

<https://doi.org/10.1038/s41746-026-02897-8>

Attention guided fair artificial intelligence modeling for skin cancer diagnosis

Check for updates

Mingcheng Zhu^{1,2,3,7}, Mingxuan Liu^{1,2,7}, Han Yuan^{1,2}, Yilin Ning^{1,2}, Zhiyao Luo³, Tingting Zhu³ & Nan Liu^{1,2,4,5,6} ✉

Artificial intelligence (AI) has shown promise in dermatology, offering accurate and non-invasive diagnosis of skin cancer. While extensive research has addressed skin-tone bias, gender bias in dermatologic AI remains underexplored, potentially perpetuating diagnostic disparities. In this study, we developed LesionAttn, an algorithm designed to mitigate gender bias by directing model attention toward lesions, thereby mirroring clinicians' diagnostic focus. Combined with Pareto Frontier optimization for dual-objective model selection, LesionAttn balances gender fairness and diagnostic performance. Validated on two large-scale dermatologic datasets for binary malignancy classification, LesionAttn significantly mitigated gender bias while maintaining high diagnostic performance, outperforming existing bias-mitigation algorithms. Our study demonstrates that explicitly guiding model attention to medically essential features provides a practical approach to advance both performance and fairness in dermatologic AI. By leveraging clinical priors to bridge the gap between human expertise and algorithmic optimization, this study demonstrates a feasible pathway for developing equitable and reliable diagnostic tools.

Skin cancer poses a major global health concern, with more than 6.7 million malignant cases reported annually worldwide. Its incidence has risen sharply over the past two decades^{1,2}, highlighting the need for efficient and accurate diagnostic tools. Conventional diagnosis often relies on invasive biopsies that are time-consuming and uncomfortable for patients³. In response, Artificial Intelligence (AI) has emerged as a transformative solution, enabling non-invasive skin cancer diagnosis directly from dermoscopic images⁴. These models now achieve diagnostic accuracy comparable to or exceeding that of expert dermatologists⁵ and are increasingly integrated into clinical practice^{6,7}.

Despite their promise of accurate diagnosis, dermatologic AI models remain vulnerable to algorithmic bias, raising concerns about unequal care across patient subgroups. Prior research has primarily focused on disparities across racial and skin tone groups, showing that models trained predominantly on lighter skin often underperform on darker tones^{8–13}. These skin tone-related biases risk exacerbating existing inequities in dermatologic outcomes¹⁴. In contrast, gender bias in AI-based skin cancer diagnosis remains relatively underexplored, perhaps because gender-specific differences, unlike visible variations in skin tone, are more subtle and less readily recognized^{15,16}.

Gender disparities in skin cancer diagnosis may stem from both biological and behavioral differences. These include variations in skin texture, hormonal and vascular characteristics^{17,18}, as well as gendered patterns in sun protection and skin care practices^{19,20}. Such factors could lead to uneven performance of AI models across genders, perpetuating diagnostic inequities. Specifically, models may learn to rely on these gender-related features present in the background skin as shortcuts, rather than focusing on the disease-specific features²¹. Yet, few studies have systematically investigated and mitigated gender bias in dermatologic AI.

The effectiveness of existing bias mitigation algorithms in skin cancer diagnosis remains uncertain. Most were developed and validated on non-medical datasets and may not generalize to dermatologic image analysis^{9,14,22}, where domain-specific challenges such as variable lighting, occlusion, and ill-defined lesion boundaries complicate bias mitigation²³. They also usually lack interpretability, modifying model parameters in ways that are difficult to trace or explain, which limits clinical acceptance and trust²⁴. Moreover, many achieve fairness at the expense of diagnostic accuracy (for example, by enforcing similarity in latent feature distributions across demographic subgroups), thereby discarding clinically meaningful information^{25–27}. Such trade-offs pose

¹Centre for Biomedical Data Science, Duke-NUS Medical School, Singapore, Singapore. ²Duke-NUS AI + Medical Sciences Initiative, Duke-NUS Medical School, Singapore, Singapore. ³Department of Engineering Science, University of Oxford, Oxford, UK. ⁴Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA. ⁵Pre-hospital & Emergency Research Centre, Health Services Research & Population Health, Duke-NUS Medical School, Singapore, Singapore. ⁶NUS Artificial Intelligence Institute, National University of Singapore, Singapore, Singapore. ⁷These authors contributed equally: Mingcheng Zhu, Mingxuan Liu.

✉ e-mail: liu.nan@duke-nus.edu.sg

a major barrier to the clinical translation of bias mitigation techniques in dermatologic AI.

To address this gap, we developed LesionAttn, a gender-equity-aware algorithm designed to direct neural network attention specifically toward clinically relevant lesion regions. This approach leverages a learnable attention mechanism that dynamically assigns spatial weights to input images, thereby prioritizing diagnostic features while actively suppressing background noise that may encode gender-related bias. In addition, we employ Pareto Frontier optimization during model training to balance diagnostic performance and fairness. By utilizing clinician-annotated lesion masks to guide spatial attention during model design, LesionAttn reduces reliance on spurious background features and mitigates gender bias without sacrificing performance. These findings offer an interpretable and clinically grounded approach to mitigating gender bias in dermatologic AI and underscore the potential of using clinical priors like annotation to advance both fairness and reliability in clinical settings.

Results

In this study, we proposed LesionAttn, a gender-equity-aware algorithm that guides attention toward clinically relevant lesion areas while optimizing both performance and fairness via Pareto Frontier optimization (Fig. 1). Further details regarding the LesionAttn algorithm are elaborated in the Methods section. LesionAttn was evaluated against existing fairness algorithms across two large-scale public dermatologic datasets—the Human Against Machine dataset (HAM)²⁸ and the BCN20000 dataset (BCN)²⁹ for the binary classification of malignant versus benign skin lesions. We further compared LesionAttn with a static lesion-masking approach, highlighting the importance of maintaining spatial context for both diagnostic performance and fairness.

Throughout this study, gender disparity (bias) was primarily quantified using equalized odds (EO), defined as the larger disparity between groups in either true positive rates (TPR) or false positive rates (FPR). We also denoted the difference in TPR between male and female groups as $EO[TP]$, and the difference in FPR as $EO[FP]$. Model performance was evaluated using the

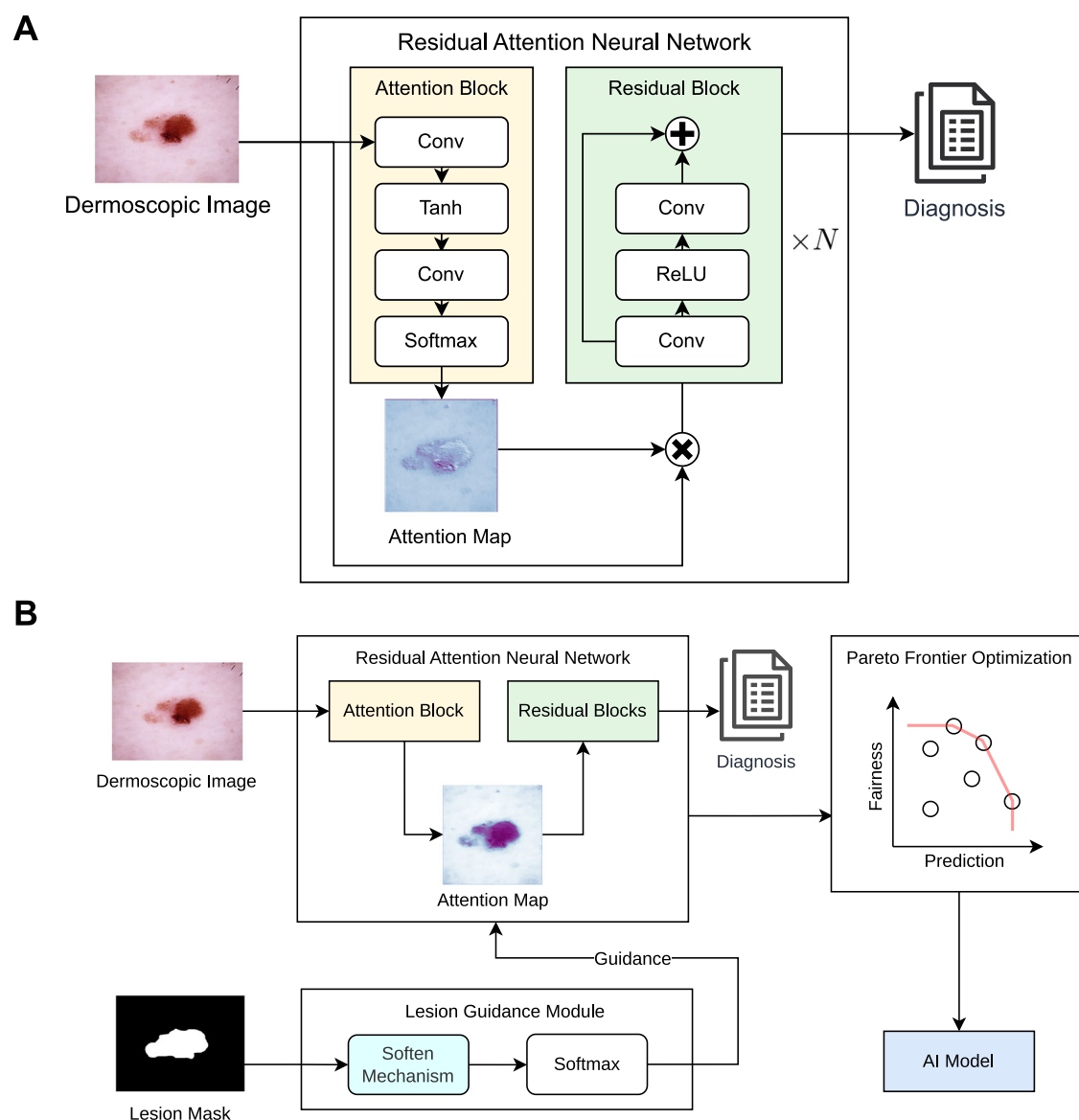


Fig. 1 | Illustration of the AI-based skin cancer diagnosis and the LesionAttn algorithm. **A** The detailed architecture of the Residual Attention Neural Network (RANN), including an attention block for generating attention maps and multiple residual blocks for inference. **B** The workflow of the LesionAttn. The Soften Attention Guidance Module guides the attention map during training. This module

encourages the AI model to focus on lesions by increasing the cosine similarity between the attention map and the softened lesion mask through the loss function. The Fairness-aware Selector is responsible for selecting the AI with optimal fairness and diagnostic performance based on the Pareto Frontier optimization.

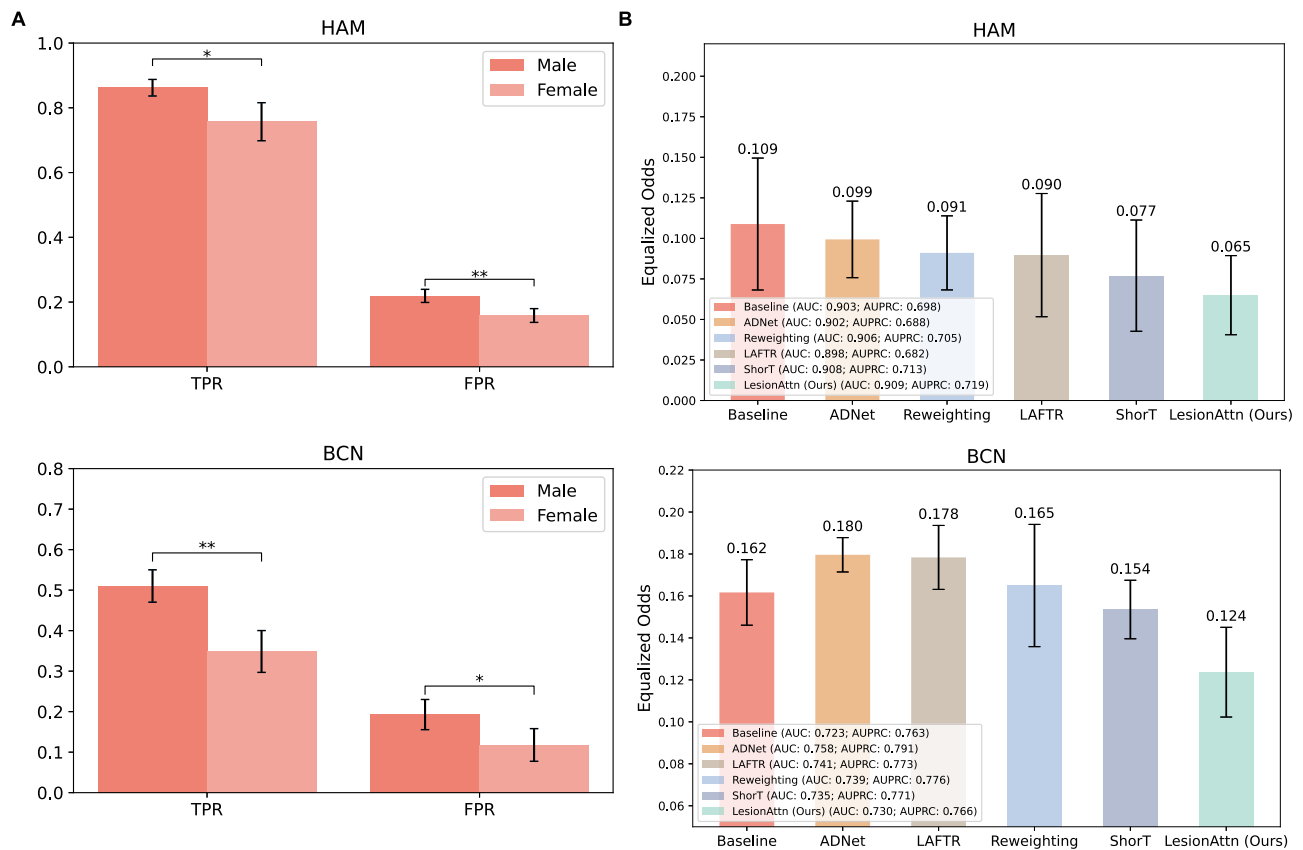


Fig. 2 | Disparity in the baseline model and comparison of five bias mitigation methods across datasets. A Gender disparity of the baseline model on HAM and BCN datasets. The fairness is evaluated by comparing the difference in TPR and FPR

between male and female groups. **B** Comparison of bias mitigation methods. Lower values of Equalized Odds indicate lower gender bias.

areas under the Receiver Operating Characteristic (AUROC) and precision-recall curves (AUPRC) for diagnostic performance and EO for fairness.

Gender bias mitigation using LesionAttn

The Residual Attention Neural Network (RANN) model, based on the widely adopted convolutional neural network (CNN) architecture in skin cancer diagnosis^{30–32}, displayed marked gender bias in both internal (HAM testing set) and external (BCN dataset) cohorts. The RANN model achieved high discrimination performance on HAM (AUROC: 0.903 [0.897–0.909]; AUPRC: 0.698 [0.668–0.727]) and moderate performance on the BCN (AUROC: 0.723 [0.696–0.750]; AUPRC: 0.763 [0.742–0.784]) datasets. The corresponding ROC and PRC curves are provided in Suppl. Fig. 1. However, as shown in Fig. 2A, in both datasets, the male subgroup exhibited a significantly higher TPR and lower TNR (i.e., 1-FPR), with statistical significance compared to the female subgroups.

Critically, the RANN model yielded a substantial clinical discrepancy in the number needed to screen (NNS), defined here as the reciprocal of positive predictive value (PPV). This metric estimates the number of alerted patients the models need to flag to identify one true positive case. In HAM dataset, the discrepancy in NNS was 0.32 (1.16 for female patients vs. 1.32 for male patients), while in BCN data, the discrepancy in NNS was 0.91 (1.96 for female patients vs. 2.87 for male patients).

In the HAM dataset, LesionAttn reduced EO by 40.4% (from 0.109 [0.068–0.150] to 0.065 [0.041–0.089]; $p = 0.041$), while achieving an AUROC of 0.909 [0.904–0.913] and AUPRC of 0.719 [0.696–0.742], higher than the baseline model. In the BCN dataset, LesionAttn reduced EO by 23.5% (from 0.162 [0.146–0.177] to 0.124 [0.102–0.145]; $p = 0.016$), and improved diagnostic performance (AUROC: 0.730 [0.712–0.747]; AUPRC: 0.766 [0.749–0.782]). Notably, LesionAttn mitigated the discrepancy in

NNS, reducing the gender gap from 0.32 to 0.08 in HAM dataset (1.15 for female patients vs. 1.23 for male patients) and from 0.91 to 0.60 in BCN dataset (1.93 for female patients vs. 2.53 for male patients).

We also compared LesionAttn against four established bias mitigation methods: Reweight³³, LAFTR²², ADNet¹⁴, and ShorT⁹, using the same model architecture of RANN. Further implementation details are provided in the Methods section. LesionAttn consistently achieved the highest diagnostic performance and fairness, across both internal (HAM testing set) and external (BCN dataset) cohorts. As shown in Fig. 2B, in the HAM dataset, LesionAttn yielded the numerically lowest mean EO of 0.065 [0.041–0.089], outperforming other bias mitigation methods on average across random seeds (Reweighting: 0.091 [0.068–0.114], $p = 0.016$; LAFTR: 0.090 [0.052–0.128], $p = 0.197$; ADNet: 0.099 [0.076–0.123], $p = 0.086$; ShorT: 0.077 [0.043–0.111], $p = 0.295$), with comparable diagnostic performance. Although the 95% CIs between LesionAttn and bias-mitigation approaches like ShorT overlap on the HAM dataset, LesionAttn consistently demonstrated an average numerical improvement in fairness across independent runs. In the BCN dataset, although all bias mitigation methods achieved better diagnostic performance relative to the baseline, only LesionAttn improved the fairness (LesionAttn: 0.124 [0.102–0.145]; Reweighting: 0.165 [0.136–0.194], $p = 0.016$; LAFTR: 0.178 [0.163–0.194], $p = 0.001$; ADNet: 0.180 [0.171–0.188], $p = 0.008$; ShorT: 0.154 [0.140–0.167], $p = 0.037$) with statistical significance.

Impacts of attention-based training

To assess the effect of attention-guided training, we examined attention maps from LesionAttn, other bias mitigation methods, and the baseline model. Spatial alignment of models' attention and the reference lesion regions was quantified using intersection over union (IoU). As shown in Fig.

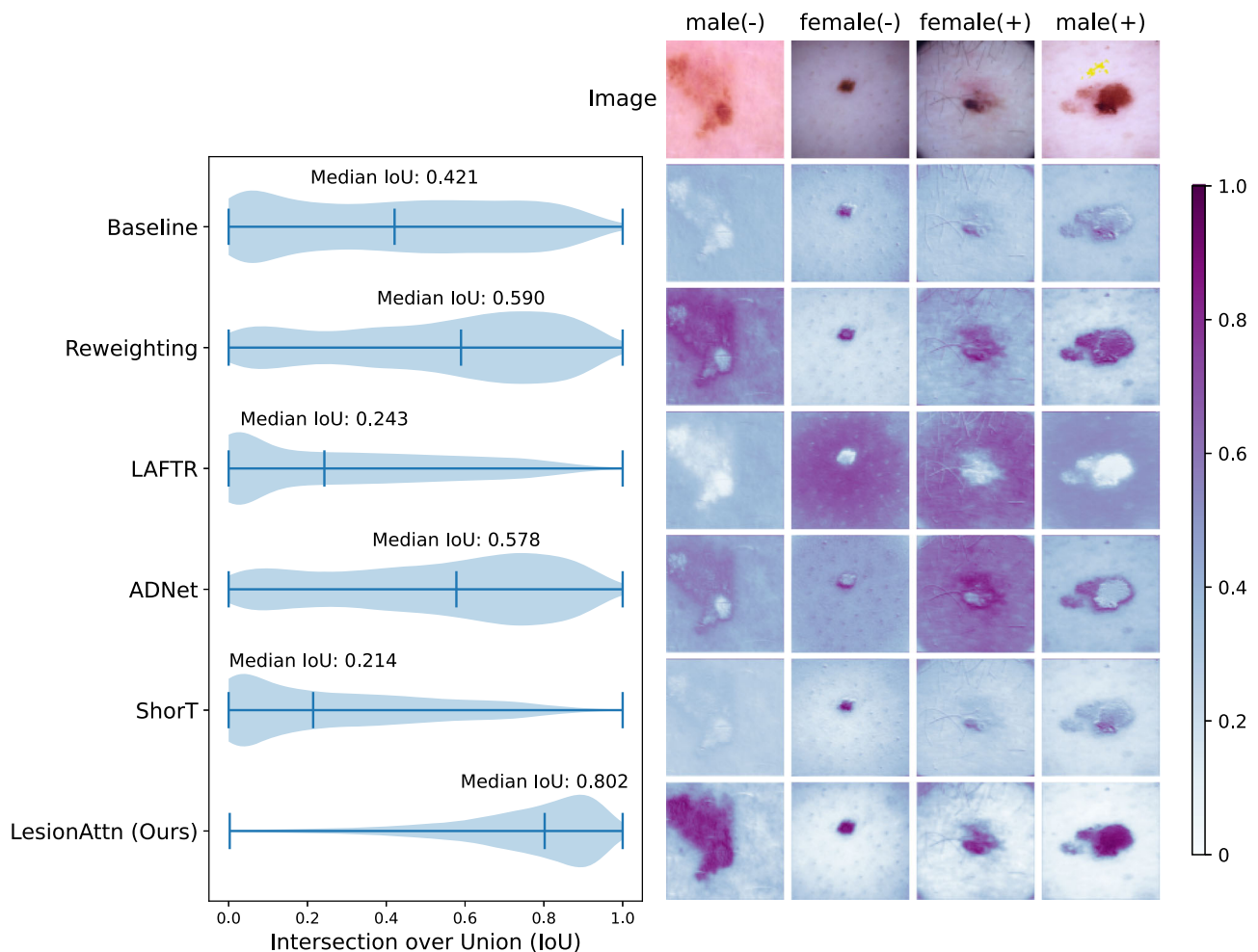


Fig. 3 | Visualization of input images and their corresponding attention maps generated by RANN models trained with Baseline, Reweighting, LAFTR, ADNet, ShorT, and LesionAttn. Higher heatmap intensity indicates greater attention to one

pixel, while lower intensity indicates less attention to one pixel. In the image columns, (–) denotes a pathologically confirmed benign lesion, and (+) denotes a pathologically confirmed malignant lesion.

3 (left panel), LesionAttn achieved the highest median IoU (0.802) and the lowest variability (standard deviation [SD]: 0.199) over the testing set, outperforming other bias mitigation methods and the baseline (Reweighting: median 0.590, SD 0.281; LAFTR: median 0.243, SD 0.270; ADNet: median 0.578, SD 0.288; ShorT: median 0.214, SD 0.249; baseline: median 0.421, SD 0.278).

As shown in Fig. 3 (right panel), attention patterns varied across bias mitigation methods. LAFTR typically highlighted non-lesion regions in its attention maps (i.e., bright areas for the lesion regions and dark areas in the non-lesion regions). Similar attention patterns also occasionally appeared in Reweighting, ADNet, ShorT, and the baseline, particularly for benign lesions in male patients (i.e., true negative cases). In contrast, LesionAttn consistently concentrated attention on lesion areas, regardless of gender or malignancy status, suggesting a more clinically aligned model reasoning.

In addition, restricting the image to lesion regions alone was not as effective as the attention guidance employed by LesionAttn. We compared LesionAttn with a variant that replaced the learnable attention map with static lesion masks, termed LesionOnly. As shown in Fig. 4, LesionOnly achieved low EO values in both datasets at the cost of markedly lower diagnostic performance. In HAM, LesionOnly yielded an AUROC of 0.621 [0.591–0.657] and an AUPRC of 0.271 [0.253–0.290], far below both the baseline and LesionAttn. Despite LesionOnly's reduction to EO (0.073 [0.064–0.081]) compared to the baseline model, it was still outperformed by LesionAttn. In BCN, LesionOnly significantly reduced gender disparity (0.063 [0.052–0.074]), but collapsed to an AUROC of 0.537 [0.450–0.490]

and an AUPRC of 0.586 [0.547–0.626]. These results indicate that hard cropping reduced gender bias but destroyed the spatial context essential for accurate diagnosis.

Sensitivity analysis for the hyperparameter ρ that controls the soft-guidance mechanism confirmed the attenuation of background noises is critical for maintaining diagnostic capability (Suppl. Fig. 2). Compared with LesionOnly ($\rho = 0$), diagnostic performance improved significantly even with a minimal weight assigned to the background (e.g., $\rho = 0.02$). The final LesionAttn model was trained using the optimal hyperparameter identified via grid-searching (i.e., $\rho = 0.7$). We observed a broad range of $\rho < 1$ consistently outperformed the baseline model regarding bias mitigation, demonstrating the algorithmic robustness of our approach.

Discussion

While AI shows promise in early skin cancer diagnosis, its fairness across gender subgroups remains insufficiently studied. In this study, we introduced LesionAttn, an attention-guided algorithm that utilizes expert lesion annotations to guide spatial focus during model training and incorporates gender fairness in model selection. Using two large-scale dermatologic datasets, LesionAttn effectively mitigated algorithmic gender bias while maintaining strong diagnostic performance, compared to the baseline model unmitigated for bias. This study also highlights that aligning model attention with annotations from clinicians can enhance both model performance and fairness in dermatologic AI, underscoring the value of domain knowledge as a key driver of equitable and reliable diagnostic tools.

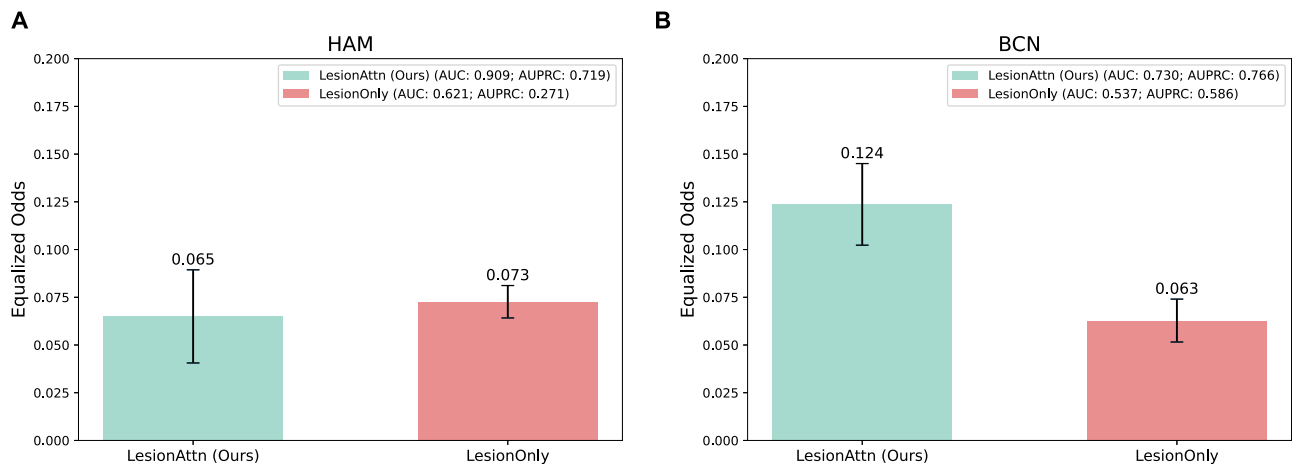


Fig. 4 | Comparison between LesionAttn and LesionOnly across datasets. **A**, comparison on the HAM dataset. **B**, comparison on the BCN dataset. Lower values of Equalized Odds indicate lower gender bias.

Table 1 | Demographics of each gender group of HAM and BCN

	Average Age	Positive Rate	Count
HAM (Internal)			
Male	54.545	0.227	5406
Female	48.712	0.160	4552
Total	51.864	0.196	10,015
BCN (External)			
Male	59.121	0.586	6499
Female	54.137	0.428	5840
Total	56.762	0.511	12,411

We found that the baseline model consistently exhibited lower TPRs in female patients across both internal and external cohorts. Because delayed diagnosis of melanoma can lead to rapid disease progression and worse prognosis, prioritizing high TPR is critical for triage and AI-assisted early diagnosis^{34,35}. These disparities suggest that female patients may suffer from a higher risk of missed or delayed diagnoses, raising concerns about the equitable deployment of AI models in dermatology. We hypothesized that this gender bias is primarily driven by misdirected model attention. As shown in Fig. 3, the baseline model frequently focuses on background healthy skin rather than the lesion itself when not explicitly guided^{36–38}. Background skin inherently encodes subtle gender-specific cues due to biological variations (e.g., hair density, skin texture, vascularity)^{17,18} and behavioral differences (e.g., sun exposure patterns, skincare products)^{19,20}. When the baseline model over-relies on these non-lesion regions, it inadvertently exploits these gender markers as predictive shortcuts, leading to the observed gender bias.

LesionAttn directly addresses this spurious correlation. By explicitly aligning model focus with the actual lesion regions used by clinicians, the framework prevents the model from accessing demographic information hidden in the background. Consequently, the model's reasoning is grounded in the morphological features of the tumor, effectively mitigating gender bias while preserving diagnostic accuracy.

This reliance on intrinsic disease morphology, rather than idiosyncratic background artifacts, may be central to the framework's superior generalizability. As illustrated in Fig. 2, existing bias-mitigation methods often failed to translate fairness gains during external evaluation due to their dependence on dataset-specific features. In contrast, the robust performance of LesionAttn across geographically distinct cohorts underscores its reliability and readiness for real-world clinical deployment.

Among the comparative bias mitigation methods, ShorT improved fairness while preserving performance by discouraging the model from exploiting visual shortcuts⁹. As shown in Fig. 3, its attention maps were more diffusely distributed, likely reducing bias by minimizing dependence on spurious background cues of demographics^{36,39}. However, ShorT offers no positive guidance on where the model should focus for its diagnosis, as the model learns only what to avoid rather than what to prioritize, thereby limiting clinical interpretability. In contrast, LesionAttn adopted a fundamentally different and clinically grounded strategy by explicitly directing model attention to lesion regions. This approach not only proved more effective but also provided clear insight into where and how the model achieved its fairness status.

This study highlights the potential of integrating clinical knowledge (e.g., the localization of disease-relevant features) into AI model design to advance both fairness and accuracy. However, our comparative analysis reveals that the method of integration is critical. While hard-cropping images to strictly isolate the lesion (i.e., the LesionOnly variant) yields better fairness, it results in a catastrophic loss of diagnostic performance. This degradation underscores the necessity of maintaining contextual features rather than enforcing rigid attention constraints⁴⁰. Clinically, the transition zone between a lesion and the surrounding normal skin is critical for assessing border irregularity and edge fading, key indicators of malignancy that are lost when the diagnostic boundary is removed^{41,42}. Computationally, binary masks introduce artificial, high-frequency gradients that might disrupt feature extraction in convolutional networks, which are highly sensitive to such stark edges²¹. LesionAttn strikes this balance through the soft-guidance of attention, preserving spatial relationships while learning to prioritize medically relevant areas⁴³. Future work may further optimize the integration of information across both background skin and lesion regions to refine these synergistic effects.

While LesionAttn reduced the EO on the external BCN dataset by 23.5% (from 0.162 to 0.124), the remaining disparity remains non-trivial. This translates to differential clinical screening burdens: the NNS to identify one malignancy was still 1.93 for female patients, compared to 2.53 for male patients. Notably, this higher screening burden for males occurred despite their significantly higher malignancy prevalence (58.6% vs. 42.8%), suggesting that male patients may undergo more “unnecessary” evaluations per case identified. These observations suggest that the residual disparity may be influenced by non-visual demographic and epidemiological factors. As shown in Table 1, in the BCN cohort, male patients are older on average (59.1 vs. 54.1 years) and likely presented with more complex skin textures²⁹. Since LesionAttn intervenes exclusively on visual spatial attention, it is not designed to correct for these non-visual, demographic imbalances. Furthermore, the persistence of bias during external validation underscores that

fairness is not a static model property but is context-dependent; therefore, local algorithmic auditing and site-specific mitigation still remain essential when deploying such frameworks across diverse clinical environments¹³.

While LesionAttn demonstrates robust performance, it currently necessitates lesion annotations during the training phase. This requirement introduces an annotation burden and inherently couples the framework's efficacy to the quality of expert-defined masks^{7,28}. Relying heavily on human-defined boundaries also carries a risk of supervision bias, where the model may inadvertently inherit human "blind spots" or subjective diagnostic patterns⁴⁴. To mitigate the potential supervision bias, our soft-guidance mechanism allows the model to retain critical contextual features rather than strictly cropping to the annotated region, thereby reducing the impact of potential edge-case inaccuracy in the masks. Looking forward, future work could explore the integration of weakly supervised learning to generate high-fidelity pseudo-masks. Transitioning toward such automated algorithms would significantly reduce the reliance on costly human expertise and facilitate the scaling of this framework to larger, more diverse clinical datasets.

There are limitations to this study. First, while we acknowledge non-binary gender identities, the datasets used were annotated only with binary gender labels, limiting the granularity of subgroup analysis. Second, while our study demonstrates the efficacy of LesionAttn using the RANN architecture, the evaluation was conducted on a single model backbone. While the proposed fairness-enhancing mechanism is designed to be architecture-agnostic, further research is required to validate its generalizability across diverse architectures. Finally, we observed fluctuation in fairness across random seeds for all models, a phenomenon likely attributable to the inherent noise within medical imaging datasets. Future work involving larger, more diverse cohorts may help further stabilize the fairness against stochastic training variations.

This study reveals gender bias in AI-based skin cancer diagnosis and demonstrates that aligning model attention with clinically relevant lesion regions can mitigate these disparities without compromising diagnostic accuracy. LesionAttn exemplifies how incorporating clinical expertise into algorithm design promotes fairness, interpretability, and clinical reliability in dermatologic AI. Furthermore, the underlying principle of attention-guidance holds significant potential for broader application across other medical imaging domains to develop more equitable and trustworthy AI diagnostic tools.

Methods

Dataset

In this study, we focused on the binary classification of malignant versus benign skin lesions from dermoscopic images^{5,45}. The HAM dataset comprises 10,015 dermoscopic images collected over a two-decade period from the Medical University of Vienna (Austria) and a dermatology clinic in Queensland (Australia)²⁸. The BCN dataset comprises 19,424 dermoscopic images acquired between 2010 and 2016 at the Hospital Clinic in Barcelona, of which 12,411 images with a confirmed outcome of interest were included in this study²⁹. Both datasets cover various skin conditions. In HAM, basal cell carcinoma (BCC), melanoma (MEL), and actinic keratoses/intraepithelial carcinoma (AKIEC) were labeled as malignant. In BCN, BCC, MEL, actinic keratoses (AK), and squamous cell carcinoma (SCC) were considered malignant. All other included conditions were categorized as benign.

To provide the clinical guidance for LesionAttn, we utilized official lesion segmentation masks from the International Skin Imaging Collaboration (ISIC) archive for both the HAM and BCN datasets^{7,28}. These masks were curated and validated by expert dermatologists associated with the original dataset releases. The ground-truth diagnoses followed a rigorous hierarchy of evidence: histopathology for malignant cases, and expert consensus or longitudinal follow-up for benign cases. To ensure high-quality annotation, images were manually reviewed by multiple dermatologists, incorporating metadata such as anatomical site and patient age to account for clinical complexity and diagnostic nuances^{7,28}.

Demographic and clinical characteristics stratified by gender are summarized in Table 1. In both datasets, male patients were older on average and had higher malignancy rates than female patients. To assess model robustness and generalizability, we conducted both internal and external validations. HAM was randomly split into 60% training, 20% validation, and 20% internal testing sets strictly at the patient level. The entire BCN dataset was reserved for external testing to evaluate performance in an independent clinical setting. As HAM and BCN were collected from entirely different institutions and geographic locations, there is no patient-level overlap between the internal training data and the external testing cohort.

Residual Attention Neural Network (RANN) for skin cancer diagnosis

The RANN model is based on the widely adopted CNN paradigms in skin cancer diagnosis^{30–32} and additionally incorporates attention maps into a residual learning framework. By embedding attention maps within a residual structure, the model can dynamically refine feature representations, providing a robust but simple-implementation demonstration of attention-driven dermatology AI. We trained the RANN model following established protocols for AI-based skin cancer classification^{30–32}.

As shown in Fig. 1A, the model processes an input dermoscopic image X , where the attention is computed via:

$$A = \text{Softmax}(\text{Conv}(X)), \quad (1)$$

with $\text{Conv}(X)$ denoting a convolution transformation applied to X and the activation function Softmax ensuring a spatial probability distribution. The attention-weighted image is then computed as:

$$X_A = A \odot X, \quad (2)$$

and passed through residual blocks for classification. This architecture allows the network's attention to be visualized, facilitating interpretability and clinical alignment³¹.

LesionAttn algorithm for bias mitigation

We proposed LesionAttn to improve gender fairness in skin cancer diagnosis while preserving diagnostic performance. This bias mitigation approach employs two key mechanisms (Fig. 1B): (1) soft-guided attention alignment with lesion annotations, and (2) Pareto Frontier optimization.

To guide the model's attention to clinically relevant lesion regions, we maximize the alignment between attention maps and lesion masks using cosine similarity. This encourages the model to prioritize features within lesion regions and reduce reliance on possibly gender-correlated background features. Given a lesion mask M and the attention map A , the alignment is quantified as:

$$\text{Cos}(M, A) = \frac{M \cdot A}{\|M\| \|A\|}, \quad (3)$$

where $\cdot$ denotes the dot product, and $\|\cdot\|$ denotes the L_2 norm. While focusing on lesions is beneficial, maintaining the contextual features can also be crucial to fully leverage the model's capability⁴³. To preserve contextual skin features, we introduce a softened mask M_s :

$$M_s = \rho + (1 - \rho) \cdot M, \quad (4)$$

where $\rho \in [0, 1]$ controls attention to the background. The attention loss is then defined as:

$$l_A = 1 - \text{Cos}(M_s, A), \quad (5)$$

as a component of the overall loss function for model training.

Model selection in fairness-aware modeling often involves balancing between competing objectives of maximizing diagnostic performance while minimizing algorithmic bias^{25,26}. To address this, LesionAttn employed the Pareto Frontier algorithm⁴⁶, a widely used approach in multi-objective optimization, to identify models representing optimal trade-offs between the objectives. Unlike approaches that use a single weighted composite loss, which forces a predetermined trade-off between gender fairness and model performance^{25,26}, the Pareto Frontier algorithm preserves the independence of both objectives⁴⁶. This also allows the training to explore models that simultaneously achieve high diagnostic performance and fairness without imposing artificial weighting.

We define the Pareto Frontier as the set of non-dominated model checkpoints during the training process, where an improvement in gender fairness cannot be achieved without a concurrent reduction in diagnostic accuracy, as visualized in Fig. 1B. The formed boundary acts like the “outer edge” of the eligible models regarding Pareto-optimality. It represents the superior trade-off between competing metrics and effectively excludes any candidate models that are objectively inferior in these two dimensions. Specifically, we simultaneously maximized diagnostic performance (P_{pred} , measured by AUROC) and fairness (P_{fair} , quantified by $1 - \text{EO}$). A model f is considered as Pareto-optimal (i.e., lying on the Pareto frontier) if there exists no other model f' that performs as well as f in both objectives and strictly better in at least one:

$$\forall f', P_{\text{pred}}(f) \geq P_{\text{pred}}(f') \text{ and } P_{\text{fair}}(f) \geq P_{\text{fair}}(f'),$$

with at least one strict inequality.

As the Pareto frontier is dynamically updated throughout training, a deterministic rule was applied to select the final model for evaluation. We selected the last model checkpoint that resided on the Pareto frontier upon the completion of training. Notably, we implemented a 10-epoch early stopping protocol, which means the training is terminated if no new non-dominated model checkpoints are discovered for 10 consecutive epochs. This deterministic approach ensures that the final selected model benefits from the most mature state of network weight convergence, while still representing a mathematically optimal trade-off between predictive performance and fairness.

Attention analysis

To examine the attention behavior, we extracted and visualized attention maps from RANN models trained with the baseline, LAFTR, ADNet, ShorT, and LesionAttn. Attention maps were obtained from the final attention layer and overlaid on the input images to compare the models' regions of focus. Quantitatively, the attention-lesion alignment was measured using the Intersection over Union (IoU). To enable a standard IoU calculation between the continuous soft-attention map and the binary ground-truth lesion mask, we applied a dynamic, proportion-based binarization procedure. For each image, we calculated the positive pixel ratio (P_r) of the ground-truth lesion mask. We then binarized the model's soft-attention map by assigning a positive label to the top P_r fraction of pixels with the highest attention weights, and a negative label to the rest. The standard IoU was then calculated between this binarized attention map and the ground-truth mask, ensuring the metric evaluated the spatial ranking of the model's focus while controlling for absolute lesion size.

For illustration, we randomly selected four cases from the HAM testing set for examples: a benign lesion in a male patient, a benign lesion in a female patient, a malignant lesion in a female patient, and a malignant lesion in a male patient, respectively.

Experiment settings

To evaluate the effectiveness of LesionAttn, we compared it with four established bias mitigation methods (Reweighting, LAFTR, ADNet, and ShorT) under the same experimental settings. All models shared the same RANN architecture to ensure consistency in comparison. Additionally, to

isolate the effect of learnable attention, we constructed a LesionOnly variant of RANN, in which input images were cropped using binary lesion masks rather than learned attention maps. This setup enabled assessment of whether static lesion focus alone could reproduce the fairness improvement achieved by LesionAttn.

All models underwent hyperparameter optimization via grid search, using AUROC on the validation set as the selection criterion. Hyperparameter ranges and final settings are detailed in the Suppl. Table 1. A sensitivity analysis of the hyperparameter ρ , which regulates the soft-guidance of attention, was conducted to evaluate the algorithmic robustness of LesionAttn.

Each experiment was repeated five times with different random seeds, and results are reported as mean values with 95% confidence intervals. To evaluate performance improvements, we conducted a paired one-sided t-test across the experimental seeds, with $p < 0.05$ defined as the threshold for statistical significance.

Data availability

The data used in this study are derived from publicly available resources within the International Skin Imaging Collaboration (ISIC) Archive (HAM10000: BCN20000).

Code availability

The code supporting the findings of this study is available in a GitHub repository at <https://github.com/nliulab/LesionAttn>.

Received: 26 November 2025; Accepted: 5 June 2026;

Published online: 19 June 2026

References

- Hasan, N. et al. Skin cancer: understanding the journey of transformation from conventional to advanced treatment approaches. *Mol. cancer* **22**, 168 (2023).
- Chandra, J. et al. Nanotechnology-empowered strategies in treatment of skin cancer. *Environ. Res.* **235**, 116649 (2023).
- Dildar, M. et al. Skin cancer detection: a review using deep learning techniques. *Int. J. Environ. Res. public health* **18**, 5479 (2021).
- Celebi, M. E., Codella, N. & Halpern, A. Dermoscopy image analysis: overview and future directions. *IEEE J. Biomed. health Inform.* **23**, 474–478 (2019).
- Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *nature* **542**, 115–118 (2017).
- Smak Gregoor, A. M. et al. An artificial intelligence based app for skin cancer detection evaluated in a population based setting. *NPJ digital Med.* **6**, 90 (2023).
- Tschandl, P. et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
- Yang, J., Soltan, A. A. S., Eyre, D. W. & Clifton, D. A. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nat. Mach. Intell.* **5**, 884–894 (2023).
- Brown, A. et al. Detecting shortcut learning for fair medical AI using shortcut testing. *Nat. Commun.* **14**, 4314 (2023).
- Celeste, C. et al. Ethnic disparity in diagnosing asymptomatic bacterial vaginosis using machine learning. *NPJ Digital Med.* **6**, 211 (2023).
- Caton, S. & Haas, C. Fairness in machine learning: a survey. *ACM Comput. Surv.* **56**, 1–38 (2024).
- Liu, M. et al. A translational perspective towards clinical AI fairness. *NPJ Digital Med.* **6**, 172 (2023).
- Liu, M. et al. A scoping review and evidence gap analysis of clinical AI fairness. *npj Digital Med.* **8**, 360 (2025).
- Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y. & Clifton, D. A. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digital Med.* **6**, 55 (2023).

15. Melarkode, N., Srinivasan, K., Qaisar, S. M. & Plawiak, P. AI-powered diagnosis of skin cancer: a contemporary review, open challenges and future research directions. *Cancers* **15**, 1183 (2023).
16. Lee, M. S., Guo, L. N. & Nambudiri, V. E. Towards gender equity in artificial intelligence and machine learning applications in dermatology. *J. Am. Med. Inform. Assoc.* **29**, 400–403 (2022).
17. Collier, V., Musicante, M., Patel, T. & Liu-Smith, F. Sex disparity in skin carcinogenesis and potential influence of sex hormones. *Ski. health Dis.* **1**, e27 (2021).
18. Rahrovan, S., Fanian, F., Mehryan, P., Humbert, P. & Firooz, A. Male versus female skin: what dermatologists and cosmeticians should know. *Int. J. women's. Dermatol.* **4**, 122–130 (2018).
19. Chen, J. et al. Gender-based differences and barriers in skin protection behaviors in melanoma survivors. *J. Skin Cancer.* **2016**, 3874572 (2016).
20. El Sharouni, M.-A. et al. Sex matters: men with melanoma have a worse prognosis than women. *J. Eur. Acad. Dermatol. Venereol.* **33**, 2062–2067 (2019).
21. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **30**, 2838–2848 (2024).
22. Madras, D. et al. Learning adversarially fair and transferable representations. In *2018 International Conference on Machine Learning (ICML)*. 3384–3393.
23. Naqvi, M., Gilani, S. Q., Syed, T., Marques, O. & Kim, H.-C. Skin cancer detection using deep learning—a review. *Diagnostics* **13**, 1911 (2023).
24. Koh, D.-M. et al. Artificial intelligence and machine learning in cancer imaging. *Commun. Med.* **2**, 133 (2022).
25. Zhao, H. & Gordon, G. J. Inherent tradeoffs in learning fair representations. *J. Mach. Learn. Res.* **23**, 2527–2552 (2022).
26. Pfohl, S. R., Foryciarz, A. & Shah, N. H. An empirical characterization of fair machine learning for clinical risk prediction. *J. Biomed. Inform.* **113**, 103621 (2021).
27. Liu, M. et al. FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare. *Patterns* **5**, 101059 (2024).
28. 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).
29. Hernández-Pérez, C. et al. Bcn20000: Dermoscopic lesions in the wild. *Scientific data* **11**, 641 (2024).
30. Hao, F., Zhang, T., Zhao, L. & Tang, Y. Efficient residual attention network for single image super-resolution. *Appl. Intell.* **52**, 652–661 (2022).
31. Wang, F. et al. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6450–6458.
32. Omeroglu, A. N., Mohammed, H. M. A., Oral, E. A. & Aydin, S. A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification. *Eng. Appl. Artif. Intell.* **120**, 105897 (2023).
33. Kamiran, F. & Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **33**, 1–33 (2012).
34. Papachristou, P. et al. Evaluation of an artificial intelligence-based decision support for the detection of cutaneous melanoma in primary care: a prospective real-life clinical trial. *Br. J. Dermatol.* **191**, 125–133 (2024).
35. Phillips, M. et al. Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions. *JAMA Netw. Open* **2**, e1913436 (2019).
36. Yan, S. et al. Towards trustable skin cancer diagnosis via rewriting model's decision. In *2023 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. 11568–11577.
37. DeGrave, A. J., Cai, Z. R., Janizek, J. D., Daneshjou, R. & Lee, S.-I. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *Nat. Biomed. Eng.* **9**, 1–13 (2023).
38. Pacheco, A. G. C. & Krohling, R. A. The impact of patient clinical information on automated skin cancer detection. *Computers Biol. Med.* **116**, 103545 (2020).
39. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **1–11** (2024).
40. Nazari, S. & Garcia, R. Going Smaller: Attention-based models for automated melanoma diagnosis. *Computers Biol. Med.* **185**, 109492 (2025).
41. Abbasi, N. R. et al. Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. *Jama* **292**, 2771–2776 (2004).
42. Azam, S. et al. An Automated Decision Support System to Analyze Malignancy Patterns of Breast Masses Employing Medically Relevant Features of Ultrasound Images. *J. Imaging Inf. Med.* **37**, 45–59 (2024).
43. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *2014 European Conference on Computer Vision (ECCV)*. 818–833.
44. Dubois, C. et al. Deep learning in medical image analysis: introduction to underlying principles and reviewer guide using diagnostic case studies in paediatrics. *BMJ.* **387**, e076703 (2024).
45. Haenssle, H. A. et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **29**, 1836–1842 (2018).
46. Ruchte, M. & Grabocka, J. Scalable Pareto Front Approximation for Deep Multi-Objective Learning. In *2021 IEEE International Conference on Data Mining (ICDM)*. 1306–1311.

Acknowledgements

This work was supported by the Duke-NUS Signature Research Programme funded by the Ministry of Health, Singapore. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Health. Tingting Zhu was supported by the Royal Academy of Engineering under the Research Fellowship scheme.

Author contributions

Conceptualization: M.Z., M.L., N.L. Method design, experiment, data analysis: M.Z., M.L. Drafting of the manuscript: M.Z., M.L. Critical revision of the manuscript: M.Z., M.L., N.L. Interpretation of the content: M.Z., M.L., H.Y., Y.N., N.L. Revisions of the manuscript: M.Z., M.L., H.Y., Y.N., N.L. Final read and approval of the completed version: all authors. Overseeing the project: N.L.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-026-02897-8>.

Correspondence and requests for materials should be addressed to Nan Liu.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026