JI-ADF: Joint–Individual Learning with Adaptive Decision Fusion for Multimodal Skin Lesion Classification

Phan Nguyen[*1,2] NHPHAN@KAIST.AC.KR
Dat Cao[*1] CTIENDAT987@KAIST.AC.KR
Quang Hien Kha[2,3] D142111015@TMU.EDU.TW
Trang Thao Quoc Pham[4] A01677@TMU.EDU.TW
Nguyen Quoc Khanh Le[†2,5] KHANHLEE@TMU.EDU.TW

[1] *Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea*
[2] *AIBioMed Research Group, Taipei Medical University, Taipei 110301, Taiwan*
[3] *International Ph.D. Program in Medicine, College of Medicine, Taipei Medical University, Taipei 110301, Taiwan*
[4] *International Ph.D. Program in Cell Therapy and Regenerative Medicine, College of Medicine, Taipei Medical University, Taipei 110301, Taiwan*
[5] *In-Service Master Program in Artificial Intelligence in Medicine, College of Medicine, Taipei Medical University, Taipei 110301, Taiwan*

Editors: Under Review for MIDL 2026

Abstract

Skin lesions encompass a wide spectrum of dermatological conditions, and their accurate classification is critical for early diagnosis and treatment planning. The development of deep learning-based Computer-Aided Diagnosis (CAD) system has shown promise in supporting dermatologists, particularly when leveraging multiple data modalities such as dermoscopic images, clinical images, and patient metadata. However, existing multimodal approaches often rely on late fusion or naive concatenation strategies, which fail to capture fine-grained cross-modal interactions. In this paper, we propose **JI-ADF** (Joint Individual with Adaptive Decision Fusion), a trimodal deep learning framework designed to address these limitations. Our approach encodes each modality independently, then unifies them via a cross-attention-based fusion mechanism to capture complementary, fine-grained interactions across dermoscopic, clinical and structured metadata views. We evaluate our method on the MILK10k dataset, a newly released multimodal benchmark spanning 11 diagnostic classes. On the hidden test set, JI-ADF achieves a high AUC (0.866) and Accuracy (0.930), along with high Specificity (0.959) and NPV (0.960). These results outperform recent baseline methods and highlight the effectiveness of our model for real-world multimodal dermatological diagnosis.

Keywords: Multimodal Learning, Skin Lesion Classification, Adaptive Fusion, Dermoscopic Imaging, Clinical Metadata, Medical Image Analysis

1. Introduction

Skin cancer remains one of the most prevalent and clinically significant malignancies worldwide, with both incidence and mortality continuing to rise. Among its subtypes, melanoma

[*] Contributed equally
[†] Corresponding author

is particularly aggressive and accounts for a disproportionate share of skin cancer deaths. Globally, over 330,000 new melanoma cases are diagnosed each year, with nearly 60,000 melanoma-related deaths annually (Bray et al., 2024). Crucially, survival outcomes are highly stage-dependent: the 5-year survival rate is 99% for localized melanoma, while this rate drops to just approximately 35% once distant metastases occur (American Cancer Society, 2024). As incidence rates is projected to increase by over 51% by 2040, early detection remains the most impactful determinant of long-term prognosis (Ferlay et al., 2024).

Dermatologists traditionally rely on structured visual heuristics such as the ABCD rule and the 7-point checklist, which summarize lesion asymmetry, border irregularity, colour and size, as well as characteristic network and vascular patterns into simple malignancy scores. Although these systems guide clinical reasoning, they remain limited by subjectivity and the ambiguity of early or atypical lesions, with performance varying by practitioner expertise. Recent advances in deep learning-based computer-aided diagnosis (CAD) offer greater scalability and consistency, with state-of-the-art models achieving dermatologist-level accuracy; meta-analyses report roughly 87% sensitivity for leading AI systems, slightly exceeding average clinician performance (Salinas et al., 2024).

However, most CAD systems still rely solely on dermoscopic images, overlooking the valuable context provided by clinical photographs and structured metadata such as demographics, lesion location, and skin tone—information routinely used by dermatologists. Multimodal learning addresses this gap by jointly modeling heterogeneous inputs, better reflecting real clinical workflows. Although early multimodal methods show potential, many depend on late fusion or simple feature concatenation, which fail to capture meaningful cross-modal interactions. Moreover, metadata remains underutilized in many dermatology pipelines despite its availability in several datasets.

In this study, we introduce JI-ADF, a deep learning framework that integrates dermoscopic images, clinical images, and patient metadata through a unified, attention-based fusion mechanism. The contributions of this paper are as follows.

- We introduce JI-ADF, a trimodal fusion framework that jointly models dermoscopic images, clinical images, and structured metadata through a unified cross-attention mechanism, enabling fine-grained and clinically meaningful interaction across modalities.
- We design an adaptive decision fusion module and a unified classification head with class-balanced optimization, allowing the model to dynamically weigh modality contributions and remain robust under severe class imbalance.
- We conduct extensive experiments and ablation studies on the MILK10k benchmark, demonstrating that JI-ADF consistently outperforms existing multimodal approaches and provides strong, balanced performance across diverse lesion types.

2. Related Work

Recent research in skin lesion analysis has increasingly explored the integration of visual and contextual information to enhance diagnostic accuracy. Early approaches relied primarily on image-based deep learning models, where CNNs such as VGG, ResNet, and DenseNet were trained on dermoscopic or clinical photographs. For example, (Shetty et al., 2022) fine-tuned CNNs on the HAM10000 dataset (Tschandl et al., 2018) and achieved over 95% accuracy with extensive augmentation, while (Thwin and Park, 2024) introduced a deep ensemble of multiple CNNs to improve robustness across lesion categories. Although effective at feature extraction, these image-only methods remain limited in real-world scenarios, where visually similar lesions often produce ambiguous predictions.

To overcome these limitations, recent studies have incorporated patient metadata alongside images, forming multimodal frameworks that better reflect clinical decision-making. (Gessert et al., 2020; Kawahara et al., 2018) demonstrated that adding tabular information such as age, sex, and anatomical site to CNN features substantially improved classification. Later works introduced more adaptive fusion strategies, including attention-based and cross-modal mechanisms. For example, (Adebiyi et al., 2025) reported notable gains when combining dermoscopic images with metadata on the HAM10000 dataset, while (Tran-Van and Le, 2025) showed that cross-attention can dynamically highlight clinically relevant cues and enhances generalization under domain shifts.

Following these advances, our approach employs a weighted fusion strategy that allows the model to learn how much each modality should contribute to the final prediction. Instead of assuming equal importance between image and metadata features, the fusion layer assigns learnable weights reflecting their relative diagnostic value. This design captures the intuition that visual cues dominate for some lesion types, whereas metadata plays a stronger role for others. By letting the model adjust these weights during training, the fusion process becomes both adaptive and interpretable, achieving more reliable predictions across diverse clinical conditions.

3. Method

3.1. Joint Fusion Structure

For convenience, we consider the fusion of a dermoscopic image, a clinical close-up image, and patient metadata for skin lesions diagnosis as a multi-class classification task. Each case contains a dermoscopic image I_{derm}, a close-up image I_{close}, patient metadata M, and a label $y \in \{1, \ldots, N\}$. M_I is the model to extract features of the 2 images and M_M is the method to extract features from patient metadata M:

$$\mathbf{f}_I = M_I\big(I_{\text{derm}}, I_{\text{close}}\big) \in \mathbb{R}^{D_I}, \qquad \mathbf{f}_M = M_M(M) \in \mathbb{R}^{D_M}. \tag{1}$$

A fusion module FM produces the joint representation

$$\mathbf{f}_{IM} = FM(\mathbf{f}_I, \mathbf{f}_M) \in \mathbb{R}^{D_{IM}}, \tag{2}$$

which a classifier C_{IM} maps to logits $\mathbf{z}_{IM} = C_{IM}(\mathbf{f}_{IM})$ and posteriors

$$\mathbf{P}_{IM} = \operatorname{softmax}(\mathbf{z}_{IM}) \in \Delta^{N-1}. \tag{3}$$

With one-hot target $\mathbf{y}$, the joint loss used in the figure is

$$\mathcal{L}_{IM} = -\sum_{c=1}^{N} y_c \log P_{IM}^{(c)}. \tag{4}$$

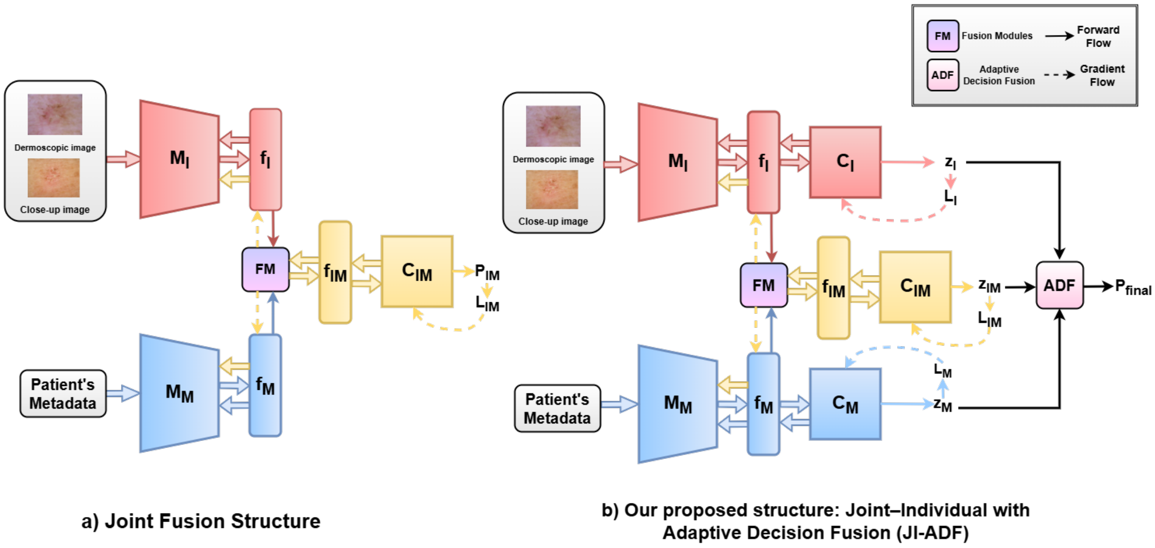
Because $\mathbf{f}_{IM}$ depends on both streams, the loss backpropagates through the fusion node:

$$\frac{\partial \mathcal{L}_{IM}}{\partial \theta_{M_I}} = \frac{\partial \mathcal{L}_{IM}}{\partial \mathbf{z}_{IM}} \frac{\partial \mathbf{z}_{IM}}{\partial \mathbf{f}_{IM}} \frac{\partial \mathbf{f}_{IM}}{\partial \mathbf{f}_{I}} \frac{\partial \mathbf{f}_{I}}{\partial \theta_{M_I}}, \tag{5}$$

$$\frac{\partial \mathcal{L}_{IM}}{\partial \theta_{M_M}} = \frac{\partial \mathcal{L}_{IM}}{\partial \mathbf{z}_{IM}} \frac{\partial \mathbf{z}_{IM}}{\partial \mathbf{f}_{IM}} \frac{\partial \mathbf{f}_{IM}}{\partial \mathbf{f}_{M}} \frac{\partial \mathbf{f}_{M}}{\partial \theta_{M_M}}. \tag{6}$$

Summary. Inputs $\big(I_{\text{derm}}, I_{\text{close}}, M\big)$ are encoded into $(\mathbf{f}_I, \mathbf{f}_M)$, fused by FM into $\mathbf{f}_{IM}$, and classified by C_{IM} to yield $\mathbf{P}_{IM}$, matching the forward and gradient flows in Figure 1.

3.2. Proposed Architecture: Joint–Individual with Adaptive Decision Fusion



a) Joint Fusion Structure

b) Our proposed structure: Joint–Individual with Adaptive Decision Fusion (JI-ADF)

Figure 1: Illustration of (a) the Joint Fusion baseline and (b) our proposed Joint–Individual architecture with Adaptive Decision Fusion. JI-ADF extends the baseline by adding individual prediction heads for each modality and an adaptive fusion module that assigns instance-dependent weights.

Backbone streams. From the two images and the metadata, the modality encoders produce features

$$\mathbf{f}_I = M_I\big(I_{\text{derm}}, I_{\text{close}}\big) \in \mathbb{R}^{D_I}, \qquad \mathbf{f}_M = M_M(M) \in \mathbb{R}^{D_M}. \tag{7}$$

A differentiable fusion module aggregates them into a joint representation

$$\mathbf{f}_{IM} = FM(\mathbf{f}_I, \mathbf{f}_M) \in \mathbb{R}^{D_{IM}}. \tag{8}$$

Branch classifiers and auxiliary supervision. Each stream has its own classifier head

$$\mathbf{z}_I = C_I(\mathbf{f}_I) \in \mathbb{R}^N, \qquad \mathbf{P}_I = \text{softmax}(\mathbf{z}_I), \tag{9}$$

$$\mathbf{z}_M = C_M(\mathbf{f}_M) \in \mathbb{R}^N, \qquad \mathbf{P}_M = \text{softmax}(\mathbf{z}_M), \tag{10}$$

$$\mathbf{z}_{IM} = C_{IM}(\mathbf{f}_{IM}) \in \mathbb{R}^N, \qquad \mathbf{P}_{IM} = \text{softmax}(\mathbf{z}_{IM}). \tag{11}$$

With one-hot target $\mathbf{y}$ we use cross-entropy on all three branches

$$\mathcal{L}_I = -\sum_{c=1}^{N} y_c \log P_I^{(c)}, \qquad \mathcal{L}_M = -\sum_{c=1}^{N} y_c \log P_M^{(c)}, \qquad \mathcal{L}_{IM} = -\sum_{c=1}^{N} y_c \log P_{IM}^{(c)}. \tag{12}$$

Adaptive Decision Fusion (ADF). Instead of a fixed average at the decision level, we learn per-sample fusion weights from the joint evidence of all heads. Let $\mathbf{s} = [\,\mathbf{z}_I \,\|\, \mathbf{z}_{IM} \,\|\, \mathbf{z}_M\,] \in \mathbb{R}^{3N}$ be the concatenated logits, where $[\cdot\|\cdot]$ denotes vector concatenation. A lightweight gating network produces simplex weights

$$\boldsymbol{\alpha} = \mathrm{softmax}\big(W_2\,\sigma(W_1\mathbf{s} + \mathbf{b}_1) + \mathbf{b}_2\big) = (\alpha_I, \alpha_{IM}, \alpha_M) \in \Delta^2, \tag{13}$$

where $\sigma(\cdot)$ denotes a pointwise (element-wise) nonlinearity and $\sum_{k\in\{I,IM,M\}} \alpha_k = 1$. The final posterior is a convex combination of branch posteriors

$$\mathbf{P}_{\text{final}} = \alpha_I \mathbf{P}_I + \alpha_{IM}\mathbf{P}_{IM} + \alpha_M \mathbf{P}_M \in \Delta^{N-1}. \tag{14}$$

We use a softmax head and take the prediction by $\hat{y} = \arg\max_{c\in\{1,\ldots,N\}} P_{\text{final}}^{(c)}$.

Training objective. We supervise the fused prediction and the auxiliary heads

$$\mathcal{L}_{\text{total}} = \mathrm{CE}\big(\mathbf{P}_{\text{final}}, \mathbf{y}\big) + \lambda_{IM}\,\mathcal{L}_{IM} + \lambda_I\,\mathcal{L}_I + \lambda_M\,\mathcal{L}_M, \tag{15}$$

We fix the auxiliary weights to $\lambda_{IM} = 0.5$ and $\lambda_I = \lambda_M = 0.25$ throughout. This keeps the total auxiliary weight at most equal to the unit weight on the final loss, emphasizes the joint branch that is closest to deployment, and treats the individual branches as regularizers that stabilize training and preserve modality-specific cues.

3.3. Multi-Modal Fusion Attention (MMFA)

We instantiate the fusion module FM as a *multimodal fusion attention* block that lets image and metadata features attend to each other while preserving self-evidence, following prior work (Tang et al., 2024).

Inputs and projections. Given $\mathbf{f}_I \in \mathbb{R}^{D_I}$ and $\mathbf{f}_M \in \mathbb{R}^{D_M}$, each head $h = 1, \ldots, H$ forms modality–specific queries, keys, and values

$$\mathbf{Q}_h = \begin{bmatrix} W_h^{Q,I}\mathbf{f}_I \\ W_h^{Q,M}\mathbf{f}_M \end{bmatrix} \in \mathbb{R}^{2\times d_h}, \quad \mathbf{K}_h = \begin{bmatrix} W_h^{K,I}\mathbf{f}_I \\ W_h^{K,M}\mathbf{f}_M \end{bmatrix} \in \mathbb{R}^{2\times d_h}, \quad \mathbf{V}_h = \begin{bmatrix} W_h^{V,I}\mathbf{f}_I \\ W_h^{V,M}\mathbf{f}_M \end{bmatrix} \in \mathbb{R}^{2\times d_h}, \tag{16}$$

where $W_h^{Q,I}, W_h^{K,I}, W_h^{V,I} \in \mathbb{R}^{d_h\times D_I}$ and $W_h^{Q,M}, W_h^{K,M}, W_h^{V,M} \in \mathbb{R}^{d_h\times D_M}$.

Two-token multi-head attention. Each head computes a 2 × 2 attention over the two modalities and mixes the values:

$$\mathbf{U}_h = \mathrm{softmax}\!\left(\frac{\mathbf{Q}_h\mathbf{K}_h^\top}{\sqrt{d_h}}\right)\mathbf{V}_h \in \mathbb{R}^{2\times d_h}, \tag{17}$$

$$\mathbf{o} = W^O\,\mathrm{Concat}\big(\mathrm{vec}(\mathbf{U}_1), \ldots, \mathrm{vec}(\mathbf{U}_H)\big) \in \mathbb{R}^{D_{IM}}. \tag{18}$$

with $W^O \in \mathbb{R}^{D_{IM}\times(2Hd_h)}$. Here $\mathrm{vec}(\cdot)$ stacks row-wise and $\mathrm{Concat}(\cdot)$ concatenates vectors.

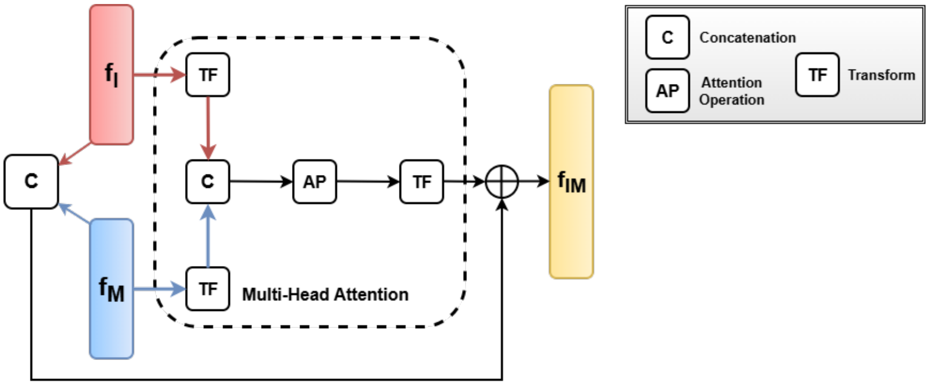


Figure 2: Multimodal Fusion Attention Module (MMFA), where image features f_I and metadata features f_M are jointly refined through cross-attention and self-attention mechanisms between the two modalities. The resulting fused representation f_{IM} captures the enhanced joint features after integration.

Output and residual path. The attention output is re-projected and merged with the raw features via a residual connection

$$\mathbf{f}_{IM} = W_{\text{skip}} \left[\mathbf{f}_I \,\|\, \mathbf{f}_M \right] \; + \; g(\mathbf{o}) \; \in \; \mathbb{R}^{D_{IM}}, \tag{19}$$

where $W_{\text{skip}} \in \mathbb{R}^{D_{IM} \times (D_I + D_M)}$, $g(\cdot)$ is a linear layer followed by a pointwise nonlinearity, and $[\,\cdot\|\cdot\,]$ denotes vector concatenation. This design explicitly models self and mutual interactions (through the 2×2 attention) while keeping a skip path that preserves modality-specific cues and maintains stable gradients back to M_I and M_M.

4. Experiments

We now first describe our experimental set-up before reporting and discussing the results.

4.1. Dataset Description

We used the MILK10k multimodal skin-lesion dataset (Tschandl et al., 2025) to train and evaluate our method. MILK10k dataset contains 10480 images from 5240 lesions, provided as paired clinical close-up and dermoscopic images collected across five centers. Most cases were biopsy-verified (95.7%), and accompanying metadata include age, sex, skin tone, anatomic site, and diagnosis mapped to the ISIC-DX hierarchy (with a simplified 11-category scheme used in this challenge). For testing, we follow the official hidden test comprising 479 lesions with paired images (958 images), provided with the metadata fields as the training set. This data can be found here: https://challenge.isic-archive.com/data/#milk10k.

4.2. Implementation

Our model was trained for 50 epochs with a batch size of 16, using AdamW optimizer with the initial value learning rate of 1e-4 and weight decay of 1e-5. The ReduceLROnPlateau scheduler is applied for the learning rate decay. We split the training set into 80% for training and 20% for validation. The model achieving the highest average Macro F1 Score on the validation set was saved for testing. The ImageNet-1K (Deng et al., 2009) pretrained EfficientNetV2 is employed as the backbone. Input images are resized to $384 \times 384 \times 3$, the length of the encoded patient's metadata is 256, and we use MMFA (Section 3.3) to fuse images and metadata features. All the experiments were performed using Python 3.12 with PyTorch 2.8.0 and run on NVIDIA A100 GPU with 40GB VRAM.

5. Results

5.1. Performance of the proposed model

Category Metric	**Mean**	**Diagnosis Category**										
		AKIEC	BCC	BEN_OTH	BKL	DF	INF	MAL_OTH	MEL	NV	SCCKA	VASC
AUC	0.866	0.903	0.905	0.837	0.834	1.000	0.757	0.521	0.933	0.932	0.912	0.992
AUC, Sens > 80%	0.734	0.816	0.832	0.540	0.632	1.000	0.449	0.214	0.895	0.862	0.848	0.985
Average Precision	0.543	0.690	0.543	0.141	0.591	1.000	0.113	0.026	0.688	0.790	0.769	0.625
Accuracy	0.930	0.881	0.835	0.983	0.850	1.000	0.956	0.979	0.950	0.946	0.866	0.985
Sensitivity	0.536	0.592	0.902	0.000	0.440	1.000	0.273	0.000	0.769	0.654	0.667	0.600
Specificity	0.959	0.931	0.825	0.998	0.958	1.000	0.972	1.000	0.966	0.981	0.931	0.989
Dice Coefficient	0.505	0.596	0.582	0.000	0.550	1.000	0.222	0.000	0.714	0.723	0.709	0.462
PPV	0.596	0.600	0.430	0.000	0.733	1.000	0.188	1.000	0.667	0.810	0.757	0.375
NPV	0.960	0.929	0.983	0.985	0.866	1.000	0.983	0.979	0.979	0.959	0.896	0.996

Table 1: Performance of the proposed JI-ADF method across all diagnostic categories. Each column reports per-class results for the evaluation metrics. The Mean Value column represents the macro-averaged score across all 11 lesion types.

The proposed JI-ADF model delivers strong and balanced performance across categories. As shown in Table 1, it reaches a mean AUC of 0.866, overall accuracy of 0.930, specificity of 0.959, and NPV of 0.960, indicating reliable discrimination with low false-negative rates. The moderate mean sensitivity (0.536) and Dice score (0.505) align with the severe class imbalance in the dataset (Appendix A). The model excels in abundant and visually distinctive classes such as BCC, NV, and DF, achieving high AUCs (0.905–1.000) and accuracies above 0.93, suggesting effective use of both image and metadata cues. In contrast, classes with very limited samples or high variability (BEN_OTH, MAL_OTH) show lower sensitivity and Dice, reflecting the difficulty of learning stable decision boundaries under data scarcity. Nevertheless, the model maintains consistent precision (mean PPV 0.596) and high specificity across almost all categories, indicating confident and reliable predictions. Overall, these results show that the weighted-fusion design generalizes well across diverse lesion types and provides a clinically meaningful foundation for multimodal skin lesion classification.

5.2. Performance comparison with other methods

To contextualize the effectiveness of the proposed JI-ADF framework, we evaluate its performance against several competitive baselines under the same experimental protocol. The results summarized in Table 2 provide a comprehensive comparison across six evaluation metrics (Appendix B).

Table 2: **Comparison with state-of-the-art skin lesion classification approaches.**

Method	**AUC**	**Precision**	**Accuracy**	**Sensitivity**	**Specificity**	**Dice**
JIF-MMFA (Tang et al., 2024)	0.750	0.328	0.896	0.406	0.939	0.306
VEMFL (Restrepo et al., 2024)	0.842	0.439	0.912	0.291	0.958	0.302
CAFFM (Tran-Van and Le, 2025)	0.787	0.306	0.898	0.233	0.954	0.232
ALBEF (Adebiyi et al., 2025)	0.760	0.294	0.889	0.208	0.945	0.222
CoscatNet-UFS (Zuo et al., 2025)	0.800	0.472	0.925	0.453	**0.959**	0.435
DualRefNet (Khurshid et al., 2025)	0.848	0.447	0.918	0.422	0.954	0.441
SkinM2Former (Zhang et al., 2025)	0.841	0.436	0.920	0.421	0.954	0.405
JI-ADF (ours)	**0.866**	**0.543**	**0.930**	**0.536**	**0.959**	**0.505**

Across state-of-the-art multimodal methods, JI-ADF achieves the strongest overall performance, achieving the highest AUC (0.866) and accuracy (0.930), demonstrating robust class separation despite the heterogeneity of MILK10k dataset. Sensitivity is particularly challenging due to severe class imbalance, yet JI-ADF attains the best score (0.536), indicating improved true-positive detection where visual cues alone may be insufficient. These improvements also result in the highest Dice coefficient (0.505) and a competitive precision (0.543), reflecting a balanced trade-off between identifying positives and minimizing false alarms. Although baseline methods already show high specificity, JI-ADF matches or exceeds them (0.959) while maintaining superior recall, suggesting that adaptive decision fusion helps prevent overconfident predictions by dynamically weighting dermoscopic images, clinical photographs, and metadata. Overall, these results indicate that JI-ADF consistently outperforms existing approaches and provides a competitive and reliable solution for multimodal skin lesion classification.

5.3. Grad-CAM Visualizations

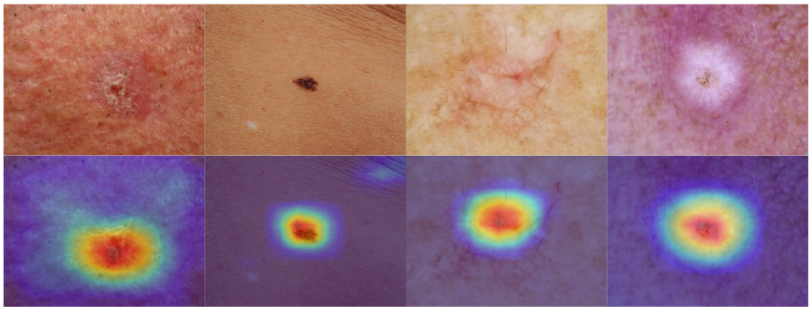


Figure 3: Comparison between the original input images and their corresponding Grad-CAM visualizations. Warmer colors (yellow–red) indicate regions with higher model attention, while cooler colors (blue–green) represent lower emphasis.

To better understand how the proposed model interprets lesion patterns, we generate Grad-CAM visualizations for the test images. As in Figure 3, the resulting heatmaps consistently concentrate on the main lesion regions. This indicates that the model bases its predictions on visually meaningful cues such as pigment distribution, localized texture variations, and boundary characteristics. The attention patterns are well aligned with areas that dermatologists typically examine, suggesting that the multimodal fusion framework encourages the network to focus on diagnostically relevant features. These qualitative visualizations provide additional insight into the model's decision process and support the interpretability of the proposed approach.

6. Ablation Study

6.1. Modality Contribution Analysis

We conducted an ablation study over all unimodal and bimodal subsets of the inputs. Based on the results shown in Table 3, models that combine modalities consistently outperform single-modality variants. In particular, the proposed JI-ADF trimodal fusion achieves the best performance across most metrics, indicating that jointly leveraging clinical images, dermoscopic images, and metadata yields the most reliable classifier.

Table 3: **Modality Configuration Ablation Study.**

Modality Config.	**AUC**	**Precision**	**Accuracy**	**Sensitivity**	**Specificity**	**Dice**
C	0.799	0.415	0.919	0.391	0.952	0.392
D	0.807	0.458	0.923	0.416	0.954	0.420
M	0.796	0.334	0.894	0.367	0.938	0.325
C + D	**0.866**	0.487	0.925	0.464	0.955	0.456
C + M	0.834	0.467	0.918	0.389	0.952	0.376
D + M	0.829	0.446	0.918	0.451	0.952	0.413
JI-ADF (ours)	**0.866**	**0.543**	**0.930**	**0.536**	**0.959**	**0.505**

C = clinical image, D = dermoscopic image, M = metadata.

6.2. Fusion Mechanism Ablation

To examine the effect of different fusion strategies, we compare a sequence of architectural variants that progressively increase the capacity for cross-modal interaction. **Late concat** simply merges the two image embeddings and metadata at the final classifier. **JF-concat** retains this linear merging but introduces three prediction heads trained with auxiliary losses. **JF-MMFA** replaces concatenation with a multimodal attention block to produce a unified representation, using a single joint head for prediction. **JI-MMFA** reinstates the three-head design on top of the attention module and combines their outputs through fixed averaging. **JI-ADF (no aux)** preserves the three-head structure but substitutes fixed averaging with a learnable adaptive fusion module that assigns instance-dependent weights.

Across the four metrics reported in Figure 4, the ablation results show a steady improvement as the fusion design becomes more expressive. MMFA strengthens joint feature learning, with JF-MMFA outperforming both Late concat and JF-concat, suggesting that attention-based fusion captures complementary information more effectively than simple

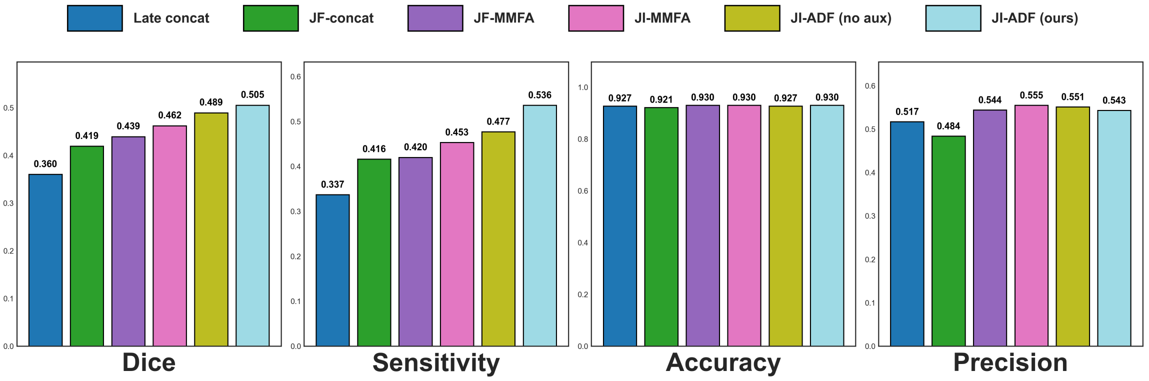


Figure 4: Fusion Architecture Ablation – Multimetrics Comparison

merging. Adding individual heads with auxiliary supervision further stabilizes the optimization process, enabling JI-MMFA to surpass the single-head JF-MMFA. Replacing fixed averaging with adaptive weighting introduces another consistent step forward, as JI-ADF (no aux) benefits from instance-dependent fusion that adjusts to the reliability of each modality. Bringing these components together in the full JI-ADF model yields the most balanced and robust performance overall, indicating that adaptive fusion and auxiliary supervision work together to produce a more stable, balanced, and reliably integrated multimodal representation.

7. Conclusions

In this work, we presented JI-ADF, a trimodal architecture for the multi-class classification of skin lesions. By introducing a unified attention-based fusion block, our model captures cross-modal interactions between dermoscopic images, clinical images, and structured metadata. Combined with class-aware optimization, JI-ADF improves diagnostic performance across both common and underrepresented conditions. Evaluations on the MILK10k benchmark confirm its strong generalization and robustness. We hope that the design principles of JI-ADF offer a scalable foundation for future diagnostic frameworks, particularly in diseases requiring multimodal diagnosis. For example, diagnosing complex conditions like endometriosis often involves identifying multiple lesion sites across different imaging sources, such as MRI and ultrasound.

References

Abdulmateen Adebiyi, Nader Abdalnabi, Emily Hoffman Smith, Jesse Hirner, Eduardo J. Simoes, Mirna Becevic, and Praveen Rao. Accurate skin lesion classification using multimodal learning on the ham10000 and isic 2017 datasets. *medRxiv*, 2025. doi: 10.1101/2024.05.30.24308213. URL https://www.medrxiv.org/content/early/2025/05/20/2024.05.30.24308213.

American Cancer Society. Cancer facts & figures 2024, 2024. URL https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2024/2024-cancer-facts-and-figures-acs.pdf.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. doi: 10.3322/caac.21834.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

J. Ferlay, M. Laversanne, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray. Global cancer observatory: Cancer tomorrow (version 1.1), 2024. URL https://gco.iarc.who.int/tomorrow.

Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta-data. *Artificial Intelligence in Medicine*, 115:102060, 2020. ISSN 0933-3657. doi: 10.1016/j.artmed.2020.101861. See also arXiv:1910.03910.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. 7-point checklist and skin lesion classification using multi-task multi-modal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2018. doi: 10.1109/JBHI.2018.2843298.

Mahapara Khurshid, Richa Singh, and Mayank Vatsa. Multimodal dual-stage feature refinement for robust skin lesion classification. *Scientific Reports*, 15(1):37775, 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-14839-7. URL https://doi.org/10.1038/s41598-025-14839-7.

David Restrepo, Chenwei Wu, Sebastián Andrés Cajas, Luis Filipe Nakayama, Leo Anthony Celi, and Diego M López. Multimodal deep learning for low-resource settings: A vector embedding alignment approach for healthcare applications, 2024. URL https://arxiv.org/abs/2406.02601.

Maria Paz Salinas, Javiera Sepúlveda, Leonel Hidalgo, Dominga Peirano, Macarena Morel, Pablo Uribe, Veronica Rotemberg, Juan Briones, Domingo Mery, and Cristian

Navarrete-Dechent. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. *npj Digital Medicine*, 7(1):125, 2024. doi: 10.1038/s41746-024-01103-x.

Bhuvaneshwari Shetty, Roshan Fernandes, Anisha P. Rodrigues, Rajeswari Chengoden, Sweta Bhattacharya, and Kuruva Lakshmanna. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Scientific Reports*, 12(1):18134, 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-22644-9. URL https://doi.org/10.1038/s41598-022-22644-9.

Peng Tang, Xintong Yan, Yang Nan, Xiaobin Hu, Bjoern H. Menze, Sebastian Krammer, and Tobias Lasser. Joint-individual fusion structure with fusion attention module for multi-modal skin cancer classification. *Pattern Recognition*, 154:110604, 2024. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2024.110604. URL https://www.sciencedirect.com/science/article/pii/S0031320324003558.

Su Myat Thwin and Hyun-Seok Park. Skin lesion classification using a deep ensemble model. *Applied Sciences*, 14(13), 2024. ISSN 2076-3417. doi: 10.3390/app14135599. URL https://www.mdpi.com/2076-3417/14/13/5599.

Nhu-Y Tran-Van and Kim-Hung Le. A multimodal skin lesion classification through cross-attention fusion and collaborative edge computing. *Computerized Medical Imaging and Graphics*, 124:102588, 2025. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2025.102588. URL https://www.sciencedirect.com/science/article/pii/S0895611125000977.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:180161, 2018. doi: 10.1038/sdata.2018.161.

Philipp Tschandl, Bengu Nisa Akay, Cliff Rosendahl, Veronica Rotemberg, Verche Todorovska, Jochen Weber, Anna Katharina Wolber, Christoph Müller, Nicholas Kurtansky, Allan Halpern, Wolfgang Weninger, and Harald Kittler. Milk10k: A hierarchical multimodal imaging-learning toolkit for diagnosing pigmented and nonpigmented skin cancer and its simulators. *Journal of Investigative Dermatology*, 2025. ISSN 0022-202X. doi: 10.1016/j.jid.2025.06.1594. URL https://doi.org/10.1016/j.jid.2025.06.1594.

Yuan Zhang, Yutong Xie, Hu Wang, Jodie C Avery, M Louise Hull, and Gustavo Carneiro. A Novel Perspective for Multi-Modal Multi-Label Skin Lesion Classification . In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3549–3558, Los Alamitos, CA, USA, March 2025. IEEE Computer Society. doi: 10.1109/WACV61041.2025.00350. URL https://doi.ieeecomputersociety.org/10.1109/WACV61041.2025.00350.

Lihan Zuo, Zizhou Wang, and Yan Wang. A multi-stage multi-modal learning algorithm with adaptive multimodal fusion for improving multi-label skin lesion classification. *Artificial Intelligence in Medicine*, 162:103091, 2025. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2025.103091. URL https://www.sciencedirect.com/science/article/pii/S0933365725000260.

Appendix A. Labels of the MILK10k dataset

Diagnostic Category	Abbreviation	Quantity (training set)
Actinic keratosis/intraepidermal carcinoma	AKIEC	242
Basal cell carcinoma	BCC	2017
Other benign proliferations including collisions	BEN_OTH	35
Benign keratinocytic lesion	BKL	435
Dermatofibroma	DF	42
Inflammatory and infectious	INF	40
Other malignant proliferations including collisions	MAL_OTH	7
Melanoma	MEL	360
Melanocytic nevus, any type	NV	597
Squamous cell carcinoma/keratoacanthoma	SCCKA	379
Vascular lesions and hemorrhage	VASC	38

Appendix B. Evaluation metrics

To assess the performance of our model, we use six standard metrics: AUC, Precision, Accuracy, Sensitivity, Specificity, and F1-Score. All metrics are computed for each class in a one-vs-rest manner and then averaged to obtain the final results.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{21}$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} \tag{22}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{23}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{24}$$

where TP, FP, TN, and FN represent True Positives, False Positives, True Negatives, and False Negatives respectively.

The Area Under the ROC Curve (AUC) quantifies the ability of the classifier to distinguish between positive and negative samples across all thresholds. In the multi-class setting, AUC is computed independently for each class using a one-vs-rest strategy and then averaged across all classes.

Appendix C. Calibration Curve Analysis

The calibration curve of the fused JI-ADF model lies close to the diagonal, indicating that predicted probabilities match observed frequencies well overall. The curve is slightly below the perfect-calibration line for mid-range probabilities, suggesting mild over-confidence in this region, but it aligns closely with the diagonal for high-confidence predictions (≥ 0.7), where clinical decisions are most critical. The low expected calibration error (ECE $= 0.032$) confirms that the model is well calibrated globally.

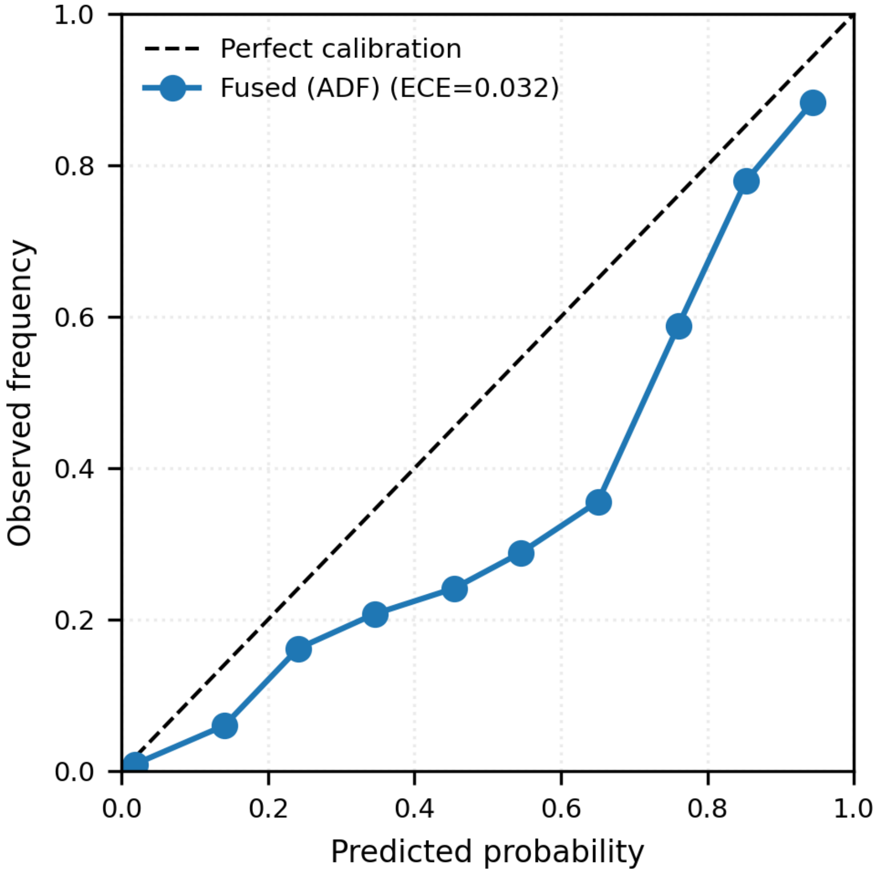


Figure 5: Calibration Curve

Appendix D. Image Preprocessing

Dermoscopic images often contain hair artifacts that obscure lesion boundaries. To address this, we built a two-pass hair removal pipeline based on vesselness filtering, morphological transforms, and inpainting. We first enhanced hair-like tubular structures with a multi-scale Frangi filter to detect and mask hairs for subsequent removal. This relies on the eigenvalues λ_1, λ_2 of the Hessian matrix $H(I_\sigma)$ of the image smoothed at scale σ. A vesselness score

$$V_\sigma(x) = \exp\left(-\frac{R_B^2}{2\beta^2}\right)\left(1 - \exp\left(-\frac{S^2}{2c^2}\right)\right), \tag{25}$$

is computed, where $R_B = \frac{|\lambda_1|}{|\lambda_2|}$ and $S^2 = \lambda_1^2 + \lambda_2^2$. This score highlights line-like patterns such as hairs. Morphological black-hat and top-hat transforms emphasized dark and bright hair structures, whose normalized responses formed a score map. Hair masks were derived via hysteresis thresholding and refined by skeletonization and dilation, then filled using Navier–Stokes inpainting to remove hairs while preserving lesion texture.