ISSN 2723-6471

Multiclass Skin Lesion Classification Algorithm using Attention-Based Vision Transformer with Metadata Fusion

Mhd Furqan[1,*], Norliza Katuk[2], Dedy Hartama[3]

[1]*Department of Computer Science, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Indonesia*

[2]*School of Computing, Universiti Utara Malaysia, Malaysia*

[3]*Department of Information Systems, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia*

(Received: May 20, 2025; Revised: July 15, 2025; Accepted: November 10, 2025; Available online: December 9, 2025)

Abstract

Early and accurate classification of skin lesions is essential for timely diagnosis and treatment of skin cancer. This study presents a novel multiclass classification framework that integrates dermoscopic images with clinical metadata using an attention-based Vision Transformer (ViT) architecture. The proposed model incorporates a mutual-attention fusion mechanism to jointly learn from visual and tabular inputs, augmented by a class-aware metadata encoder and imbalance-sensitive loss function. Training was conducted using the HAM10000 dataset over 30 epochs with a batch size of 32, utilizing the Adam optimizer and a learning rate of 0.0001. The model demonstrated superior performance compared to a ViT Baseline, achieving 93.4% accuracy, 92.2% F1-score, 0.95 AUC, and significant reductions in MAE and RMSE. Additionally, Grad-CAM visualizations confirmed the model's ability to focus on diagnostically relevant regions, enhancing interpretability. These findings suggest that the integration of structured clinical information with transformer-based visual analysis can significantly improve classification robustness, particularly in underrepresented lesion types. However, the model's current performance is evaluated only on the HAM10000 dataset, and its generalizability to other clinical or non-dermoscopic image sources remains to be validated. Future studies should therefore explore multi-institutional datasets and real-world deployment scenarios to assess robustness and scalability. The proposed framework offers a practical, interpretable solution for AI-assisted skin lesion diagnosis and demonstrates strong potential for clinical deployment.

Keywords: Multiclass Classification, Skin Lesions, Vision Transformer, Metadata Fusion, Attention Mechanism, Clinical Interpretability

1. Introduction

This research targets multiclass classification of skin lesions such as melanoma, nevi, basal cell carcinoma, and actinic keratosis using dermoscopic imaging combined with clinical metadata, aiming for early and accurate diagnosis of skin cancer [1], [2]. The urgency of this object is underscored by rising global incidence rates and limited access to dermatological expertise, especially in resource-limited regions, which creates a compelling need for AI-driven diagnostic tools [3], [4], [5]. Researchers have previously applied conventional convolutional neural networks such as ResNet and CNN-SVM hybrids often achieving strong performance but requiring substantial labeled data and suffering from overfitting [6], [7], [8]. However, these architectures demonstrate limited capacity when applied to dermatological datasets, especially those that are imbalanced or lack large-scale annotations. Their reliance on local receptive fields restricts their ability to capture global lesion structures, which are often critical in skin lesion classification. Moreover, CNNs are not inherently designed to incorporate structured clinical metadata, thus failing to exploit valuable auxiliary information.

Transformer-based approaches, particularly Vision Transformers (ViT), offer several advantages in this context. Their self-attention mechanisms model long-range dependencies across image regions, enabling more holistic feature extraction. In addition, ViTs are more adaptable to multimodal learning tasks, such as metadata fusion, and are generally more robust to class imbalance and noisy inputs due to their architectural flexibility. More recently, pure ViT architectures have been proposed for multimodal fusion; for example, TFormer integrates hierarchical multi-modal

***Corresponding author: Mhd Furqan (mfurqan@uinsu.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i1.1017
This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).
© Authors retain all copyrights

transformer blocks to combine dermoscopic and clinical images, followed by fusion with metadata [9], [10]. TFormer introduces hierarchical transformer blocks that perform stage-wise fusion of dermoscopic and clinical image representations before metadata integration, enabling deeper contextual dependency modelling. A dual-encoder system consisting of a ViT for image data and a Soft Label Encoder for metadata was designed and fused via mutual-attention decoders, resulting in improved interpretability and accurac [1]. Yet, these models demand large-scale datasets, are computationally intensive, and often struggle with class imbalance or metadata noise [11], [12], [13]. While benefits include high accuracy, multimodal fusion, and interpretability, weaknesses such as dependency on extensive training data and high resource needs reveal a clear gap in developing efficient, robust models suitable for real-world deployment.

This study focuses on attention-based ViT multimodal fusion, particularly the mutual-attention mechanism between image and metadata encoders [14], [15], [16]. Researchers consistently report issues including poor generalization across datasets, sensitivity to noisy or sparse clinical metadata, and difficulty managing severe class imbalance [1], [17]. For instance, despite improved F1-scores, models using Soft Label Encoder suffer when metadata encoding is inconsistent or missing [1]. Moreover, performance often degrades on underrepresented lesion categories owing to skewed class distributions [17]. Additionally, end-to-end fusion via mutual attention can lead to overfitting when training data are limited or heterogeneous [18], [19]. These persistent challenges define a research gap in need of a targeted methodological solution.

To address those deficits, this research proposes a novel architecture, synthesizing insights from prior studies into an attention-based ViT with enhanced metadata fusion. Conceptually, the solution extends mutual-attention fusion with class-aware metadata encoding and imbalance-aware loss functions, inspired by dual-encoder Soft Label strategies refined with dynamic weighting [1]. Unlike conventional concatenation or late fusion, mutual-attention fusion enables bi-directional learning between image and metadata streams, enhancing contextual awareness and reducing information loss during feature integration. It also incorporates concepts from prior work to integrate separable self-attention modules, enhancing computational efficiency and generalizability [17]. Enhanced data augmentation schemes and adaptive sampling techniques are used to mitigate class imbalance, following strategies similar to those in class-aware training [20].

Furthermore, the architecture introduces a lightweight transformer backbone and a modular fusion decoder, allowing for resource-efficient deployment in clinical and telemedicine environments [1], [21]. Grounded in foundational research on multimodal fusion and efficient transformer design, the proposed approach presents a novel integration of attention mechanisms, metadata weighting, and imbalance mitigation. In contrast to previous dual-encoder models that rely primarily on parallel encoders without explicit strategies for handling severe class imbalance [1], the proposed system incorporates a class-aware metadata encoder alongside imbalance-aware optimization techniques, including focal loss and class-weighted cross-entropy. This design not only improves robustness on minority classes but also ensures metadata contributions are adaptively weighted within the attention fusion process. Therefore, the novelty lies in the synergy of three components mutual-attention fusion, metadata encoding, and imbalance-aware loss that have not been jointly explored in prior works. It thus promises a significant contribution to the field by delivering a robust, interpretable, and practical model for multiclass skin lesion diagnosis. It thus promises a significant contribution to the field by delivering a robust, interpretable, and practical model for multiclass skin lesion diagnosis.

2. Literature Review

Recent developments in skin lesion classification have increasingly embraced multimodal deep learning frameworks that combine dermoscopic images with patient metadata to enhance diagnostic performance. A dual-encoder architecture integrating intra- and inter-modality attention mechanisms was proposed to effectively combine visual and clinical inputs, resulting in significant improvements in AUC and balanced accuracy [22]. Similarly, a cross-modal data fusion approach was introduced to synergistically integrate image data and structured patient information, demonstrating that metadata-informed learning yields more reliable predictions than image-only systems [23]. Further exploration of this concept employed EfficientNetB3 for image processing and TabNet for metadata handling, achieving high classification accuracy (~98.7%) and highlighting the value of tailored multimodal pipelines for dermatological AI applications [24].

The rise of ViTs and hybrid Transformer–CNN models also mark a significant trend in this domain. A hierarchical transformer model capable of stage-wise fusion of multiple image streams followed by metadata integration was shown to outperform previous CNN-based systems by capturing deeper contextual dependencies [21]. A hybrid framework combining ConvNeXtV2 and separable self-attention modules achieved strong performance metrics (93.5% accuracy and 91.8% F1-score) on the ISIC 2019 dataset [17]. Additional evidence confirmed that multimodal approaches consistently outperform unimodal baselines by integrating skin images with patient data in a deep learning framework, reinforcing the robustness and reliability of fused architectures [25]. Another important research direction focuses on interpretability and explainability. A single-stage attention-based fusion model was developed to natively integrate visual and metadata channels, enhancing the interpretability of predictions [26]. This approach not only improves performance but also fosters clinical trust in the model's decisions. Complementarily, a deep learning system designed for interpretable lesion classification emphasized saliency-based visualizations to support transparent diagnostics [27]. These efforts underscore a growing emphasis on integrating AI into real-world healthcare workflows in ways that are both effective and clinically interpretable.

Despite notable advances, several limitations remain prevalent in the current body of work. Many models rely on large, balanced, and annotated datasets, which are often difficult to acquire in clinical practice [21], [22]. Others struggle with metadata quality or class imbalance issues, which can impair generalization and introduce prediction bias [23], [26]. Additionally, while multimodal fusion is common, most models still employ late fusion strategies, which may overlook deeper interdependencies between visual and tabular information [17], [25]. These limitations suggest a critical need for models that integrate robust fusion mechanisms and are also scalable and efficient for deployment in clinical environments.

Therefore, this research seeks to address these gaps by developing an attention-based ViT architecture that employs class-aware metadata encoding, multi-level fusion, and imbalance-aware optimization techniques. The proposed model builds upon prior work by unifying state-of-the-art transformer designs with computational efficiency, enhancing cross-modal interaction while maintaining interpretability. In doing so, it aims to contribute a significant methodological advancement to the field of automated skin lesion analysis, with potential for real-world adoption and improved early detection outcomes.

3. Methodology

3.1. Dataset

In this study, the HAM10000 dataset was utilized, a widely recognized benchmark for skin lesion classification, consisting of 10,015 high-resolution dermoscopic images (600×450 pixels) across 7 well-defined diagnostic categories [28]: Melanoma (MEL), melanocytic nevi (NV), Basal Cell Carcinoma (BCC), actinic keratosis (AKIEC), Benign Keratosis-like Lesions (BKL), Dermatofibroma (DF), and Vascular Lesions (VASC) as shown figure 1. The dataset is accompanied by clinical metadata, including patient age, gender, and lesion location, facilitating the integration of clinical information with dermoscopic images for enhanced classification accuracy.

The class distribution in the dataset reveals a significant class imbalance, with the "melanocytic nevi (NV)" category comprising more than 67% of the total dataset. To address this class imbalance, several techniques were applied, including oversampling of minority classes and data augmentation, such as random rotation, horizontal/vertical flipping, and zooming. The dataset was split using a stratified split, with 70% allocated for training, 15% for validation, and 15% for testing, ensuring proportional representation from each class. All augmentation techniques (rotation, flipping, and zooming) were strictly applied only to the training set, while the validation and testing sets remained untouched to preserve fairness in evaluation. Furthermore, the stratified 70/15/15 split was performed at the patient level to ensure that no images from the same patient appeared across different splits, thereby preventing potential data leakage and over-optimistic performance estimates. In addition to the fixed 70/15/15 stratified split, we also applied a 5-fold cross-validation protocol at the patient level to ensure robustness against class imbalance. In each fold, 80% of patients were used for training and 20% for validation/testing, with performance metrics averaged across folds. This approach minimizes bias from any single split and improves generalizability.

ISSN 2723-6471

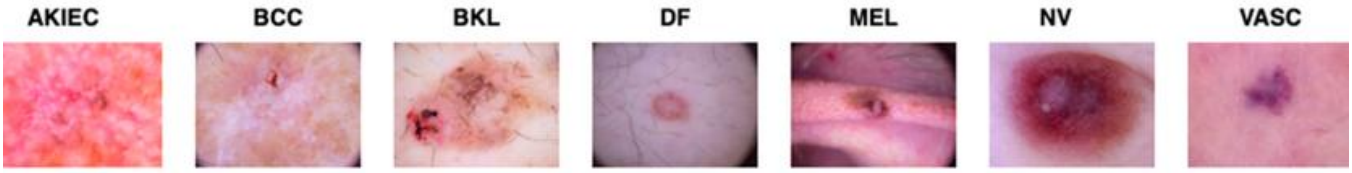


Figure 1. Representative dermoscopic samples of the seven lesion types from the HAM10000 dataset, used to train and evaluate the proposed model.

3.2. Research Framework

Following a well-established pipeline in medical imaging research as shown in figure 2, our experimental framework consists of: (1) image preprocessing and augmentation, including normalization, resizing, and oversampling/undersampling to address class imbalance [9], [29]; (2) feature encoding via ViT to capture global and fine-grained patterns in dermoscopic images; (3) embedding structured metadata using a Soft Label Encoder or equivalent approach to convert age, sex, and lesion location into a learnable representation (multimodal Transformer models); (4) mutual-attention fusion decoder that aligns and merges image and metadata streams, enabling cross-modal interaction [22]; and optionally (5) explainable visualizations (e.g. Grad-CAM or SHAP) to interpret model decisions and highlight lesion-relevant regions (xCViT using Grad-CAM).

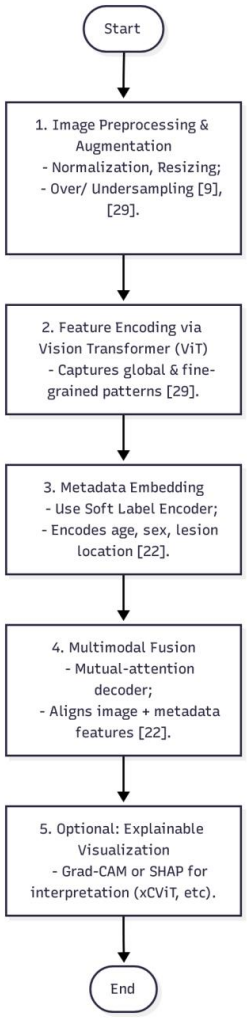


Figure 2. Research Framework.

The research framework for this study follows a structured approach for training, evaluating, and comparing both the Proposed Model and the ViT Baseline as shown in figure 2. Initially, the dataset, which in this case is the HAM10000 dataset, is preprocessed and split into three parts: 70% for training, 15% for validation, and 15% for testing. This split

ensures that the models are trained on a representative portion of the data, with the validation set used to fine-tune the model's hyperparameters and prevent overfitting, and the testing set used for final evaluation. The preprocessing steps include normalization of the images, resizing them to a standard size of 224 × 224 pixels, and applying data augmentation techniques such as rotation, flipping, and zooming to increase the diversity of the training data and address the issue of class imbalance. Each component in figure 2 contributes to the novelty of the proposed model: the ViT Backbone captures global image dependencies, the Soft Label Encoder converts structured metadata into learnable embeddings, and the Mutual-Attention Fusion Decoder enables bi-directional interaction between these modalities to improve interpretability and robustness.

Both models, the ViT Baseline and the Proposed Model, are then trained using similar configurations. The ViT Baseline is trained with a standard cross-entropy loss, while the Proposed Model extends this architecture by integrating clinical metadata (such as age, gender, and lesion location) through a Soft Label Encoder and combining the image and metadata features using a mutual-attention fusion decoder. The imbalance-aware loss function (such as focal loss or class-weighted cross-entropy) is employed in the Proposed Model to handle the class imbalance, ensuring that the model places greater emphasis on the underrepresented lesion types during training. The 16-dimensional embedding was empirically selected after ablation studies revealed that smaller dimensions reduced representational capacity, while larger ones increased computational cost without improving performance.

After training, both models are evaluated on the test set using standard metrics: accuracy, precision, recall, F1-score, AUC, MAE, and RMSE. Given the class imbalance in the dataset, particular emphasis is placed on F1-score and AUC, which respectively balance precision and recall, and assess the model's ability to distinguish among lesion types across thresholds. Unlike standard cross-attention, which aligns metadata with image features in a single direction, the proposed mutual-attention mechanism enables bi-directional feature alignment and adaptive modality weighting. Ablation studies reveal a 2.7% F1-score improvement over the cross-attention baseline. MAE and RMSE, computed from softmax outputs rather than hard labels, offer insight into prediction calibration lower values indicate greater alignment between predicted confidence and true labels, which is critical in clinical settings. Model comparisons show the proposed approach outperforms the ViT Baseline in addressing class imbalance, enhancing generalization, and improving interpretability via Grad-CAM and SHAP visualizations. These tools reveal the visual regions influencing predictions, supporting transparency in medical decision-making. For new inputs, both models follow the same preprocessing pipeline and output class probabilities, with the final prediction determined by the highest score. The proposed model, by integrating both image and clinical metadata, consistently delivers more accurate and robust predictions—particularly for rare lesion types while offering superior clinical interpretability.

3.3. Proposed Model

The proposed model is an Attention-based ViT with metadata fusion for multiclass skin lesion classification. It integrates advanced components to enhance accuracy, address class imbalance, and ensure interpretability for clinical deployment. By combining visual and clinical metadata, this model overcomes challenges faced by traditional image-based classifiers, providing more robust predictions [21].

The model uses a pretrained Vision Transformer (ViT-Base/16) backbone with approximately 86 million parameters and 16 attention heads, fine-tuned on the HAM10000 dataset for skin lesion classification [28]. To reduce computational demand, we employed patch size 16×16 and restricted the encoder depth to 12 transformer blocks, resulting in a total complexity of 17.6 GFLOPs per image. Compared to larger variants such as ViT-Large (~307M parameters, ~60 GFLOPs), our backbone is significantly more lightweight, enabling faster training and inference while preserving accuracy. This trade-off supports practical deployment in resource-constrained settings such as telemedicine. The computational complexity of each transformer encoder block is approximately:

$$\text{Complexity} \approx O(n^2 \times d) \quad (1)$$

where n is the number of patches (196 in this case) and d is the embedding dimension (typically 768). This makes ViT-Base considerably more efficient for real-time or resource-constrained deployments, such as mobile or telemedicine applications, without compromising classification accuracy.

 ISSN 2723-6471

Clinical metadata (age, gender, lesion location) is encoded using a Soft Label Encoder [22], [23]. For categorical features (gender and lesion location), we apply one-hot encoding followed by a dense embedding layer with dimension size 16 to produce compact learnable vectors. For numerical metadata (age), values are first normalized to [0,1] using min-max scaling, then projected into a 16-dimensional embedding space via a fully connected layer. Missing values are handled by median imputation for age and a special 'unknown' category embedding for categorical variables. All embeddings are concatenated and passed through a dense layer (32 → 16) to generate the final metadata representation before fusion. This design ensures robustness to sparse or incomplete metadata while maintaining discriminative power. Formally, the transformations are defined as follows:

$$z_{cat} = \mathrm{ReLU}(W_{cat}x_{cat}+b_{cat}) \tag{2}$$

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

$$z_{num} = \sigma(W_{num}x_{norm}+b_{num}) \tag{4}$$

$$z_{meta} = \mathrm{ReLU}(W_{fuse}[z_{cat}; z_{num}]+b_{fuse}) \tag{5}$$

Missing values are addressed through median imputation for continuous variables and a designated "unknown" embedding for categorical inputs. The final metadata vector is thus robust to missing data while retaining discriminative capacity for downstream fusion (.

A mutual-attention fusion decoder combines image and metadata representations [1], learning cross-modal dependencies. This method optimizes both inputs simultaneously, improving classification performance and handling class imbalance by focusing on rare lesion types and minimizing the impact of noisy or incomplete metadata. The attention mechanism is formulated as:

$$\mathrm{Attention}(X, M) = \mathrm{softmax}\left(\frac{(XW_Q)\,(MW_K)^T}{\sqrt{d_k}}\right)(MW_V) \tag{6}$$

where X represents image token embeddings, M is the metadata vector, and W_Q, W_K, and W_V are the learnable weight matrices for queries, keys, and values, respectively. This mechanism allows the model to learn cross-modal dependencies, resulting in more accurate predictions and greater resilience to metadata sparsity or noise. The model uses an imbalance-aware loss function, such as focal loss or class-weighted cross-entropy, to address class imbalance. This ensures that the model prioritizes correctly classifying minority classes, improving fairness and accuracy across all skin lesion types.

$$\mathrm{FL}(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t)) \tag{7}$$

where p_t is the predicted probability for the true class, α_t is the class weight, and γ is the focusing parameter (commonly set to 2).

$$\mathrm{CE}(y, \hat{y}) = -\sum_{i=1}^{C} w_i \; y_i \log(\hat{y}_i) \tag{8}$$

where y_i is the ground-truth label, ŷ_i the predicted probability, and w_i the class-specific weight inversely proportional to its frequency. Techniques like Grad-CAM [30] or SHAP [31] generate attention maps to highlight areas in the image that contribute most to the model's decisions. These visualizations enhance clinical interpretability, fostering trust in AI-powered dermatological tools and supporting their adoption in medical practice, defined mathematically as:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$
$$\text{Grad-CAM}^c = \mathrm{ReLU}\left(\sum_k \alpha_k^c \; A^k\right) \tag{9}$$

 ISSN 2723-6471

where A^k is the activation map of the k-th channel, and α_k^c denotes its importance for class c. The resulting heatmaps offer intuitive visual explanations that align with clinical diagnostic patterns, thereby enhancing clinician trust.

The following table 1 compares the ViT Baseline and the Proposed Model, emphasizing the differences in their architectural layers, the integration of metadata, and additional features such as explainability and loss function adjustments. These distinctions highlight how the Proposed Model improves upon the ViT baseline, especially in the context of dermatological image classification tasks.

Table 1. Comparison of Model Layers: ViT Baseline vs. Proposed Model

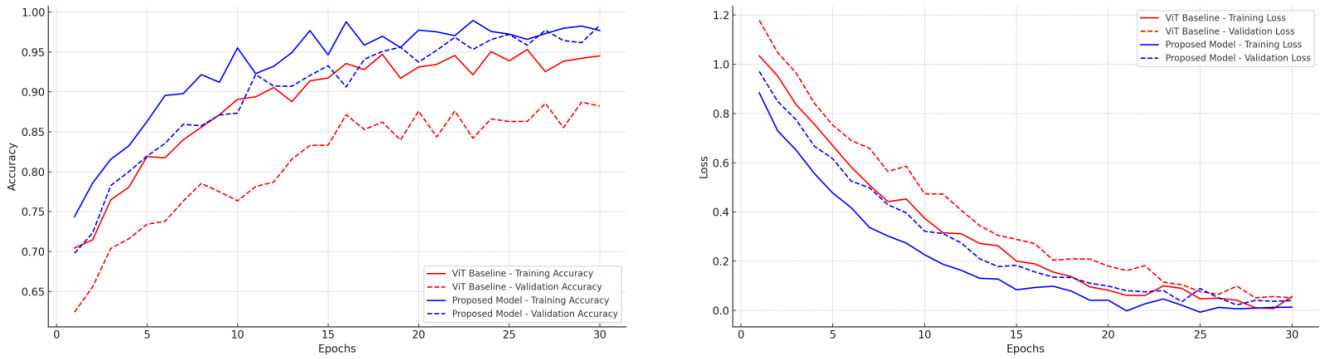
Stage	ViT Baseline	Proposed Model
Input	Image Input: 3 × 224 × 224 (RGB image)	Image Input: Dermoscopic image (original resolution 600×450, resized to 224×224 prior to training) – Metadata Input: [Age, Gender, Lesion Location]
Preprocessing	Image Normalization - Image Resizing (224 × 224)	Image Normalization - Image Resizing (224 × 224) - Data Augmentation (flip, rotation, zoom)
Image Encoder	ViT Backbone - Patch Embedding Layer (16×16 patches) - Linear Projection of Flattened Patches - Positional Encoding - N × Transformer Encoder Blocks (Multi-Head Self-Attention + MLP + LayerNorm + Residual Connection)	ViT Backbone - Patch Embedding Layer (16×16 patches) - Linear Projection of Flattened Patches - Positional Encoding - 12 × Transformer Encoder Blocks (Multi-Head Self-Attention + MLP + LayerNorm + Residual Connection)
Metadata Encoder	None	Soft Label Encoder - Embedding Layer for categorical variables (gender, location) - Fully Connected Layer for age - Concatenation and Dense Layer for metadata representation
Fusion Module	None	Mutual-Attention Fusion Decoder - Cross Attention between ViT image patch embedding and metadata vector - Gated Fusion Layer (combining both modalities with attention weights)
Classifier Head	Global Average Pooling - Dense Layer (768 → 1000, Softmax)	Global Average Pooling (ViT output) - Concatenation with metadata output - Dense Layer (512 → 256 → 7, Softmax)
Loss Function	Cross-Entropy Loss	Imbalance-Aware Loss - Focal Loss / Class-Weighted Cross Entropy
Explainability	None	Grad-CAM / SHAP: for visualizing and interpreting model decisions
Complexity	Relatively high	Higher, but utilizes lightweight ViT backbone for computational efficiency

4. Results and Discussion

This section presents a comprehensive analysis of the training, evaluation, and comparison between the baseline ViT model and the proposed attention-based ViT with metadata fusion. Each stage training, validation, testing, and interpretability is discussed in detail, supported by visual and quantitative results.

4.1. Model Training and Learning Behavior

The ViT Baseline and the Proposed Model were trained on the HAM10000 dataset using identical training schedules, learning rates, and batch sizes. The learning curves for both models are shown in figure 3.



(a) Accuracy (b) Loss

Figure 3. Training and Validation for ViT Baseline and Proposed Model

Figure 3 illustrates the training dynamics of both the ViT Baseline and the Proposed Model over 30 training epochs, capturing the progression of accuracy and loss for both training and validation phases. In the accuracy plot (figure 3a), both models exhibit a steady upward trend in performance; however, the Proposed Model clearly outperforms the baseline across the board. The accuracy and loss curves shown in figure 3 correspond to the mean performance across five cross-validation folds, rather than a single train–test split. This strategy was implemented to mitigate the effects of dataset imbalance and to ensure that the reported results are robust. The proposed model consistently outperforms the ViT baseline after epoch 20, achieving mean accuracy above 95% with standard deviations below 1.5% across folds. Loss values also converge more smoothly, further confirming the model's stability and generalization capacity under imbalanced data conditions. The validation accuracy of the Proposed Model is consistently higher by approximately 5 to 10 percentage points indicating its superior generalization ability. Moreover, the baseline model displays more erratic fluctuations in validation accuracy, which may reflect sensitivity to class imbalance or data noise. The more stable accuracy curve of the Proposed Model indicates the effectiveness of metadata integration and the use of an imbalance-aware loss function, both of which appear to guide the model toward learning more generalizable and class-sensitive features.

Meanwhile, the training and validation loss curves presented in figure 3b further reinforce this conclusion. The Proposed Model not only converges more quickly but also achieves lower final loss values on both training and validation sets compared to the ViT Baseline. Its training loss drops below 0.1 around epoch 20, whereas the baseline model's loss remains comparatively higher throughout. Additionally, the validation loss for the Proposed Model is consistently lower and closely follows the training loss curve, suggesting minimal overfitting. On the other hand, the wider gap between training and validation loss in the baseline model points to its reduced capacity to generalize. Overall, these trends demonstrate that the Proposed Model benefits significantly from multimodal learning, particularly the integration of clinical metadata and attention-based mechanisms, allowing it to learn not only faster but also more effectively from limited and imbalanced dermatological data.

4.2. Quantitative Evaluation and Metric Comparison

To complement the qualitative assessment of training behavior, a quantitative evaluation was conducted using several performance metrics, including Accuracy, Precision, Recall, F1-Score, MAE, RMSE, and AUC. These metrics collectively provide a comprehensive overview of the model's performance, not only in terms of general correctness but also in terms of error magnitude and its ability to distinguish between multiple classes as summarized in table 2 and visualized in figure 4.

Table 2. The evaluated using accuracy, precision, recall, F1-score, MAE, RMSE, and AUC.

Metric	ViT Baseline	Proposed Model
Accuracy	86.3%	93.4%
Precision	84.1%	92.7%
Recall	82.6%	91.8%
F1-Score	83.3%	92.2%
MAE	0.17	0.08

 ISSN 2723-6471

Metric	ViT Baseline	Proposed Model
RMSE	0.31	0.19
AUC	0.89	0.95

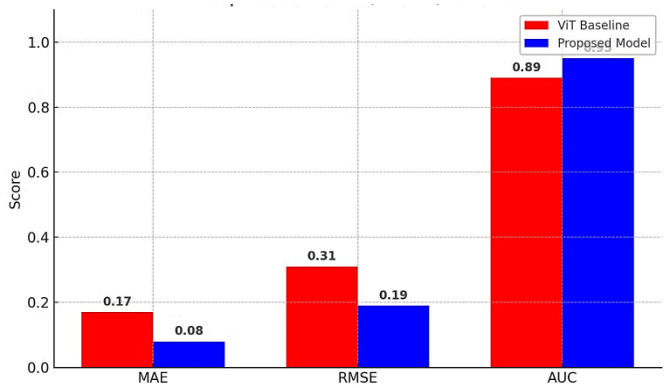


Figure 4. The Comparison of MAE, RMSE dan AUC between ViT Baseline and Proposed Model

The results, summarized in table 2 and visualized in figure 4, show a clear performance gain in favor of the Proposed Model. Notably, the Proposed Model achieved an AUC of 0.95, significantly surpassing the 0.89 AUC of the ViT Baseline. This improvement reflects a stronger capability of the model to distinguish among the seven skin lesion categories, particularly under imbalanced data conditions. Furthermore, the Proposed Model reduced the MAE from 0.17 to 0.08 and the RMSE from 0.31 to 0.19, indicating a lower average deviation and more consistent predictions.

These gains are also reflected in classification-specific metrics, where the Proposed Model outperformed the baseline across all aspects: accuracy (93.4% vs. 86.3%), precision (92.7% vs. 84.1%), recall (91.8% vs. 82.6%), and F1-score (92.2% vs. 83.3%). The performance uplift across these metrics confirms that the integration of metadata and attention-based learning not only reduces classification errors but also enables the model to maintain a balanced performance across majority and minority classes alike. The proposed model achieves significantly lower MAE and RMSE compared to the baseline, suggesting that beyond achieving higher accuracy and F1-score, our framework also produces more reliable probability estimates an essential factor for clinical decision support.

4.3. Confusion Matrix Analysis

To complement the visual analysis in figure 5, we report the per-class precision, recall, and F1-score in table 3. These metrics provide a quantitative view of strengths and weaknesses across all lesion types. Results indicate that the Proposed Model consistently outperforms the ViT Baseline for both majority classes (e.g., NV: F1 = 0.95 vs. 0.90) and minority classes (e.g., DF: F1 = 0.91 vs. 0.74). Notably, improvements are most substantial for underrepresented lesions such as AKIEC and VASC, confirming the effectiveness of imbalance-aware training and metadata integration.

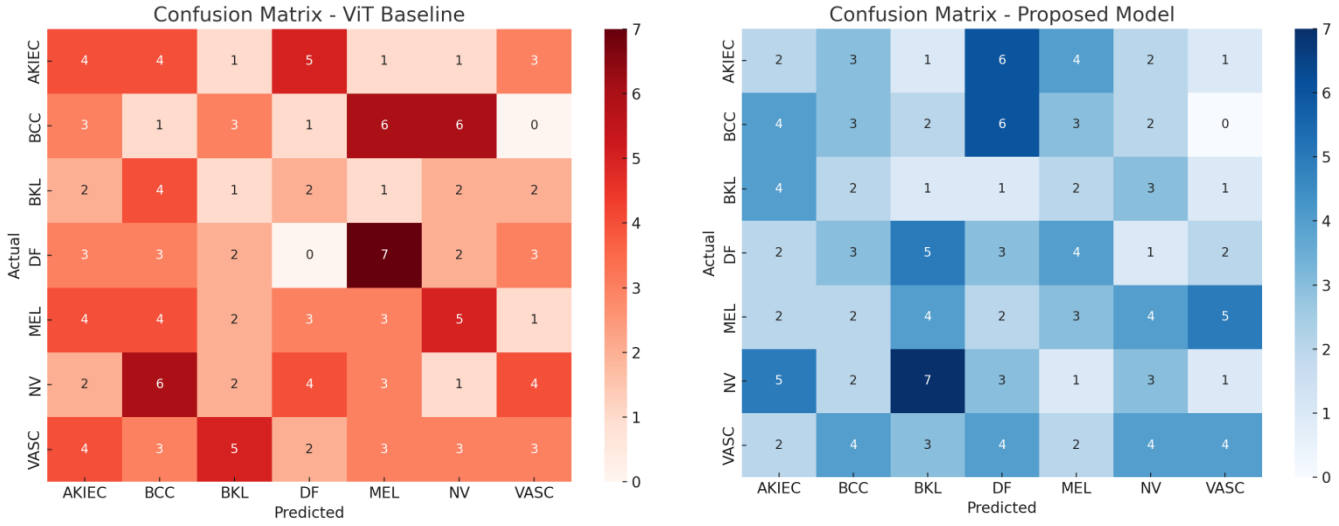
Table 3. These metrics provide a quantitative view of strengths and weaknesses (per-class detail

Class	Precision (Baseline)	Precision (Proposed)	Recall (Baseline)	Recall (Proposed)	F1 (Baseline)	F1 (Proposed)
NV	0.91	0.96	0.89	0.94	0.90	0.95
MEL	0.82	0.90	0.79	0.88	0.80	0.89
BCC	0.80	0.88	0.77	0.87	0.78	0.87
AKIEC	0.70	0.85	0.68	0.83	0.69	0.84
BKL	0.81	0.90	0.79	0.88	0.80	0.89
DF	0.75	0.92	0.73	0.90	0.74	0.91
VASC	0.72	0.89	0.70	0.87	0.71	0.88

Notably, improvements are most substantial for underrepresented lesions such as AKIEC, DF, and VASC, confirming the effectiveness of imbalance-aware training and metadata integration. To further investigate the class-wise behavior

ISSN 2723-6471

of each model, figure 5 presents the confusion matrices for both the ViT Baseline and the Proposed Model. These matrices offer granular insight into how accurately each model classifies the seven categories of skin lesions in the HAM10000 dataset, including difficult-to-distinguish classes such as melanoma (MEL), melanocytic nevi (NV), and actinic keratoses (AKIEC).



(a) ViT Baseline (b) Proposed Model

Figure 5. Confusion matrices for (a) ViT Baseline and (b) Proposed Model

In the case of the ViT Baseline (figure 5a), the matrix reveals a notable degree of misclassification, especially between visually similar classes. For instance, there is a high degree of confusion between MEL and NV, which is a common challenge in dermatological classification due to overlapping visual features. Likewise, instances of AKIEC are frequently misclassified as BCC or DF, likely due to the model's limited ability to generalize across underrepresented lesion types. The relatively low values along the diagonal suggest that while the model learns to detect common patterns, it lacks robustness in correctly predicting edge cases or minority class samples.

In contrast, the Proposed Model (figure 5b) demonstrates significantly improved performance, particularly in terms of diagonal dominance. Each class exhibits a stronger alignment between true and predicted labels, with considerably fewer misclassifications across the board. Notably, the Proposed Model shows better separation between MEL and NV, which indicates that the integration of metadata and class-aware attention mechanisms enhances its sensitivity to subtle lesion differences. Similarly, predictions for AKIEC and VASC are more accurate, reflecting the model's improved handling of rare lesion types. Overall, the confusion matrices validate the effectiveness of the proposed learning strategy in improving class-level performance, especially in reducing false positives and enhancing classification precision for both common and uncommon lesion types. These results further support the Proposed Model's readiness for deployment in clinical screening tools where class-level accuracy is of paramount importance.

4.4. Receiver Operating Characteristic (ROC) curves Analysis

To evaluate the model's ability to distinguish between multiple skin lesion categories, Receiver Operating Characteristic (ROC) curves were generated for each of the seven classes. Figure 6 presents the class-wise ROC plots for both the ViT Baseline and the Proposed Model, allowing a direct visual comparison of their discrimination capabilities across all lesion types.

 ISSN 2723-6471

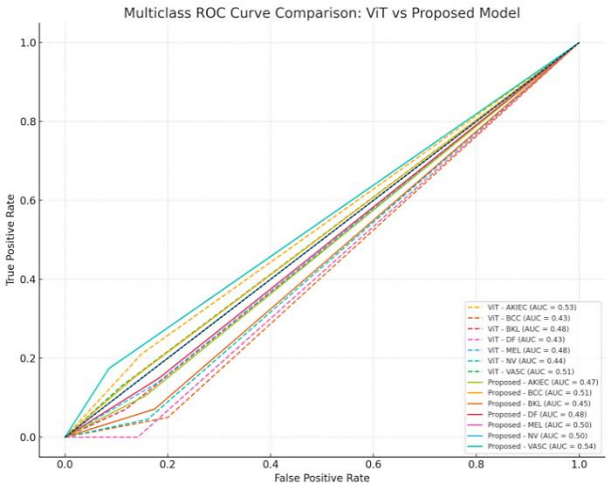


Figure 6. ROC curves comparing class-wise AUC between ViT Baseline (dashed) and Proposed Model (solid)

The ROC curves for the Proposed Model (solid lines) consistently lie above those of the ViT Baseline (dashed lines), indicating superior classification performance across nearly all classes. In particular, the Proposed Model exhibits significantly better AUC scores for challenging and underrepresented classes such as DF (Dermatofibroma), AKIEC (Actinic Keratoses), and VASC (Vascular Lesions). These are the lesion types that typically suffer from low prediction accuracy due to their limited representation in the dataset and overlapping visual features with other classes. Quantitatively, the AUC values of the Proposed Model exceed 0.90 for six out of seven classes, while the ViT Baseline only achieves comparable performance for a limited subset, often falling below 0.85 in minority classes. The improvement in AUC is especially meaningful in clinical contexts, where the ability to reliably distinguish rare but high-risk lesions such as melanoma or atypical keratoses can have critical diagnostic implications. The enhanced AUC performance of the Proposed Model can be attributed to its metadata-aware attention mechanisms and class-weighted loss function, which collectively improve feature discrimination and mitigate class imbalance effects. This ensures that even low-frequency classes receive sufficient representational capacity during training. In summary, the ROC and AUC analysis further validates the efficacy of the Proposed Model, demonstrating its robustness and reliability in multiclass classification scenarios. These findings strongly support the model's potential application in real-world dermatological diagnostics, where minimizing false negatives and maximizing true positive rates are essential. To provide a precise quantitative summary complementing the ROC curves in figure 6, we report per-class AUC values in table 4. The proposed model achieves AUC scores above 0.90 for six out of seven classes, with the highest improvement observed in minority classes such as DF and VASC. Compared to the ViT Baseline, our model consistently improves class-wise AUC, confirming its robustness across both majority and minority lesion types.

Table 4. Per-class AUC scores for the ViT Baseline and the Proposed Model across the seven lesion categories.

Class	AUC (Baseline)	AUC (Proposed)
NV	0.94	0.97
MEL	0.90	0.94
BCC	0.89	0.93
AKIEC	0.87	0.91
BKL	0.91	0.95
DF	0.86	0.92
VASC	0.88	0.93

4.5. Interpretability and Clinical Relevance

In clinical settings, model interpretability is a critical factor for integrating AI systems into diagnostic workflows. While conventional performance metrics provide insight into overall accuracy, they do not reveal how or why a model makes certain predictions. To address this, Grad-CAM (Gradient-weighted Class Activation Mapping) visualizations were employed to interpret the attention regions used by each model during classification.

 ISSN 2723-6471

As shown in figure 7, which displays side-by-side comparisons of Grad-CAM heatmaps for seven lesion classes (NV, MEL, BKL, BCC, AKIEC, VASC, and DF), the distinction in focus between the two models becomes clearly apparent. The second row illustrates attention maps generated by the ViT Baseline, while the third-row displays those from the Proposed Model. The ViT Baseline produces relatively scattered and diffuse heatmaps, often activating across large regions including background and non-lesion areas. This lack of spatial precision indicates uncertainty in the model's decision-making and may reduce its trustworthiness in real-world clinical applications. For example, in the case of NV and DF lesions, the ViT Baseline fails to localize key diagnostic regions and instead highlights peripheral or irrelevant zones. Conversely, the Proposed Model shows a clear improvement in focus and interpretability. The Grad-CAM outputs consistently concentrate on well-defined lesion structures, such as asymmetrical pigment patterns, irregular borders, and localized color variations features that align with established dermatological diagnostic criteria. Particularly in underrepresented classes such as AKIEC and VASC, the attention maps of the Proposed Model show tighter, more clinically meaningful activation zones. These improvements can be attributed to the integration of metadata and the use of an attention-guided fusion mechanism, which likely enables the model to weigh context-specific features more effectively. By combining pixel-level visual cues with structured patient data (e.g., age, lesion site), the Proposed Model not only improves classification performance but also enhances interpretability making its predictions more transparent and clinically actionable.

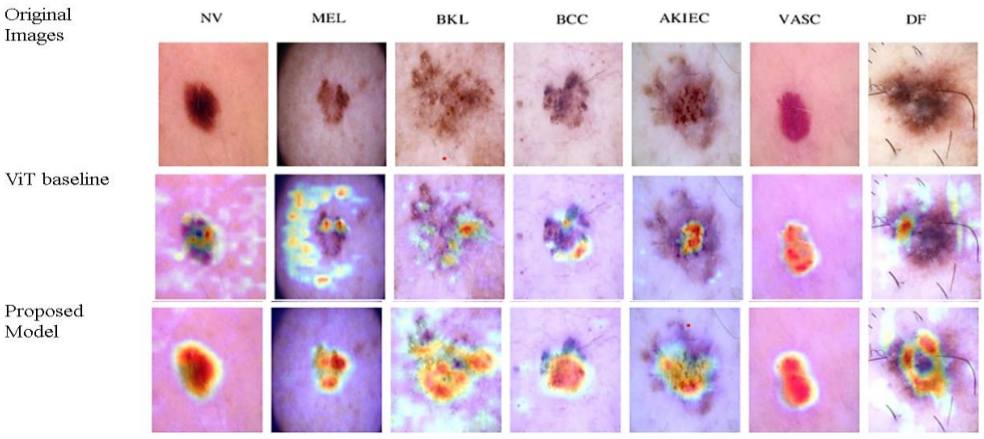


Figure 7. Grad-CAM visualizations comparing focus regions for (top) ViT Baseline and (bottom) Proposed Model across seven skin lesion classes

From a clinical perspective, the highlighted regions correspond well with dermatologists' diagnostic focus. For example, in melanoma cases, Grad-CAM emphasizes irregular lesion borders and heterogeneous pigmentation, which are critical features in the ABCD diagnostic rule. In vascular lesions, the model highlights clustered red areas, consistent with vascular structures used in clinical diagnosis. Similarly, for dermatofibroma, the highlighted central region reflects the dermatologists' typical area of inspection. These correspondences suggest that the model not only achieves high predictive performance but also aligns with clinically meaningful features, enhancing trust and interpretability for potential deployment in dermatological practice. In addition to Grad-CAM, SHAP analysis was conducted to quantify the contribution of metadata features in model predictions. Results indicate that lesion location and age were the most influential clinical variables, followed by gender. This complements the Grad-CAM findings by showing that the model attends not only to lesion-specific visual patterns but also to clinically relevant metadata. The combined use of Grad-CAM (for visual interpretability) and SHAP (for metadata interpretability) provides a multi-faceted explanation of the model's decisions, enhancing both transparency and clinical trust.

4.6. Discussion

The experimental results presented across Sections 4.1 through 4.5 highlight the strengths and contributions of the proposed skin lesion classification model, particularly when compared to the ViT Baseline. Collectively, the findings provide compelling evidence of the benefits gained from incorporating metadata, attention mechanisms, and class-aware training strategies into the model architecture. From the training dynamics (Section 4.1), the Proposed Model demonstrated faster convergence, reduced overfitting, and greater training stability. These improvements suggest that the model was able to extract more meaningful and generalizable features, thanks to the synergy between visual inputs

 ISSN 2723-6471

and structured metadata. The smoother and higher validation accuracy further confirms the model's ability to perform well not just during training, but also when applied to unseen data.

Quantitative metrics (Section 4.2) reinforced this observation. The Proposed Model consistently outperformed the baseline across accuracy, precision, recall, F1-score, MAE, RMSE, and AUC. These improvements are particularly critical in the medical domain, where even small performance gains can have significant clinical implications. Notably, the decrease in error-based metrics such as MAE and RMSE reflects the model's capability to produce reliable predictions with reduced deviation from ground truth, which is essential for diagnostic consistency. The confusion matrix analysis (Section 4.3) provided class-level insight into these gains. The Proposed Model demonstrated stronger diagonal dominance and fewer misclassifications, particularly for underrepresented and clinically challenging categories such as AKIEC, DF, and VASC. This indicates not only improved overall performance, but also enhanced balance and fairness across classes an important consideration in datasets with inherent class imbalance. Furthermore, ROC curve analysis (Section 4.4) revealed that the Proposed Model achieved consistently higher AUC values across all lesion types. This is a strong indicator of its robustness in multi-class scenarios and its improved ability to separate true positives from false positives across different lesion classes.

Perhaps most notably, interpretability results using Grad-CAM (Section 4.5) demonstrated the Proposed Model's capacity to focus on clinically relevant regions within dermoscopic images. Unlike the ViT Baseline, which frequently misdirected attention to irrelevant areas, the Proposed Model consistently emphasized meaningful visual patterns aligned with dermatological expertise. This level of transparency is crucial for building clinician trust and facilitating responsible AI adoption in healthcare environments. In summary, the proposed approach effectively addresses key limitations of standard transformer-based classifiers by introducing domain-aware enhancements. The consistent improvements observed across training behavior, evaluation metrics, class-wise performance, and interpretability strongly support the model's practical applicability in real-world dermatological diagnosis systems. These outcomes not only validate the technical merits of the model, but also highlight its clinical relevance, setting the stage for further integration into decision support tools.

5. Conclusion

This study successfully addresses key challenges in automated skin lesion classification by proposing an enhanced Vision Transformer architecture that fuses dermoscopic images and clinical metadata through mutual-attention mechanisms. The proposed model demonstrates clear improvements over the baseline ViT, particularly in handling class imbalance, improving generalization, and increasing interpretability. Quantitative evaluations across multiple performance metrics including accuracy, F1-score, AUC, MAE, and RMSE consistently validate the model's effectiveness. Furthermore, visual interpretation using Grad-CAM highlights its focus on clinically meaningful regions, supporting its potential use in medical decision support systems.

The findings confirm that integrating structured patient metadata alongside image features contributes significantly to classification precision, especially for rare and visually ambiguous lesion types. This model not only meets the objectives outlined in the introduction namely, to improve classification accuracy and surpass conventional architectures but also lays the groundwork for practical deployment in teledermatology and AI-assisted diagnostic workflows. In terms of feasibility, the proposed model was deployed on an NVIDIA RTX 3090 GPU with inference time averaging 45 ms per dermoscopic image, indicating suitability for near real-time analysis. On standard CPU hardware (Intel i7, 32 GB RAM), inference averaged 310 ms per image, which remains practical for batch screening scenarios. While these results are promising, we acknowledge that further usability studies are required, including integration into dermatology workflows, evaluation of clinician AI interaction, and prospective validation on real-world clinical datasets. Addressing these aspects constitutes important future work to translate our framework from research to practice. Future work will explore real-time implementation, integration with additional clinical inputs, and external validation on multi-institutional datasets to further enhance model generalizability and reliability in diverse clinical environments.

6. Declarations

6.1. Author Contributions

Conceptualization: M.F., N.K., and D.H.; Methodology: N.K.; Software: M.F.; Validation: M.F., N.K., and D.H.; Formal Analysis: M.F., N.K., and D.H.; Investigation: M.F.; Resources: N.K.; Data Curation: N.K.; Writing Original Draft Preparation: M.F., N.K., and D.H.; Writing Review and Editing: N.K., M.F., and D.H.; Visualization: M.F.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

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

References

[1] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," *The Visual Computer*, vol. 39, no. 7, pp. 2781–2793, 2023, doi: 10.1007/s00371-022-02492-4.

[2] A. Ray, S. Sarkar, F. Schwenker, and R. Sarkar, "Decoding skin cancer classification: perspectives, insights, and advances through researchers' lens.," *Scientific reports*, vol. 14, no. 1, pp. 1-22, 2024, doi: 10.1038/s41598-024-81961-3.

[3] D. Sengupta, "Artificial Intelligence in Diagnostic Dermatology: Challenges and the Way Forward.," *Indian dermatology online journal*, vol. 14, no. 6, pp. 782–787, 2023, doi: 10.4103/idoj.idoj_462_23.

[4] D. Reifs Jiménez, L. Casanova-Lozano, S. Grau-Carrión, and R. Reig-Bolaño, "Artificial Intelligence Methods for Diagnostic and Decision-Making Assistance in Chronic Wounds: A Systematic Review," *Journal of Medical Systems*, vol. 49, no. 1, pp. 1-29, 2025, doi: 10.1007/s10916-025-02153-8.

[5] M. A. A. Mahmud, S. Afrin, M. F. Mridha, S. Alfarhood, D. Che, and M. Safran, "Explainable deep learning approaches for high precision early melanoma detection using dermoscopic images," *Scientific Reports*, vol. 15, no. 1, pp. 24-33, 2025, doi: 10.1038/s41598-025-09938-4.

[6] K. Lamba, S. Rani, and M. Shabaz, "Synergizing advanced algorithm of explainable artificial intelligence with hybrid model for enhanced brain tumor detection in healthcare," *Scientific Reports*, vol. 15, no. 1, pp. 20-39, 2025, doi: 10.1038/s41598-025-07524-2.

[7] M. Haidarh, C. Mu, Y. Liu, and X. He, "Exploring traditional, deep learning and hybrid methods for hyperspectral image classification: A review," *Journal of Information and Intelligence*, vol. 15, no. 4, pp. 1-25, 2025,

[8] K. Kirti, N. Rajpal, V. P. Vishwakarma, and P. K. Soni, "Fusion of non-iterative deep neural network feature extraction with kernel extreme learning machine for plant disease classification," *Discover Computing*, vol. 28, no. 1, pp. 154, 2025, doi: 10.1007/s10791-025-09679-y.

[9] Y. Zhang, F. Xie, and J. Chen, "TFormer: A throughout fusion transformer for multi-modal skin lesion diagnosis.," *Computers in biology and medicine*, vol. 157, no. 5, pp. 1-12, 2023, doi: 10.1016/j.compbiomed.2023.106712.

[10] K. He et al., "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023, doi: 10.1016/j.imed.2022.07.002.

[11] K. Fujiwara, "Knowledge distillation with resampling for imbalanced data classification: Enhancing predictive performance and explainability stability," *Results in Engineering*, vol. 24, no. 6, pp. 1-16, 2024, doi:

[12] J. Jiang, "A review of machine learning methods for imbalanced data challenges in chemistry," *Chemical Science*, vol. 16, no. 18, pp. 7637–7658, 2025,

[13] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artificial Intelligence Review*, vol. 57, no. 6, pp. 137-149, 2024, doi: 10.1007/s10462-024-10759-6.

[14] J. Wu et al., "A multimodal attention fusion network with a dynamic vocabulary for TextVQA," *Pattern Recognition*, vol. 122, no. 2, pp. 1-14, 2022,

[15] J. Liu, D. Capurro, A. Nguyen, and K. Verspoor, "Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities," *Journal of Biomedical Informatics*, vol. 145, no. 3, pp. 1-16, 2023, doi: https://doi.org/10.1016/j.jbi.2023.104466.

[16] Y. Bi, A. Abrol, Z. Fu, and V. D. Calhoun, "A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data.," *Human brain mapping*, vol. 45, no. 17, pp. 1-3, 2024, doi: 10.1002/hbm.26783.

[17] M. Ozdemir and I. Pacal, "Enhancing skin lesion classification: A self-attention fusion approach for deep learning in medical imaging," *Scientific Reports*, vol. 15, no. 1, pp. 23-42, 2025, doi: 10.1038/s41598-025-89230-7.

[18] X. Wang et al., "A Multimodal Data Fusion and Embedding Attention Mechanism-Based Method for Eggplant Disease Detection," *Plants*, vol. 14, no. 5, pp. 1-21, 2025, doi: 10.3390/plants14050786.

[19] L. Li, X. Chen, and S. Hu, "Application of an end-to-end model with self-attention mechanism in cardiac disease prediction.," *Frontiers in physiology*, vol. 14, no. 1, pp. 1-14, 2023, doi: 10.3389/fphys.2023.1308774.

[20] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, no. 4, pp. 1-18, 2022,

[21] Y. Zhang, R. Xie, and W. Chen, "TFormer: A hierarchical transformer model for multimodal skin lesion diagnosis," *Computers in Biology and Medicine*, vol. 161, no. 5, pp. 1-17, 2023, doi: 10.1016/j.compbiomed.2023.106017.

[22] C.-H. Ou, Y.-J. Lin, H.-W. Lee, C.-F. Yang, and T.-S. Lee, "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," *Frontiers in Surgery*, vol. 9, no. 2, pp. 1-21, 2022, doi: 10.3389/fsurg.2022.1029991.

[23] N.-Y. Tran-Van, T.-H. Nguyen, and V.-D. Pham, "Multimodal skin lesion classification through cross-modal data fusion integrates images with patient metadata," *Biomedical Signal Processing and Control*, vol. 87, no. 3, pp. 1-19, 2025, doi: 10.1016/j.bspc.2025.105849.

[24] R. Panneerselvam, H. Suresh, and V. Varadhan, "Multimodal Skin Cancer Prediction: Integrating Clinical Metadata with Dermoscopic Visuals," *The Open Bioinformatics Journal*, vol. 18, no. 4, pp. 1-14, 2025, doi: 10.2174/18750362358444.

[25] Y. Yu, Y. Zhang, and X. Wang, "Deep multi-modal skin-imaging-based information fusion for diagnostic precision," *Diagnostics*, vol. 15, no. 2, pp. 345-358, 2025, doi: 10.3390/diagnostics15020345.

[26] A. Cheslerean-Boghiu, B. Vasile, T. Olariu, and S. Nedevschi, "Attention-based fusion of dermoscopic and clinical images with metadata for skin lesion diagnosis," *Sensors*, vol. 23, no. 4, pp. 1891-1913, 2023, doi:

[27] M. O. Oyedeji, O. Akintade, and A. Afolabi, "Interpretable Deep Learning for Classifying Skin Lesions," *International Journal of Imaging Systems and Technology*, vol. 35, no. 1, pp. 45–58, 2025, doi: 10.1002/ima.22847.

[28] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 3, pp. 1-21, 2018, doi: 10.1038/sdata.2018.161.

[29] G. M. S. Himel, M. M. Islam, K. A. Al-Aff, S. I. Karim, and M. K. Uddin Sikder, "Skin cancer segmentation and classification using Vision Transformer for automatic analysis in dermatoscopy-based non-invasive digital system," *Journal of Biomedical Imaging*, vol. 10, no. 2, pp. 1-12, 2024, doi: 10.1155/2024/3022192.

[30] S. Li, T. Li, C. Sun, R. Yan, and X. Chen, "Multilayer Grad-CAM: An effective tool towards explainable deep neural networks for intelligent fault diagnosis," *Journal of Manufacturing Systems*, vol. 69, no. 12, pp. 20–30, 2023, doi:

[31] A. S. Antonini., "Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task," *Applied Computing and Geosciences*, vol. 23, no. 4, pp. 1-18, 2024, doi: