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

Anonymous Authors[1]

Abstract

Skin lesion classification is essential for early dermatological diagnosis, yet many existing computer-aided systems rely primarily on dermoscopic images and underutilize the multimodal evidence routinely available in clinical practice. To address this gap, we propose **JI-ADF**, a trimodal deep learning framework that integrates dermoscopic images, clinical photographs, and structured patient metadata for clinically grounded skin lesion classification. The proposed architecture combines joint multimodal representation learning with modality-specific auxiliary supervision and an adaptive decision fusion mechanism that dynamically calibrates modality contributions on a per-sample basis. To enhance cross-modal reasoning while preserving modality-specific evidence, we further introduce a multimodal fusion attention (MMFA) module. We evaluate JI-ADF on the large-scale MILK10k benchmark. Extensive analyses, including modality ablation, calibration evaluation, and Grad-CAM visualization, further confirm the robustness and clinically meaningful behavior of the model. The results indicate that JI-ADF provides a reliable and practical foundation for multimodal skin lesion classification in real-world clinical settings.

1. Introduction

Skin cancer is one of the most prevalent malignancies worldwide, with melanoma accounting for a disproportionate number of deaths. While the 5-year survival rate for localized melanoma is 99%, it drops to approximately 35% at metastatic stages, making early diagnosis critical (Bray et al., 2024; American Cancer Society, 2024; Ferlay et al., 2024). Deep learning has significantly advanced skin lesion classification, with CNN-based models reaching dermatologist-level accuracy on ISIC benchmarks in some studies (Yu et al., 2017; Tang et al., 2020; Xie et al., 2020; Esteva et al., 2017; Haenssle et al., 2018). However, most existing systems rely solely on dermoscopic images, limiting the ability to reflect real clinical practice where diagnosis depends on both visual and contextual information.

Recent work has explored multimodal learning to integrate images with clinical context and metadata (Atrey et al., 2010; Huang et al., 2020). In dermatology, combining dermoscopic images, clinical photographs, and patient information has been shown to improve classification performance (Yap et al., 2018; Liu et al., 2020; Pacheco & Krohling, 2021; Bi et al., 2020). Nevertheless, many approaches rely on simple concatenation or late fusion, which fail to capture complex cross-modal interactions (Pacheco & Krohling, 2020; Li et al., 2020). While attention-based methods partially address this issue (Pacheco & Krohling, 2021; Cai et al., 2023), most designs still compress multimodal information into a single representation and treat metadata primarily as auxiliary guidance, limiting their ability to model complementary and instance-specific modality contributions (He et al., 2021; Hu et al., 2018).

To address these limitations, we propose JI-ADF, a trimodal framework that integrates dermoscopic images, clinical images, and structured metadata through attention-based interaction and adaptive decision fusion. Our contributions are as follows:

- We propose JI-ADF, a trimodal framework that jointly models dermoscopic images, clinical images, and structured metadata through attention-based interaction.
- We introduce an adaptive decision fusion strategy that dynamically combines joint and modality-specific predictions, improving robustness when modalities provide unequal diagnostic evidence.
- We conduct extensive experiments on the MILK10k benchmark, showing that JI-ADF outperforms existing multimodal approaches and achieves balanced performance across lesion types.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

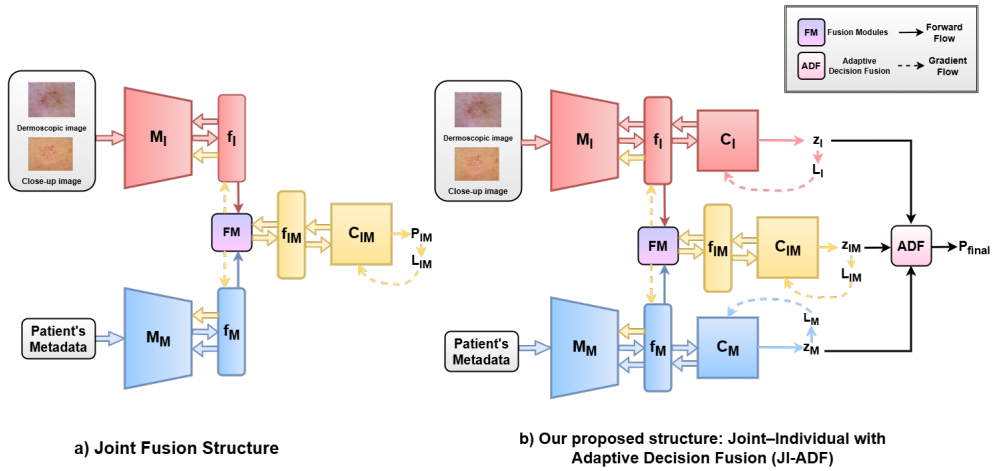

Figure 1. Illustration of (a) the Joint Fusion Structure and (b) our proposed Joint–Individual architecture with Adaptive Decision Fusion. JI-ADF extends the baseline by adding individual prediction heads for each modality and an adaptive fusion module that assigns instance-dependent weights.

2. Method

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

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

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

A differentiable fusion module aggregates them into a joint representation

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

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

$$\begin{aligned} \mathbf{z}_I &= C_I(\mathbf{f}_I) \in \mathbb{R}^N,\ \mathbf{P}_I = \text{softmax}(\mathbf{z}_I) \\ \mathbf{z}_M &= C_M(\mathbf{f}_M) \in \mathbb{R}^N,\ \mathbf{P}_M = \text{softmax}(\mathbf{z}_M) \\ \mathbf{z}_{IM} &= C_{IM}(\mathbf{f}_{IM}) \in \mathbb{R}^N,\ \mathbf{P}_{IM} = \text{softmax}(\mathbf{z}_{IM}) \end{aligned} \tag{3}$$

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

$$\begin{aligned} \mathcal{L}_I &= -\sum_{c=1}^{N} y_c \log P_I^{(c)} \\ \mathcal{L}_M &= -\sum_{c=1}^{N} y_c \log P_M^{(c)} \\ \mathcal{L}_{IM} &= -\sum_{c=1}^{N} y_c \log P_{IM}^{(c)} \end{aligned} \tag{4}$$

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

$$\begin{aligned} \boldsymbol{\alpha} &= \text{softmax}\big(W_2\,\sigma(W_1\mathbf{s} + \mathbf{b}_1) + \mathbf{b}_2\big) \\ &= (\alpha_I, \alpha_{IM}, \alpha_M) \end{aligned} \tag{5}$$

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

$$\mathbf{P}_{\text{final}} = \alpha_I \mathbf{P}_I + \alpha_{IM}\mathbf{P}_{IM} + \alpha_M \mathbf{P}_M \in \Delta^{N-1} \tag{6}$$

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

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

$$\mathcal{L}_{\text{total}} = \text{CE}(\mathbf{P}_{\text{final}}, \mathbf{y}) + \lambda_{IM}\mathcal{L}_{IM} + \lambda_I \mathcal{L}_I + \lambda_M \mathcal{L}_M \tag{7}$$

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.

2.2. 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).

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 & 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
Fine-tuned PanDerm (Yan et al., 2025)	0.840	0.476	0.920	0.447	0.953	0.434
JI-ADF (ours)	**0.866**	**0.543**	**0.930**	**0.536**	**0.959**	**0.505**

Table 1. Comparison with state-of-the-art skin lesion classification approaches.

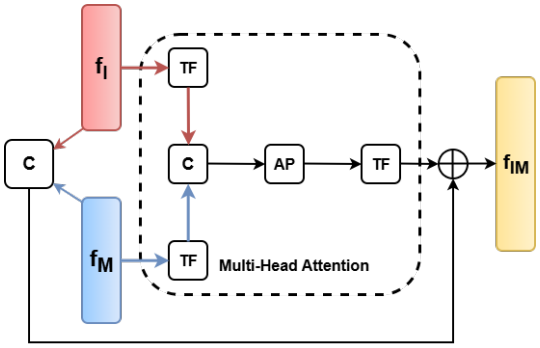

Figure 2. Multimodal Fusion Attention Module (MMFA), where image features f_I and metadata features f_M are jointly refined through cross-attention and self-attention mechanisms between the two modalities. The resulting fused representation f_{IM} captures the enhanced joint features after integration.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

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

$$\mathbf{Q}_h = \begin{bmatrix} W_h^{Q,I}\mathbf{f}_I \\ W_h^{Q,M}\mathbf{f}_M \end{bmatrix} \in \mathbb{R}^{2\times d_h}$$
$$\mathbf{K}_h = \begin{bmatrix} W_h^{K,I}\mathbf{f}_I \\ W_h^{K,M}\mathbf{f}_M \end{bmatrix} \in \mathbb{R}^{2\times d_h} \quad (8)$$
$$\mathbf{V}_h = \begin{bmatrix} W_h^{V,I}\mathbf{f}_I \\ W_h^{V,M}\mathbf{f}_M \end{bmatrix} \in \mathbb{R}^{2\times d_h}$$

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

$$\mathbf{U}_h = \mathrm{softmax}\left(\frac{\mathbf{Q}_h\mathbf{K}_h^\top}{\sqrt{d_h}}\right)\mathbf{V}_h \in \mathbb{R}^{2\times d_h} \quad (9)$$

$$\mathbf{o} = W^O \mathrm{Concat}(\mathrm{vec}(\mathbf{U}_1), \ldots, \mathrm{vec}(\mathbf{U}_H)) \in \mathbb{R}^{D_{IM}} \quad (10)$$

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

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

$$\mathbf{f}_{IM} = W_{\text{skip}}\,[\,\mathbf{f}_I \,\|\, \mathbf{f}_M\,] \;+\; g(\mathbf{o}) \;\in\; \mathbb{R}^{D_{IM}} \quad (11)$$

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.

3. Results

Across competitive multimodal baselines (Table 1), JI-ADF consistently achieves the best overall performance, with the highest AUC (0.866) and accuracy (0.930), indicating strong class separability and stable predictions despite the heterogeneity and long-tailed nature of the MILK10k dataset. Sensitivity remains particularly challenging under severe class imbalance; nevertheless, JI-ADF attains the top sensitivity (0.536), suggesting improved true-positive detection, especially in ambiguous cases where visual cues alone may be insufficient. This gain is achieved without sacrificing precision, as reflected in the highest Dice (0.505) and leading precision (0.543), indicating a better balance between recall and false positives. In addition, JI-ADF matches or slightly exceeds the best specificity (0.959), further demonstrating well-calibrated predictions. Overall, these results suggest that adaptive, per-sample fusion of multimodal inputs provides a more balanced and reliable decision strategy than fixed fusion approaches.

4. Ablation Study

4.1. Modality Contribution Analysis

We conducted an ablation study over all unimodal and bimodal subsets of the inputs. Based on the results shown in Table 2, models that combine modalities consistently outperform single-modality variants. In particular, the proposed JI-ADF trimodal fusion achieves the best performance

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
C+D+M	**0.866**	**0.543**	**0.930**	**0.536**	**0.959**	**0.505**

Table 2. Modality Configuration Ablation Study. (C: clinical image, D: dermoscopic image, M: metadata).

across most metrics, indicating that jointly leveraging clinical images, dermoscopic images, and metadata yields the most reliable classifier.

4.2. Ablation of Attention and Residual Branches

The attention and residual paths in MMFA serve complementary roles. The attention path explicitly models cross-modal dependencies, enabling metadata to modulate image features when strong semantic correlations exist. In contrast, the residual path preserves modality-specific evidence by allowing the original image and metadata features to directly contribute to the fused representation. When using attention alone, the fused representation relies entirely on learned cross-modal correlations, which can be unstable under weak or noisy inter-modal relationships. Conversely, the skip-only variant lacks the capacity to capture fine-grained interactions across modalities. Combining both paths allows the model to dynamically balance cross-modal reasoning and modality-specific evidence, leading to more stable and expressive fusion. The superior Sensitivity and Dice achieved by the full MMFA confirm that the residual branch is not merely an optimization shortcut, but a structural component that complements attention by safeguarding reliable unimodal signals.

Fusion Variant	AUC	Precision	Accuracy	Sensitivity	Specificity	Dice
Skip-only	0.880	0.508	0.927	0.421	0.960	0.419
Attention-only	0.882	0.542	0.930	0.438	0.963	0.457
Attention + Skip	0.866	0.543	0.930	0.536	0.959	0.505

Table 3. Ablation of attention and residual branches in the Multimodal Fusion Attention (MMFA) module.

4.3. 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.

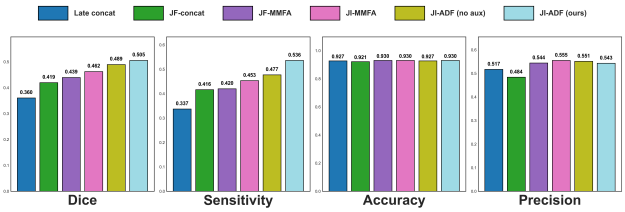

Figure 3. Fusion Architecture Ablation – Multimetrics Comparison

Across the four metrics reported in Figure 3, 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 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.

5. 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

Adebiyi, A., Abdalnabi, N., Smith, E. H., Hirner, J., Simoes, E. J., Becevic, M., and Rao, P. 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.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6): 345–379, Nov 2010. ISSN 1432-1882. doi: 10.1007/s00530-010-0182-0. URL https://doi.org/10.1007/s00530-010-0182-0.

Bi, L., Feng, D. D., Fulham, M., and Kim, J. Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network. *Pattern Recognition*, 107:107502, 2020. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107502. URL https://www.sciencedirect.com/science/article/pii/S0031320320303058.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. 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.

Cai, G., Zhu, Y., Wu, Y., Jiang, X., Ye, J., and Yang, D. A multimodal transformer to fuse images and metadata for skin disease classification. *The Visual Computer*, 39(7): 2781–2793, Jul 2023. ISSN 1432-2315. doi: 10.1007/s00371-022-02492-4. URL https://doi.org/10.1007/s00371-022-02492-4.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. ISSN 1476-4687. doi: 10.1038/nature21056. URL https://doi.org/10.1038/nature21056.

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

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., Uhlmann, L., Level-I, R. S., and Groups, L.-I. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, Aug 2018. ISSN 0923-7534. doi: 10.1093/annonc/mdy166. URL https://doi.org/10.1093/annonc/mdy166.

He, X., Deng, Y., Fang, L., and Peng, Q. Multi-modal retinal image classification with modality-specific attention network. *IEEE Transactions on Medical Imaging*, 40(6): 1591–1602, 2021. doi: 10.1109/TMI.2021.3059956.

Hu, J., Lu, J., and Tan, Y.-P. Sharable and individual multi-view metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(9):2281–2288, 2018. doi: 10.1109/TPAMI.2017.2749576.

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(1):136, 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00341-z. URL https://doi.org/10.1038/s41746-020-00341-z.

Khurshid, M., Singh, R., and Vatsa, M. 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.

Li, W., Zhuang, J., Wang, R., Zhang, J., and Zheng, W.-S. Fusing metadata and dermoscopy images for skin disease diagnosis. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1996–2000, 2020. doi: 10.1109/ISBI45749.2020.9098645.

Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G. S., Peng, L. H., Webster, D. R., Ai, D., Huang, S. J., Liu, Y., Dunn, R. C., and Coz, D. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26 (6):900–908, Jun 2020. ISSN 1546-170X. doi: 10.1038/

s41591-020-0842-3. URL https://doi.org/10.1038/s41591-020-0842-3.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Pacheco, A. G. and Krohling, R. A. The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine*, 116:103545, 2020. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2019.103545. URL https://www.sciencedirect.com/science/article/pii/S0010482519304019.

Pacheco, A. G. C. and Krohling, R. A. An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. *IEEE Journal of Biomedical and Health Informatics*, 25 (9):3554–3563, 2021. doi: 10.1109/JBHI.2021.3062002.

Restrepo, D., Wu, C., Cajas, S. A., Nakayama, L. F., Celi, L. A., and López, D. M. Multimodal deep learning for low-resource settings: A vector embedding alignment approach for healthcare applications, 2024. URL https://arxiv.org/abs/2406.02601.

Tang, P., Liang, Q., Yan, X., Xiang, S., and Zhang, D. Gp-cnn-dtel: Global-part cnn model with data-transformed ensemble learning for skin lesion classification. *IEEE Journal of Biomedical and Health Informatics*, 24(10): 2870–2882, 2020. doi: 10.1109/JBHI.2020.2977013.

Tang, P., Yan, X., Nan, Y., Hu, X., Menze, B. H., Krammer, S., and Lasser, T. 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.

Tran-Van, N.-Y. and Le, K.-H. 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.

Tschandl, P., Akay, B. N., Rosendahl, C., Rotemberg, V., Todorovska, V., Weber, J., Wolber, A. K., Müller, C., Kurtansky, N., Halpern, A., Weninger, W., and Kittler, H. 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.

Xie, Y., Zhang, J., Xia, Y., and Shen, C. A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Transactions on Medical Imaging*, 39 (7):2482–2493, 2020. doi: 10.1109/TMI.2020.2972964.

Yan, S., Yu, Z., Primiero, C., Vico-Alonso, C., Wang, Z., Yang, L., Tschandl, P., Hu, M., Ju, L., Tan, G., Tang, V., Ng, A. B., Powell, D., Bonnington, P., See, S., Magnaterra, E., Ferguson, P., Nguyen, J., Guitera, P., Banuls, J., Janda, M., Mar, V., Kittler, H., Soyer, H. P., and Ge, Z. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31(8):2691–2702, August 2025. doi: 10.1038/s41591-025-03747-y. URL https://doi.org/10.1038/s41591-025-03747-y.

Yap, J., Yolland, W., and Tschandl, P. Multimodal skin lesion classification using deep learning. *Experimental Dermatology*, 27(11):1261–1267, Nov 2018. ISSN 0906-6705. doi: 10.1111/exd.13777.

Yu, L., Chen, H., Dou, Q., Qin, J., and Heng, P.-A. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2017. doi: 10.1109/TMI.2016.2642839.

Zhang, Y., Xie, Y., Wang, H., Avery, J. C., Hull, M. L., and Carneiro, G. A Novel Perspective for Multi-Modal Multi-Label Skin Lesion Classification . In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 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.

Zuo, L., Wang, Z., and Wang, Y. 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.

A. Experiments

A.1. Dataset Description

We used the MILK10k multimodal skin-lesion dataset (Tschandl et al., 2025) to train and evaluate our method. The MILK10k dataset contains 10480 images from 5240 lesions, provided as paired clinical close-up and dermoscopic images collected across five centers. 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.

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

Table 4. Distribution of diagnostic categories in the training set.

A.2. Implementation

Our model was trained for 50 epochs with a batch size of 16, using AdamW optimizer (Loshchilov & Hutter, 2019) with the initial learning rate value 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 2.2) 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.

B. Joint Fusion Structure

For convenience, we consider the fusion of a dermoscopic image, a clinical close-up image, and patient metadata for skin lesion 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 2 images and M_M is the method to extract features from patient metadata M:

$$\begin{aligned} \mathbf{f}_I &= M_I(I_{\text{derm}}, I_{\text{close}}) \in \mathbb{R}^{D_I} \\ \mathbf{f}_M &= M_M(M) \in \mathbb{R}^{D_M} \end{aligned} \tag{12}$$

A fusion module FM produces the joint representation

$$\mathbf{f}_{IM} = FM(\mathbf{f}_I, \mathbf{f}_M) \in \mathbb{R}^{D_{IM}} \tag{13}$$

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

$$\mathbf{P}_{IM} = \text{softmax}(\mathbf{z}_{IM}) \in \Delta^{N-1} \tag{14}$$

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

$$\mathcal{L}_{IM} = -\sum_{c=1}^{N} y_c \log P_{IM}^{(c)} \tag{15}$$

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

$$\frac{\partial \mathcal{L}_{IM}}{\partial \theta_{M_I}} = \frac{\partial \mathcal{L}_{IM}}{\partial \mathbf{z}_{IM}} \frac{\partial \mathbf{z}_{IM}}{\partial \mathbf{f}_{IM}} \frac{\partial \mathbf{f}_{IM}}{\partial \mathbf{f}_I} \frac{\partial \mathbf{f}_I}{\partial \theta_{M_I}} \tag{16}$$

$$\frac{\partial \mathcal{L}_{IM}}{\partial \theta_{M_M}} = \frac{\partial \mathcal{L}_{IM}}{\partial \mathbf{z}_{IM}} \frac{\partial \mathbf{z}_{IM}}{\partial \mathbf{f}_{IM}} \frac{\partial \mathbf{f}_{IM}}{\partial \mathbf{f}_M} \frac{\partial \mathbf{f}_M}{\partial \theta_{M_M}} \tag{17}$$

Summary. Inputs $(I_{\text{derm}}, I_{\text{close}}, M)$ 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.

C. 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 5. 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 5, 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 (Table 4). 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.

D. Grad-CAM Visualizations

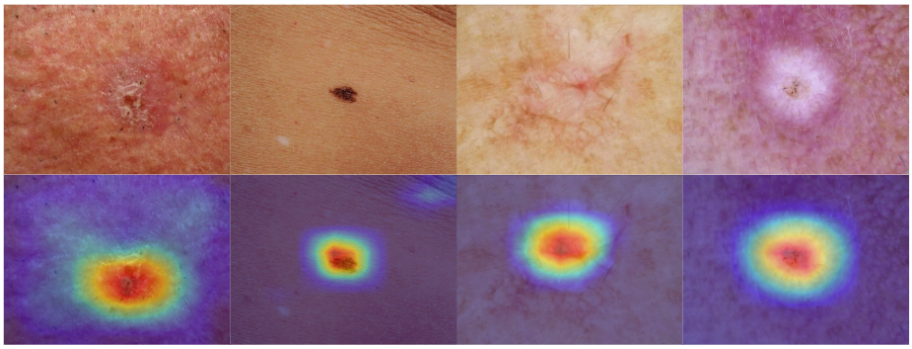

Figure 4. Comparison between the original input images and their corresponding Grad-CAM visualizations. Warmer colors (yellow–red) indicate regions with higher model attention, while cooler colors (blue–green) represent lower emphasis.

To better understand how the proposed model interprets lesion patterns, we generate Grad-CAM visualizations for the test images. As in Figure 4, 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.

E. Calibration Curve Analysis

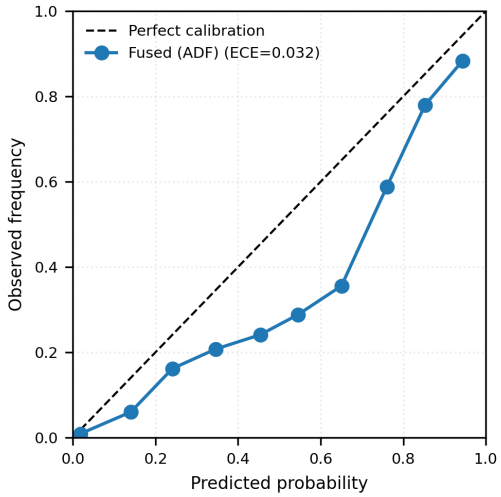

Figure 5. Calibration Curve.

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.