RG-DermNet: A Multimodal Attention-Based Model with Residual Block Usage for Skin Lesion Classification

Wyctor F. da Rocha[1], Pedro H. G. Bouzon[1], Lucas A. Ramos[2], André G. C. Pacheco[1], Luis A. Souza Jr.[1]
[1]*Graduate Program of Informatics, Federal University of Espírito Santo, Vitória, Brazil*
{wyctor.rocha, pedro.bouzon}@edu.ufes.br, {apacheco, la.souza}@inf.ufes.br
[2]*Computer Vision and Data Science, NHL Stenden University of Applied Sciences, Leeuwarden, The Netherlands*
lucas.ramos@nhlstenden.com

***Abstract*—Skin cancer accounts for nearly one-third of all diagnosed tumors worldwide, making early and accurate recognition critical for improving patient outcomes. In this work, we propose RG-DermNet, a multimodal deep learning framework that integrates skin lesion images with structured clinical metadata through a residual gated-attention (RG-ATT) fusion mechanism. The architecture combines CNN- and Transformer-based visual backbones with a lightweight one-hot encoding pipeline for metadata, enabling effective cross-modal interaction. The proposed model is evaluated using a patient-wise cross-validation protocol across four dermatological datasets with heterogeneous metadata. On PAD-UFES-20, using Caformer-B36 as the visual backbone, RG-DermNet achieves an accuracy of 0.75 ± 0.05, balanced accuracy of 0.78 ± 0.03, F1-score of 0.77 ± 0.04, and AUC of 0.95 ± 0.01, outperforming existing multimodal baselines under the same evaluation setting. In addition, a SHAP-based analysis provides insights into the contribution of clinical metadata to the model's predictions, supporting both performance gains and interpretability.**

***Index Terms*—Skin lesion, Convolutional Neural Network, Attention Mechanism, Residual Connection, Interpretability.**

I. Introduction

SKIN cancer remains a major global health concern, with both melanoma and non-melanoma types contributing substantially to worldwide cancer incidence and mortality [1]. Recent projections indicate a continued growth in melanoma incidence, with approximately 510,000 new cases expected globally by 2040 [2]. Non-melanoma skin cancers (NMSC), although typically less aggressive, account for an even larger disease burden, with more than one million new cases annually and a considerable number of associated deaths [3]. These trends highlight the increasing clinical and public health relevance of automated diagnostic support systems.

The growing demand for accurate and scalable skin lesion assessment has driven rapid advances in computer-aided diagnosis (CAD) systems. Numerous studies demonstrate that combining medical imaging with machine learning techniques can significantly enhance lesion detection and classification, supporting clinicians in complex diagnostic scenarios [4].

In recent years, Transformer-based architectures have gained increasing attention in medical image analysis, particularly in hybrid designs that combine convolutional feature extraction with global attention mechanisms. Vision Transformers (ViTs) [5] integrated with CNN backbones have shown improved performance in skin lesion classification and segmentation by capturing long-range contextual dependencies while preserving local texture information [6], [7].Attention and residual connections further improve training stability and generalization [8].

Beyond image-only approaches, multimodal fusion strategies that incorporate clinical metadata have demonstrated notable performance gains. Early fusion methods based on feature concatenation or multiplication already showed improvements in diagnostic accuracy [9]. More recent attention-based modules, such as MetaBlock [10] and its extensions (e.g., MetaBlock-SE [11]), inject metadata into deeper network layers, enabling adaptive feature refinement and increased robustness under missing or incomplete patient information.

Despite these advances, most attention-based fusion strategies aggregate modalities without explicitly regulating how clinical metadata should influence visual representations on a per-sample basis. Cross-attention mechanisms provide a natural way to address this limitation by allowing one modality to conditionally guide another. However, their application to multimodal dermatological diagnosis remains relatively underexplored. Moreover, maintaining stable and interpretable multimodal representations as network depth increases requires mechanisms that preserve original feature information, motivating the use of residual pathways [12].

In this work, we propose a novel multimodal diagnostic framework that integrates visual features and structured clinical metadata through a Cross-Attention–based fusion strategy enhanced by a *Residual–Gated Attention* (RG-ATT) module. By combining a strong visual feature extractor with metadata-aware gating and residual connections, the proposed approach aims to improve diagnostic robustness across diverse clinical settings. Unlike other attention mechanisms, the proposed RG-ATT module introduces a query-conditioned gating controller that regulates cross-modal influence at the sample level, making the block an applicable approach for multimodal neural architectures in dermatological imaging.

The main contributions of this work are as follows:

- We introduce a Cross-Attention–based multimodal architecture with a *Residual–Gated Attention* (RG-ATT) module for principled fusion of visual features and clinical metadata.
- We conduct a comprehensive patient-wise evaluation across four heterogeneous dermatological datasets, analyzing the impact of metadata quality and dataset characteristics on multimodal performance.
- We achieve state-of-the-art balanced accuracy on PAD-UFES-20 using an identical evaluation protocol and publicly available code.[1]
- We offer an interpretability analysis on the metadata branch to identify which contextual variables most contribute to predictions, supporting clinical plausibility while acknowledging the limitations of surrogate explainability methods

II. Related Works

A. Machine Learning for Skin Lesion Recognition

Early skin lesion classifiers relied on handcrafted descriptors and conventional classifiers [13]. Although these methods enabled early progress, their performance often depended on careful feature design and tended to be sensitive to dataset-specific conditions [14]. Over time, this pushed the community toward more data-driven solutions that reduce manual intervention [15].

The adoption of deep learning, especially Convolutional Neural Networks (CNNs), substantially changed automated skin lesion analysis.CNNs learn hierarchical representations directly from images and have shown strong dermoscopic classification results [16], [17]. Reports in the literature also suggest that, with sufficiently large and diverse training data, these systems can approach the diagnostic accuracy of experienced dermatologists [18].

Even so, image-only models still face practical barriers to clinical deployment. Class imbalance, acquisition artifacts, and the lack of patient context can reduce robustness and hurt sensitivity for clinically relevant categories [19], [20]. These limitations have motivated multimodal designs that incorporate complementary clinical information.

B. Multimodality-Based Approaches

Multimodal learning has therefore become a common strategy for combining visual evidence with structured patient metadata, such as demographic variables and clinical history [21], [22]. By leveraging information that is not visible in the image alone, these models often improve performance in skin lesion classification.

Prior work consistently shows the benefit of adding metadata. Lyakhov *et al.* [19], Pacheco *et al.* [20], Souza *et al.* [23], and Li *et al.* [24] reported gains over image-only baselines, reinforcing the clinical value of contextual patient variables. In addition to accuracy, metadata can also improve interpretability, since these variables are closer to how clinicians reason than latent visual embeddings [23].

However, many fusion strategies implicitly treat metadata as uniformly reliable. In practice, real-world dermatological datasets often include missing entries, noisy annotations, or weakly informative attributes, which can limit the benefit of metadata-driven fusion. This motivates mechanisms that can adjust the impact of metadata during inference instead of always injecting it with the same strength.

III. Methodology

A. Attention-Based Feature Fusion

Multimodal fusion integrates complementary information from heterogeneous sources. While early approaches relied on feature concatenation, more expressive fusion strategies have been proposed to better capture cross-modal interactions [23], including MetaBlock [10], MetaNet [24], and MD-Net [25].Among these, attention-based mechanisms are particularly relevant because they improve the interaction between heterogeneous feature spaces and support more informative multimodal representations [22], [26].

Formally, given query, key, and value representations (Q, K, V), the scaled dot-product attention is defined as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \qquad (1)$$

Multi-head attention extends this formulation by jointly attending to information from multiple representation subspaces.

1) Cross-Attention for Multimodal Integration: Cross-attention mechanisms generalize self-attention by allowing one feature set to conditionally attend to another, enabling directed interactions between distinct modalities [26]. In the context of skin lesion analysis, cross-attention has been shown to improve robustness and predictive performance when fusing visual and non-visual features [27], [28].This is especially useful in clinical settings, where metadata can guide visual interpretation.

B. Proposed Architecture

We propose a multimodal architecture designed to improve the integration of visual features and structured clinical metadata.The model supports different visual backbones and multiple fusion strategies, enabling fair experimental comparisons.

1) Feature Projection and Fusion Overview: As illustrated in Fig. 1, visual features extracted by a CNN- or Transformer-based backbone and clinical metadata are first projected into a shared latent space. The projection layers, denoted as **VFeatProj** and **TFeatProj**, map both modalities to 512-dimensional representations, enabling unified processing regardless of the underlying visual encoder. This projection-based design decouples the fusion module from the visual feature extractor, allowing RG-ATT to function as a backbone-agnostic component without requiring architectural modifications to the extractor model. Simple concatenation followed by an MLP is used as a baseline fusion strategy, while the proposed

[1]https://github.com/life-ufes/rg-dermnet

Residual–Gated Attention (RG-ATT) block enables adaptive multimodal integration.

2) Residual–Gated Attention Block (RG-ATT): The RG-ATT block (Fig. 2) integrates visual features and clinical metadata by modulating the attention output through a learnable gate while preserving the original query representation via a residual connection. This design extends standard attention-based fusion by enabling sample-wise control over the influence of attended features.

Metadata features are used as queries (Q), while visual features provide keys (K) and values (V). Figure 1 shows the overall architecture, while Figure 2 provides a detailed view of the RG-ATT module.

Given (Q, K, V), an intermediate attention output is first computed using multi-head attention:

$$att = \mathrm{MHA}(Q, K, V).$$

A gating vector is then derived directly from the queries:

$$\text{gate} = \sigma(W_g Q),$$

where W_g is a learnable projection and $\sigma(\cdot)$ denotes the sigmoid function.The gate is bounded in [0,1] and modulates the attended signal element-wise. Dropout is applied to the attention output as a standard regularization strategy.

The final RG-ATT output is obtained through a residual aggregation followed by normalization:

$$\text{output}_{\text{RG-ATT}} = \mathrm{LayerNorm}\big(Q + (\text{gate} \cdot att)\big). \tag{2}$$

Eq. 2 formalizes the RG-ATT output and its underlying behavior. When the learned gate approaches zero, the model relies predominantly on the original query representation, effectively attenuating the attended cross-modal signal. Conversely, when the gate approaches one, the attended signal is more strongly incorporated, allowing metadata to exert greater influence on visual feature refinement.

In multimodal clinical settings, metadata may be noisy or only partially informative. The RG-ATT mechanism learns when to amplify or attenuate metadata-conditioned attention, yielding a flexible fusion strategy across samples and lesion classes.

The final classification head is implemented as a multi-layer perceptron (MLP) to capture non-linear interactions between features.

IV. Experiments and Results

This section describes the experimental setup adopted to evaluate the proposed RG-DermNet architecture, including the datasets, evaluation protocol, model training configuration, comparison methodology, and results. All experiments follow a consistent patient-wise validation strategy to ensure clinically realistic performance assessment.

A. Experimental Setup

We evaluate RG-DermNet on public skin lesion datasets with paired images and metadata. The selected datasets differ in image modality, metadata richness, number of classes, and class imbalance, enabling evaluation under heterogeneous and realistic clinical scenarios.

PAD-UFES-20 [29] is used as the primary benchmark due to its rich multimodal structure, comprising clinical images paired with 21 patient-level metadata attributes and biopsy-confirmed diagnostic labels. This dataset is employed for architecture optimization and visual backbone selection.

Generalization is assessed on PAD-UFES-20+ [11], MILK10K [30], and ISIC-2019 [31]. PAD-UFES-20+ extends PAD-UFES-20 by increasing the number of samples while preserving the same metadata schema. ISIC-2019 provides a large-scale dermatoscopic benchmark with limited metadata, enabling evaluation under reduced non-visual information. MILK10K is a large-scale multimodal dermatological dataset characterized by a highly imbalanced class distribution across 11 diagnostic categories.

All experiments are conducted using 5-fold stratified cross-validation at the *patient level*, ensuring that images from the same patient do not appear in both training and validation folds. This protocol mitigates information leakage and provides a more clinically meaningful assessment of generalization performance.

All visual backbones are initialized with ImageNet-1K pretrained weights and fine-tuned end-to-end. CNN-based backbones include MobileNet-V2 [32], DenseNet169 [16], and ResNet50 [12], while Transformer-based models include MViTv2 [33], DaViT-Tiny [34], and Caformer-B36 [35]. These models were selected to cover a wide spectrum of architectural paradigms, computational costs, and representativeness in the skin lesion literature. Categorical variables were encoded using one-hot encoding, while continuous variables (age, diameter_1, diameter_2) were normalized and kept in numerical form. The final metadata representation is obtained by concatenating binary and normalized continuous features.

Training is performed using weighted Cross-Entropy loss to mitigate class imbalance. Optimization uses Adam with learning rate 5×10^{-5}, weight decay 1×10^{-4}, batch size 32, and early stopping based on validation loss (patience = 10). Standard data augmentation (random flips, rotations, and color jitter) is applied during training. Batch Normalization was replaced with Layer Normalization to ensure stable training under the small, irregular mini-batches induced by the patient-wise protocol. Full implementation details are available in the public repository.

The experimental evaluation follows three main steps. First, multiple visual backbones are compared under a fixed multimodal configuration on PAD-UFES-20 to identify an effective feature extractor. Backbone selection is guided by patient-wise cross-validation results, following the same protocol used in prior multimodal dermatology studies.

Second, different fusion strategies are compared to isolate the contribution of metadata and residual gating. These include

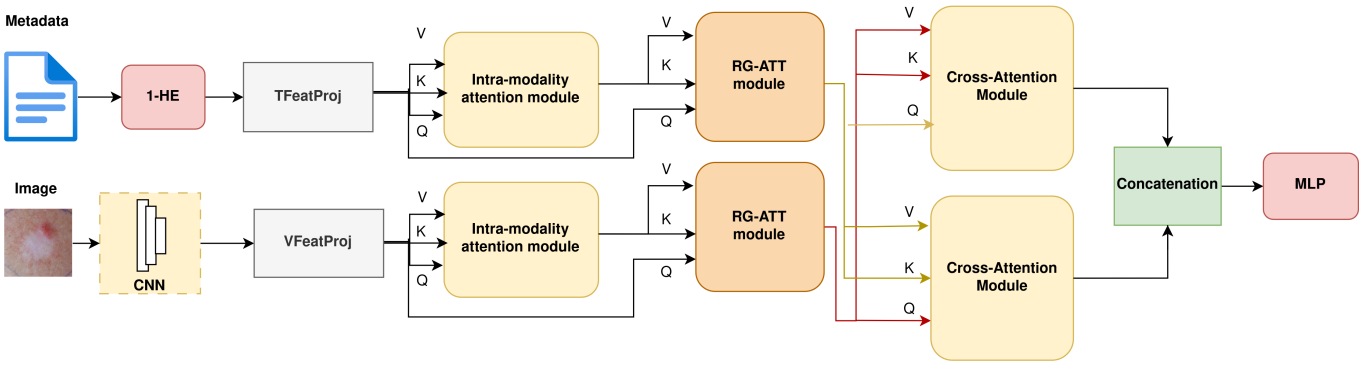


Fig. 1: Overview of the proposed RG-DermNet architecture. Visual features extracted by a CNN/Transformer backbone and clinical metadata are projected into a shared 512-dimensional space through VFeatProj and TFeatProj, respectively. Fusion is performed using the proposed Residual–Gated Attention (RG-ATT) block.

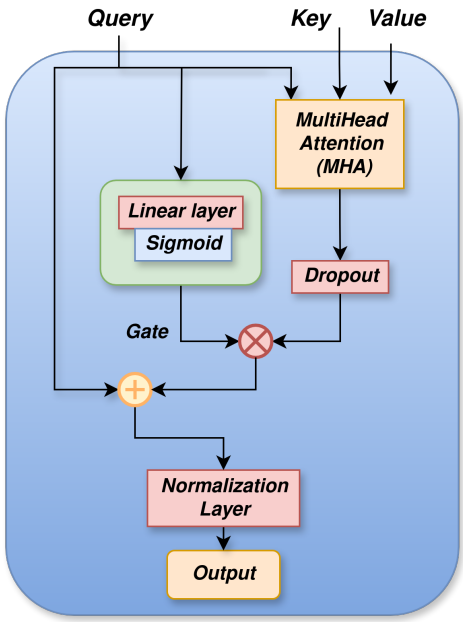


Fig. 2: Structure of the Residual–Gated Attention (RG-ATT) block introduced in Figure 1. The query pathway is reused to generate the gating vector, enabling adaptive modulation of the attended signal through a residual connection.

image-only models, simple concatenation, attention-based fusion methods from the literature, and the proposed Residual–Gated Attention (RG-ATT) block.

Third, RG-DermNet is compared against state-of-the-art multimodal approaches on PAD-UFES-20 under identical patient-wise evaluation protocols. Statistical significance is assessed using non-parametric tests, as detailed in the Results section.

To analyze metadata influence, we use SHAP (SHapley Additive exPlanations) [36]. Since direct attribution on the full multimodal network is non-trivial, a metadata-only surrogate model is used to approximate the metadata-driven component of RG-DermNet predictions.

B. Results

This section presents the experimental results obtained with the proposed RG-DermNet architecture. We first analyze the impact of different visual feature extractors on PAD-UFES-20 under a fixed multimodal configuration. We then compare RG-DermNet with state-of-the-art multimodal approaches using an identical patient-wise evaluation protocol. Finally, we assess cross-dataset generalization and analyze error patterns through confusion matrices.

a) Visual Backbone Performance on PAD-UFES-20: We begin by evaluating the influence of different visual feature extractors within the same multimodal architecture on PAD-UFES-20. Table I reports patient-wise cross-validation results expressed as mean ± standard deviation across five stratified folds.

Among the evaluated backbones, Caformer-B36 yields strong results on PAD-UFES-20, including the highest mean Balanced Accuracy (BACC) under our protocol. Nevertheless, the *official backbone selection criterion* in this study is the A-TOPSIS multi-criteria decision-making procedure, which aggregates Accuracy (ACC), BACC, F1-Score, and Area Under Curve (AUC) [2] while accounting for both mean and

[2]Precision, Recall, and F1-score were computed using weighted averaging to account for class imbalance, while Accuracy and Balanced Accuracy were computed directly.

TABLE I: Cross-dataset performance of different visual backbones under patient-wise cross-validation (mean ± sd). Best results per dataset are highlighted in bold.

Backbone	ACC	BACC	F1-Score	AUC
PAD-UFES-20				
Caformer-B36	0.75 ± 0.05	**0.78 ± 0.03**	**0.77 ± 0.04**	**0.95 ± 0.01**
DaViT-Tiny	0.66 ± 0.04	0.75 ± 0.02	0.69 ± 0.04	0.94 ± 0.01
MViTv2-Small	0.66 ± 0.09	0.74 ± 0.05	0.68 ± 0.09	0.94 ± 0.01
MobileNet-V2	0.62 ± 0.09	0.73 ± 0.04	0.63 ± 0.10	0.93 ± 0.01
DenseNet169	0.64 ± 0.06	0.73 ± 0.04	0.66 ± 0.06	0.93 ± 0.01
ResNet-50	0.59 ± 0.17	0.73 ± 0.06	0.59 ± 0.18	0.93 ± 0.02
PAD-UFES-20+				
Caformer-B36	0.76 ± 0.04	0.71 ± 0.02	0.77 ± 0.03	0.93 ± 0.01
DaViT-Tiny	0.78 ± 0.02	**0.72 ± 0.02**	0.79 ± 0.02	**0.94 ± 0.01**
MViTv2-Small	0.78 ± 0.03	**0.72 ± 0.02**	**0.79 ± 0.02**	0.94 ± 0.01
MobileNet-V2	0.76 ± 0.02	0.69 ± 0.02	0.77 ± 0.02	0.93 ± 0.01
DenseNet169	0.75 ± 0.01	0.69 ± 0.01	0.76 ± 0.01	0.93 ± 0.01
ResNet-50	0.78 ± 0.01	0.69 ± 0.02	0.78 ± 0.01	0.93 ± 0.01
ISIC-2019				
Caformer-B36	0.82 ± 0.04	0.81 ± 0.04	0.83 ± 0.03	0.97 ± 0.01
DaViT-Tiny	0.84 ± 0.01	0.82 ± 0.01	0.83 ± 0.01	0.98 ± 0.01
MViTv2-Small	0.83 ± 0.02	**0.83 ± 0.01**	**0.84 ± 0.02**	**0.98 ± 0.01**
MobileNet-V2	0.76 ± 0.01	0.76 ± 0.02	0.77 ± 0.01	0.96 ± 0.01
DenseNet169	0.80 ± 0.01	0.81 ± 0.01	0.81 ± 0.01	0.97 ± 0.01
ResNet-50	0.77 ± 0.02	0.77 ± 0.02	0.78 ± 0.02	0.96 ± 0.01
MILK10K				
Caformer-B36	0.73 ± 0.01	**0.50 ± 0.01**	**0.74 ± 0.01**	**0.89 ± 0.01**
DaViT-Tiny	0.73 ± 0.02	**0.50 ± 0.02**	0.72 ± 0.02	**0.90 ± 0.02**
MViTv2-Small	0.74 ± 0.02	0.49 ± 0.03	0.73 ± 0.02	0.89 ± 0.02
MobileNet-V2	0.69 ± 0.01	0.49 ± 0.02	0.69 ± 0.01	0.88 ± 0.02
DenseNet169	0.70 ± 0.02	0.49 ± 0.02	0.70 ± 0.02	0.88 ± 0.02
ResNet-50	0.70 ± 0.01	0.48 ± 0.01	0.73 ± 0.01	0.87 ± 0.04

standard deviation (weighted by 0.75 and 0.25, respectively). The resulting A-TOPSIS ranking on PAD-UFES-20 selects **Caformer-B36** as the most favorable trade-off between performance and stability. Therefore, Caformer-B36 is adopted as the default visual feature extractor in all subsequent experiments.

b) Comparison with State-of-the-Art Methods: Table II presents a quantitative comparison between RG-DermNet and representative state-of-the-art multimodal approaches evaluated on PAD-UFES-20 under identical patient-wise cross-validation conditions.

A Friedman test indicates statistically significant differences among the evaluated methods ($p = 6.04 \times 10^{-12}$). Subsequent pairwise Wilcoxon signed-rank tests confirm that the improvements achieved by RG-DermNet over all competing approaches are statistically significant ($p < 0.05$).

c) Cross-Dataset Generalization: Using the Caformer-B36 as a visual feature extractor model, selected by the A-TOPSIS usage, RG-DermNet is further evaluated on PAD-UFES-20+, ISIC-2019, and MILK10K (Table I). Datasets with richer, clinically meaningful metadata tend to benefit more from multimodal fusion, whereas datasets with limited metadata show smaller improvements in balanced accuracy, despite consistently high AUC values.

d) Analysis on MILK10K: On MILK10K, all evaluated backbones achieve consistently high AUC values, whereas Balanced Accuracy (BACC) remains comparatively low (≈ 0.50). This discrepancy is largely due to the extreme class imbalance across its 11 diagnostic categories, which biases learning toward the majority classes and limits per-class recall. Overall, these results indicate that, although RG-DermNet effectively ranks predictions, multimodal learning under highly long-tailed distributions remains challenging, motivating future use of strategies such as focal loss, re-sampling, and cost-sensitive learning.

e) Confusion matrices and error patterns: Fig. 3 shows confusion matrices for the best-trained multimodal model on PAD-UFES-20, PAD-UFES-20+, MILK10K, and ISIC-2019 (validation). Misclassifications concentrate among *SCC*, *BCC*, and *ACK*, a trend also reported previously [20], [23], likely influenced by the comparatively smaller number of *SCC* samples.

Misclassifications are primarily concentrated among SCC, BCC, and ACK classes, consistent with prior dermatological studies. This behavior reflects visual similarity and class imbalance effects rather than systematic model bias.

f) Metadata Interpretability: To complement the predictive evaluation of RG-DermNet, we analyze how structured clinical metadata contributes to its probabilistic outputs on PAD-UFES-20, which was selected due to its relatively rich and complete annotations.

Direct SHAP attribution on the full multimodal pipeline is challenging because predictions arise from joint interactions between visual and non-visual streams. Therefore, we adopt a metadata-focused surrogate approach: a Random Forest regressor is trained on normalized metadata features to approximate RG-DermNet's predicted class probabilities (from the validation fold data). This surrogate achieves $R^2 \approx 0.59$, indicating that a substantial fraction of the variance in the predicted probabilities can be explained by metadata alone, enabling meaningful feature attribution. However, these explanations should be interpreted as an approximation of the metadata-driven component rather than a complete representation of the end-to-end multimodal decision process.

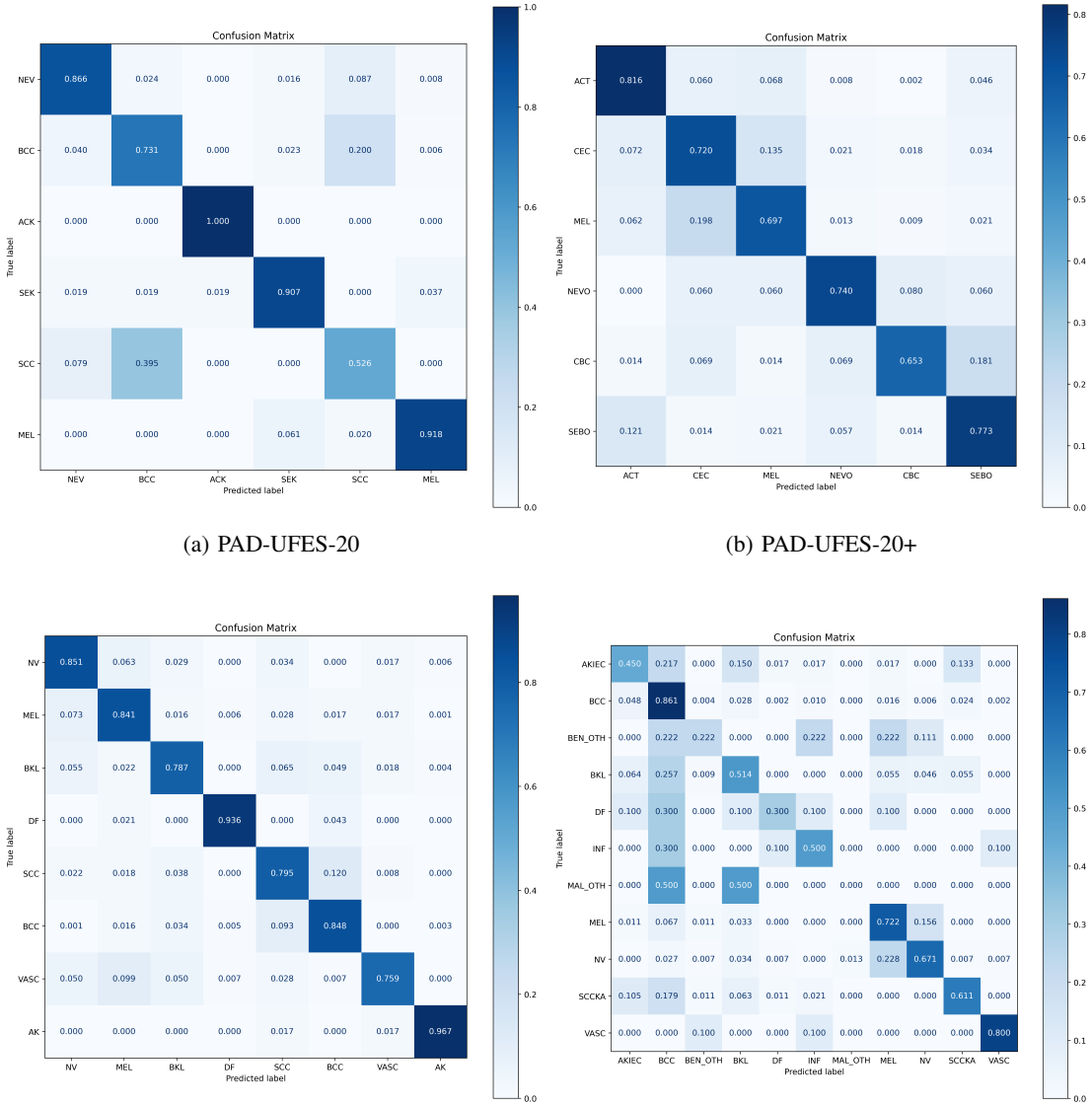
Table III lists the ten metadata features with the highest global mean absolute SHAP values, aggregated across diagnostic categories. These values quantify each feature's overall influence on the predicted probabilities, regardless of direction.

Overall, this analysis provides a clinically interpretable view of metadata-driven behavior in RG-DermNet. While it does not explicitly explain cross-modal interactions, it highlights the complementary role of non-visual information and supports the plausibility of the learned decision patterns on PAD-UFES-20.

Across all diagnostic categories, the most influential metadata features consistently include patient age, lesion growth indicators (*grew*), pruritus (*itch*), lesion elevation, anatomical region (notably face and forearm), and lesion diameter measurements. Also, the found attributes closely match the factors commonly emphasized by dermatologists during clinical assessment, particularly in distinguishing malignant from benign lesions [37]. This alignment indicates that the RG-DermNet model internalizes clinically meaningful patterns rather than relying on spurious correlations within the metadata.

TABLE II: Main performance comparison with state-of-the-art methods on PAD-UFES-20 (mean±sd across 5 patient-wise cross-validation folds). The table highlights the comparative advantage of the proposed RG-DermNet under an identical evaluation protocol. The "$p < 0.05$" column reports the p-value from the Wilcoxon signed-rank test for our method.

Comparative method	**Visual feature extractor**	**ACC**	**BACC**	**F1-Score**	**AUC**	$p < 0.05$
No metadata [9]	ResNet-50	0.44 ± 0.03	0.61 ± 0.03	0.43 ± 0.04	0.88 ± 0.02	yes
Concatenation [20]	ResNet-50	0.54 ± 0.06	0.70 ± 0.04	0.56 ± 0.07	0.92 ± 0.01	yes
MetaBlock [20]	EfficientNet-B0	0.66 ± 0.05	0.73 ± 0.03	0.68 ± 0.05	0.93 ± 0.01	yes
MetaNet [20]	ResNet-50	0.40 ± 0.05	0.58 ± 0.02	0.39 ± 0.08	0.89 ± 0.01	yes
MD-Net [25]	DenseNet169	0.63 ± 0.07	0.71 ± 0.01	0.65 ± 0.07	0.93 ± 0.01	yes
CrossAttention [27]	ResNet-50	0.55 ± 0.04	0.70 ± 0.04	0.55 ± 0.05	0.93 ± 0.01	yes
LiwTERM [23]	ViT-Large-Patch16-224	0.70 ± 0.05	0.74 ± 0.04	0.72 ± 0.05	0.93 ± 0.01	yes
RG-DermNet (Ours)	Caformer-B36	$\mathbf{0.75 \pm 0.05}$	$\mathbf{0.78 \pm 0.03}$	$\mathbf{0.77 \pm 0.04}$	$\mathbf{0.95 \pm 0.01}$	–



(a) PAD-UFES-20

(b) PAD-UFES-20+

(c) ISIC-2019

(d) MILK10K

Fig. 3: Confusion matrices for the best-trained multimodal model on PAD-UFES-20, PAD-UFES-20+, ISIC-2019, and MILK10K (validation).

V. Discussion

The experimental results show that combining residual connections with the proposed Residual–Gated Attention (RG-ATT) block improves multimodal skin lesion classification under a unified patient-wise evaluation protocol. On PAD-UFES-20, the RG-ATT-based model achieves higher BACC and AUC

TABLE III: The ten greatest global mean absolute SHAP values for metadata features, indicating their overall influence on the multimodal model's predicted class probabilities across all samples and diagnostic categories.

Feature	Global mean absolute SHAP value
age	0.0406
grew_False	0.0248
itch_True	0.0237
diameter_2	0.0205
itch_False	0.0171
diameter_1	0.0152
elevation_False	0.0111
elevation_True	0.0109
region_FACE	0.0087
region_FOREARM	0.0081

than previously reported multimodal baselines evaluated under comparable settings (Table II). These gains are supported by the Friedman test ($p = 6.04 \times 10^{-12}$) and pairwise Wilcoxon signed-rank tests ($p < 0.05$), indicating that the improvements are unlikely to be due to chance. However, comparisons with prior studies should be interpreted cautiously, since patient-wise validation reduces optimistic estimates that may arise from intra-patient correlation.

A key observation is that the residual-gated formulation helps regulate the contribution of clinical metadata relative to ungated fusion strategies, such as simple concatenation or standard cross-attention. By modulating metadata-conditioned attention through a residual pathway, RG-ATT is designed to amplify informative metadata while attenuating less reliable signals. This is particularly relevant in dermatological datasets, where metadata quality varies across samples. Missing clinical data is inherent to PAD-UFES-20, exposing the model to incomplete metadata during training. However, controlled robustness analyses under systematically simulated missing or noisy conditions were not explored and remain an important direction for future work. In addition, RG-ATT introduces only lightweight gating and residual operations, resulting in limited overhead compared with conventional multimodal attention-based fusion schemes.

Backbone analysis also reveals differences between CNN-based and Transformer/Metaformer-based architectures. Across datasets, attention-based backbones generally yield higher balanced accuracy than conventional CNNs, likely because they capture both local patterns and broader context. Although no single backbone dominates all datasets, Caformer-B36 and DaViT-Tiny consistently rank among the strongest performers, suggesting that hybrid convolutional–attention designs provide a practical balance between capacity and robustness. This observation suggests that future benchmarking efforts in multimodal dermatological classification should prioritize patient-wise evaluation protocols to ensure fair and clinically meaningful comparisons.

From an interpretability perspective, a SHAP-based analysis on PAD-UFES-20 provides additional insight into how metadata contributes to the multimodal decision process. Using a surrogate Random Forest regressor to approximate the metadata-driven behavior of RG-DermNet, we estimate the relative influence of clinical variables on predicted class probabilities. Patient age, lesion growth indicators, and lesion diameter consistently emerge among the most influential attributes, in agreement with clinical reasoning. Although limited to the metadata pathway, this analysis supports RG-ATT's motivation by showing that clinically relevant metadata signals are effectively leveraged.

Finally, cross-dataset evaluation indicates that the benefits of multimodal fusion depend strongly on metadata quality. Datasets with richer metadata, such as PAD-UFES-20 and PAD-UFES-20+, show clearer gains over image-only baselines, whereas datasets with sparse or weakly informative metadata yield smaller improvements in balanced accuracy, even when AUC remains high. This highlights both the importance of class-aware evaluation metrics in imbalanced clinical scenarios and the dependence of metadata-aware fusion on the quality of non-visual information.

VI. Conclusion

In this paper, we presented RG-DermNet, a multimodal architecture that integrates clinical images and structured metadata through cross-attention and a Residual–Gated Attention (RG-ATT) block. Under a strict patient-wise evaluation protocol across heterogeneous datasets, RG-DermNet consistently matched or outperformed prior multimodal approaches when clinically meaningful metadata was available. On PAD-UFES-20, the proposed model achieved higher BACC and AUC than state-of-the-art fusion baselines evaluated under identical conditions, with improvements supported by Friedman and Wilcoxon statistical tests and confirmed by the A-TOPSIS ranking.

In summary, combining a robust Transformer-based visual backbone (Caformer-B36) with the proposed RG-ATT fusion and an explainability analysis yields a multimodal model that is accurate and clinically interpretable under a patient-wise evaluation protocol. The modular, backbone-agnostic design of RG-DermNet enables adaptation to various visual encoders and metadata configurations. Future work will prioritize validation on additional external cohorts, tighter patient-wise evaluation protocols, and multimodal explainability methods that more directly connect visual evidence with structured clinical variables in practical clinical settings.

Acknowledgment

The authors thank the Espírito Santo Research Foundation (FAPES); the Capixaba Institute of Health Education, Research, and Innovation (ICEPi); the National Council for Scientific and Technological Development (CNPq); the Department of Science and Technology of the Ministry of Health (Decit/SECTICS/MS); Brazilian National Program of Genomics and Precision Health (Genomas Brasil); and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

[1] M. Santos, F. Lima, L. F. L. Martins, J. F. P. Oliveira, L. Almeida, and M. Cancela, "Estimated cancer incidence in brazil, 2023-2025," *Rev Bras Cancerol*, vol. 69, no. 1, pp. e–213 700, 2023.

[2] G. De Pinto *et al.*, "Global trends in cutaneous malignant melanoma incidence," *Melanoma Research*, 2024.

[3] M. Wang, X. Gao, and L. Zhang, "Recent global patterns in skin cancer incidence, mortality, and prevalence," *Chinese Medical Journal*, vol. 138, no. 2, pp. 185–192, 2025. [Online]. Available: https://doi.org/10.1097/CM9.0000000000003416

[4] D. Adla, G. V. R. Reddy, P. Nayak, and G. Karuna, "Deep learning-based computer aided diagnosis model for skin cancer detection and classification," *Distrib. Parallel Databases*, vol. 40, no. 4, p. 717–736, Dec. 2022. [Online]. Available: https://doi.org/10.1007/s10619-021-07360-z

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] F. H. Najjar *et al.*, "Transformer-assisted deep learning for skin lesion classification via hybrid vision transformer models," *Scientific Reports*, 2025, demonstrates improved performance of CNN+ViT hybrid models.

[7] A. Alrabai, A. Echtioui, and F. Kallel, "Transformer and cnn-based deep learning models for skin lesion segmentation," in *2025 IEEE International Conference on Advanced Systems and Emergent Technologies*, 2025.

[8] Q. Pu, Z. Xi, and L. Zhao, "Advantages of transformer and its application for medical image segmentation: a survey," *BioMedical Engineering OnLine*, 2024.

[9] A. G. Pacheco and R. A. Krohling, "The impact of patient clinical information on automated skin cancer detection," *Computers in Biology and Medicine*, vol. 116, p. 103545, 2020.

[10] W. Li, J. Zhuang, R. Wang, J. Zhang, and W.-S. Zheng, "Fusing metadata and dermoscopy images for skin disease diagnosis," in *IEEE International Symposium on Biomedical Imaging*, 2020, pp. 1996–2000.

[11] P. H. Bouzon, W. F. Rocha, L. A. Souza Jr, and A. G. Pacheco, "Metablock-se: a method to deal with missing metadata in multimodal skin cancer classification," *IEEE journal of biomedical and health informatics*, In press 2025.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[13] R. Javed, M. S. M. Rahim, T. Saba, and A. Rehman, "A comparative study of features selection for skin lesion detection from dermoscopic images," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, p. 4, 2020.

[14] A. Adebiyi, N. Abdalnabi, E. J. Simoes, M. Becevic, E. H. Smith, and P. Rao, "Transformers in skin lesion classification and diagnosis: A systematic review," *medRxiv*, 2024. [Online]. Available: https://www.medrxiv.org/content/early/2024/09/30/2024.09.19.24314004

[15] D. Bisla, A. Choromanska, R. S. Berman, J. A. Stein, and D. Polsky, "Towards automated melanoma detection with deep learning: Data purification and augmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 2720–2728.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[17] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nature medicine*, vol. 26, no. 6, pp. 900–908, 2020.

[18] Z. Mirikharaji, K. Abhishek, A. Bissoto, C. Barata, S. Avila, E. Valle, M. E. Celebi, and G. Hamarneh, "A survey on deep learning for skin lesion segmentation," *Medical Image Analysis*, vol. 88, p. 102863, 2023.

[19] P. A. Lyakhov, U. A. Lyakhova, and D. I. Kalita, "Multimodal analysis of unbalanced dermatological data for skin cancer recognition," *IEEE Access*, vol. 11, pp. 131 487–131 507, 2023.

[20] A. G. Pacheco and R. A. Krohling, "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*, vol. 25, no. 9, pp. 3554–3563, 2021.

[21] R. B. Oliveira, J. P. Papa, A. S. Pereira, and J. M. R. Tavares, "Computational methods for pigmented skin lesion classification in images: review and future trends," *Neural Computing and Applications*, vol. 29, no. 3, pp. 613–636, 2018.

[22] 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," *Vis. Comput.*, vol. 39, no. 7, p. 2781–2793, may 2022. [Online]. Available: https://doi.org/10.1007/s00371-022-02492-4

[23] L. A. Souza, A. G. Pacheco, G. G. de Angelo, T. Oliveira-Santos, C. Palm, and J. P. Papa, "Liwterm: A lightweight transformer-based model for dermatological multimodal lesion detection," in *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2024, pp. 1–6.

[24] W. Li, J. Zhuang, R. Wang, J. Zhang, and W.-S. Zheng, "Fusing metadata and dermoscopy images for skin disease diagnosis," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1996–2000.

[25] W. long Yin, J. Huang, J. Chen, and Y. Ji, "A study on skin tumor classification based on dense convolutional networks with fused metadata," *Frontiers in Oncology*, vol. 12, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254808062

[26] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 347–356.

[27] C. Ou, S. Zhou, R. Yang, W. Jiang, H. He, W. Gan, W. Chen, X. Qin, W. Luo, X. Pi *et al.*, "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," *Frontiers in Surgery*, vol. 9, p. 1029991, 2022.

[28] N.-Y. Tran-Van and K.-H. Le, "A multimodal skin lesion classification through cross-attention fusion and collaborative edge computing," *Computerized Medical Imaging and Graphics*, vol. 124, p. 102588, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611125000977

[29] A. G. Pacheco, G. R. Lima, A. S. Salomão, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro *et al.*, "PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in Brief*, vol. 32, pp. 1–10, 2020.

[30] P. Tschandl *et al.*, "Milk10k: A hierarchical multimodal imaging-learning toolkit for diagnosing pigmented and nonpigmented skin cancer and its simulators," *Journal of Investigative Dermatology*, 2025. [Online]. Available: https://doi.org/10.1016/j.jid.2025.06.1594

[31] "Isic 2019: Skin lesion analysis towards melanoma detection challenge dataset," ISIC Archive, 2019, available at: https://challenge.isic-archive.com/data/.

[32] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[33] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4804–4814.

[34] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," in *European conference on computer vision*. Springer, 2022, pp. 74–92.

[35] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, and X. Wang, "Metaformer baselines for vision," *arXiv preprint arXiv:2210.13452*, 2022.

[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[37] J. Höhn, A. Hekler, E. Krieghoff-Henning, J. N. Kather, J. S. Utikal, F. Meier, F. F. Gellrich, A. Hauschild, L. French, J. G. Schlager, K. Ghoreschi, T. Wilhelm, H. Kutzner, M. Heppt, S. Haferkamp, W. Sondermann, D. Schadendorf, B. Schilling, R. C. Maron, M. Schmitt, T. Jutzi, S. Fröhling, D. B. Lipka, and T. J. Brinker, "Integrating patient data into skin cancer classification using convolutional neural networks: Systematic review," *J Med Internet Res*, vol. 23, no. 7, p. e20708, Jul 2021. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/34255646