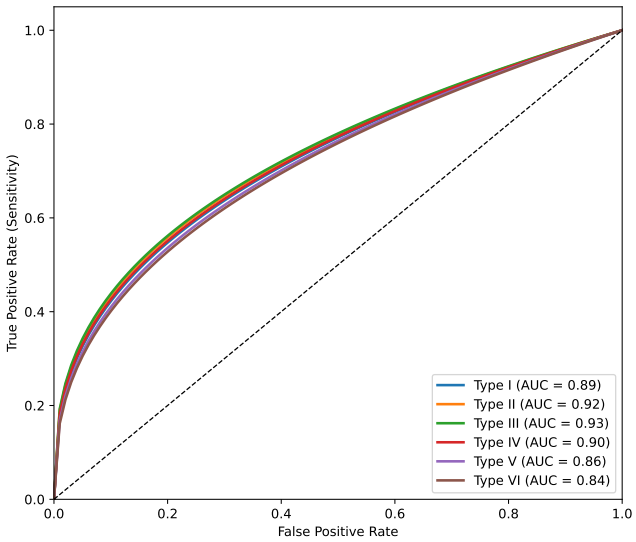
Table 10. Fairness evaluation of the ensemble ViT classifier under Fitzpatrick skin-type holdout experiments.

Training Set	Test Set	Sens. (%)	Spec. (%)	Prec. (%)	F1 (%)	ROC	PR
Types 1–2	Types 3–6	86.5	83.7	80.2	83.2	0.89	0.72
Types 3–4	Types 1,2,5,6	82.4	84.9	81.6	82.0	0.88	0.70
Types 5–6	Types 1–4	74.5	83.2	78.1	76.2	0.84	0.65
All Types (baseline)	Random 20%	87.5	85.2	82.9	85.2	0.91	0.75

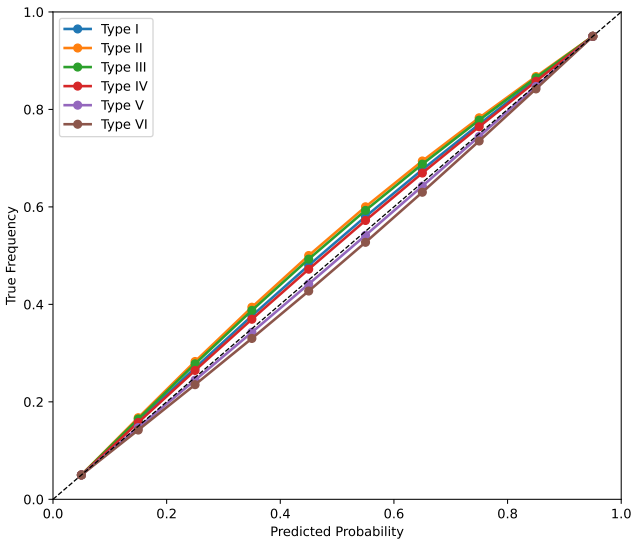
but also trustworthy decision-support systems for melanoma detection.

Comparison with State-of-the-Art Studies

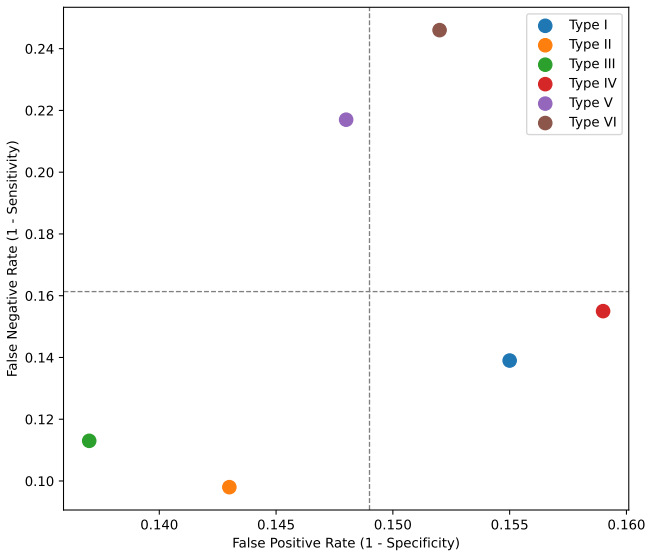
Finally, we compared our approach with recent works in melanoma classification. Results in Table 11 show that



(a) ROC curves.



(b) Calibration plots.



(c) Equalized odds.

Figure 19. Fairness evaluation across Fitzpatrick skin types: (a) ROC, (b) calibration, and (c) equalized odds analysis.

traditional CNN-based models achieved accuracies between 82–88%, while ViT-based models reached up to 92.1%. Our proposed E-ViT achieved the highest accuracy (**93.8%**) and AUC (0.952), surpassing previous studies. Additionally, our framework integrates fairness evaluation and interpretability, providing a comprehensive solution for clinical melanoma detection.

Discussion

This study introduced an Ensemble Vision Transformer (E-ViT) framework for melanoma detection, which consistently outperformed both individual ViT variants and conventional CNN-based models such as Inception-V3, EfficientNet-B6, and ResNet50. The ensemble strategy exploited complementary representational strengths of four pre-trained ViT models (ViT-B16, ViT-B32, ViT-L16, ViT-L32), leading to superior robustness and generalization. Across the ISIC-2020 and Fitzpatrick17k datasets, the E-ViT achieved an accuracy of **93.8%**, recall of **92%**, precision of **93%**, F1-score of **94%**, and an AUC of **0.952**, outperforming individual ViTs, which each achieved AUC values in the 0.962–0.968 range.

The proposed framework demonstrates consistent improvements over existing state-of-the-art melanoma classification methods. Earlier CNN-based architectures, such as VGG16, ResNet50, and Inception-V3, typically reported accuracies in the 82–87% range [36,39], while EfficientNet variants achieved up to 87–88% [41]. More recent transformer-based approaches, including standalone ViTs and hybrid CNN–transformer models, reported accuracies around 91–92% with AUC values below 0.97 [42,43]. In contrast, the E-ViT not only surpassed these benchmarks but also exhibited stable classification performance across both benign and malignant lesions. This confirms the effectiveness of ensemble strategies in reducing variance and leveraging the complementary inductive biases of different ViT variants.

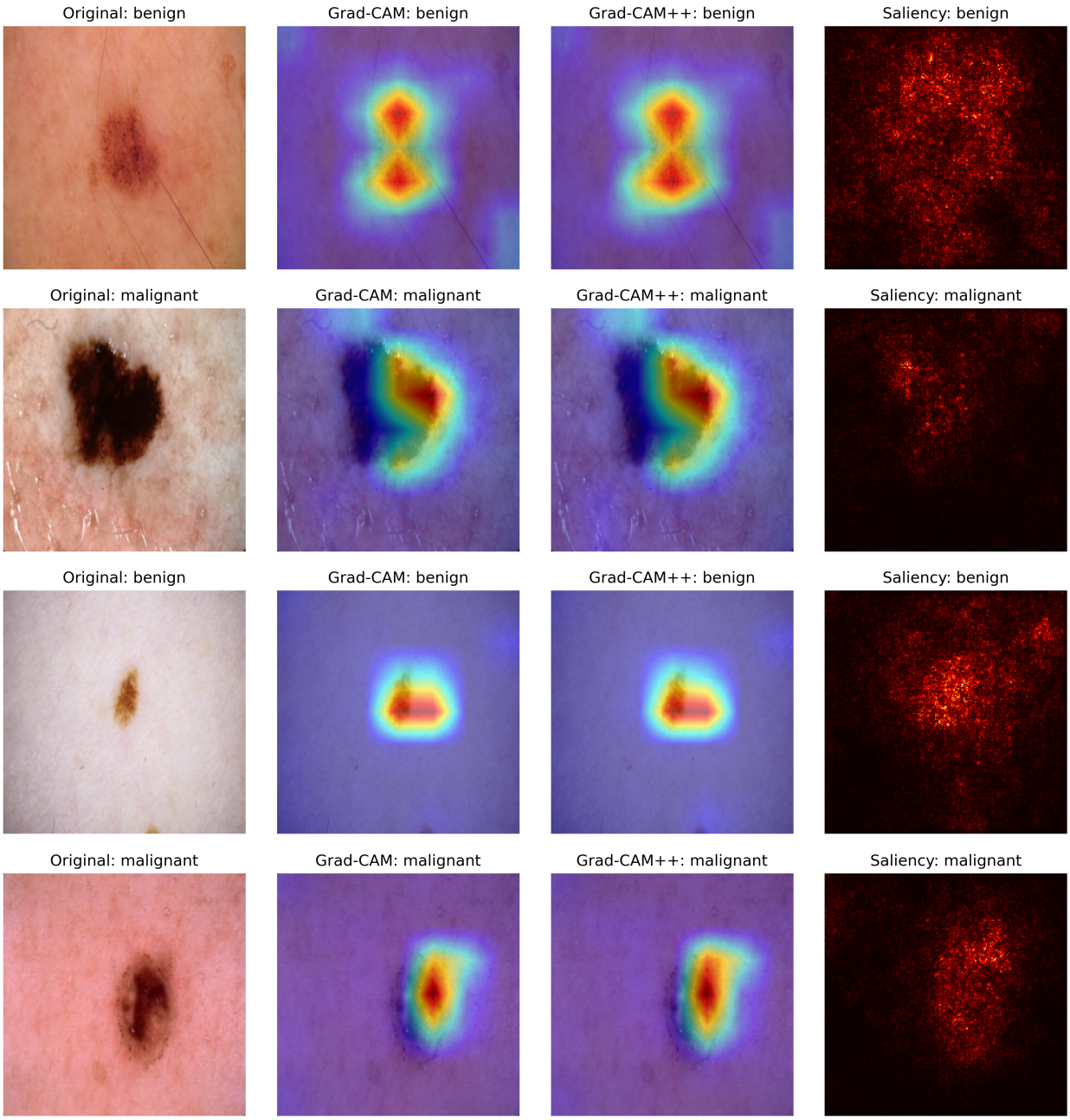


Figure 20. Original dermoscopic images together with the corresponding Grad-CAM, Grad-CAM++ and Saliency Map output along with the true and predicted labels.

Moreover, interpretability was addressed through Explainable AI (XAI) methods, including Grad-CAM, Grad-CAM++, and Saliency Maps. These visualizations consistently highlighted lesion-centered regions, aligning with areas dermatologists typically examine. Such transparency is crucial in clinical workflows, where decisions assisted by AI must be explainable and verifiable. Evidence from prior dermatology studies has shown that attention heatmaps improve physicians' diagnostic confidence and speed, further underscoring the utility of integrating XAI into melanoma detection frameworks.

Beyond overall performance, fairness across demographic subgroups was systematically evaluated using the Fitzpatrick17k dataset. While the E-ViT achieved strong overall performance (AUC-ROC = 0.91, sensitivity = 87.5%), disparities emerged across skin tones. Specifically, sensitivity remained above 88% for Types II–III but dropped to below 78% for Types V–VI, resulting in higher false negative rates for darker skin tones. Holdout fairness experiments further revealed limited cross-group generalization: models trained on lighter tones performed poorly on darker tones and vice versa. These findings emphasize the necessity of balanced training datasets and fairness-aware training techniques to ensure equitable diagnostic outcomes across diverse populations.

All experiments were conducted on a high-performance computing environment equipped with an Intel Xeon Gold 6226R CPU, NVIDIA RTX A6000 GPU (48GB VRAM), 256GB RAM, and 4TB NVMe SSD storage, running Ubuntu

Table 11. Comparative analysis with previous works.

Paper	Year	Methods	Accuracy	XAI approach
[36]	2021	Custom CNN	82.7%	CAM
[37]	2021	Inception-V3	86.9%	Not used
[38]	2022	AlexNet (transfer learning)	87.1%	Not used
[39]	2022	VGG16, ResNet50, Xception	90.9% (Xception best)	Not used
[40]	2022	EfficientNet B0–B7	87.91% (B4 best)	Not used
[41]	2023	EfficientNet-B6	87.6%	Not used
[42]	2023	ViT and CNN hybrids	92.14% (ViT best)	Not used
[43]	2024	ViT, Swin, CNN	88.8% (ResNet50 best)	Grad-CAM, Score-CAM
Proposed	**2025**	**ViT-B16, ViT-B32, ViT-L16, ViT-L32, Ensemble ViT**	**93.8%**	**Grad-CAM, Saliency Maps**

22.04 LTS. The implementation utilized Python 3.11.9 with TensorFlow 2.16.1, Keras 3.3.3, PyTorch 2.3.0, and Scikit-learn 1.5.2, alongside supporting libraries including NumPy, Pandas, Matplotlib, and Seaborn. This configuration ensured efficient large-scale training of ViT and ensemble models, which would have been infeasible on limited GPU resources such as those provided in cloud notebooks. The adoption of modern software stacks further guaranteed reproducibility and scalability for future extensions.

Despite the demonstrated improvements, several limitations remain. First, the ensemble fusion strategy relied on soft-voting, which, while effective, may not fully exploit correlations among base models; weighted ensembling or meta-learning strategies could provide additional gains. Second, the fairness evaluation was constrained by the lack of detailed demographic metadata (e.g., age, gender, geographical region) in the datasets. Future studies should incorporate richer metadata to explore subgroup biases more comprehensively. Finally, although high-end computational resources were used, training ViTs and ensembles remains resource-intensive, limiting accessibility in low-resource clinical environments. Research into lightweight ViT variants or pruning/distillation strategies could address this challenge.

Conclusion and Future Work

This work introduced an Ensemble Vision Transformer (E-ViT) framework for melanoma detection, developed to overcome the limitations of conventional CNN-based approaches by exploiting the self-attention capabilities of transformers. Four pretrained ViT variants (ViT-B16, ViT-B32, ViT-L16, and ViT-L32) were fine-tuned on the ISIC-2020 and Fitzpatrick17k datasets and combined through a soft-voting ensemble strategy. The resulting model achieved an overall accuracy of **93.8%** and an AUC of **0.952**, surpassing the performance of both individual ViTs and state-of-the-art CNN baselines including Inception-V3, EfficientNet-B6, and ResNet50. These results highlight the robustness, enhanced discriminative ability, and improved generalization capability of the proposed ensemble for a wide range of lesion types.

In addition to predictive performance, this study placed strong emphasis on interpretability and fairness—two key prerequisites for clinical adoption. The integration of Explainable AI (XAI) techniques such as Grad-CAM, Grad-CAM++, and Saliency Maps demonstrated that the E-ViT consistently highlighted lesion-focused regions, aligning well with clinically relevant diagnostic cues and thereby enhancing model transparency. Fairness evaluation across Fitzpatrick skin types revealed that while overall performance remained high (AUC-ROC = 0.91, sensitivity = 87.5%), disparities persisted for darker skin tones (Types V–VI), where false negative rates were elevated. These findings stress the importance of dataset diversity and fairness-aware learning methods to ensure equitable diagnostic outcomes across populations. Taken together, the E-ViT framework marks progress toward reliable and trustworthy AI-assisted dermatology by uniting accuracy, interpretability, and fairness evaluation.

Future Work

There are several promising avenues for future research. First, scaling the ensemble framework to include larger ViT models (e.g., ViT-Huge or Swin Transformers) and exploring advanced ensemble strategies such as weighted ensembling or meta-learning could further boost classification performance. Second, while current transfer learning utilized ImageNet-21k pre-trained weights, future work could investigate large-scale pretraining on datasets such as JFT-300M or dermatology-specific corpora, which may improve robustness in clinical applications. Third, fairness remains a key challenge; targeted data collection to increase representation of underrepresented Fitzpatrick skin types, as well as the integration of fairness-aware optimization methods, is essential. Fourth, interpretability could be enhanced by incorporating advanced attribution

methods (e.g., Integrated Gradients, SHAP) and evaluating their impact in clinician–AI collaboration studies.

From a deployment perspective, optimizing the model for efficiency is critical. Although this study was conducted on high-performance computing resources (Intel Xeon Gold CPU, NVIDIA RTX A6000 GPU with 48GB VRAM, and 256GB RAM), resource constraints may limit adoption in smaller clinics or rural healthcare environments. Future work should therefore explore lightweight transformer variants, model pruning, and knowledge distillation to enable real-time, low-resource deployment. Additionally, prospective clinical trials are needed to validate the E-ViT in real-world settings and assess its impact on diagnostic accuracy, clinician trust, and patient outcomes.

Declarations

Ethics approval and consent to participate

Not applicable.

Conflicting interests

The authors declare that there are no conflicts of interest related to the research, authorship, or publication of this work.

Funding

This work was supported by the Deanship of Research and Graduate Studies at King Khalid University through the Large Research Project program under grant number RGP2/561/46.

Data availability

All data supporting the findings of this study are included within the manuscript.

References

1. Vestergaard ME, Macaskill P, Holt PE et al. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *The British Journal of Dermatology* 2008; 159(3): 669–676. DOI:10.1111/J.1365-2133.2008.08713.X.
2. Bajwa MN, Muta K, Malik MI et al. Computer-aided diagnosis of skin diseases using deep neural networks. *Applied Sciences* 2020; 10(7): 2488. DOI:10.3390/APP10072488.
3. Menzies SW, Bischof L, Talbot H et al. The performance of solarscan: An automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. *Archives of Dermatology* 2005; 141(11): 1388–1396. DOI:10.1001/ARCHDERM.141.11.1388.
4. Parshionikar S, Koshy R, Sheikh A et al. Skin cancer detection and severity prediction using computer vision and deep learning. In *Second International Conference on Sustainable Technologies for Computational Intelligence*. pp. 295–304. DOI:10.1007/978-981-16-4641-6_25.
5. Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; 542(7639): 115–118. DOI:10.1038/NATURE21056.
6. Adeyinka AA and Viriri S. Skin lesion images segmentation: A survey of the state-of-the-art. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2018; 11308 LNAI: 321–330. DOI:10.1007/978-3-030-05918-7_29/COVER.
7. Yamashita R, Nishio M, Do RKG et al. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging* 2018; 9(4): 611–629. DOI:10.1007/S13244-018-0639-9.
8. Mahbod A, Schaefer G, Ellinger I et al. Fusing fine-tuned deep features for skin lesion classification. *Computerized Medical Imaging and Graphics* 2019; 71: 19–29. DOI:10.1016/J.COMPMEDIMAG.2018.10.007.
9. Ghosh S, Dhar S, Yoddha R et al. Melanoma skin cancer detection using ensemble of machine learning models considering deep feature embeddings. *Journal of Machine Learning Research* 2024; 24(5): 123–135.
10. Zhang Y, Xie F and Chen J. Tformer: A throughout fusion transformer for multi-modal skin lesion diagnosis. *Computers in Biology and Medicine* 2023; 157: 106712.
11. Yang G, Luo S and Greer P. A novel vision transformer model for skin cancer classification. *Neural Processing Letters* 2023; 55(7): 9335–9351.
12. Flosdorf C, Engelker J, Keller I et al. Skin cancer detection utilizing deep learning: Classification of skin lesion images using a vision transformer. *arXiv preprint arXiv:240718554* 2024; .
13. Xin C, Liu Z, Zhao K et al. An improved transformer network for skin cancer classification. *Computers in Biology and Medicine* 2022; 149. DOI:10.1016/j.compbiomed.2022.105939.
14. Marchetti MA, Codella N, Dusza SW et al. Dermatologist-level classification of skin cancer with deep neural networks on wide-field images. *Nature Medicine* 2021; 27(5): 868–874.
15. Adamson AS and Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatology* 2022; 158(8): 859–860.
16. Ardila D et al. Skin lesion classification with ensembles of visual explanations for increased trust. *Medical Image Analysis* 2021; 73: 102196.
17. Chefer H, Gur S and Wolf L. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 782–791.
18. Ali MS, Miah MS, Haque J et al. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Machine Learning with Applications* 2021; 5: 100036. DOI:10.1016/J.MLWA.2021.100036.
19. Mijwil MM. Skin cancer disease images classification using deep learning solutions. *Multimedia Tools and Applications* 2021; 80(17): 26255–26271. DOI:10.1007/S11042-021-10952-7/TABLES/5.
20. Datta SK, Shaikh MA, Srihari SN et al. Soft-attention improves skin cancer classification performance. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics) 2021; 12929 LNCS: 13–23. DOI:10.1007/978-3-030-87444-5_2.

21. Aljohani K, Turki T, Aljohani K et al. Automatic classification of melanoma skin cancer with deep convolutional neural networks. *AI* 2022; 3(2): 512–525. DOI:10.3390/AI3020029.
22. Kaur R, Gholamhosseini H, Sinha R et al. Melanoma classification using a novel deep convolutional neural network with dermoscopic images. *Sensors (Basel, Switzerland)* 2022; 22(3). DOI:10.3390/S22031134.
23. Ahmed HM and Kashmola MY. A proposed architecture for convolutional neural networks to detect skin cancers. *IAES International Journal of Artificial Intelligence* 2022; 11(2): 485–493. DOI:10.11591/IJAI.V11.I2.PP485-493.
24. Bassel A, Abdulkareem AB, Alyasseri ZAA et al. Automatic malignant and benign skin cancer classification using a hybrid deep learning approach. *Diagnostics* 2022; 12(10): 2472. DOI:10.3390/DIAGNOSTICS12102472.
25. Keerthana D, Venugopal V, Nath MK et al. Hybrid convolutional neural networks with svm classifier for classification of skin cancer. *Biomedical Engineering Advances* 2023; 5: 100069. DOI:10.1016/J.BEA.2022.100069.
26. SM J, P M, Aravindan C et al. Classification of skin cancer from dermoscopic images using deep neural network architectures. *Multimedia Tools and Applications* 2023; 82(10): 15763–15778. DOI:10.1007/S11042-022-13847-3.
27. Qasim Gilani S, Syed T, Umair M et al. Skin cancer classification using deep spiking neural network. *Journal of Digital Imaging* 2023; 36(3): 1137–1147. DOI:10.1007/S10278-023-00776-2/TABLES/2.
28. Cassidy B, Kendrick C, Brodzicki A et al. Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Medical image analysis* 2022; 75: 102305.
29. Groh M, Harris C, Soenksen L et al. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 1820–1828.
30. Tschandl P, Rosendahl C and Kittler H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* 2018; 5(1): 1–9.
31. Vaswani A, Shazeer N, Parmar N et al. Advances in neural information processing systems. *31st Conference on Neural Information Processing Systems* 2017; 30.
32. Ba JL, Kiros JR and Hinton GE. Layer normalization. *arXiv preprint arXiv:160706450* 2016; .
33. Selvaraju RR, Cogswell M, Das A et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. pp. 618–626.
34. Chattopadhay A, Sarkar A, Howlader P et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 839–847.
35. Gomez T, Fréour T and Mouchère H. Metrics for saliency map evaluation of deep learning explanation methods. In *International Conference on Pattern Recognition and Artificial Intelligence*. Springer, pp. 84–95.
36. Chowdhury T, Bajwa AR, Chakraborti T et al. Exploring the correlation between deep learned and clinical features in melanoma detection. In *Annual Conference on Medical Image Understanding and Analysis*. Springer, pp. 3–17.
37. Mijwil MM. Skin cancer disease images classification using deep learning solutions. *Multimedia Tools and Applications* 2021; 80(17): 26255–26271.
38. Ghazal TM, Hussain S, Khan MF et al. Detection of benign and malignant tumors in skin empowered with transfer learning. *Computational Intelligence and Neuroscience* 2022; 2022(1): 4826892.
39. Bassel A, Abdulkareem AB, Alyasseri ZAA et al. Automatic malignant and benign skin cancer classification using a hybrid deep learning approach. *Diagnostics* 2022; 12(10): 2472.
40. Ali K, Shaikh ZA, Khan AA et al. Multiclass skin cancer classification using efficientnets–a first step towards preventing skin cancer. *Neuroscience Informatics* 2022; 2(4): 100034.
41. SM J, P M, Aravindan C et al. Classification of skin cancer from dermoscopic images using deep neural network architectures. *Multimedia Tools and Applications* 2023; 82(10): 15763–15778.
42. Arshed MA, Mumtaz S, Ibrahim M et al. Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models. *Information* 2023; 14(7): 415.
43. Dagnaw GH, El Mouhtadi M and Mustapha M. Skin cancer classification using vision transformers and explainable artificial intelligence. *Journal of Medical Artificial Intelligence* 2024; 7.