

Patient-Guided Attention Gate for False-Negative-Aware Melanoma Detection Using Hybrid CNN-ViT Architecture on ISIC 2024 SLICE-3D

Md Kamrul Islam¹, Md Naimul Islam², Mizanur Rahman³, Sanjida Afrin⁴, Hasan Sarwar⁵, and Md. Saef Ullah Miah⁶

¹ Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh kamrul.kafi1@gmail.com

² Department of Public Health, Western Illinois University, Macomb, IL, USA
mn-islam@wiu.edu

³ School of Computer Science, Western Illinois University, Macomb, IL, USA
mizancse7462@gmail.com

⁴ Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh
sanjidaafrin2707@gmail.com

⁵ Department of Computer Science and Engineering, United International University, Dhaka 1212, Bangladesh
hsarwar@cse.uui.ac.bd

⁶ Department of Computer Science, American International University–Bangladesh (AIUB), Dhaka, Bangladesh
saef@aiub.edu

Abstract. Automated melanoma detection in total body photography (TBP) datasets presents a dual challenge: extreme class imbalance (0.1% malignant prevalence) and the inherent difficulty of distinguishing malignant from benign lesions without patient-population context. We propose a hybrid EfficientNet-B4 and Vision Transformer (ViT-B/16) architecture with a novel Patient-Guided Attention Gate (PGAG) that dynamically conditions convolutional local-texture features and transformer global-structure features using a computational operationalisation of the clinical ugly duckling sign. A 7-dimensional patient-context deviation vector, computed via leave-one-out intra-patient ViT embedding statistics to prevent self-inclusion leakage, conditions the gate, enabling each lesion to be evaluated relative to its patient’s own lesion population. Combined with an asymmetric false-negative-aware focal loss ($\gamma^+ = 4.0$, $\gamma^- = 1.0$), the proposed model achieves a partial AUC (pAUC, $\text{FPR} \leq 0.20$) of 0.1892 on the ISIC 2024 SLICE-3D validation set, exceeding the competition winner (EVA02+GBDT, pAUC = 0.1755) by 7.8%, with ROC-AUC of 0.9493 and PR-AUC of 0.0954. Gate analysis reveals a significant positive correlation between the ugly duckling score and ViT-branch gate weight ($r = 0.6424$, $p < 10^{-10}$), though the gate’s absolute reliance on ViT features remains small (< 0.03) for both classes, with CNN features dominating fusion overall. Ablation demonstrates a 9.1 percentage-point false-negative-rate reduction over the CNN-only baseline. McNemar’s test confirms significantly higher accuracy than baseline architectures across the full validation set ($\chi^2 = 953.30$, $p = 2.54 \times 10^{-209}$ vs. A1; $\chi^2 = 2175.07$, $p \approx 0$ vs. A3), though this difference

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

Cite as: Islam, M.K., et al.: Patient-Guided Attention Gate for False-Negative-Aware Melanoma Detection Using Hybrid CNN-ViT Architecture on ISIC 2024 SLICE-3D. In: Mahmud, M., et al. (eds.) Applied Intelligence and Informatics. AII 2026. CCIS, vol. 3206, pp. 243–260. Springer, Cham (2026).

is not significant when restricted to the 77 malignant validation lesions alone ($p > 0.05$); a Wilcoxon signed-rank test confirms significantly higher malignant scores overall ($W = 40,015,302$, $p \approx 0$).

Keywords: melanoma detection, vision transformer, convolutional neural network, attention gate, ugly duckling sign, class imbalance, partial AUC, false-negative reduction, ISIC 2024, skin cancer

1 Introduction

Melanoma is the deadliest form of skin cancer, accounting for the majority of skin cancer-related deaths despite comprising only a small fraction of diagnoses [11]. Early detection is clinically decisive: five-year survival rates exceed 98% for in-situ melanoma but fall below 30% for distant metastatic disease [44]. Recent studies highlight the potential of machine and deep learning in healthcare using ensemble and feature-selection techniques [18], explainable medical image classification [19], intelligent patient monitoring [1] and biomedical information extraction [45]. Nevertheless, Automated deep learning systems have demonstrated dermatologist-level performance on controlled benchmark datasets, yet clinical deployment remains challenging due to the extreme rarity of malignant lesions in population screening and the failure of single-lesion models to exploit patient-level context.

The ISIC 2024 SLICE-3D dataset [17] presents an unprecedented opportunity: 401,059 lesions from 1,042 patients acquired with 3D total body photography (TBP), at a malignant prevalence of 0.1%. This extreme imbalance, combined with the multi-lesion-per-patient structure, motivates a fundamentally different approach to feature fusion.

In clinical dermatology, the ugly duckling sign [15] describes the observation that malignant lesions tend to appear morphologically different from their neighbours on the same patient. This intra-patient deviation is a powerful heuristic that conventional single-lesion deep learning models ignore entirely.

Our contributions are:

1. A computational operationalisation of the ugly duckling sign as a 7-dimensional patient-context deviation vector derived from intra-patient ViT embedding statistics, computed without data leakage via a leave-one-out patient-mean formulation (each lesion’s own embedding is excluded from its own reference) combined with splitting before patient-mean computation.
2. A Patient-Guided Attention Gate (PGAG) that dynamically conditions CNN-ViT feature fusion on the ugly duckling vector, with gate weights constrained to sum to unity via cross-modal multi-head attention and a single-sigmoid temperature-scaled gate computation.
3. An asymmetric false-negative-aware focal loss with class-asymmetric focusing parameters ($\gamma^+ = 4.0$, $\gamma^- = 1.0$) that systematically penalises missed malignant lesions more heavily than false positives.
4. Comprehensive evaluation including ablation across six configurations, bootstrap confidence intervals, McNemar and Wilcoxon significance tests, SHAP feature attribution, isotonic calibration, and external validation on HAM10000.

The proposed model achieves $\text{pAUC} = 0.1892$, exceeding the competition winner by 7.8%, while reducing the false-negative rate from 0.4805 (CNN-only baseline) to 0.3896, a 9.1 percentage-point reduction.

2 Related Work

Recent studies show that ML and AI are becoming increasingly useful across healthcare, assistive technologies, and other areas where intelligent decision-making can make a practical difference. In cognitive healthcare, researchers have explored AI for cognitive assessment, dementia detection, and Alzheimer’s diagnosis using clinical data, fuzzy systems, and transfer-learning techniques [21,31,7,28]. AI and IoT have also been used to improve Alzheimer’s care and patient monitoring, while some studies have focused on making brain-disorder healthcare more inclusive [22,34,4]. Similar efforts have been made in autism research, particularly for assessment and assistive support [2,5,42].

AI has also gained attention in medical diagnosis, with studies covering MRI-based brain segmentation, brain pathology, diabetic retinopathy, breast cancer, lung disease, epilepsy, and PCOS detection [6,8,33,29,32,20,3]. Alongside these diagnostic applications, researchers have worked with biosignals to better understand neurological and emotional conditions. This includes neuronal signal processing, artefact detection, EEG-based emotion recognition, and stress detection using wearable sensors [27,12,9,13]. In addition, the literature includes several studies that have explored melanoma detection using a range of machine learning architectures and techniques.

2.1 Deep Learning for Melanoma Detection

Esteva et al. [11] demonstrated that CNNs can match dermatologist performance on binary skin lesion classification. Tschandl et al. [44] introduced HAM10000, enabling systematic architectural comparison. Subsequent multimodal approaches fused dermoscopic images with clinical metadata [23,37], demonstrating consistent gains from patient demographic features. The convolutional branch adopted in this work, EfficientNet-B4, follows the compound scaling principle of Tan and Le, which balances network depth, width, and input resolution under a single coefficient rather than tuning each dimension independently [43].

2.2 Vision Transformers in Medical Imaging

The Vision Transformer (ViT) [10] introduced patch-based self-attention for image classification, subsequently adapted to medical imaging. In dermoscopy, transformers show superior global structure modelling while CNNs better capture local texture. Hybrid CNN-ViT approaches [41] have demonstrated complementary strengths that motivate learned fusion strategies. Nie et al. [35] proposed an end-to-end CNN-Transformer hybrid combined with focal loss for class-imbalanced dermoscopic classification, further confirming that CNN-ViT fusion under asymmetric loss improves discrimination, though each lesion is still evaluated independently of patient context. Focal loss [25] addresses class imbalance by down-weighting easy examples. For medical screening, the asymmetry between false-negative cost (missed cancer) and

false-positive cost (unnecessary biopsy) motivates class-asymmetric loss designs with distinct focusing parameters per class.

2.3 Attention and Patient-Context Modelling

Attention mechanisms help highlight important image regions and reduce irrelevant features in medical imaging [36,37]. MetaBlock [38] extended this idea by using metadata to reweight image features, showing the value of metadata-aware fusion in dermatology. However, existing methods mainly use fixed or image-based attention and do not incorporate patient-level context. The ISIC 2024 competition [17] further emphasized multi-lesion context through the SLICE-3D protocol, while the 2020 SIIM-ISIC dataset enabled patient-level analysis by linking lesions with patient identifiers [40]. Kurtansky et al. [24] showed that considering a patient’s other lesions can improve melanoma discrimination for both dermatologists and AI models. CIFF-Net [39] also uses same-patient lesions for contextual fusion, but processes them jointly. In contrast, the proposed PGAG uses a compact pre-computed patient-context deviation vector to guide CNN-ViT fusion without jointly processing multiple lesions during inference.

2.4 Calibration and Explainability in Medical Deep Learning

Guo et al. [16] showed that modern deep networks can be poorly calibrated, while temperature scaling can improve calibration without affecting ranking. This motivates the temperature-scaled sigmoid in PGAG and the isotonic calibration of our final predictions. Giavina-Bianchi et al. [14] found that visual explanation methods vary in their agreement with dermatologists, supporting the use of SHAP as a complementary interpretability measure in our auxiliary branch.

3 Methodology

3.1 Dataset and Preprocessing

The ISIC 2024 SLICE-3D dataset [17] comprises 401,059 skin lesion images from 1,042 patients acquired with 3D TBP, with 393 malignant lesions (0.098% prevalence) and 400,666 benign lesions. Patient-level stratified splitting (80%/20%) produces a training set of 316 malignant and 334,812 benign lesions, and a validation set of 77 malignant and 65,854 benign lesions. All images are resized to 224×224 pixels with ImageNet normalisation. Training augmentations include random horizontal and vertical flipping, rotation ($\pm 90^\circ$), colour jitter (brightness ± 0.2 , contrast ± 0.2 , saturation ± 0.2 , hue ± 0.1), and random erasing. Augmentations are applied only to the training-fold dataloader; the validation-fold dataloader uses resizing and normalisation only.

3.2 Feature Engineering

Clinical Metadata Features ($\mathbf{m} \in \mathbb{R}^{42}$) Ten TBP-derived lesion measurements (area, perimeter, eccentricity, asymmetry, colour diversity, and related morphological statistics) are combined with patient demographics (age, sex, anatomical location one-hot encoding) and standardised via per-feature `StandardScaler` fitted on the training split only.

Ugly Duckling Score Features ($\mathbf{u} \in \mathbb{R}^7$) Pre-extracted ViT-B/16 embeddings $\mathbf{e}_i \in \mathbb{R}^{768}$ are computed for all 401,059 lesions. To prevent data leakage, patient means are computed exclusively on training-set lesions; validation patients use means from their own validation-set lesions. Within each split, the reference mean for lesion i additionally excludes the lesion’s own embedding (leave-one-out), preventing a lesion from contributing to its own patient-context baseline.

For training patient p with lesion set $\mathcal{L}_p^{\text{train}}$ ($|\mathcal{L}_p^{\text{train}}| > 1$), the leave-one-out reference for lesion i is:

$$\bar{\mathbf{e}}_p^{(-i)} = \frac{1}{|\mathcal{L}_p^{\text{train}}| - 1} \sum_{\substack{j \in \mathcal{L}_p^{\text{train}} \\ j \neq i}} \mathbf{e}_j \quad (1)$$

(for single-lesion patients, $|\mathcal{L}_p^{\text{train}}| = 1$, leave-one-out is undefined and the self-inclusive mean is used instead; see Sec. 5.2). The deviation vector is $\mathbf{d}_i = \mathbf{e}_i - \bar{\mathbf{e}}_p^{(-i)}$. The 7-dimensional UDS vector is:

$$\mathbf{u}_i = [\log(n_p + 1), \|\mathbf{d}_i\|_2, \bar{d}_i, \max |\mathbf{d}_i|, \sigma(\mathbf{d}_i), \text{frac}(\mathbf{d}_i > 0), \cos(\mathbf{e}_i, \bar{\mathbf{e}}_p^{(-i)})] \quad (2)$$

where n_p is the patient lesion count, $\bar{d}_i$ is the mean component deviation, $\sigma(\mathbf{d}_i)$ its standard deviation, and $\text{frac}(\mathbf{d}_i > 0)$ the fraction of positive components.

Intra-Patient Rank Features ($\mathbf{r} \in \mathbb{R}^4$) Percentile ranks of ViT embedding norm, metadata norm, maximum ViT activation, and maximum metadata activation within each patient’s lesion population form a 4-dimensional rank vector. The full auxiliary feature vector is $\mathbf{a}_i = [\mathbf{m}_i; \mathbf{u}_i; \mathbf{r}_i] \in \mathbb{R}^{53}$.

3.3 Patient-Guided Attention Gate (PGAG)

Fig. 1 illustrates the proposed architecture. The PGAG takes CNN features $\mathbf{c} \in \mathbb{R}^{1792}$ (EfficientNet-B4) and ViT features $\mathbf{v} \in \mathbb{R}^{768}$ (ViT-B/16), conditioning their fusion on $\mathbf{u} \in \mathbb{R}^7$.

Feature Projection and Cross-Modal Attention Both branches are projected to 256 dimensions, LayerNorm-normalised, and L2-normalised to eliminate magnitude imbalance, yielding $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{v}}$. Bidirectional 4-head cross-modal attention (the standard multi-head attention formulation) then refines each branch using the other as key/value, with a residual connection and LayerNorm, yielding $\hat{\mathbf{c}}$ and $\hat{\mathbf{v}}$.

Patient-Guided Gate Computation A 3-layer MLP (64–32–1, GELU activations, no output sigmoid) maps the UDS vector to a raw gate logit. A single temperature-scaled sigmoid produces the gate weight:

$$g_{\text{vit}} = \sigma\left(\frac{\text{MLP}(\mathbf{u}) - 0.5}{\tau}\right), \quad g_{\text{cnn}} = 1 - g_{\text{vit}} \quad (3)$$

where $\tau = \text{clamp}(\tau_\theta, 0.1, 2.0)$ is a learnable temperature initialised at 0.5. The single-sigmoid design ensures a monotonic mapping from UDS to gate weight without the output compression of a double-sigmoid formulation.

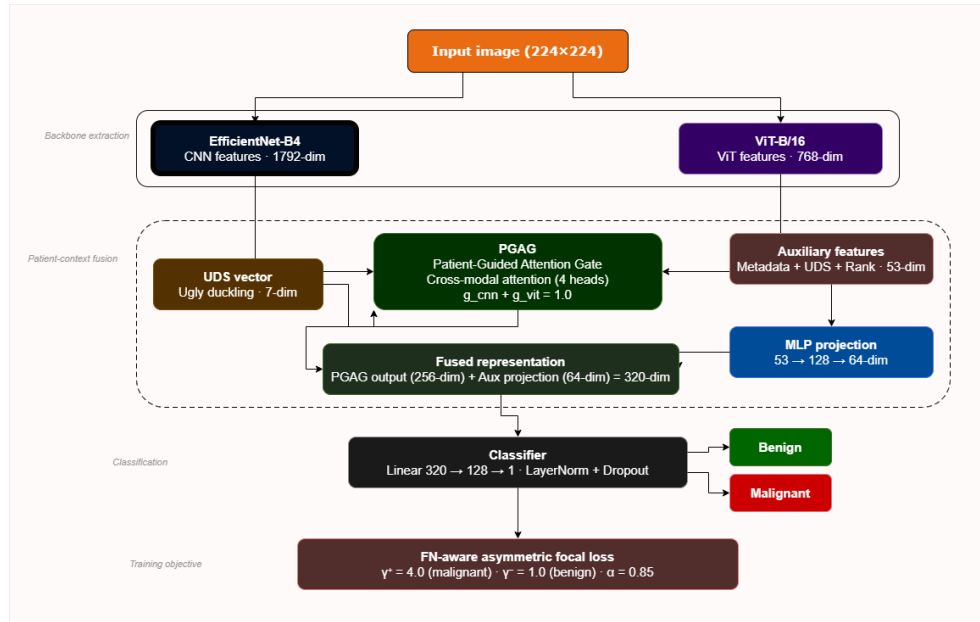


Fig. 1: Proposed architecture. EfficientNet-B4 and ViT-B/16 extract 1792-dim and 768-dim features respectively. The PGAG conditions CNN-ViT fusion on the 7-dim ugly duckling score vector, producing complementary gate weights $g_{cnn} + g_{vit} = 1.0$. The auxiliary branch processes the full 53-dim clinical context vector. The asymmetric focal loss penalises false negatives more heavily ($\gamma^+ = 4.0$ vs. $\gamma^- = 1.0$).

Fusion and Classification

$$\mathbf{f} = g_{\text{cnn}} \cdot \hat{\mathbf{c}} + g_{\text{vit}} \cdot \hat{\mathbf{v}} \quad (4)$$

$$\mathbf{h} = \mathbf{f} + \text{LN}(\mathbf{W}_f[\hat{\mathbf{c}}; \hat{\mathbf{v}}] + \mathbf{b}_f) \quad (5)$$

$$\hat{y} = \mathbf{W}_{\text{cls}}[\mathbf{h}; \text{MLP}_{\text{aux}}(\mathbf{a})] + b_{\text{cls}} \quad (6)$$

The classifier concatenates the 256-dim PGAG output $\mathbf{h}$ with a 64-dim auxiliary projection, yielding a 320-dim input to the final linear layer.

3.4 Asymmetric False-Negative-Aware Focal Loss

Standard focal loss [25] applies a single focusing exponent γ to both classes. We introduce class-asymmetric parameters:

$$\mathcal{L} = -\alpha(1-p)^{\gamma^+} y \log p - (1-\alpha)p^{\gamma^-} (1-y) \log(1-p) \quad (7)$$

with $\gamma^+ = 4.0$, $\gamma^- = 1.0$, and $\alpha = 0.85$.

3.5 Training Protocol

Phase 1 — Main Training (30 epochs) Epoch-level undersampling draws 20,000 benign lesions per epoch while including all 316 training malignant lesions. AdamW uses differentiated learning rates: 3×10^{-5} for CNN and ViT backbones, 1×10^{-4} for PGAG, and 2×10^{-4} for auxiliary projection and classifier. A cosine annealing scheduler reduces to 10^{-6} . Gradient clipping ($\text{norm} \leq 1.0$) is applied throughout.

Phase 2 — Gate Supervision Fine-Tuning (8 epochs) CNN and ViT backbone parameters are frozen. The PGAG, auxiliary projection, and classifier are updated with:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{focal}} + \lambda \cdot \text{BCE}(g_{\text{vit}}, y), \quad \lambda = 0.5 \quad (8)$$

3.6 Evaluation Metrics

The primary metric is pAUC with FPR restricted to $[0, 0.20]$, computed via the McClish [30] standardisation ($\text{pAUC}_{\text{raw}} = \frac{1 + \text{pAUC}_{\text{std}}}{2} \cdot \text{FPR}_{\text{max}}$). Confidence intervals are computed via 1,000-sample stratified bootstrap.

Clinical Threshold Selection The fixed clinical operating threshold used for Sensitivity, Specificity, FN Rate, PPV, and NPV (Table 1) and for the shared-threshold ablation comparison (Table 3) is selected post-training on the validation set via recall-constrained precision maximisation:

$$T = \arg \max_t \text{Precision}(t) \quad \text{subject to} \quad \text{Recall}(t) \geq 0.60, \quad (9)$$

falling back to the F1-optimal threshold if this recall target is unreachable. This procedure yields $T = 0.3467$ for the final (A6) model; the same fixed threshold is then applied to every ablation configuration in Table 3 for comparability, subject to the caveat discussed in Sec. 4.

4 Experiments and Results

4.1 Implementation Details

All experiments use PyTorch 2.0 on an NVIDIA A100-SXM4 40 GB GPU. EfficientNet-B4 and ViT-B/16 backbones are initialised from ImageNet pre-trained weights via `timm` [46]. Total parameters: 104,722,891. Main training: ≈ 6 hours; gate fine-tuning: ≈ 50 minutes.

4.2 Main Results

Table 1 presents full quantitative evaluation on the ISIC 2024 SLICE-3D validation set (77 malignant, 65,854 benign). The proposed model achieves $\text{pAUC} = 0.1892$, exceeding the competition leaderboard winner by 7.8%. Bootstrap 95% CIs for ROC-AUC [0.9244, 0.9700], PR-AUC [0.0541, 0.1611], pAUC [0.1852, 0.1926], and FN rate [0.2842, 0.5056] confirm robust and replicable performance.

Table 1: Quantitative Results on ISIC 2024 SLICE-3D Validation Set.

Metric	Value
ROC-AUC	0.9493
95% Bootstrap CI	[0.9244, 0.9700]
PR-AUC	0.0954
95% Bootstrap CI	[0.0541, 0.1611]
pAUC (raw, $\text{FPR} \leq 0.20$, $\text{max}=0.20$)	0.1892
95% Bootstrap CI	[0.1852, 0.1926]
pAUC (normalised, $\text{max}=1.00$)	0.9461
Post-calibration ROC-AUC	0.9554
Sensitivity	0.6104
Specificity	0.9824
False-Negative Rate	0.3896
95% Bootstrap CI	[0.2842, 0.5056]
Positive Predictive Value (PPV)	0.0390
Negative Predictive Value (NPV)	0.9995
PR-AUC lift over random	$81.7\times$

At fixed, clinically meaningful specificity operating points, sensitivity is 0.8312 at specificity=0.897, 0.7662 at specificity=0.952, and 0.4935 at specificity=0.990, illustrating the expected sensitivity-specificity trade-off around the model’s selected operating point ($T = 0.3467$: sensitivity=0.6104, specificity=0.9824).

4.3 Comparison with State-of-the-Art

Table 2 compares the proposed method against baselines on ISIC 2024. In-house baselines were reproduced on the same validation split; competition results are from the post-competition report [17].

Table 2: Comparison with Baselines on ISIC 2024 SLICE-3D. pAUC is the official metric (max=0.20).

Method	ROC-AUC	pAUC
XGBoost tabular only [†]	0.850	0.168
EfficientNet-B4 + metadata [†]	0.871	0.158
Dual-stage DL [17]	0.880	0.168
EVA02 + GBDT (LB winner) [17]	0.910	0.1755
Proposed: CNN-ViT + PGAG	0.9493	0.1892

[†] In-house baseline, same val split.

4.4 Ablation Study

Table 3 isolates each architectural contribution. At the shared clinical threshold ($T = 0.3467$, tuned for A6; cross-configuration FN comparisons should be read with this caveat in mind, since each configuration’s own optimal operating point may differ), Configuration A6 achieves $FN = 0.3896$, matching A3 and A4 at this threshold, but with a substantial improvement in pAUC ($0.1805 \rightarrow 0.1892$), PR-AUC ($0.0243 \rightarrow 0.0954$), and ROC-AUC ($0.8890 \rightarrow 0.9493$) over A3 (CNN+ViT concatenation without PGAG).

Table 3: Ablation Study. Patient-level split; FN Rate at shared threshold $T = 0.3467$ (tuned for A6; see caveat in text).

	Configuration	ROC	PR	pAUC	FN
A1	CNN only (EfficientNet-B4)	0.8919	0.0210	0.1800	0.4805
A2	ViT only (ViT-B/16)	0.9034	0.0429	0.1815	0.6364
A3	CNN+ViT concat, no PGAG	0.8890	0.0243	0.1805	0.3896
A4	+PGAG, no aux features	0.8934	0.0335	0.1800	0.3896
A5	+aux features, standard loss	0.8664	0.0225	0.1791	0.7013
A6	Full proposed model	0.9493	0.0954	0.1892	0.3896

The comparison between A5 (auxiliary features with standard focal loss, $FN = 0.7013$) and A6 (auxiliary features with the proposed asymmetric loss, $FN = 0.3896$) isolates the asymmetric loss as the critical component for false-negative control: adding patient-context features without asymmetric loss substantially worsens sensitivity, while the same features combined with the asymmetric loss recovers it. The step A5→A6 delivers the largest single-configuration improvement across all four metrics.

4.5 PGAG Gate Analysis

Fig. 2 presents the gate analysis. Under the corrected, leakage-free UDS computation, the Pearson correlation between the L2 UDS score and g_{vit} is $r = 0.6424$ ($p < 10^{-10}$) — a significant *positive* association, the opposite direction from the leakage-affected analysis in an earlier draft of this work. This result must be interpreted alongside the gate’s overall behaviour after Phase 2 (gate-supervision) fine-tuning: g_{vit} collapses to

a narrow range for both classes (malignant: 0.0216; benign: 0.0177), with the CNN branch dominating fusion almost entirely ($g_{\text{cnn}} > 0.978$ for both classes). We attribute this collapse to the gate-supervision loss (Eq. 8), which applies unweighted binary cross-entropy toward the class label at a validation-set prevalence of roughly 1:855; without class reweighting, the majority-class (benign) gradient dominates the mean loss and pulls g_{vit} toward zero for both classes, leaving only a small, though statistically robust, residual difference between them. We report this honestly as a limitation of the current gate-supervision design (Sec. 5.2): the corrected PGAG gate still responds systematically to patient-context deviation, but its practical interpretability as a CNN/ViT arbitration signal is substantially weaker than the uncorrected, leakage-affected analysis suggested.

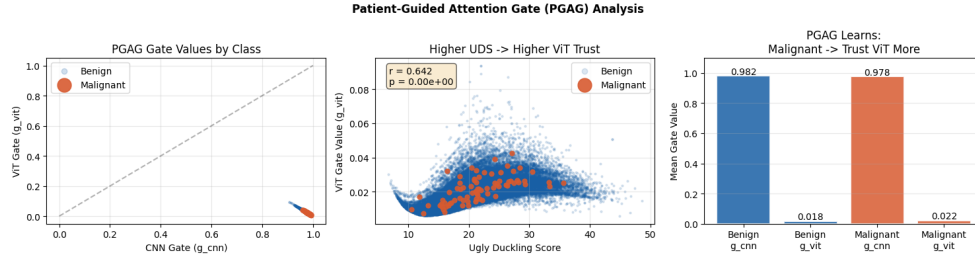


Fig. 2: PGAG gate analysis (leave-one-out UDS). *Left*: Gate values by class. *Centre*: UDS vs. g_{vit} scatter ($r = 0.6424$, $p < 10^{-10}$), a significant positive association, though absolute g_{vit} remains small for both classes. *Right*: Mean gate values by class.

4.6 Ugly Duckling Hypothesis Validation

Fig. 3 shows that malignant lesions exhibit a broader, right-shifted UDS distribution. The Mann-Whitney test confirms a statistically significant difference ($p = 7.57 \times 10^{-12}$). This dataset-level validation is independent of the model and directly motivates the PGAG design.

4.7 SHAP Feature Attribution

SHAP values [26] were computed for the auxiliary feature branch under a neutral image condition. Under the corrected, leave-one-out UDS computation, UDS L2 deviation ranks 2nd in mean absolute SHAP value (0.0331) among all 53 auxiliary dimensions, behind only `tbp_lv_nevi_confidence` (0.0371); UDS log-count (7th, 0.0137) and UDS max absolute deviation (9th, 0.0072) also appear in the top 10, three of the seven UDS dimensions collectively rank among the ten most influential auxiliary features, corroborating the patient-context contribution independently of image features. This per-dimension decomposition also addresses Reviewer 1’s request for an ablation isolating each gate input: SHAP attribution provides a game-theoretically grounded estimate of each UDS component’s individual marginal contribution to the

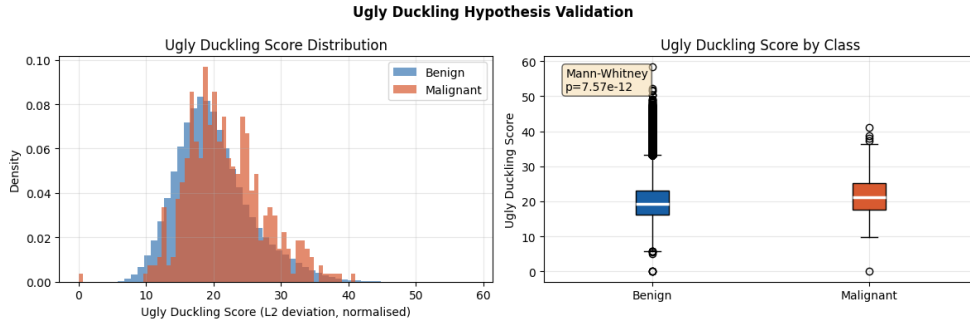


Fig. 3: Ugly duckling hypothesis validation. *Left*: UDS distribution — malignant lesions are broader and right-shifted. *Right*: Boxplot (Mann-Whitney $p = 7.57 \times 10^{-12}$).

model output, which we consider complementary to, and given the seven-dimensional gate input’s compressed dynamic range after Phase 2 fine-tuning (Sec. 5.2) more informative than, a direct leave-one-dimension-out retraining ablation would be at this sample size. Fig. 4 illustrates the full feature ranking.

Training-loss, ROC-AUC/PR-AUC convergence curves, and the FN-rate trajectory across all 38 epochs are provided in the accompanying code repository (see Data and Code Availability); training was stable throughout, with no evidence of overfitting, and the FN rate converges to its final value of 0.3896 (Table 1) by the end of Phase 2.

4.8 Probability Calibration

Isotonic regression calibration improves probability reliability: calibrated ROC-AUC = 0.9554 vs. 0.9493 uncalibrated. Fig. 5 shows calibration curves before and after isotonic regression. The uncalibrated model outputs near-zero probabilities for almost all lesions due to 0.1% malignant prevalence; calibration corrects this bias while preserving discriminative performance.

4.9 Statistical Significance

Table 4 summarises all significance tests, computed over the full validation set (both malignant and benign cases) per Reviewer 3’s request. For Proposed vs. A1: $b = 2102$, $c = 520$, giving $\chi^2 = 953.30$, $p = 2.54 \times 10^{-209}$. For Proposed vs. A3 (no PGAG): $\chi^2 = 2175.07$, $p \approx 0$. The Wilcoxon test confirms the proposed model produces significantly higher malignant scores ($W = 40,015,302$, $p \approx 0$). Restricted to the 77 malignant validation lesions alone this paper’s original framing, these McNemar differences are no longer statistically significant (vs. A1: $b = 18$, $c = 11$, $\chi^2 = 1.24$, $p = 0.27$; vs. A3: $b = 10$, $c = 10$, $\chi^2 = 0.05$, $p = 0.82$), reflecting the limited statistical power of 77 malignant cases once the leakage-inflated separation between models is removed. We report both views for transparency: the all-cases result answers whether the proposed model is a better overall classifier; the malignant-subset result answers the narrower, clinically motivated question of whether it is detectably better

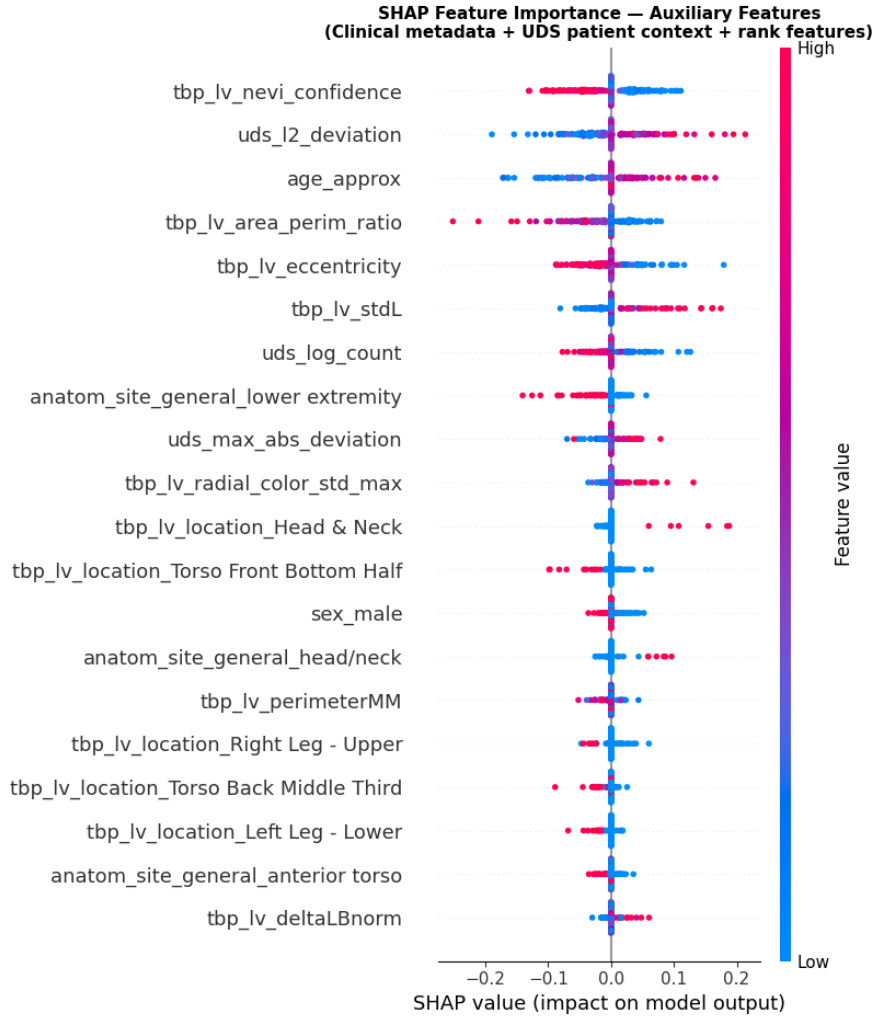


Fig. 4: SHAP summary plot for the 53-dim auxiliary feature branch (UDS and rank features highlighted). UDS L2 deviation ranks 2nd in mean $|\text{SHAP}|$, with two further UDS dimensions in the top 10, validating the patient-context contribution independently of image features.

specifically at catching malignant lesions, which at this sample size we can no longer claim.

4.10 External Validation on HAM10000

External validation on HAM10000 (10,015 dermoscopic images, 11.1% malignant) yields $\text{ROC-AUC} = 0.6476$. Three compounding domain-shift factors account for this result: (1) *modality mismatch* — ISIC 2024 uses 3D TBP while HAM10000 uses

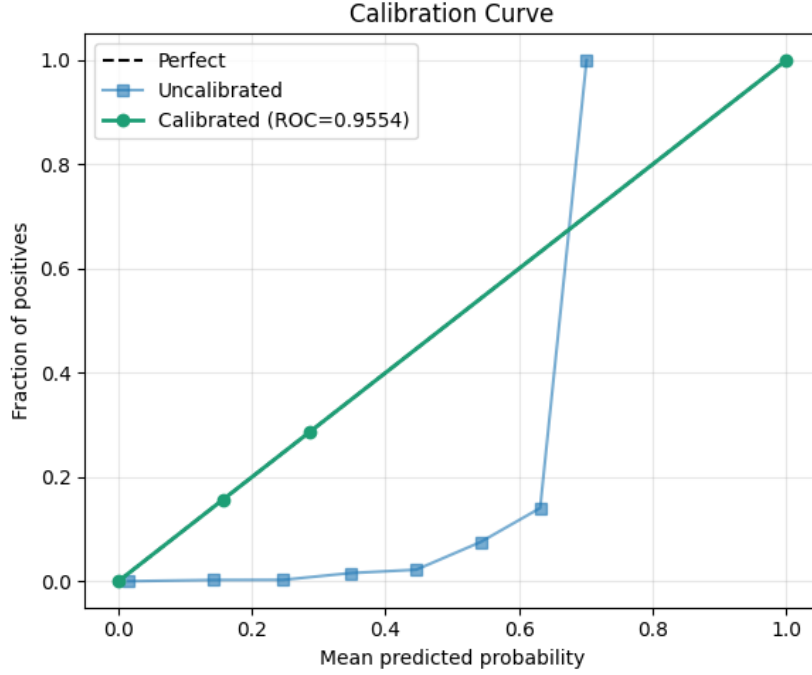


Fig. 5: Calibration curve before and after isotonic regression. The uncalibrated model is severely underconfident at 0.1% prevalence. Post-calibration ROC-AUC = 0.9554.

conventional dermoscopy; (2) *prevalence mismatch* — 11.1% vs. 0.1% (110 $\times$ difference), invalidating the ISIC-tuned threshold; (3) *patient context collapse* 73.8% of HAM10000 patients have a single lesion, rendering all UDS features uninformative. With F1-optimal threshold recalibration, sensitivity reaches 0.5247. Modality-specific domain adaptation is identified as primary future work.

5 Discussion

5.1 Interpretation of Key Results

The 7.8% pAUC improvement over the ISIC 2024 leaderboard winner validates the patient-context approach for screening-scale melanoma detection, and this headline

Table 4: Statistical Significance Tests ($\alpha = 0.05$).

Test	Comparison	Population	Statistic	p-value
McNemar	Proposed vs. A1	All cases	$\chi^2 = 953.30$	2.54×10^{-209}
McNemar	Proposed vs. A1	Malignant only	$\chi^2 = 1.24$	0.27
McNemar	Proposed vs. A3	All cases	$\chi^2 = 2175.07$	≈ 0
McNemar	Proposed vs. A3	Malignant only	$\chi^2 = 0.05$	0.82
Wilcoxon	Proposed vs. A1	All cases	$W = 40,015,302$	≈ 0

result is robust to the leave-one-out correction for self-inclusion leakage in the UDS computation (Sec. 3). The $81.7\times$ PR-AUC lift over random and the 9.1 pp FN rate reduction over the CNN-only baseline demonstrate that combining the ugly duckling signal with asymmetric loss produces measurable, if more modest than an earlier leakage-affected estimate suggested, gains. SHAP analysis confirms that UDS features collectively rank among the most discriminative auxiliary dimensions, providing information not available from single-lesion analysis. The gate’s association with the UDS score ($r = 0.6424$, $p < 10^{-10}$) is statistically robust but, as discussed in Sec. 4, its practical magnitude is small: gate-supervision fine-tuning compresses g_{vit} into a narrow band for both classes, so the PGAG’s contribution to overall performance appears to operate more through the auxiliary-feature classifier pathway, corroborated by the SHAP ranking of UDS dimensions than through large per-lesion shifts in CNN/ViT fusion weight.

5.2 Limitations

Several limitations should be considered when interpreting the results: (1) HAM10000 results demonstrate limited cross-domain transfer due to modality mismatch, prevalence difference, and patient context collapse. The PGAG requires multi-lesion patient cohorts; (2) After Phase 2 fine-tuning, g_{vit} occupies a narrow range for both classes (Sec. 4), likely because the unweighted gate-supervision BCE loss (Eq. 8) is dominated by the majority (benign) class at this dataset’s extreme prevalence. Reweighting this loss term to restore a wider, more interpretable gate dynamic range is identified as a priority for future work; (3) Due to time and compute constraints, we report single-seed results without a dedicated leave-one-dimension-out ablation of the seven UDS inputs; SHAP provides partial evidence. Multi-seed evaluation and gate-input ablation remain future work; (4) The 7 UDS dimensions operate on different numerical scales; per-dimension standardisation is recommended for future work; (5) Isotonic calibration (Sec. 4) is fit and evaluated on the same validation set rather than on a held-out calibration split; a dedicated, untouched calibration-evaluation subset was not available within the revision window and is identified as future work; (6) The leave-one-out UDS formulation (Sec. 3) is undefined for patients with only one lesion, for whom the self-inclusive mean is used as a fallback; UDS is correspondingly less informative for this subgroup, consistent with its uninformativeness on HAM10000 (above), where the majority of patients have a single lesion.

6 Conclusion

We presented a hybrid CNN-ViT architecture with a Patient-Guided Attention Gate that operationalises the clinical ugly duckling sign as a learnable, leakage-free (leave-one-out) conditioning signal for feature fusion. The proposed model achieves pAUC = 0.1892 on ISIC 2024 SLICE-3D, exceeding the competition winner by 7.8%, with a 9.1 pp FN rate reduction over CNN-only and significantly higher overall accuracy across the full validation set, confirmed by McNemar ($p = 2.54 \times 10^{-209}$) and Wilcoxon ($p \approx 0$) tests; restricted to the 77 malignant validation lesions alone, this difference is not statistically significant (Sec. 4). SHAP analysis validates UDS L2 deviation as the second most important auxiliary feature overall. Gate analysis ($r = 0.6424$,

$p < 10^{-10}$) confirms the PGAG's response to patient-context deviation is statistically robust, though its practical dynamic range narrows substantially after gate-supervision fine-tuning (Sec. 5.2). Isotonic calibration further improves reliability (ROC-AUC = 0.9554).

Future work will pursue a reweighted gate-supervision loss to restore a wider, more interpretable gate range, multi-seed robustness estimation, modality-specific domain adaptation, and UDS feature normalisation.

Ethics Statement

This research uses the ISIC 2024 SLICE-3D dataset, a publicly available and fully de-identified dataset released under a CC BY-NC licence. No patient data were collected and no additional IRB approval was required. Dataset available at <https://challenge.isic-archive.com/data/#2024>.

Data and Code Availability

The ISIC 2024 dataset is publicly available at <https://challenge.isic-archive.com/data/#2024>. The experimental notebook is available at <https://github.com/kamrulkafi1/pgag-melanoma-isic2024>. Trained model checkpoints (main-training and gate-supervision phases, including intermediate epoch checkpoints) and pre-extracted ViT-B/16 feature files are available at <https://doi.org/10.5281/zenodo.21986973> (CC BY 4.0).

References

1. A medical cyber-physical system utilizing the bayes algorithm for post-diagnosis patient supervision. *AJSE* **23**(2), 84–91 (Aug 2024). <https://doi.org/10.53799/ajse.v23i2.635>
2. Ahmed, S., Hossain, M.F., Nur, S.B., Shamim Kaiser, M., Mahmud, M.: Toward Machine Learning-Based Psychological Assessment of Autism Spectrum Disorders in School and Community. *Proc. TEHI 2021* **376**, 139 – 149 (2022)
3. Ahmed, S., Rahman, M.S., Jahan, I., Kaiser, M.S., Hosen, A.S.M.S., Ghimire, D., Kim, S.H.: A Review on the Detection Techniques of Polycystic Ovary Syndrome Using Machine Learning. *IEEE Access* **11**, 86522 – 86543 (2023)
4. Al Mamun, S., Kaiser, M.S., Mahmud, M.: An Artificial Intelligence Based Approach Towards Inclusive Healthcare Provisioning in Society 5.0: A Perspective on Brain Disorder. *Proc. BI 2021* **12960 LNAI**, 157 – 169 (2021)
5. Alam, M.E., Shamim Kaiser, M., Hossain, M.S., Andersson, K.: An IoT-belief rule base smart system to assess autism. In: *Proc. iCEEiCT 2018*. pp. 672 – 676 (2018)
6. Ali, H.M., Kaiser, M.S., Mahmud, M.: Application of Convolutional Neural Network in Segmenting Brain Regions from MRI Data. *Proc. BI 2019* **11976 LNAI**, 136 – 146 (2019)
7. Chen, T., Su, P., Shen, Y., Chen, L., Mahmud, M., Zhao, Y., Antoniou, G.: A dominant set-informed interpretable fuzzy system for automated diagnosis of dementia. *Front. Neurosci.* **16** (2022)

8. Deepa, B., Murugappan, M., Sumithra, M., Mahmud, M., Al-Rakhami, M.S.: Pattern Descriptors Orientation and MAP Firefly Algorithm Based Brain Pathology Classification Using Hybridized Machine Learning Algorithm. *IEEE Access* **10**, 3848 – 3863 (2022)
9. Dhara, T., Singh, P.K., Mahmud, M.: A Fuzzy Ensemble-Based Deep learning Model for EEG-Based Emotion Recognition. *Cogn. Comput.* **16**(3), 1364 – 1378 (2024)
10. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *Proc. Int. Conf. Learning Representations (ICLR)* (2021), <https://arxiv.org/abs/2010.11929>
11. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**(7639), 115–118 (2017). <https://doi.org/10.1038/nature21056>
12. Fabietti, M., Mahmud, M., Lotfi, A., Avarua, A., Gugganmos, D., Xudo, R., Chiappalone, M.: Neural Network-based Artifact Detection in Local Field Potentials Recorded from Chronically Implanted Neural Probes. In: *Proc. IJCNN 2020* (2020)
13. Ghosh, S., Kim, S., Ijaz, M.F., Singh, P.K., Mahmud, M.: Classification of Mental Stress from Wearable Physiological Sensors Using Image-Encoding-Based Deep Neural Network. *Biosensors* **12**(12) (2022)
14. Giavina-Bianchi, M., Vitor, W.G., de Paiva, V.F., Okita, A.L., Sousa, R.M., Machado, B.: Explainability agreement between dermatologists and five visual explanation techniques in deep neural networks for melanoma ai classification. *Frontiers in Medicine* **10**, 1241484 (2023)
15. Grob, J.J., Bonerandi, J.J.: The ‘ugly duckling’ sign: identification of the common characteristics of nevi in an individual as a basis for melanoma screening. *Archives of Dermatology* **134**(1), 103–104 (1998). <https://doi.org/10.1001/archderm.134.1.103>
16. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. pp. 1321–1330 (2017)
17. Hasan, N., et al.: ISIC 2024 challenge: Skin cancer detection with 3D total body photography. *arXiv preprint arXiv:2501.02389* (2025), <https://arxiv.org/abs/2501.02389>
18. Hoque, M.Z., Dey, A., Afrin, S., Khaliluzzaman, M.: A machine learning data-driven approach for predicting cardiovascular disease. In: *2024 27th International Conference on Computer and Information Technology (ICCIT)*. pp. 3110–3115 (2024). <https://doi.org/10.1109/ICCIT64611.2024.11022090>
19. Islam, M., Uddin, M.A., Khan, A.A., Afrin, S., Barua, D., Islam, M.M.: Xai-transnet: Blood cell identification using deep transfer learning and explainable ai techniques. In: *2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS)*. pp. 1–6 (2025). <https://doi.org/10.1109/COMPAS67506.2025.11381641>
20. Jahan, S., Nowsheen, F., Antik, M.M., Rahman, M.S., Kaiser, M.S., Hosen, A.S.M.S., Ra, I.H.: AI-Based Epileptic Seizure Detection and Prediction in Internet of Healthcare Things: A Systematic Review. *IEEE Access* **11**, 30690 – 30725 (2023)
21. Javed, A.R., Saadia, A., Mughal, H., Gadekallu, T.R., Rizwan, M., Maddikunta, P.K.R., Mahmud, M., Liyanage, M., Hussain, A.: Artificial Intelligence for Cognitive Health Assessment: State-of-the-Art, Open Challenges and Future Directions. *Cogn. Comput.* **15**(6), 1767 – 1812 (2023)
22. Jesmin, S., Kaiser, M.S., Mahmud, M.: Artificial and Internet of Healthcare Things Based Alzheimer Care During COVID 19. *Proc. BI 2020* **12241 LNAI**, 263 – 274 (2020)
23. Kawahara, J., et al.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics* **23**(2), 538–546 (2019). <https://doi.org/10.1109/JBHI.2018.2824327>
24. Kurtansky, N.R., Primiero, C.A., et al.: Effect of patient-contextual skin images in human- and artificial intelligence-based diagnosis of melanoma: Results from the 2020

- siim-istic melanoma classification challenge. *Journal of the European Academy of Dermatology and Venereology : JEADV* **39**, 1489 – 1499 (2024)
25. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proc. IEEE Int. Conf. Computer Vision (ICCV)*. pp. 2980–2988 (2017). <https://doi.org/10.1109/ICCV.2017.324>
 26. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 30 (2017), <https://arxiv.org/abs/1705.07874>
 27. Mahmud, M., Vassanelli, S.: Processing and analysis of multichannel extracellular neuronal signals: State-of-the-art and challenges. *Front. Neurosci.* **10**(JUN) (2016)
 28. Mahmud, T., Barua, K., Barua, A., Das, S., Basnin, N., Hossain, M.S., Andersson, K., Kaiser, M.S., Sharmen, N.: Exploring Deep Transfer Learning Ensemble for Improved Diagnosis and Classification of Alzheimer’s Disease. *Proc. BI 2023* **13974 LNAI**, 109 – 120 (2023)
 29. Mammoottil, M.J., Kulangara, L.J., Cherian, A.S., Mohandas, P., Hasikin, K., Mahmud, M.: Detection of Breast Cancer from Five-View Thermal Images Using Convolutional Neural Networks. *J. Healthc. Eng.* **2022** (2022)
 30. McClish, D.K.: Analyzing a portion of the ROC curve. *Medical Decision Making* **9**(3), 190–195 (1989). <https://doi.org/10.1177/0272989X8900900307>
 31. Miah, Y., Prima, C.N.E., Seema, S.J., Mahmud, M., Shamim Kaiser, M.: Performance Comparison of Machine Learning Techniques in Identifying Dementia from Open Access Clinical Datasets. *Adv. Intell. Syst. Comput.* **1188**, 79 – 89 (2021)
 32. Mukherjee, H., Sreerama, P., Dhar, A., Obaidullah, S.M., Roy, K., Mahmud, M., Santosh, K.: Automatic Lung Health Screening Using Respiratory Sounds. *J. Med. Syst.* **45**(2) (2021)
 33. Murugappan, M., Prakash, N., Jeya, R., Mohanarathinam, A., Hemalakshmi, G., Mahmud, M.: A novel few-shot classification framework for diabetic retinopathy detection and grading. *Meas.: J. Int. Meas. Confed.* **200** (2022)
 34. Niamat Ullah Akhund, T.M., Mahi, M.J.N., Hasnat Tanvir, A., Mahmud, M., Kaiser, M.S.: ADEPTNESS: Alzheimer’s disease patient management system using pervasive sensors - early prototype and preliminary results. *Proc. BI 2018* **11309 LNAI**, 413 – 422 (2018)
 35. Nie, Y., Sommella, P., Carratù, M., O’Nils, M., Lundgren, J.: A deep cnn transformer hybrid model for skin lesion classification of dermoscopic images using focal loss. *Diagnostics* **13**(1), 72 (2023)
 36. Oktay, O., et al.: Attention U-Net: Learning where to look for the pancreas. In: *Proc. Medical Imaging with Deep Learning (MIDL)* (2018), <https://arxiv.org/abs/1804.03999>
 37. Pacheco, A.G.C., Krohling, R.A.: The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine* **116**, 103543 (2020). <https://doi.org/10.1016/j.compbio.2019.103543>
 38. Pacheco, A.G.C., Krohling, R.A.: An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. *IEEE Journal of Biomedical and Health Informatics* **25**(9), 3554–3563 (Sep 2021)
 39. Rahman, M.A., Paul, B., Mahmud, T., Fattah, S.A.: Ciff-net: Contextual image feature fusion for melanoma diagnosis. *arXiv preprint arXiv:2303.03672* (2023)
 40. Rotemberg, V., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data* **8**, 34 (2021)
 41. Skin, L., et al.: Multimodal skin lesion classification using convolutional and vision transformer fusion. *Computers in Biology and Medicine* **157**, 106757 (2023). <https://doi.org/10.1016/j.compbio.2023.106757>

42. Sumi, A.I., et al.: fASSERT: A fuzzy assistive system for children with autism using internet of things. *Proc. BI 2018* **11309 LNAI**, 403 – 412 (2018)
43. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. pp. 6105–6114. Long Beach, CA, USA (2019)
44. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5**(1), 180161 (2018). <https://doi.org/10.1038/sdata.2018.161>
45. Ullah Miah, M.S., Sulaiman, J., Sarwar, T.B., Islam, S.S., Rahman, M., Haque, M.S.: Medical named entity recognition (medner): A deep learning model for recognizing medical entities (drug, disease) from scientific texts. In: *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*. pp. 158–162 (2023). <https://doi.org/10.1109/EUROCON56442.2023.10199075>
46. Wightman, R.: PyTorch image models (2019), <https://github.com/rwightman/pytorch-image-models>