scientific reports

Check for updates

OPEN

Multimodal dual-stage feature refinement for robust skin lesion classification

Mahapara Khurshid, Richa Singh & Mayank Vatsa✉

Skin cancer continues to pose a formidable global health challenge, where expedient detection is paramount to diminishing mortality. However, the inherent heterogeneity of skin lesions, exacerbated by class imbalance, frequently undermines automated classification efforts, particularly in unconstrained environments such as smartphone imagery that lacks dermoscopic clarity. In this research, we present *DualRefNet*, a novel multimodal deep learning paradigm that employs a dual-stage feature refinement strategy. First, an auxiliary super-resolution task augments visual representations; subsequently, a class-frequency-based regularization of the final fully connected layers refines the fused features, thus mitigating errors induced by high intra-class and low inter-class variability. Concurrently, a weighted cross-entropy loss deftly addresses class imbalance. Empirical evaluations on the PAD-UFES20 and ISIC-2019 datasets demonstrate balanced accuracies of 0.845 and 0.815, respectively, attesting to *DualRefNet*'s prowess under varied conditions. Furthermore, the confusion matrix and class-wise analyses highlight its equitable performance across all categories, rendering it a potential candidate for widespread, resource-constrained deployments.

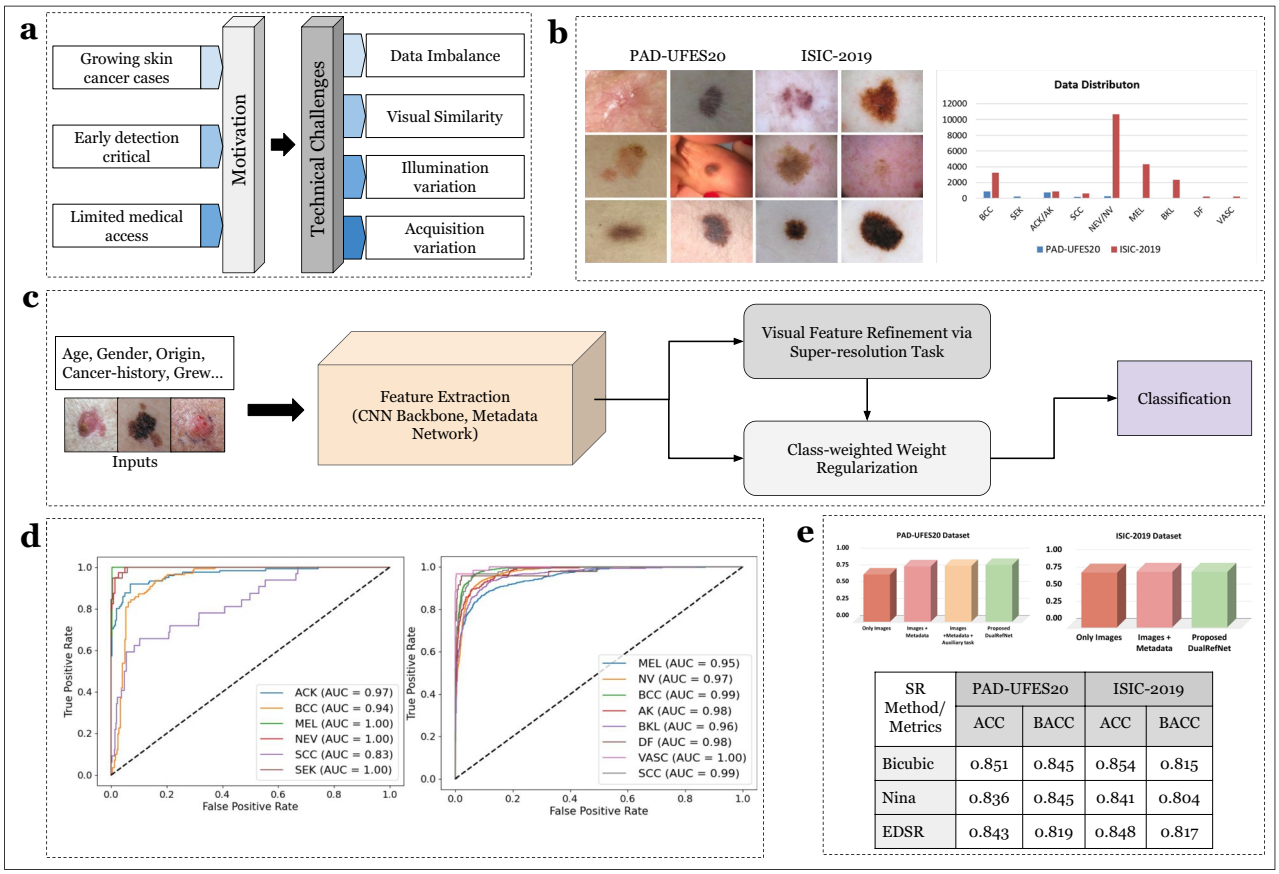
Among the most prevalent malignancies worldwide, skin cancer stands out for its extensive impact on individuals of every age and background. According to the World Health Organization (WHO), approximately 330,000 new melanoma cases were recorded globally in 2022, culminating in nearly 60,000 deaths[1]. By 2040, projections indicate over 500,000 novel melanoma diagnoses and 100,000 associated fatalities annually, thus emphasizing the paramount importance of early detection in reducing mortality rates[2,3].

Dermatoscopy, guided by the ABCD rule (Asymmetry, Border, Color, Diameter), remains a widely embraced clinical modality for diagnosing pigmented skin lesions[4,5]. While efficacious, this practice can be susceptible to diagnostic subjectivity, laborious screening procedures, and inconsistencies in interpretation, particularly in regions bereft of sufficient dermatological expertise[6,7]. As illustrated in Fig. 1(a), developing automated solutions for skin lesion classification entails multiple technical challenges, including robust feature extraction and balancing class distributions. Consequently, there is a growing need for automated classification systems that can handle class imbalance and lesion variability, thereby enabling clinicians to make more accurate and efficient diagnoses[8].

Automated skin lesion classification has advanced significantly through the use of dermoscopic images[9–11]. Nevertheless, as depicted in Fig. 1(b), these methods frequently encounter difficulties such as low inter-class variation, high intra-class variation, inconsistent illumination, and significant class imbalance[12]. These issues become even more pronounced in unconstrained scenarios where smartphone-captured images, despite their broader accessibility, generally lack the level of detail captured by dermoscopic equipment. Addressing these gaps requires solutions that effectively manage data imbalance, capture nuanced class variations, and function reliably in resource-limited settings.

Skin lesion classification research can be broadly categorized into *unimodal* and *multimodal* approaches, distinguished by the type or variety of input data. Unimodal methods rely on a single data source (e.g., images or text) and concentrate on feature extraction, class imbalance mitigation, and addressing image-centric variability. Broadly, these approaches can be classified into deep feature extraction[13–15], attention mechanisms[16–19], feature enhancement via additional blocks[20], or ensemble learning[21,22]. Although unimodal techniques have demonstrated significant progress in extracting robust features and optimizing model architectures[23], their reliance on image data alone can limit diagnostic comprehensiveness-particularly when lesion classes appear visually similar. In contrast, multimodal approaches integrate images with additional metadata (e.g., clinical or demographic information), aligning more closely with the diagnostic workflow employed by dermatologists, who often combine visual inspection with patient-specific details.

Department of CSE, Indian Institute of Technology Jodhpur, Jodhpur, India. ✉email: mvatsa@iitj.ac.in



SR Method/ Metrics	PAD-UFES20		ISIC-2019	
	ACC	BACC	ACC	BACC
Bicubic	0.851	0.845	0.854	0.815
Nina	0.836	0.845	0.841	0.804
EDSR	0.843	0.819	0.848	0.817

Fig. 1. Overview of the proposed DualRefNet (**a**) motivation and technical challenges in skin lesion classification (**b**) Dataset analysis showing variations in image illumination and lesion classes and the histogram showing the class distribution to highlight data imbalance (**c**) the high-level view of the proposed DualRefNet, where the visual features are first extracted from input images, then refined through an auxiliary super-resolution task, followed by weight regularization of the final fully connected layers to enhance fused features (visual and metadata) and accuracy (**d**) the AUC curves showing the classwise performance for each dataset (left: PAD-UFES20, right: ISIC-2019) and (**e**) the ablation results presenting the improvement in performance by integrating metadata, auxiliary task and the class-weighted weight regularization, accompanied by the table presenting the effect of varying super-resolution methods (SR) in the proposed approach.

Multimodal strategies integrate image data with complementary metadata (e.g., clinical or demographic information) to provide a holistic diagnostic perspective that closely mirrors the workflow of dermatologists, who routinely combine visual inspection with patient-specific details[24]. By jointly leveraging both visual features and metadata, these approaches more effectively address challenges such as class imbalance and visually similar lesions[25]. Representative paradigms include feature-level fusion[26–29], multi-task learning[30–32], attention-based mechanisms[33–35], and joint fusion models[36–39]. Despite their advantages, multimodal methods can still encounter difficulties when aligning and merging disparate data types, and simplistic fusion techniques may underperform if fused representations are not carefully optimized. Moreover, many of the existing approaches rely on simple concatenation or fusion methods and do not focus on improving the quality of the extracted features. Hence, there is a scope to refine the extracted features before classification to improve model performance.

Research contributions: This paper introduces *DualRefNet*, a novel dual-stage feature refinement framework specifically designed to tackle persistent challenges in skin lesion classification, such as class imbalance, high intra-class variation, and low inter-class variation. As illustrated in Fig. 1(c) and 2, *DualRefNet* systematically refines both *visual* and *fused* feature spaces through two key mechanisms:

1. Auxiliary task for visual feature refinement. A super-resolution prediction task is jointly trained with the primary classification objective using a shared encoder. This setup prompts the visual features to capture richer, class-specific details and mitigates challenges arising from limited visual quality or significant intra-class variation[32,40–42].
2. Class-frequency-based weight regularization for fused features. In the second stage, the model integrates metadata (e.g., clinical or demographic attributes) into the fused feature space. A class-frequency-driven regularization term is applied to the weights of the final fully connected layers using class-specific weights. By including these class weights in the regularization term, we impose amplified penalties on weights associated

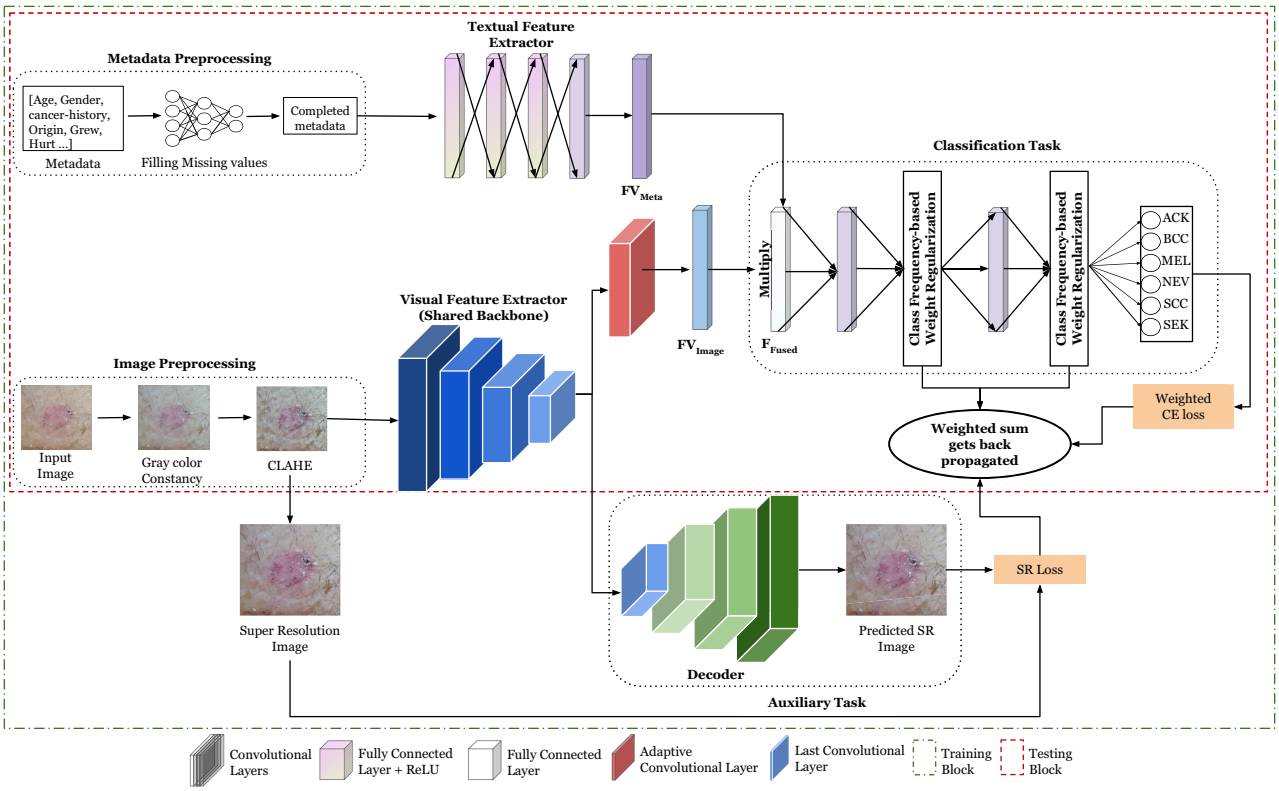


Fig. 2. Showcasing the architecture of the proposed approach, illustrating the training and testing process of the proposed approach. The auxiliary task acts as a guide to refine the visual features, and class-frequency-based weight regularisation ensures balanced learning to handle the class imbalance.

with rarer lesion types. This ensures balanced learning by reducing bias toward the majority classes and is further reinforced by a weighted cross-entropy loss function.

Unlike methods that optimize only visual representations or rely on simplistic data fusion strategies, *DualRefNet* provides a synergistic combination of super-resolution guidance and frequency-based weight regularization. Figure 3 presents the comparison of the proposed DualRefNet with the existing architectures for multimodal skin lesion classification. Building on our preliminary work[32], which focused on refining visual features via an auxiliary super-resolution task, this enhanced version additionally optimizes fused (visual + textual) representations through class-frequency-based weight regularization. By strengthening class separation and prioritizing underrepresented classes, *DualRefNet* enables more accurate and equitable performance-particularly in real-world, resource-constrained scenarios where high-fidelity imaging and comprehensive metadata may be scarce.

We evaluate our approach on two multimodal benchmark datasets-PAD-UFES20 [43] (smartphone-captured images) and ISIC-2019 [44,45,46] (dermoscopic images)-under unconstrained conditions that include variability in image quality and metadata availability. Experimental results confirm that *DualRefNet* surpasses state-of-the-art methods, achieving more robust and balanced classification across different lesion types.

Results

This section presents the datasets used to evaluate *DualRefNet*, followed by a detailed discussion of the experimental outcomes.

Dataset details

We evaluated the proposed framework on two well-established multimodal skin lesion datasets: **PAD-UFES20**[43] and **ISIC-2019**[44–46]. These datasets were specifically chosen due to their diversity, multimodal nature, and strong relevance to practical clinical scenarios. Beyond supporting multi-class classification, they incorporate a range of metadata that aids in evaluating the generalization capacity of deep learning models for lesion diagnosis. Table 1 summarizes the train, test, and validation splits for each dataset.

PAD-UFES20[43]. This dataset contains 2298 images captured by smartphones, categorized into six lesion types: Melanoma (MEL), Melanocytic nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Squamous Cell Carcinoma (SCC), and Seborrheic Keratosis (SEK). Each image is accompanied by 26 metadata features, including patient demographics (age, gender), lesion characteristics (e.g., location, itch, growth), and

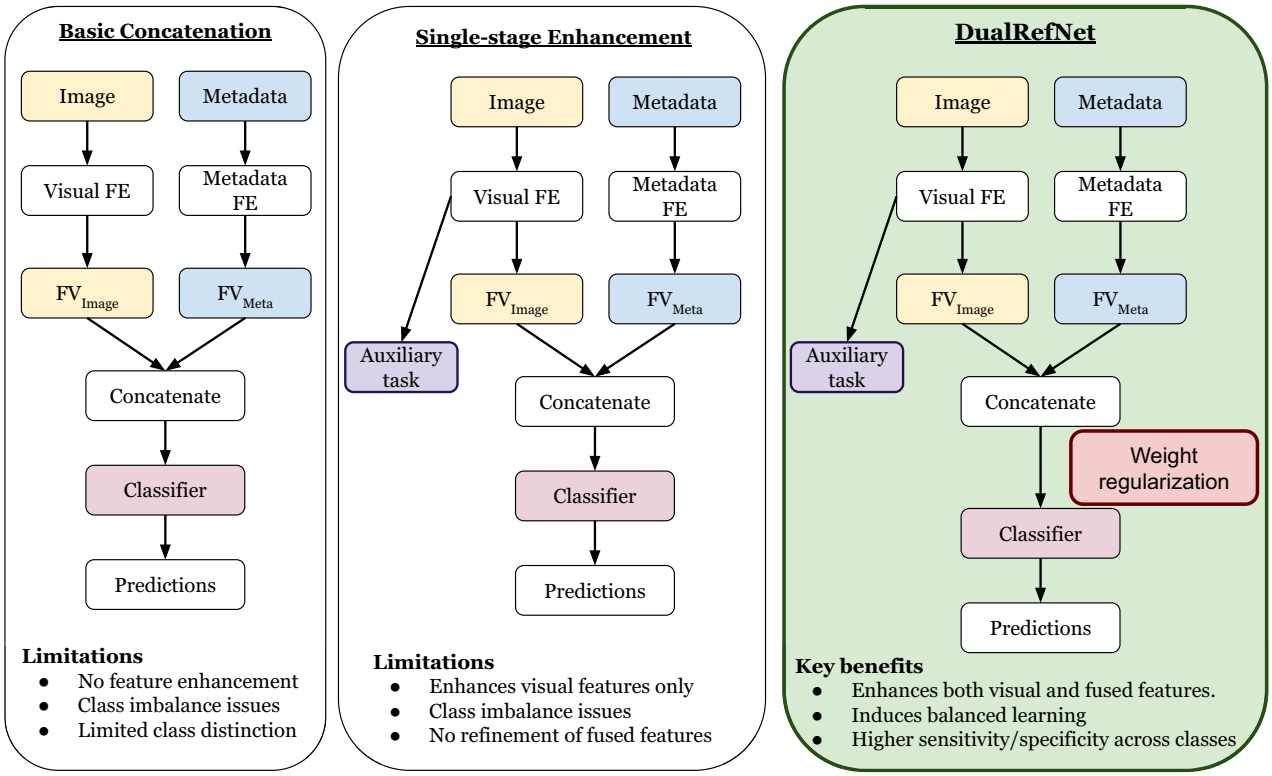


Fig. 3. Showcasing the comparison of the proposed DualRefNet with the existing architectures for multimodal skin lesion classification.

Dataset	Total images	Split	BCC	SEK	ACK/AK	SCC	NEV/NV	MEL	BKL	DF	VASC
PAD-UFES20	2298	Train	563	157	486	128	163	35	–	–	–
		Test	141	39	122	32	40	9	–	–	–
		Val	141	39	122	32	41	8	–	–	–
ISIC-2019	22480	Train	2337	–	591	458	7642	3171	1667	168	151
		Test	646	–	182	121	2119	832	489	47	60
		Val	262	–	72	43	871	343	177	20	11

Table 1. Dataset details showcasing the number of samples from each class in each split with individual classes being Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic keratosis (AK), Benign keratosis (BKL), Dermatofibroma (DF), Vascular lesion (VASC), Squamous cell carcinoma (SCC), and Seborrheic Keratosis (SEK).

unique identifiers (e.g. patient ID). Following prior studies[32,34], we employ an 80:20 split for training and testing. The training split is further divided into training and validation subsets.

ISIC-2019 [44–46]. Comprising 25,331 dermoscopic images for training and 8238 for testing, this dataset covers eight lesion classes: MEL, NV, BCC, AK, Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), and SCC. Each image is supplemented with metadata on the patient age, gender, and anatomical site. To avoid biases, we excluded approximately 3051 samples lacking metadata. Since the official test set is not publicly available, we partitioned the remaining data into a new training and testing set in an 80:20 ratio. The training set is then further divided into training and validation subsets for model selection.

Classification results

We evaluated *DualRefNet* using five CNN architectures: VGG-13[47], ResNet50[48], MobileNet-V2[49], EfficientNet-B4[50], and DenseNet-121[51], all pre-trained on ImageNet[52]. These models vary in depth, width, and connection patterns (e.g. sequential vs. residual), making them an ideal suite of feature extractors. Each CNN was fine-tuned on the respective dataset (PAD-UFES20 or ISIC-2019) to account for different capture devices and imaging conditions. For the final classification, we added two fully connected layers on top of each backbone network, with the first layer mapping the extracted features to an intermediate latent space, and the second layer producing the final class probabilities.

For metadata processing (clinical and demographic), we employed a fully connected four-layer network with dimensions of [64, 128, 256, 512]. We also incorporated an auxiliary super-resolution task-implemented via bilinear or bicubic interpolation, as well as deep-learning-based techniques[53,54]-to enhance visual feature learning. The images were uniformly resized to 224×224×3 and then upscaled by a factor of 2 to 448×448×3. Model training continued for 70 epochs with a batch size of 32, using stochastic gradient descent (SGD) as the optimizer. We set the learning rate, weight decay, and patience hyperparameters to [0.01, $1e^{-3}$, 5] for PAD-UFES20 and [0.001, $1e^{-5}$, 20] for ISIC-2019, respectively. In addition, standard data augmentations (e.g., flips, random scaling, brightness/contrast changes, saturation adjustments, and noise) were applied to increase the diversity of training samples. All experiments were performed in PyTorch on an Nvidia DGX station.

Tables 3a and 3b present the classification results and comparative analyses for PAD-UFES20 and ISIC-2019, respectively. We report balanced accuracy (BACC), classification accuracy (ACC), and area under the ROC curve (AUC) to assess overall performance. To provide a deeper clinical insight, we also computed precision, F1 score, sensitivity, specificity, and class-wise accuracy for the best-performing model, along with Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Kappa Statistic (K) given in Tables 2a and 2b. This comprehensive evaluation demonstrates both predictive accuracy and reliability across different lesion categories, offering a robust understanding of *DualRefNet*'s capabilities.

Results on PAD-UFES20 Dataset[43]: For this dataset, our best performing model (VGG13) achieves an ACC of 0.851, a BACC of 0.845, and an AUC of 0.997. Notably, similar gains are observed with other CNN backbones, emphasizing the overall effectiveness of the proposed approach. Table 2 presents additional evaluation metrics for VGG13, including precision, recall, sensitivity, specificity, and class-wise accuracy, which are critical in assessing clinical reliability for medical imaging applications.

The results in Table 2 a indicates that specificity exceeds 90% for all classes in PAD-UFES20, highlighting the strong ability of the model to correctly identify non-lesion or benign cases. Except for class SCC, sensitivity values for all other classes surpass 80%, reflecting robust detection of true positives and affirming the clinical utility of the approach for assisting in reliable diagnoses. Although SCC shows comparatively lower sensitivity, it generally warrants a confirmatory biopsy due to its pigmentation, minimizing any adverse clinical impact.

In addition to sensitivity and specificity, we report PPV and NPV values, both of which are high across most classes. PPV scores greater than 0.87 demonstrate strong confidence in positive predictions, while NPV scores emphasize the accuracy of the method in ruling out negatives. The Kappa statistic (0.80) and overall accuracy of 0.85 (95% CI) provide further evidence of consistency and reliability.

Results on ISIC-2019 Dataset[44–46]. For ISIC-2019, DenseNet-121 emerges as the best-performing backbone, achieving ACC, BACC, and AUC of 0.854, 0.815, and 0.986, respectively. Table 2b reports additional metrics, showing that specificity remains above 90% across all classes and sensitivity stays above 70%. PPV and NPV values are similarly high (above 0.70), emphasizing the method's reliable classification of both positive and negative instances. The Kappa statistic (0.79) and overall accuracy of 0.85 (95% CI) confirm the model's strong agreement with ground truth and its robustness across varying lesion types.

Collectively, these results confirm that *DualRefNet* maintains a balanced performance by effectively managing trade-offs between sensitivity and specificity. The consistently high AUC values-especially pronounced with VGG13 on PAD-UFES20 (0.997)-highlight the role of dual-stage feature refinement in creating well-separated, discriminative feature representations. By boosting both visual and fused features, the proposed approach delivers enhanced clarity in lesion classification.

Metrics	ACK	BCC	MEL	NEV	SCC	SEK		
(a)								
Precision	0.91	0.87	0.89	0.88	0.47	0.95		
F1 score	0.89	0.86	0.89	0.90	0.53	0.95		
Sensitivity	0.88	0.84	0.89	0.93	0.59	0.95		
Specificity	0.96	0.93	0.99	0.98	0.94	0.99		
Positive predictive value (PPV)	0.91	0.87	0.89	0.88	0.47	0.95		
Negative predictive value (NPV)	0.96	0.93	1.00	0.99	0.94	0.99		
Classwise accuracy	0.88	0.84	0.89	0.93	0.59	0.95		

Metrics	MEL	NV	BCC	AK	BKL	DF	VASC	SCC
(b)								
Precision	0.83	0.91	0.87	0.67	0.75	0.73	0.88	0.71
F1 score	0.79	0.92	0.88	0.71	0.75	0.75	0.91	0.72
Sensitivity	0.76	0.92	0.89	0.75	0.76	0.77	0.95	0.73
Specificity	0.96	0.91	0.97	0.98	0.96	0.98	0.98	0.99
Positive predictive value (PPV)	0.83	0.91	0.87	0.67	0.75	0.73	0.88	0.71
Negative predictive value (NPV)	0.96	0.92	0.98	0.98	0.97	1.00	1.00	0.99
Classwise accuracy	0.76	0.92	0.89	0.75	0.76	0.77	0.95	0.73

Table 2. Showcasing (**a**) Evaluation metrics of the proposed approach on PAD-UFES20 dataset[43] using the best-performing model- VGGNet13 and (**b**) Evaluation metrics of the proposed approach on ISIC-2019 dataset[44–46] using the best-performing model- DenseNet121.

Comparative analysis with existing approaches

We have compared the performance of the proposed algorithm with multiple existing algorithms as given in Table 3. The algorithms were chosen for their relevance to multimodal medical image analysis and their application on similar skin lesion datasets. These methods use diverse fusion strategies and represent recent state-of-the-art techniques, making them suitable baselines for comparison. We followed similar dataset protocols to ensure consistency and used the performance metrics reported in the original publications to avoid redundant computation. The key characteristics of the selected approaches are summarized below:

- Pacheco et al.[26] studied the impact of metadata on the classification performance of skin cancer detection. The authors used various CNN architectures and showcased an increase of 7% in balanced accuracy
- Li et al.[38] proposed a multiplication-based data fusion approach for skin cancer detection, where metadata features modulate the importance of visual feature channels by enhancing focus on relevant regions. Experimental results show improved performance, especially for small-sample classes, highlighting the importance of selective metadata usage for optimal results.

Model	Metrics	VGG13	ResNet50	EfficientNet-B4	MobileNet-V2	DenseNet-121
(a)						
No metadata[33]	ACC	0.709	0.616	0.656	0.655	0.636
	BACC	0.654	0.651	0.640	0.637	0.640
	AUC	0.901	0.901	0.911	0.898	0.893
Concat[26]	ACC	0.712	0.741	0.765	0.738	0.742
	BACC	0.720	0.728	0.758	0.741	0.747
	AUC	0.929	0.929	0.945	0.927	0.932
MetaNet[38]	ACC	0.749	0.732	0.744	0.700	0.745
	BACC	0.754	0.742	0.737	0.717	0.745
	AUC	0.937	0.936	0.931	0.922	0.933
MetaBlock[33]	ACC	0.728	0.735	0.748	0.724	0.723
	BACC	0.736	0.765	0.770	0.754	0.746
	AUC	0.933	0.935	0.944	0.938	0.931
Visual[34]	ACC	0.807	0.812	0.772	0.806	0.709
	BACC	0.770	0.806	0.784	0.789	0.779
	AUC	0.952	0.953	0.953	0.954	0.950
AuxNet[32]	ACC	0.849	0.848	0.833	0.836	0.822
	BACC	0.832	0.811	0.797	0.811	0.794
	AUC	0.960	0.967	0.967	0.953	0.962
Proposed	ACC	**0.851**	**0.843**	**0.846**	**0.836**	**0.846**
	BACC	**0.845**	**0.811**	**0.816**	**0.804**	**0.812**
	AUC	**0.997**	**0.995**	**0.994**	**0.981**	**0.982**
Model	**Metrics**	**VGG13**	**ResNet50**	**EfficientNet-B4**	**MobileNet-V2**	**DenseNet-121**
(b)						
Concat[26]	ACC	0.724	0.729	0.784	0.716	0.738
	BACC	0.729	0.726	0.768	0.723	0.737
	AUC	0.949	0.948	0.960	0.946	0.952
MetaNet[38]	ACC	0.767	0.753	0.766	0.742	0.725
	BACC	0.746	0.746	0.756	0.731	0.723
	AUC	0.959	0.956	0.959	0.955	0.949
MetaBlock[33]	ACC	0.753	0.804	0.807	0.777	0.800
	BACC	0.740	0.771	0.762	0.760	0.769
	AUC	0.955	0.966	0.962	0.958	0.965
Visual[34]	ACC	**0.846**	0.827	0.826	0.804	0.839
	BACC	**0.818**	0.791	0.782	0.790	0.807
	AUC	0.976	0.971	0.968	0.968	0.967
Proposed	ACC	0.824	**0.844**	**0.835**	**0.838**	**0.854**
	BACC	0.812	**0.806**	**0.800**	**0.815**	**0.815**
	AUC	**0.987**	**0.982**	**0.982**	**0.983**	**0.986**

Table 3. Showcasing (a) performance comparison with existing approaches on PAD-UFES20 dataset[43] and (b) performance comparison with existing approaches on ISIC-2019 dataset[44–46]. (ACC: Accuracy, BACC: Balanced Accuracy, AUC: Area Under Curve).

- MetaBlock[33] introduces an LSTM-inspired architecture where metadata features act as attention weights to reweight intermediate visual feature maps. Experimental results demonstrate improved classification performance on dermoscopic and clinical skin lesion datasets.
- Pundhir et al.[34] proposed a deep learning-based skin lesion classification approach that enhances visual context using a visual attention mechanism over CNNs. It integrates skin lesion images with patient demographics to guide attention toward clinically relevant regions. This multimodal attention improves the overall classification performance.
- Vachmanus et al.[35] proposed DeepMetaForge, a deep learning framework for multimodal skin cancer detection. It uses BEiT, a vision transformer, for image encoding and a custom Deep Metadata Fusion Module that merges visual and metadata features while blending them simultaneously. The authors also analysed scalability to showcase its applicability to other relevant paradigms.
- Pham et al.[29] proposed a multimodal skin lesion classification framework that combines smartphone-captured images with patient metadata. It uses an ensemble of three models, each integrating a pre-trained image encoder and a 1D CNN for metadata processing. The concatenated features are classified jointly, and the final prediction is obtained via weighted averaging of the models' softmax outputs.
- Tang et al.[39] proposed a novel fusion strategy for multimodal skin cancer classification, introducing the Joint-Individual Fusion (JIF) structure to jointly learn shared and modality-specific features and the Multimodal Fusion Attention (MMFA) module to enhance salient features through self and mutual attention mechanisms. Experimental results demonstrate that the JIF-MMFA framework consistently outperforms existing fusion methods across various CNN backbones.
- Khurshid et al.[32] introduced a method that enhances visual features through an auxiliary super-resolution prediction task. The approach effectively improves skin lesion classification performance by jointly optimizing the main classification task and the auxiliary task. Experimental results demonstrate superior performance over state-of-the-art methods across multiple evaluation metrics.

Unlike prior models that fuse image and metadata features through early, late, or hybrid strategies without enhancing the input quality, DualRefNet introduces a visual refinement stage via an auxiliary super-resolution task. This encourages the shared encoder to capture fine-grained, lesion-specific details often overlooked by standard classification pipelines. Furthermore, to address the prevalent class imbalance issue in skin lesion datasets, DualRefNet incorporates a second-stage class-frequency-based weight regularization, which applies stronger penalties to under-represented classes in the final fully connected layers. This promotes balanced learning across all lesion types.

Tables 3(a),(b) compare *DualRefNet* with leading methods on PAD-UFES20 and ISIC-2019, respectively. For a fair comparison, we adopt established protocols [32,34] and baseline approaches with published metrics [26,32–34,38]. For PAD-UFES20, our approach achieves a BACC of 0.845-exceeding the previous state-of-the-art of 0.832[32]- and an AUC of 0.997, outperforming the earlier benchmark of 0.960. We have added a comparative analysis with other recent approaches over various evaluation metrics in Table 4. Our proposed model outperforms most existing approaches on all metrics. Although the Ensemble[29] achieves slightly higher precision and F1-score by 0.02 and 0.008, respectively, it relies on an ensemble of at least three visual feature extractors along with corresponding metadata encoders, resulting in significantly higher computational cost. In contrast, the proposed DualRefNet uses a single feature extractor per modality while maintaining competitive performance.

On ISIC-2019 *DualRefNet*, enhances state of the art in terms of ACC and AUC, and achieves a BACC of 0.815 with DenseNet-121, closely matching the best-reported values in[34]. Minor differences stem from the limited clinical metadata available in the ISIC-2019 dataset and the exclusion of images with missing metadata. These findings emphasize the importance of rich metadata for optimized performance. Overall, our method's strong results across both datasets highlight its ability to tackle class imbalances and variations in skin lesions, highlighting its suitability for clinical scenarios with diverse imaging conditions.

Qualitative analysis

Figure 4(a) provides the confusion matrices for both datasets, illustrating that *DualRefNet* successfully increases the classification accuracy for the SCC class (from 0.50 to 0.59) in PAD-UFES20, without significantly compromising performance for other classes. This improvement showcases the model's balanced learning objectives. Meanwhile, Fig. 4(a-ii) highlights class-wise misclassifications in ISIC-2019, where sensitivity remains above 70% for all classes. Most errors occur between lesions that share morphological or color-based similarities, presenting a recognized challenge even for clinical experts. Also evident in Fig. 4(b-ii) are t-SNE

Model	Accuracy	Precision	Recall	F1 score
MetaBlock[33]	0.726	0.677	0.726	0.696
DeepMetaForge[35]	0.769	0.773	0.769	0.769
JIF[39]	0.830	–	–	–
AuxNet[32]	0.849	0.825	0.832	0.828
Ensemble[29]	0.844	0.849	0.844	0.845
Proposed	0.851	0.828	0.847	0.837

Table 4. Showcasing performance comparison with existing approaches using various evaluation metrics on the PAD-UFES20 dataset[43].

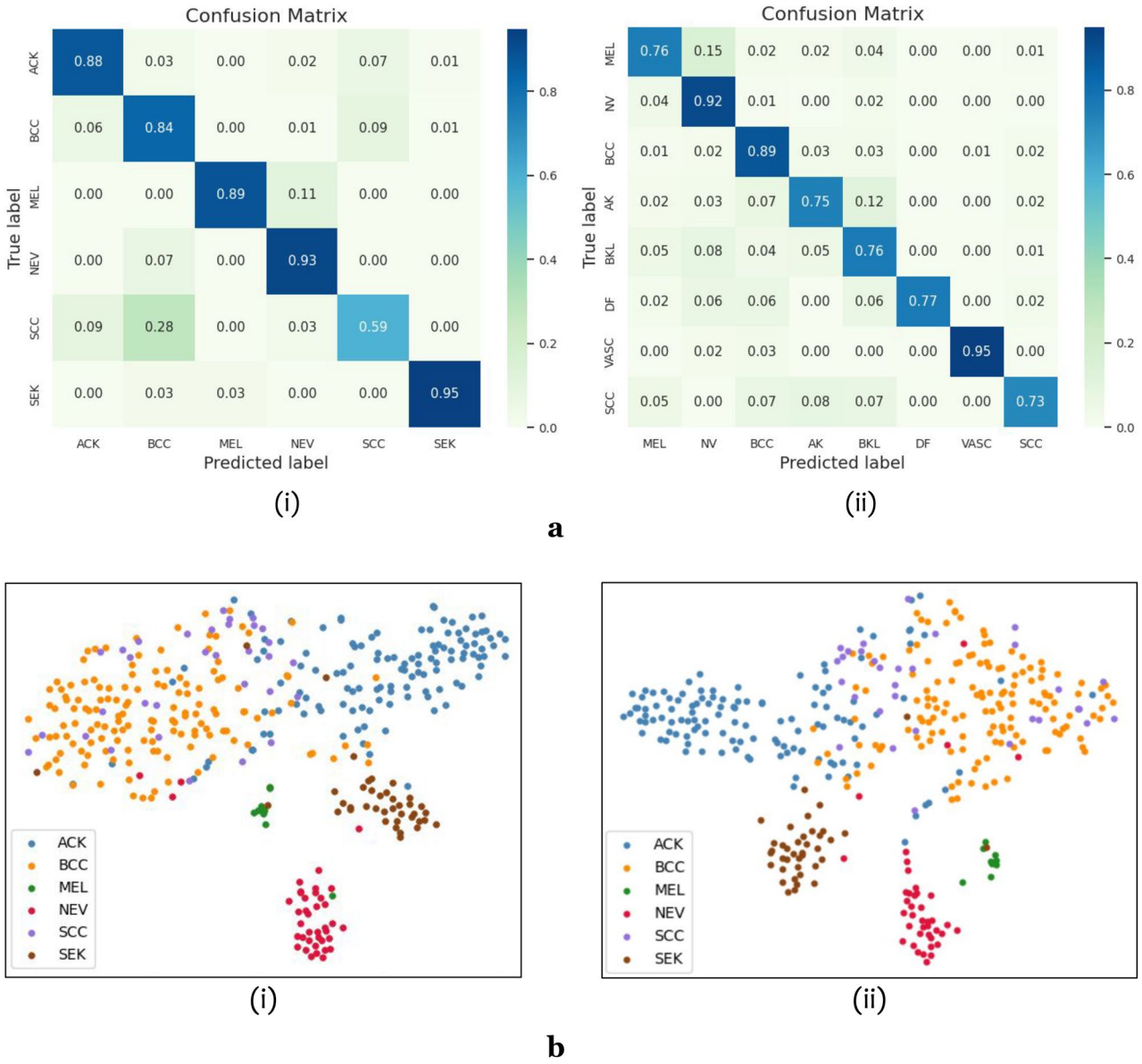


Fig. 4. Showcasing the (**a**) confusion matrix of the (i) PAD-UFES20 dataset[43] and (ii) ISIC-2019 dataset[44–46] on their respective best-performing models and (**b**) Comparing tSNE plots of the PAD-UFES20 using (i) SOTA[32] and (ii) the proposed approach on the best-performing model. The proposed approach achieves more compact and well-distinguished clusters, which showcases the ability of the model to learn more discriminating features of the dataset.

plots comparing embeddings from *DualRefNet* with those of a prior state-of-the-art (SOTA) approach. The tighter, more distinct clusters suggest that our dual-stage refinement strategy yields more discriminative feature representations.

Figure 5(a) further illustrates both correctly classified and misclassified samples from PAD-UFES20 and ISIC-2019, along with the highest predicted probability for each lesion. As shown, classes that are visually alike in boundary characteristics, color, or illumination often lead to misclassifications-an issue even experienced dermatologists can encounter. Nonetheless, these findings also suggest avenues for future improvement, such as incorporating additional metadata features or refined data augmentation, to better distinguish visually similar lesion types. To further analyze the model predictions, we used gradient-weighted class activation mapping (GradCAM)[55]. This visualization highlights the important regions which are focused on when making a prediction. It is evident in Fig. 5(b) that the model focuses on the colour variations or lesion borders when making decisions. These visualizations demonstrate the models' ability to identify clinically relevant features without explicit annotation of these regions during training. To analyze and interpret the influence of metadata on model predictions, we used SHAP (SHapley Additive exPlanations). We generated a SHAP summary plot to quantify the global importance of metadata features in the proposed model. The plot aggregates the SHAP values across a random subset of the test set, highlighting the metadata features that consistently influence the

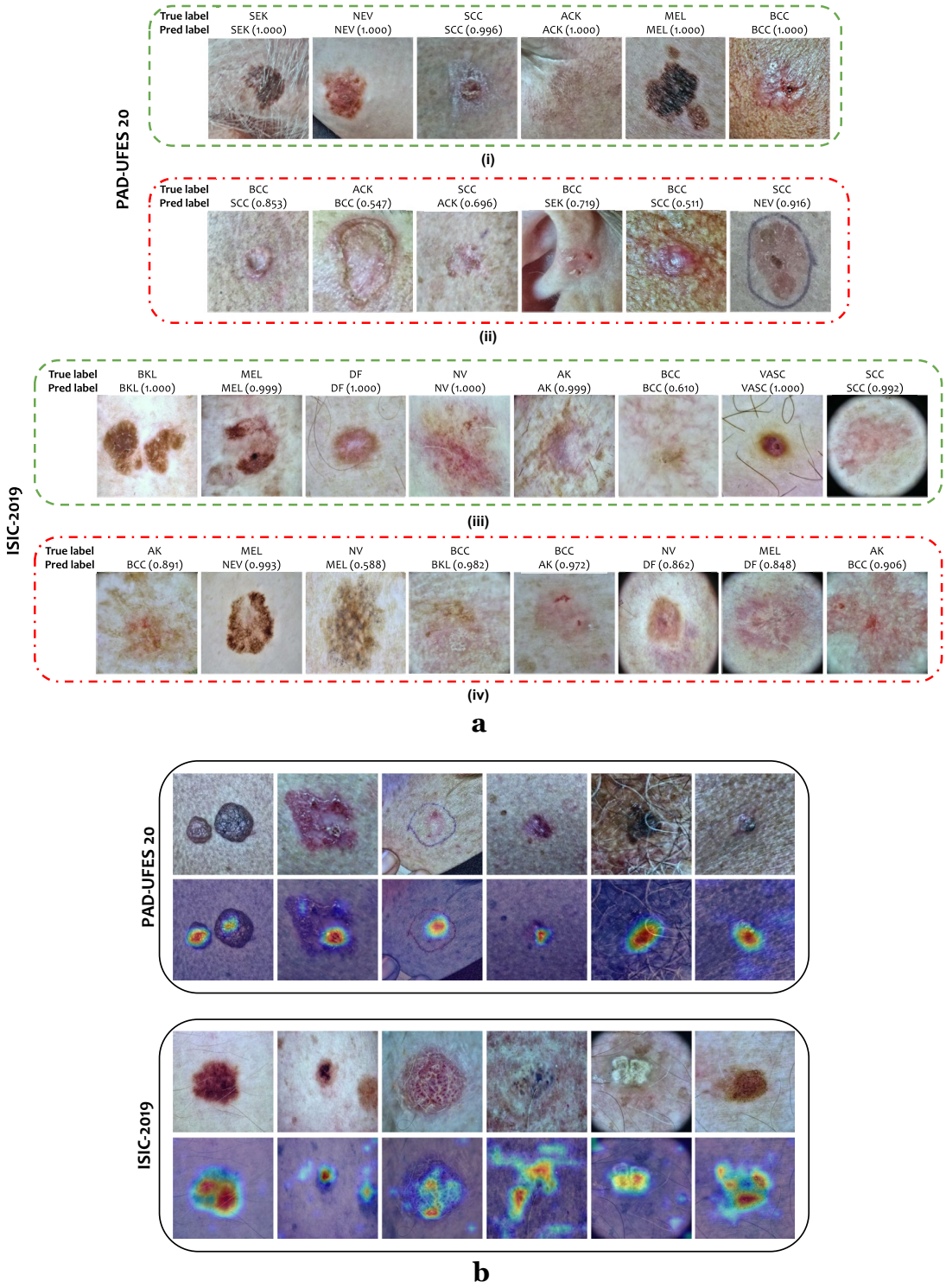


Fig. 5. Showcasing the (**a**) classification results from the PAD-UFES20[43] and ISIC-2019 datasets[44–46], highlighting cases with the highest prediction probability score. Rows (i) and (ii) showcase samples from PAD-UFES20, while rows (iii) and (iv) show results from ISIC-2019. The visualization highlights a key insight into misclassification patterns that the errors mostly occur when lesions share visual similarities with other diagnostic categories, making them particularly challenging to distinguish based on the appearance alone. (**b**) Visualization of the best-model attention using GradCAM for PAD-UFES20 and ISIC-2019 datasets. The heatmaps highlight important regions for model predictions, where red indicates higher importance and blue indicates lower importance. The top row is the original inputs, and the bottom row presents the GradCAM activation maps overlaid on the original images, demonstrating that the model focuses primarily on the regions affected for classification.

predictions. Figure 6 presents the SHAP summary plot of the PAD-UFES20 and ISIC-2019 datasets, highlighting the key features affecting the predictions of the lesion classes.

For the PAD-UFES20 dataset, the following points summarise the key insights of the metadata features.

1. norm_D2 (Normalized secondary diameter of the lesion) and norm_age (normalized age) are the most influential metadata features, contributing significantly to predictions for multiple classes, particularly classes SEK and NEV.
2. Features such as background_mother_POMERANIA, elevation_0 (elevated lesion), and norm_D1 (Normalized primary diameter of the lesion) also show notable contributions, reflecting the ability of the model to learn from continuous and categorical metadata.
3. The stacked SHAP values allow us to observe the class-specific contributions of each feature. For instance, norm_D2 is most associated with Class SEK predictions, whereas norm_age influences Classes ACK and NEV more prominently.
4. Less influential features, such as bleed_1, pesticide_0, and has_piped_water_1, exhibit relatively minor impact, suggesting the model assigns lower importance to these variables.

The following points highlight the key insights of the metadata features for the ISIC-2019 dataset.

1. norm_age (normalized age) is the most influential metadata feature, with consistently high SHAP values across multiple lesion classes, particularly NV, BKL, and SCC. This highlights the strong correlation between patient age and lesion type.
2. Anatomical location features such as anatom_site_general_anterior torso, lower extremity, and upper extremity contribute notably to the model's output. Their class-specific contributions highlight that the model uses regional lesion distribution patterns effectively.
3. Sex-related features (sex_male, sex_female) provide moderate predictive value, with visible influence across classes like MEL and BCC, indicating gender-related variation in lesion occurrence.
4. The stacked bar representation of SHAP values enables the interpretation of the importance of class-wise features. For example, the anatom_site_general_anterior torso shows greater relevance for class AK, while norm_age is more impactful for NV and SCC.
5. Features such as anatom_site_general_palms/soles, oral/genital, and lateral torso show minimal SHAP values, revealing limited utility in class discrimination within the dataset.

Comparisons with Transformer-based Approaches: We evaluated our approach against several transformer backbones—including the Swin transformer (swin_base_patch4_window7_224) [56], the Vision transformer (vit_base_patch16_224) [57] and the Convolutional vision transformer (convit_base) [58]—on the PAD-UFES20 and ISIC-2019 datasets. As shown in Fig. 7, our method outperforms these transformer models on PAD-UFES20,

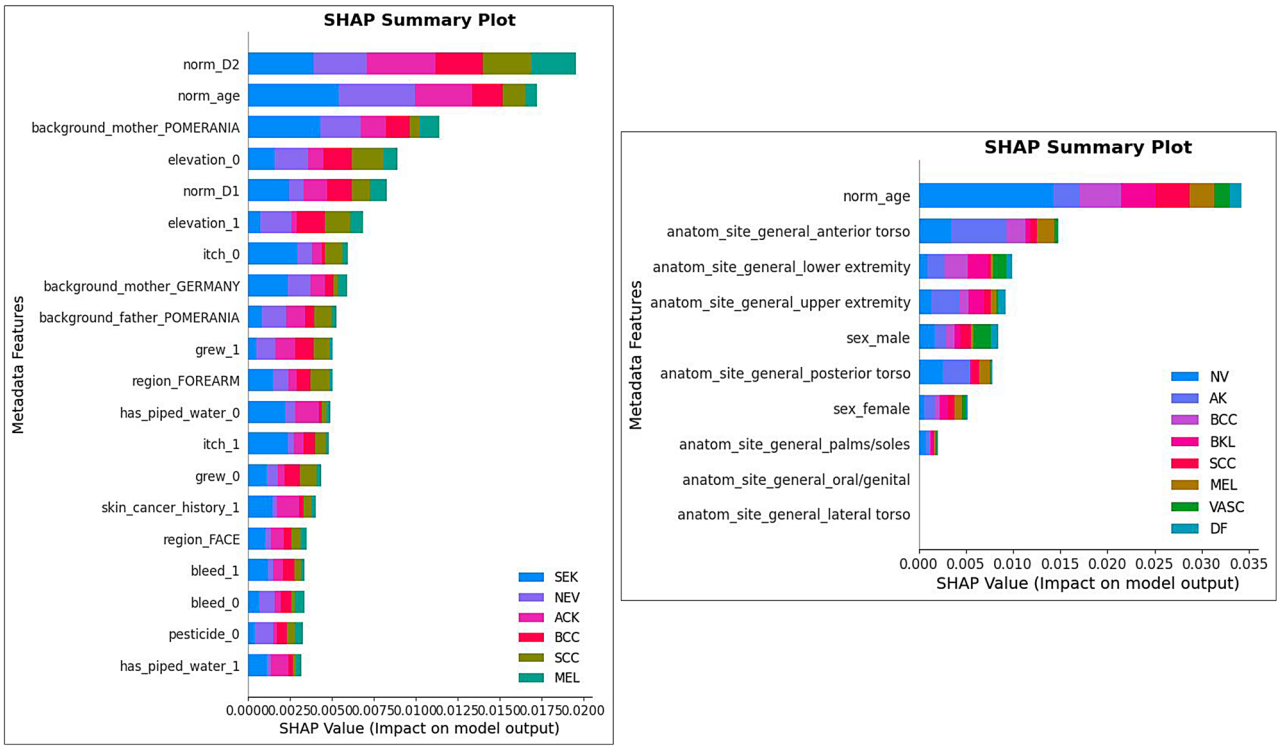


Fig. 6. SHAP summary plot showcasing the impact of metadata features on the model predictions on the PAD-UFES20 (left) [43] ISIC-2019 (right) [44–46] datasets.

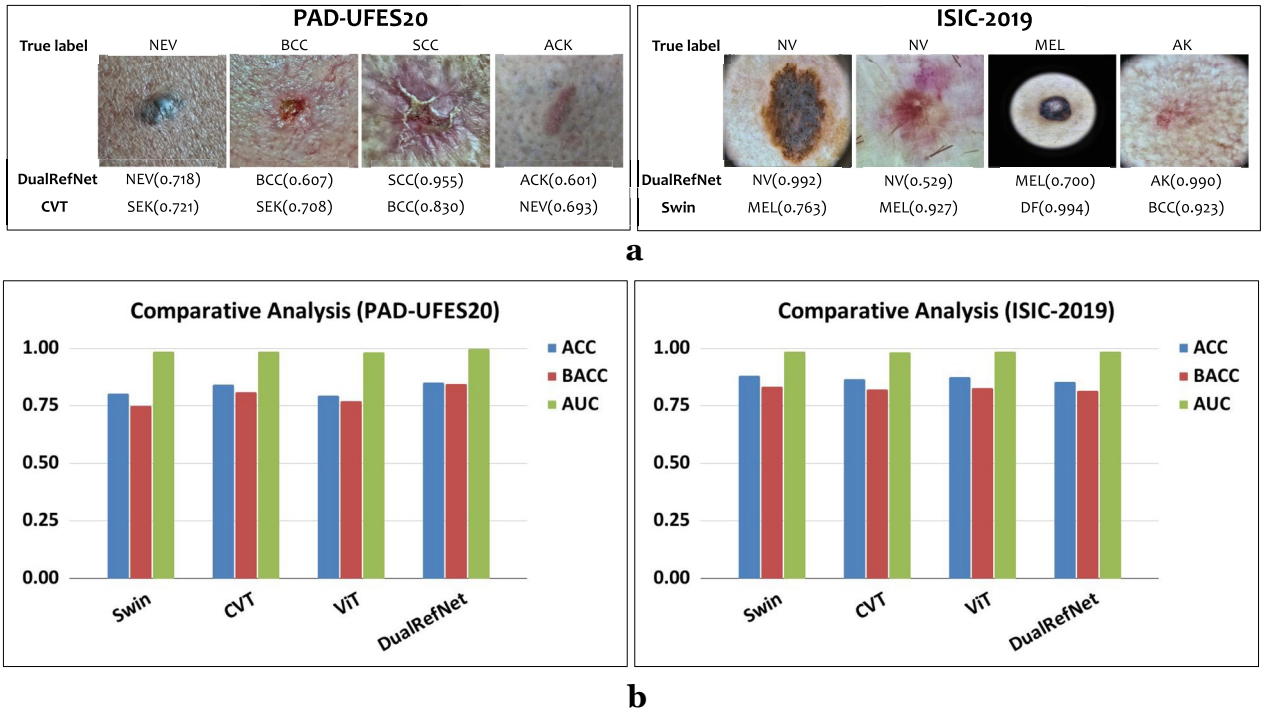


Fig. 7. Illustration of (**a**) correct and incorrect classifications with the probability values by the proposed DualRefNet and the best Transformer model on PAD-UFES20 [43] and ISIC-2019 [44–46], and (**b**) a comparative analysis of both datasets in terms of classification accuracy (ACC), balanced accuracy (BACC), and area under the ROC curve (AUC). For PAD-UFES20, the top row shows predicted labels correctly classified by DualRefNet's best model, while the bottom row displays the classes predicted for the same samples misclassified by the Convolutional Vision Transformer (CVT). For ISIC-2019, the top row presents predicted classes correctly classified by DualRefNet, whereas the bottom row highlights the predicted values of the same images misclassified by the Swin Transformer.

primarily due to the enriched metadata feature set and dual-stage feature refinement that collectively enhance classification accuracy. The metadata provides complementary lesion information, further boosting the system's performance. Conversely, on the ISIC-2019 dataset-which provides only three metadata features-transformer-based models generally perform on par with, or slightly better than, CNN-based approaches. Moreover, as illustrated in Fig. 7(i), the Swin Transformer occasionally misclassifies visually dissimilar images belonging to the same class. By contrast, our approach correctly classifies these challenging samples, owing to its dual-stage refinement mechanism that effectively integrates features beyond mere visual similarity.

Ablation study

To evaluate the individual contributions of each component in the proposed approach, we conducted an extensive ablation study under consistent hyperparameter settings across the PAD-UFES20 and ISIC-2019 datasets. This ablation study is done to validate the effectiveness of each design choice, ranging from metadata integration to adding feature refinement. We also explored fusion strategies and feature weighting, ensuring that every component meaningfully contributes to the balanced and generalizable performance of the model.

Figure 1e presents the ablation study results for the best-performing model under consistent hyperparameter settings across both PAD-UFES20 and ISIC-2019. These findings confirm that each component of *DualRefNet* contributes meaningfully to overall performance. Notably, incorporating metadata leads to substantial gains on both datasets, mirroring real-world diagnostic workflows that combine image observations with clinical details. Further integrating the auxiliary super-resolution task and class-frequency-based regularization (i.e., the dual-stage pipeline) produces marked improvements in balanced accuracy, particularly on PAD-UFES20. We attribute this to PAD-UFES20's richer metadata, which provides a more comprehensive clinical context, ultimately enhancing feature representation and generalizability.

Figure 1e also includes a table comparing various super-resolution techniques, showing that *DualRefNet* consistently achieves higher balanced accuracy-a robust metric for imbalanced datasets-than existing methods. This highlights the model's ability to handle minority classes effectively while maintaining strong overall classification metrics. Collectively, the ablation study demonstrates that each design element of *DualRefNet*, from metadata utilization to super-resolution guidance and frequency-based weight regularization, is integral to achieving accurate, balanced, and clinically meaningful performance on diverse skin lesion classification tasks. In addition, we explored the dynamic weighting of metadata features on the ISIC-2019 and PAD-UFES20 datasets using an attention-based feature weighting mechanism within the metadata encoder of the best-performing

model. Specifically, we used a two-layer neural network with a sigmoid activation to assign learnable weights to each metadata feature. The performance with dynamic weighting on the ISIC-2019 dataset, which contains only three metadata features, was comparable to that with equal weighting. We observed a balanced accuracy (BACC) of 81.07% using dynamic weighting, close to the 81.5% achieved with equal feature weighting. This similarity may be due to the lower risk of overfitting and the relatively compact nature of the metadata. In contrast, the PAD-UFES20 dataset includes many metadata features, some of which may be less informative. In this case, the attention mechanism struggled to learn optimal feature weights, resulting in diminished performance. We observed a BACC of 71.34% with dynamic weighting, which is notably lower than the 84.5% achieved using equal weighting. This observation highlights that while dynamic weighting is effective, its benefits are influenced by the quality and dimensionality of the metadata.

We also experimented by replacing the original feed-forward metadata encoder with a transformer encoder. However, we observed a lower performance than the proposed setting by getting a BACC of 80.11%, lower than 84.5% with the feed-forward network. This outcome can be attributed to the nature of the available metadata, which consists of categorical and non-sequential features, such as age, gender, origin of the mother and father, and presence of bleeding. Such features lack inherent temporal or positional relationships and thus do not benefit from the inductive biases that transformer architectures offer for modelling sequential or contextually dependent data. Transformer-based encoders have demonstrated effectiveness in cases where metadata exhibits rich inter-feature dependencies or follows an ordered structure. However, the available metadata is low-dimensional and unordered, making transformer-based modelling less suitable. We also explored replacing the fusion strategy with cross-attention via multi-head attention, with four heads and an embedding dimension of 512, to align metadata and image features. However, empirical results did not show performance improvements. A BACC of 68.99% and 16.27% was achieved for the PAD-UFES20 and ISIC-2019 datasets, respectively. Attention-based fusion introduced additional parameters and computational overhead without yielding gains in accuracy. This could be due to the relatively low dimensionality and sparsity of the metadata, which may not benefit significantly from the complexity of attention mechanisms.

Discussion

The proposed *DualRefNet* introduces a significant advancement in handling both class imbalance and lesion variability, as evidenced by its strong balanced accuracy scores of 0.845 (PAD-UFES20) and 0.815 (ISIC-2019). By achieving robust performance on both smartphone-captured and dermoscopic images, *DualRefNet* demonstrates its suitability for real-world, often unconstrained settings. This improvement arises from a dual-stage feature refinement strategy wherein the model progressively enhances visual and fused features through an auxiliary super-resolution task and class-frequency-based weight regularization. Unlike earlier methods that rely on early fusion or modality-specific architectures, *DualRefNet* refines both visual and fused representations, thereby minimizing misclassifications linked to low inter-class and high intra-class variations. Additionally, the use of a weighted cross-entropy loss aligns the training process towards minority classes, ensuring balanced performance across all lesion categories.

Tables 3a,b compare *DualRefNet* with leading methods, highlighting the value of integrating metadata into the classification pipeline. This holistic perspective on a patient's condition enhances diagnostic effectiveness, resulting in superior class separability, high AUC values, and consistently elevated balanced accuracy. Whereas previous approaches often relied on feature concatenation[26,38] or visual attention mechanisms[32–34], their efficacy diminished for lesions exhibiting close visual similarity. By contrast, *DualRefNet*'s dual-stage refinement, encompassing an auxiliary task and class-frequency-based weight adjustments, mitigates class overlap and reliably reduces misclassification of lesions such as SCC. While the PAD-UFES20 dataset benefits notably from richer metadata, the ISIC-2019 dataset also sees considerable gains, despite having fewer clinical features.

Quantitative metrics (e.g., specificity, sensitivity, class-wise accuracy) further reinforce *DualRefNet*'s balanced performance. Specificity exceeds 90% for most classes, while sensitivity surpasses 80% in nearly all categories; the SCC class in PAD-UFES20 is a mild exception, although its lower score remains clinically manageable since pigmented lesions typically undergo confirmatory biopsy. Qualitative insights from t-SNE embeddings and confusion matrices also validate improved class separability, demonstrating the model's robustness in tackling visually similar lesions. Meanwhile, high PPV and NPV-often above 0.87 and consistently high, respectively-highlight reliable positive and negative classification. The Kappa statistic exceeds 0.79 for both datasets, signifying substantial agreement with ground truth. These strong clinical metrics, coupled with overall accuracies near 85%, showcase *DualRefNet*'s reliability for assisting in practical diagnostic use.

Although smartphone-based images generally lack the resolution of traditional dermoscopic imaging, our results reveal *DualRefNet*'s ability to adapt through its dual-stage refinement pipeline. By enhancing both visual features and fused representations, the model maintains high specificity, sensitivity, and class-wise accuracy, making it particularly suitable for resource-limited settings where specialized imaging equipment may be scarce. This capability suggests wide-ranging applications in teledermatology, enabling earlier screenings and more equitable patient care through smartphone-based submissions.

In conclusion, we propose *DualRefNet*, a multimodal, dual-stage feature refinement framework that effectively addresses class imbalance and lesion variability by jointly optimizing visual and fused features. Our approach employs weighted cross-entropy loss to manage class imbalance, while t-SNE visualizations confirm improved class separability. Comprehensive evaluations of smartphone-captured and dermoscopic images demonstrate the versatility and practicality of the proposed approach in unconstrained environments. *DualRefNet* consistently achieves robust performance across multiple CNN backbones and datasets, highlighting its potential for reliable assistance in clinical diagnosis. The effectiveness of the proposed approach is notable when processing rich metadata features, resulting in enhanced diagnostic accuracy. In scenarios where metadata is sparse or lacks sufficient descriptive information, the impact of the second-stage refinement may be reduced. Given its proven

success with smartphone images and ability to effectively use available metadata, *DualRefNet* shows particular promise for advancing early detection and treatment strategies in resource-constrained healthcare environments, mainly benefiting regions with limited dermatological expertise. Future research includes developing adaptive mechanisms to maintain robust performance across varying levels of metadata availability. This work represents a significant step forward in making reliable skin cancer diagnosis more accessible while maintaining high diagnostic standards.

Methods

Figure 2 outlines the overall architecture of our proposed *DualRefNet*. Central to this research is the concept of refining both visual and fused features through two complementary mechanisms: an auxiliary supervision task and class-frequency-based weight regularization. The auxiliary task encourages the extraction of more fine-grained, lesion-specific information, thereby enhancing the primary classification task. Simultaneously, class-frequency-based weight regularization ensures balanced treatment of underrepresented classes, mitigating bias toward majority classes.

Given a set of skin lesion images X_{img} and corresponding metadata X_{meta} (e.g., demographic and clinical attributes), along with class labels $y \in [1, \ldots, N_{\text{class}}]$, the objective is to train a model that accurately predicts y for unseen test data $(X_{img\text{-test}}, X_{meta\text{-test}})$.

Preprocessing

To minimize the impact of illumination differences and uneven contrast, each image undergoes a gray color constancy algorithm [59] followed by contrast-limited adaptive histogram equalization (CLAHE)[60,61]. These steps improve color uniformity and enhance contrast. For metadata, the PAD-UFES20 dataset uses a neural network-based imputation technique[28] to fill missing entries, while the ISIC-2019 dataset excludes samples lacking metadata to prevent potential biases given the limited number of metadata features.

Feature extraction

Visual feature extractor: We employ various CNN backbones for extracting features from input images. These architectures differ in breadth, depth, and connectivity (e.g., residual connections), capturing both low-level edges/textures and high-level lesion patterns. Denoting the visual feature extractor by ϕ_{visual}, its output for an input image x_{img} is given by

$$FV_{\text{Image}} = \phi_{\text{visual}}(x_{\text{img}}). \tag{1}$$

Textual feature extractor: To process patient metadata (e.g., age, gender, lesion history), we use a four-layer fully connected network with ReLU activations after each layer except the final output layer. This yields a learned vector representation:

$$FV_{\text{Meta}} = \phi_{\text{meta}}(x_{\text{meta}}). \tag{2}$$

Combining this textual embedding with the visual embedding forms the basis of our multimodal fusion strategy.

Auxiliary task: super-resolution

To address the challenge of visually similar lesions, we introduce a super-resolution (SR) auxiliary task. Specifically, we generate higher-resolution images from the CNN-encoded feature maps. Experimental variations include bilinear/bicubic interpolation as well as deep-learning-based SR methods[53,54]. By learning to predict SR images, the encoder is guided to capture finer lesion details. This auxiliary branch shares the encoder's parameters with the primary classification task, effectively serving as a regularizer that encourages more discriminative representation learning.

After the encoder extracts visual features, these features are passed to both a classification head and a decoder. The decoder is a 6-layer CNN whose goal is to reconstruct the SR image from the latent representation. Training jointly for classification and SR prediction helps the model focus on nuances that can differentiate lesions.

Feature-level fusion and classification

Once we obtain FV_{Image} from the visual backbone and FV_{Meta} from the metadata network, these are concatenated or otherwise combined into a fused feature vector:

$$FV_{\text{Fused}} = [FV_{\text{Image}} \,\|\, FV_{\text{Meta}}]. \tag{3}$$

This fused representation then proceeds through the final fully connected layers for classification. We tackle class imbalance and low inter-class variability through two key strategies:

Class weight-based feature refinement

Because certain lesion classes are underrepresented, the model may overfit to majority classes. To mitigate this, we introduce class-frequency-based regularization in the last two fully connected layers. Each class is assigned a weight inversely proportional to its frequency, placing stronger penalties on lower-sampled classes and thus encouraging the network to focus more on minority classes. Formally, if $W_{\text{beforeclassifier},i}$ and $W_{\text{classifier},j}$ represent the weight parameters in these layers, the regularization term R is defined as:

$$R = \lambda \left(\sum_{i=1}^{256} f_i \| W_{\text{beforeclassifier},i} \|_1 + \sum_{j=1}^{N_{\text{class}}} f_j \| W_{\text{classifier},j} \|_1 \right), \tag{4}$$

where λ scales the overall strength of the regularization, and f_i, f_j are class-specific penalty coefficients inversely proportional to class frequency. Empirically, we set $\lambda = 0.3$ for PAD-UFES20 and $\lambda = 0.2$ for ISIC-2019.

Loss function
The final training loss L_{DualRef} combines classification loss, super-resolution (auxiliary) loss, and class-frequency-based regularization:

- Classification loss We use a weighted cross-entropy to address class imbalance more directly. For each class *i*, let w_i be the class weight inversely proportional to the sample count:

$$L_{\text{class}} = -\sum_{i=1}^{N} w_{y_i} \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} \tag{5}$$

- Auxiliary loss The auxiliary (SR) loss measures the difference between the predicted super-resolved image $\hat{SR}_i$ and its ground-truth SR_i, using mean squared error:

$$L_{\text{aux}} = \frac{1}{n} \sum_{i=1}^{n} (SR_i - \hat{SR}_i)^2. \tag{6}$$

- Regularization term As defined above, *R* penalizes weights to improve class balance in the final layers.

Thus, the overall loss is given by:

$$L_{\text{DualRef}} = \alpha\, L_{\text{class}} + \beta\, L_{\text{aux}} + \gamma\, R, \tag{7}$$

where $\alpha = 0.5$, $\beta = 1.0$, and $\gamma = 1.0$ balance the relative contributions of classification, auxiliary supervision, and regularization.

By integrating these components-auxiliary super-resolution for finer feature extraction and class-weight-based regularization for balanced classification-*DualRefNet* is designed to effectively handle both the data imbalance and the subtle inter-class similarities often observed in skin lesion classification.

Received: 20 January 2025; Accepted: 4 August 2025
Published online: 29 October 2025

References

1. WHO. Skin cancer. https://www.iarc.who.int/news-events/melanoma-awareness-month-2022/ (2022).
2. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature***542**, 115–118 (2017).
3. Shetty, B. et al. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Sci. Rep.* **12**, 18134 (2022).
4. Camacho-Gutiérrez, J. A., Solorza-Calderón, S. & Álvarez-Borrego, J. Multi-class skin lesion classification using prism-and segmentation-based fractal signatures. *Expert Syst. Appl.* **197**, 116671 (2022).
5. Saleh, N., Hassan, M. A. & Salaheldin, A. M. Skin cancer classification based on an optimized convolutional neural network and multicriteria decision-making. *Sci. Rep.* **14**, 17323 (2024).
6. Murzaku, E. C., Hayan, S. & Rao, B. K. Methods and rates of dermoscopy usage: a cross-sectional survey of us dermatologists stratified by years in practice. *J. Am. Acad. Dermatol.* **71**, 393–395 (2014).
7. Mahmoud, N. M. & Soliman, A. M. Early automated detection system for skin cancer diagnosis using artificial intelligent techniques. *Sci. Rep.* **14**, 9749 (2024).
8. Umirzakova, S., Muksimova, S., Baltayev, J. & Cho, Y. I. Force map-enhanced segmentation of a lightweight model for the early detection of cervical cancer. *Diagnostics* **15**, 513 (2025).
9. Shimizu, K., Iyatomi, H., Celebi, M. E., Norton, K.-A. & Tanaka, M. Four-class classification of skin lesions with task decomposition strategy. *IEEE Trans. Biomed. Eng.* **62**, 274–283 (2014).
10. Verdelho, M. R. & Barata, C. On the impact of self-supervised learning in skin cancer diagnosis. In *International Symposium on Biomedical Imaging* 1–5 (2022).
11. Khurshid, M., Vatsa, M. & Singh, R. Multi-task explainable skin lesion classification. *arXiv preprint* arXiv:2310.07209 (2023).
12. Wang, L., Zhang, L., Shu, X. & Yi, Z. Intra-class consistency and inter-class discrimination feature learning for automatic skin lesion classification. *Med. Image Anal.* **85**, 102746 (2023).
13. Yu, Z. et al. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. *IEEE Trans. Biomed. Eng.* **66**, 1006–1016 (2018).
14. Ali, M. S., Miah, M. S., Haque, J., Rahman, M. M. & Islam, M. K. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Mach. Learn. Appl.* **5**, 100036 (2021).
15. Sulthana, R., Chamola, V., Hussain, Z., Albalwy, F. & Hussain, A. A novel end-to-end deep convolutional neural network based skin lesion classification framework. *Expert Syst. Appl.* **246**, 123056 (2024).
16. Gessert, N. et al. Skin lesion classification using CNNS with patch-based attention and diagnosis-guided loss weighting. *IEEE Trans. Biomed. Eng.* **67**, 495–503 (2019).
17. Zhang, J., Xie, Y., Xia, Y. & Shen, C. Attention residual learning for skin lesion classification. *IEEE Trans. Med. Imaging* **38**, 2092–2103 (2019).

18. Nakai, K., Chen, Y.-W. & Han, X.-H. Enhanced deep bottleneck transformer model for skin lesion classification. *Biomed. Signal Process. Control* **78**, 103997 (2022).
19. Yang, G., Luo, S. & Greer, P. A novel vision transformer model for skin cancer classification. *Neural Process. Lett.* **55**, 9335–9351 (2023).
20. Huang, R. *et al.* Melanomanet: An effective network for melanoma detection. In *International Conference of the IEEE Engineering in Medicine and Biology Society* 1613–1616 (2019).
21. Harangi, B., Baran, A. & Hajdu, A. Classification of skin lesions using an ensemble of deep neural networks. In *IEEE Engineering in Medicine and Biology Society* 2575–2578 (2018).
22. Tang, P., Liang, Q., Yan, X., Xiang, S. & Zhang, D. Gp-cnn-dtel: Global-part cnn model with data-transformed ensemble learning for skin lesion classification. *IEEE J. Biomed. Health Inform.* **24**, 2870–2882 (2020).
23. He, X. et al. Fully transformer network for skin lesion analysis. *Med. Image Anal.* **77**, 102357 (2022).
24. Islam, S. et al. Leveraging AI and patient metadata to develop a novel risk score for skin cancer detection. *Sci. Rep.* **14**, 20842 (2024).
25. Cai, G. et al. A multimodal transformer to fuse images and metadata for skin disease classification. *Vis. Comput.* **39**, 2781–2793 (2023).
26. Pacheco, A. G. & Krohling, R. A. The impact of patient clinical information on automated skin cancer detection. *Comput. Biol. Med.* **116**, 103545 (2020).
27. Tang, P. et al. Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. *Med. Image Anal.* **76**, 102307 (2022).
28. Pundhir, A., Dadhich, S., Agarwal, A. & Raman, B. Towards improved skin lesion classification using metadata supervision. In *International Conference on Pattern Recognition* 4313–4320 (2022).
29. Pham, N. L. T., Pham, D. D., Le, T. D. & Huynh, K. T. A multimodal deep ensemble framework for skin lesion classification. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making* 100–111 (2025).
30. Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE J. Biomed. Health Inform.* **23**, 538–546 (2018).
31. Lin, Q. et al. A novel multi-task learning network for skin lesion classification based on multi-modal clues and label-level fusion. *Comput. Biol. Med.* **175**, 108549 (2024).
32. Khurshid, M., Vatsa, M. & Singh, R. Optimizing skin lesion classification via multimodal data and auxiliary task integration. In *IEEE International Symposium on Biomedical Imaging* 1–5 (2024).
33. Pacheco, A. G. & Krohling, R. A. An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. *IEEE J. Biomed. Health Inform.* **25**, 3554–3563 (2021).
34. Pundhir, A., Agarwal, A., Dadhich, S. & Raman, B. Visually aware metadata-guided supervision for improved skin lesion classification using deep learning. *Ethic. Philos. Issues Med. Imaging* **13755**, 65–76 (2022).
35. Vachmanus, S., Noraset, T., Piyanonpong, W., Rattananukrom, T. & Tuarob, S. Deepmetaforge: A deep vision-transformer metadata-fusion network for automatic skin lesion classification. *IEEE Access* **11**, 145467–145484 (2023).
36. Bi, L., Feng, D. D., Fulham, M. & Kim, J. Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network. *Pattern Recogn.* **107**, 107502 (2020).
37. Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A. & Garnavi, R. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In *Medical Image Computing and Computer Assisted Intervention*, 250–258 (2017).
38. Li, W., Zhuang, J., Wang, R., Zhang, J. & Zheng, W.-S. Fusing metadata and dermoscopy images for skin disease diagnosis. In *International Symposium on Biomedical Imaging*, 1996–2000 (2020).
39. Tang, P. et al. Joint-individual fusion structure with fusion attention module for multi-modal skin cancer classification. *Pattern Recogn.* **154**, 110604 (2024).
40. He, X., Zhou, Y., Wang, B., Cui, S. & Shao, L. Dme-net: Diabetic macular edema grading by auxiliary task learning. In *Medical Image Computing and Computer Assisted Intervention*, 788–796 (2019).
41. Liu, L., Tsui, Y. Y. & Mandal, M. Skin lesion segmentation using deep learning with auxiliary task. *J. Imaging* **7**, 67 (2021).
42. Chen, H., Wang, X., Guan, C., Liu, Y. & Zhu, W. Auxiliary learning with joint task and data scheduling. In *International Conference on Machine Learning*, 3634–3647 (2022).
43. Pacheco, A. G. et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data Brief* **32**, 106221 (2020).
44. Gutman, D. *et al.* Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint* arXiv:1605.01397 (2016).
45. Tschandl, P., Rosendahl, C. & Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 1–9 (2018).
46. Codella, N. C. *et al.* Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *International Symposium on Biomedical Imaging*, 168–172 (2018).
47. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations* (2015).
48. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
49. Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A. & Chen, L. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520 (2018).
50. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114 (2019).
51. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *IEEE conference on Computer Vision and Pattern Recognition*, 4700–4708 (2017).
52. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009).
53. Gouvine, G. Ninasr: Efficient small and large convnets for super-resolution. https://github.com/Coloquinte/torchSR/blob/main/doc/NinaSR.md, https://doi.org/10.5281/zenodo.4868308 (2021).
54. Lim, B., Son, S., Kim, H., Nah, S. & Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 136–144 (2017).
55. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 618–626 (2017).
56. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision*, 10012–10022 (2021).
57. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations* (2021).
58. d'Ascoli, S. *et al.* Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, 2286–2296 (2021).
59. Finlayson, G. D. & Trezzi, E. Shades of gray and colour constancy. *Color Imaging Conf.* **2004**, 37–41 (2004).

60. Pizer, S. M. et al. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **39**, 355–368 (1987).
61. Jaisakthi, S. M., Mirunalini, P. & Aravindan, C. Automated skin lesion segmentation of dermoscopic images using grabcut and k-means algorithms. *IET Comput. Vision* **12**, 1088–1095 (2018).

Author contributions
All the authors contributed to the problem formulation. M.K. conceived the methodology and conducted the experiments. All the authors analyzed the results and contributed towards drafting, reviewing, and editing the manuscript.

Funding
Vatsa and Singh have been partly supported by the Srijan Center for Generative AI at IIT Jodhpur, supported by IndiaAI Mission and Meta.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.