STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–19.
Published online in International Academic Press (www.IAPress.org)

Benchmarking Attention Mechanisms in ConvNeXt-Tiny for Dermoscopic Skin Lesion Classification Under Stratified and Lesion-Level Leakage-Safe Cross-Validation.

Asfan Muqtadir [1,3,*], Kusworo Adi [2], Vincencius Gunawan [2]

[1]*Doctoral Program of Information Systems, School of Postgraduate Studies, Universitas Diponegoro, Indonesia*
[2]*Department of Physics, Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia*
[3]*Department of Informatics Engineering, Faculty of Engineering, Universitas PGRI Ronggolawe, Indonesia*

Abstract Attention mechanisms are widely added to convolutional backbones to improve skin lesion classification, yet their comparative benefit is rarely assessed under evaluation protocols that prevent lesion-level data leakage. We present a controlled benchmark of four canonical attention modules—Channel Attention, Spatial Attention, the Convolutional Block Attention Module (CBAM), and Coordinate Attention—each inserted as a post-backbone plug-in on a shared ConvNeXt-Tiny feature extractor for seven-class dermoscopic classification on the HAM10000 dataset. Every configuration is evaluated twice: under conventional image-level stratified five-fold cross-validation, and under lesion-aware group five-fold cross-validation that ensures all images sharing the same `lesion_id` remain within a single fold. Across all configurations, replacing stratified with group validation reduces the macro-averaged F1-score by an average of 10.82 percentage points, an inflation concentrated mainly in the minority classes—the rarest class loses more than 22 points of F1—and shows an exploratory negative association with class rarity ($r = -0.66$). The ranking of attention mechanisms is not preserved across protocols: Channel Attention attains the highest observed macro-F1 under image-level stratified validation, whereas Spatial Attention attains the highest observed macro-F1 under lesion-level group validation and is the only attention mechanism with a numerically higher mean than the corresponding attention-free baseline. However, none of the eight fold-level comparisons against the protocol-matched baselines remains statistically significant after Holm correction. Peak GPU-memory requirements are broadly comparable across configurations, although Coordinate Attention requires moderately longer training under lesion-level group validation. These results show that the cross-validation protocol influences reported performance more strongly than the choice of attention mechanism, and that image-level evaluation on HAM10000 can provide optimistic estimates of generalisation to unseen lesions. We recommend lesion-level group partitioning as the preferred protocol for trustworthy evaluation of skin lesion classifiers.

Keywords Skin lesion classification, HAM10000, attention mechanisms, ConvNeXt, lesion-level data leakage, cross-validation, dermoscopy, deep learning

DOI: 10.19139/soic-2310-5070-4134

1. Introduction

Melanoma accounts for only a small proportion of all confirmed skin cancer cases, yet causes the vast majority of skin cancer-related deaths [1]. Detecting and identifying lesions early therefore matters a great deal for patient prognosis. In practice, however, reading dermoscopy images still depends heavily on the physician's experience, and different experts often disagree [2]. Deep learning has pushed this field forward, making it possible to build automated tools that are both objective and scalable [3]. Most of this progress has been benchmarked

*Correspondence to: Asfan Muqtadir (Email: asfanme@unirow.ac.id). Doctoral Program of Information Systems, School of Postgraduate Studies, Universitas Diponegoro, Pleburan, Central Java, Indonesia (50241).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

on HAM10000 [4], a widely used dataset in the field [5]: it contains 10,015 dermoscopic images across seven diagnostic categories. Two core problems remain, though. The first is severe class imbalance — melanocytic nevi make up more than 66% of the samples, while rare categories such as dermatofibroma fall below 2%. The second is data leakage. When several images of the same lesion from the same patient are split carelessly across folds, the model is partly tested on what it has already seen, and its scores end up inflated [6]. Yet most existing studies still rely on traditional stratified cross-validation, without isolating data at the patient or lesion level — which leaves the generalisability of their reported results in doubt.

In recent years, rapid advances in computer-vision architectures have reshaped how deep learning develops for medical imaging. Vision Transformers (ViTs) [7] proved effective at modeling global attention features, and inspired a wave of transformer-style designs. ConvNeXt [8] is one such design: it modernizes the convolutional network while keeping its efficiency, and its lightweight ConvNeXt-Tiny variant suits resource-constrained medical scenarios well. The model has one clear limitation, though — during inference, it treats all feature channels and spatial positions equally. Four attention mechanisms can compensate for this: channel attention (CA) [9], spatial attention (SA) [10], the convolutional block attention module (CBAM) [10], and coordinate attention (CoordAtt) [11]. Prior work [12] has shown that embedding such modules improves medical image classification. Even so, how these lightweight modules behave when placed after the backbone of a modern convolutional network is still not fully understood — and for dermoscopy classification under leakage-safe evaluation, relevant studies are scarcer still.

Driven by existing challenges in dermoscopic skin lesion classification, this study proposes a clear benchmarking framework for attention-augmented architectures, developed on the HAM10000 dataset with ConvNeXt-Tiny as the backbone. The framework embeds four attention mechanisms—CA, SA, CBAM, and CoordAtt—and isolates the independent contribution of each mechanism under uniform training conditions.

This study is positioned as a methodological benchmarking study rather than as a proposal of a new attention architecture. It compares conventional image-level stratified cross-validation with lesion-aware group cross-validation to determine how lesion-level overlap affects reported performance and the relative ranking of attention mechanisms. The main contributions are threefold: (i) a controlled post-backbone comparison of four canonical attention mechanisms using an identical backbone and training configuration; (ii) direct quantification of the performance gap between image-level and lesion-level validation; and (iii) analysis of how the evaluation protocol affects minority-class performance and model selection.

1.1. Convolutional Backbone Architectures for Medical Imaging

For a long time, convolutional neural networks (CNNs) were the default choice for medical image classification. The lineage is familiar — AlexNet [13], VGGNet [14], GoogLeNet/Inception [15], ResNet [16], DenseNet [17], and EfficientNet [18] — each pushing feature representation and efficiency a little further, and each finding its way into dermoscopic analysis through transfer learning. EfficientNet, for example, performed well on HAM10000 by scaling depth, width, and input resolution together. Newer dermoscopic pipelines still lean on these backbones: some fuse localized, contextual, and hierarchical features [3], while others combine images with patient metadata for multimodal classification [19].

A different idea emerged with Vision Transformers (ViTs) [7], which models global structure through self-attention; hierarchical versions, such as the Swin Transformer [20], then improved scalability. Recent work published in SOIC has likewise investigated Vision Transformers and attention mechanisms for mammographic image classification, illustrating the growing role of attention-based global representation in medical image analysis [35]. The catch is cost—transformers usually require more data and more compute, which limits their use in many medical settings.

ConvNeXt [8] took a middle path. It borrows several design ideas from transformers but keeps the efficiency and inductive biases that make CNNs practical. With refinements such as large-kernel depthwise convolutions, inverted bottlenecks, LayerNorm, and GELU activations, it matches modern transformers across several vision benchmarks. Its lightweight ConvNeXt-Tiny variant strikes a good balance between accuracy and cost, which makes it a sensible fit for dermoscopic classification. Recent attention-augmented ConvNeXt models have also shown promise on both skin lesion and brain tumor tasks [21, 22, 23] — which is why we chose it as the backbone

here. The practical importance of computationally efficient medical-image classifiers is also reflected in recent SOIC research, including a lightweight CNN developed to balance predictive performance and computational requirements in histopathological lung-cancer classification [37].

1.2. Attention Mechanisms in Convolutional Networks

Not every feature in a convolutional map is equally useful. Attention mechanisms exploit this idea: they amplify the responses that matter and dampen the ones that do not. In CNNs, this reweighting is usually applied along one of three axes—channel, spatial, or a hybrid of both.

Channel Attention (CA) operates along the channel axis, learning the relative importance of each channel. The Squeeze-and-Excitation (SE) framework [9] introduced this idea and showed that even a lightweight channel recalibration can lift representation quality at almost no extra cost.

Spatial Attention (SA) asks a different question—not which channel, but which region. It builds a spatial map over the feature tensor to mark the discriminative areas. In Woo *et al.*'s formulation [10], channel information is first pooled by average and max operations, then turned into a spatial mask that steers feature refinement. Related lightweight variants, such as the bottleneck attention module, run a comparable spatial branch alongside channel attention instead of after it [24].

CBAM [10] simply does both, one after the other: it refines channels first, then space, so the two cues complement each other. Its mix of simplicity and effectiveness has made it one of the most widely used attention modules in medical image analysis, and lightweight channel–spatial variants continue to appear for this domain [25, 26].

Coordinate Attention (CoordAtt) [11] takes channel attention a step further by folding in directional position. It encodes the horizontal and vertical axes separately, so unlike plain global pooling, it keeps track of where features sit while staying cheap to compute — a useful trait for lightweight models.

These four mechanisms cover genuinely different attention strategies, which is exactly what makes them a fair set to compare within one ConvNeXt-Tiny framework.

1.3. Attention-Augmented Networks for Skin Lesion Classification

Attention has found its way into more and more dermoscopic classification systems, mostly to sharpen feature discrimination and pinpoint the lesion. Several recent studies report that pairing attention modules with CNN or hybrid backbones lifts performance, largely by foregrounding the patterns that matter diagnostically [22, 27, 21]. The same holds when attention is wired into hybrid segmentation–classification pipelines for skin lesions [5], and when channel and spatial attention are embedded across several CNN backbones for medical imaging at large [12]. Related work published in SOIC has also demonstrated the value of controlled multi-model evaluation in medical imaging. Ali and Sadeeq evaluated five pretrained CNN backbones within a unified transfer-learning pipeline using stratified five-fold cross-validation, focal loss for multi-class dental-disease classification [36]. Although conducted in a different medical-imaging domain, this study illustrates the importance of holding the training and evaluation settings constant when comparing alternative architectures.

This paper sorts out two core gaps in existing research on AI attention mechanisms in clinical contexts: First, most current studies test only pre-packaged, integrated attention designs and fail to disentangle the independent contributions of individual mechanisms. Second, commonly used evaluation protocols cannot prevent patient-level data leakage, leaving the academic community unable to clearly ascertain the actual performance advantages and disadvantages of different mechanisms.

To the best of our knowledge, no previous study has systematically compared Channel Attention, Spatial Attention, CBAM, and Coordinate Attention as post-backbone modules within a unified ConvNeXt-Tiny framework while simultaneously examining leakage-safe and leakage-unaware validation protocols on HAM10000.

1.4. Cross-Validation Strategies and Data Leakage in HAM10000

Cross-validation is widely used to estimate model generalisation performance in medical imaging. Stratified k-fold cross-validation preserves class distributions across folds and is therefore commonly adopted for the highly imbalanced HAM10000 dataset. However, stratified splitting operates at the sample level and ignores the grouping structure created by multiple images originating from the same lesion or patient.

Because HAM10000 contains visually related images acquired from identical lesions under different conditions, random fold assignment can place highly similar samples in both training and evaluation sets. This introduces a subtle form of data leakage that may lead to overly optimistic performance estimates and reduced external validity [6, 28]. Such leakage is now recognised as a recurring source of methodological failure in medical imaging machine learning [29], and has been shown to inflate even meta-analytic estimates of predictive performance [30], prompting systematic efforts to catalogue and mitigate its many forms across the learning pipeline [31].

Group k-fold cross-validation addresses this problem by keeping each image from a given lesion or patient in a single fold [32]. The trade-off is that estimates become more conservative and vary more from fold to fold — but in return they reflect what the model would actually do on new patients.

Scenarios. In this study, we use both lesion-aware group k-fold validation and stratified validation to evaluate all attention configurations in parallel. We measure the optimistic bias caused by data leakage using the performance gap between the two methods, providing a reliable basis for comparing attention mechanisms for dermoscopic image classification.

2. Methodology

This section lays out the full experimental setup used to benchmark the four attention mechanisms on the HAM10000 dermoscopy dataset. Figure 1 provides an overview of the pipeline, tracing the data from raw images through feature extraction, attention modulation, and classification, and showing the two cross-validation protocols. We then describe the dataset, data preparation, cross-validation procedures, the attention-augmented ConvNeXt-Tiny architecture, training configuration, and evaluation metrics. The leakage-safe protocol is defined at the lesion level using the `lesion_id` field available in the HAM10000 metadata.

2.1. Dataset Description

The HAM10000 (*Human Against Machine with 10000 training images*) dataset [4] is a large-scale, multi-source collection of dermoscopic images designed to benchmark machine learning models for pigmented skin lesion analysis. The dataset comprises 10,015 images gathered from two clinical centers — the Medical University of Vienna and the Rosendahl clinic in Queensland, Australia — using different imaging devices and acquisition protocols, thereby introducing inter-site variability that enhances the benchmark's representativeness.

Each image is annotated with one of seven diagnostic categories: actinic keratosis and intraepithelial carcinoma (*akiec*), basal cell carcinoma (*bcc*), benign keratosis-like lesions (*bkl*), dermatofibroma (*df*), melanoma (*mel*), melanocytic nevi (*nv*), and vascular lesions (*vasc*). The class distribution is severely imbalanced: melanocytic nevi account for 6,705 images (66.95%) while the rarest category, dermatofibroma, contains only 115 images (1.15%), motivating the use of Focal Loss during training and macro-averaged F1-score as the primary evaluation metric.

A structurally important — and often overlooked — property of HAM10000 is the presence of multiple images per lesion. A single lesion may be captured from different angles, under varying lighting conditions, or across multiple clinical visits, resulting in clusters of visually near-identical images sharing the same label and patient identity. Across the full dataset, 10,015 images correspond to only 7,470 unique lesions, implying that 2,545 images (25.41%) are duplicate views of lesions already represented elsewhere in the dataset. Table 1 reports the per-class breakdown of this structure, including the number of unique lesions, the number of duplicate images, and the average images-per-lesion ratio.

The duplication rate varies considerably across classes, as illustrated in Figure 2(b) and quantified in Table 1: melanoma (*mel*) exhibits the highest images-per-lesion ratio of 1.81, indicating that nearly half of all melanoma images are duplicate views of already-represented lesions, while melanocytic nevi (*nv*) — despite being the largest

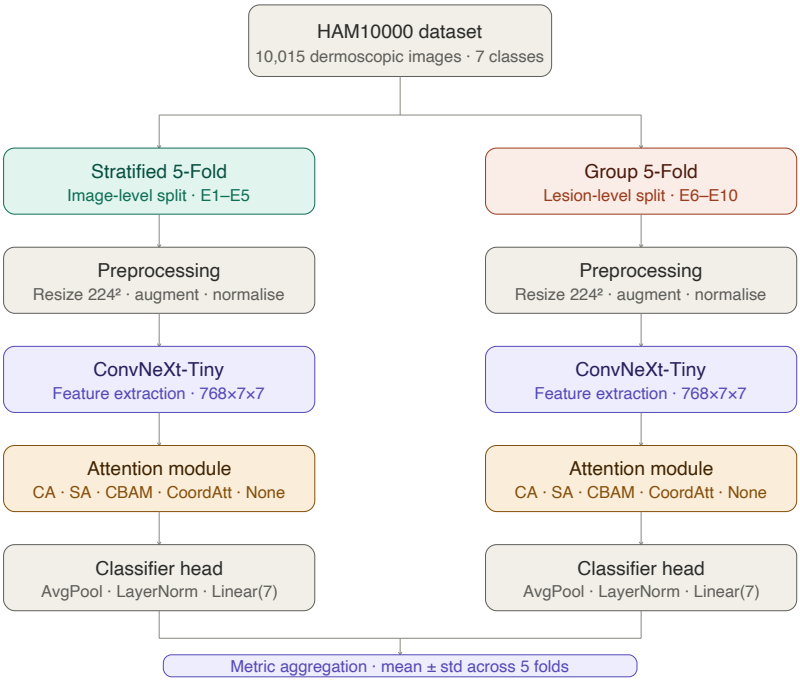

Figure 1. Overview of the experimental pipeline used to evaluate attention-enhanced ConvNeXt-Tiny models under image-level stratified and lesion-level leakage-safe validation protocols.

Table 1. Lesion-level statistics of HAM10000. Duplicates are images sharing a `lesion_id` with at least one other; Images/Lesion is the total-to-unique-lesion ratio per class.

Class	Images	Unique Lesions	Duplicates	Images/Lesion
akiec	327	228	99	1.43
bcc	514	327	187	1.57
bkl	1099	727	372	1.51
df	115	73	42	1.58
mel	1113	614	499	1.81
nv	6705	5403	1302	1.24
vasc	142	98	44	1.45
Total	**10015**	**7470**	**2545**	**1.34**

class by far — has the lowest ratio at 1.24. This heterogeneity implies that the leakage problem is not uniformly distributed across classes; a standard stratified split is particularly prone to leaking melanoma images between folds, which is especially problematic given that melanoma is the clinically most critical category. The `lesion_id` field in the HAM10000 metadata uniquely identifies each physical lesion and is therefore used as the grouping variable for constructing the lesion-level Group 5-fold cross-validation protocol described in Section 2.2.

2.2. Data Preparation and Cross-Validation Protocol

Image loading and class encoding. All images are loaded from disk in RGB format using PIL and mapped to integer class indices according to the ordering: *akiec*=0, *bcc*=1, *bkl*=2, *df*=3, *mel*=4, *nv*=5, *vasc*=6. Image paths are constructed at runtime by concatenating the base data directory, the class-named subdirectory, and the `image_id` field with a `.jpg` extension, following the standard HAM10000 directory structure.

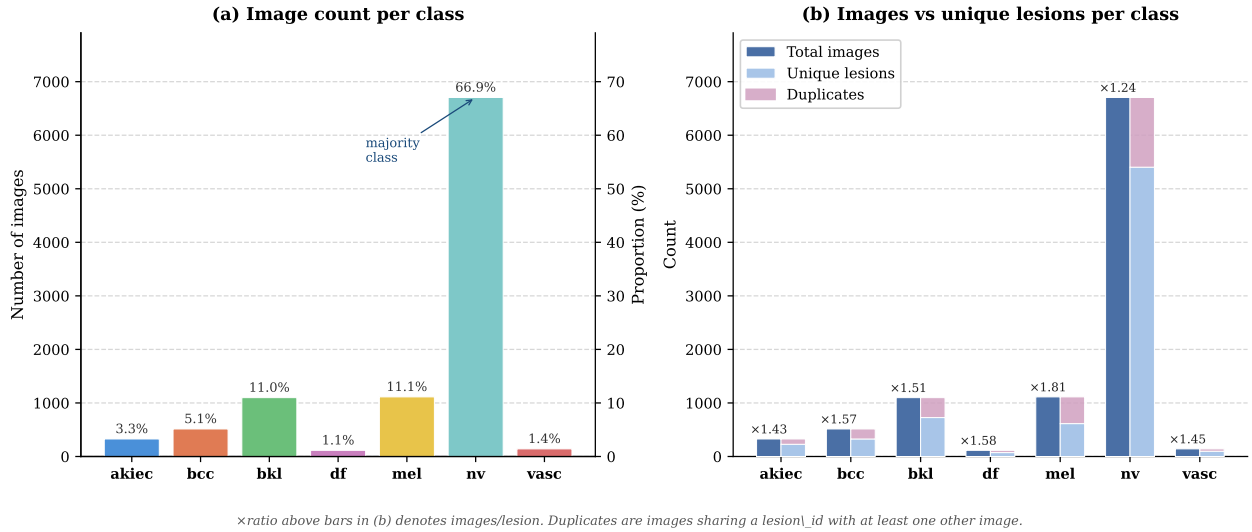

Figure 2. Class distribution of the HAM10000 dataset. (a) Total image count per class with proportion (%). (b) Grouped comparison of total images versus unique lesions, with duplicate images shown as stacked portions. The ×ratio above each bar in (b) denotes the images-per-lesion ratioper class.

Preprocessing and augmentation. Two distinct transformation pipelines are defined for the training and validation subsets. The training pipeline applies the following sequence: (i) resize to 224×224 pixels; (ii) random horizontal flip with probability 0.5; (iii) random vertical flip with probability 0.5; (iv) random rotation by a uniformly sampled angle in $[-20°, +20°]$; (v) conversion to tensor; and (vi) normalisation using the ImageNet channel statistics (mean $= [0.485, 0.456, 0.406]$, std $= [0.229, 0.224, 0.225]$). The validation pipeline omits all stochastic augmentations and applies only steps (i), (v), and (vi), ensuring that evaluation metrics are computed on unmodified images. The geometric augmentations — horizontal flip, vertical flip, and rotation — were selected because dermoscopic images carry no canonical orientation, and the corresponding lesion appearance is invariant to these transformations.

The resize operation maps every image directly to 224×224 pixels using the default bilinear interpolation of `torchvision.transforms.Resize`. Consequently, the original aspect ratio is not explicitly preserved. This transformation was applied identically to all ten configurations to maintain a controlled comparison. Random rotation used the default settings of `torchvision.transforms.RandomRotation`, with no expansion and zero-valued fill outside the rotated image. The potential effect of aspect-ratio distortion on absolute classification performance is acknowledged as a limitation.

Cross-validation protocols. Two cross-validation schemes are implemented and applied independently to the same set of model configurations.

The first, *Stratified 5-Fold* (S-KFold), uses `StratifiedKFold` [32] with `n_splits=5`, `shuffle=True`, and `random_state=42`. Stratification on the class label vector preserves the seven-class distribution in each fold. This protocol mirrors the most common evaluation practice in the HAM10000 literature and is replicated here to enable direct comparison with published results.

The second, *Lesion-Level Group 5-Fold* (G-KFold), uses `StratifiedGroupKFold` [32] with `n_splits=5`, `shuffle=True`, and `random_state=42`, with the `lesion_id` column as the grouping variable. This ensures that all images belonging to the same lesion are assigned to a single fold, preventing any image in the validation set from sharing the same lesion with an image in the training set. The algorithm enforces non-overlapping lesion groups while approximating the class distribution across folds; consequently, exactly identical class proportions are not guaranteed when the grouping and stratification constraints conflict.

Let $\mathcal{G}$ be the set of unique lesion identifiers and let $g(x)$ denote the lesion identifier of sample x. The validation partition at fold t is defined as

$$\mathcal{D}_{\text{val}}^{(t)} = \{\, x \in \mathcal{D} \mid g(x) \in \mathcal{G}_t \,\}, \tag{1}$$

where $\{\mathcal{G}_1, \ldots, \mathcal{G}_5\}$ is a disjoint partition of $\mathcal{G}$. This construction eliminates the lesion-level data leakage described in Section 1.4 and produces performance estimates that generalise to entirely unseen lesions. A fixed `random_state=42` was used to generate the partitions for both protocols.

2.3. Experimental Framework

A total of ten experimental configurations are evaluated, organised into two groups of five according to the cross-validation protocol, as summarised in Table 2. Within each group, one configuration uses the ConvNeXt-Tiny backbone without any attention module (baseline), and four configurations augment the same backbone with one of the four attention mechanisms under study. The backbone architecture, classifier head, loss function, optimiser, preprocessing pipeline, and all training hyperparameters are held constant across all ten configurations; the only experimental factors are the attention module and the cross-validation protocol.

Table 2. Experimental configurations evaluated in this study.

Attention Module	Stratified 5-Fold	Lesion-Level Group 5-Fold
Baseline	✓	✓
CA	✓	✓
SA	✓	✓
CBAM	✓	✓
CoordAtt	✓	✓

The complete training configuration, loss function, and resource profiling procedures are described in Section 2.6.

2.4. ConvNeXt-Tiny Backbone

ConvNeXt-Tiny [8] serves as the shared feature extractor across all ten configurations. The model is initialised with ImageNet-1K pretrained weights (`ConvNeXt_Tiny_Weights.DEFAULT`) and its `features` module — comprising four hierarchical stages with progressive channel expansion from 96 to 768 dimensions — is used as the backbone. The backbone is not frozen; all parameters including those of the pretrained stages are updated during training, allowing the network to adapt its representations to the dermoscopy domain.

The terminal output of the `features` module is a feature map $\mathbf{F} \in \mathbb{R}^{768\times7\times7}$ for a 224×224 input, which constitutes the input to the attention module. After the attention module, an `AdaptiveAvgPool2d(1)` layer collapses the spatial dimensions to produce a 768-dimensional vector, which is forwarded to the classifier head comprising a `LayerNorm`(768) layer followed by a linear projection to the seven output logits. This architecture is illustrated in Figure 3.

ImageNet-1K pretraining was used to provide a common and well-established initialisation for all configurations, thereby avoiding differences caused by training each variant from randomly initialised weights. The same pretrained backbone was used in every experiment. Although natural-image pretraining is widely used in medical-image transfer learning, domain-specific pretraining may produce different absolute results and is outside the controlled scope of this benchmark.

2.5. Attention Modules

Each attention module is inserted as a post-backbone plug-in between the output of `convnext.features` and the global average pooling layer, operating on the feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, where $C = 768$ and $H = W = 7$ for a 224×224 input. The module produces a recalibrated map $\widetilde{\mathbf{F}}$ of the same dimensions, which is then passed to the pooling and classifier stages unchanged. The baseline configuration (Exp. 1 and 6) omits the attention module

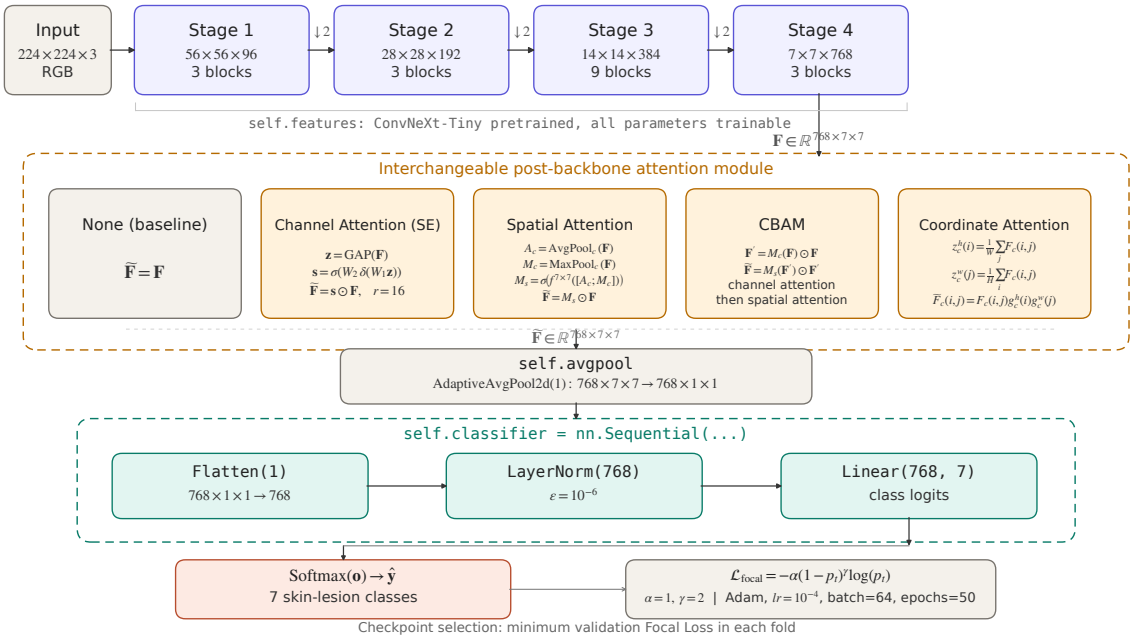

Figure 3. Architecture of the attention-augmented ConvNeXt-Tiny model. Feature maps extracted by the pretrained backbone are processed by one of four attention mechanisms prior to global average pooling and linear classification. The dashed block indicates the attention component evaluated in this study.

entirely, passing $\mathbf{F}$ directly to global average pooling. All attention modules were evaluated at the same post-backbone location under an identical training configuration. Canonical module settings were used without module-specific hyperparameter optimisation to preserve a controlled comparison.

2.5.1. Channel Attention Channel Attention (CA) recalibrates the feature map along the channel dimension by learning a vector of per-channel importance weights from the global statistics of $\mathbf{F}$. In this study, CA was implemented as a Squeeze-and-Excitation (SE) block following the formulation of Hu et al. [9]. A global average pooling (squeeze) operation first aggregates spatial information into a channel descriptor $\mathbf{z} \in \mathbb{R}^C$:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_c(i,j), \quad c = 1, \ldots, C. \tag{2}$$

An excitation network—two bias-free fully connected layers with a bottleneck ratio of $r = 16$ and ReLU/sigmoid activations—then maps $\mathbf{z}$ to a channel-wise scaling vector:

$$\mathbf{s} = \sigma(W_2\,\delta(W_1\,\mathbf{z}))\,, \tag{3}$$

where $W_1 \in \mathbb{R}^{C/r\times C}$, $W_2 \in \mathbb{R}^{C\times C/r}$, $\delta(\cdot)$ is the ReLU activation, and $\sigma(\cdot)$ is the sigmoid function. The recalibrated feature map is obtained by channel-wise scaling:

$$\widetilde{F}_c = s_c \cdot F_c, \quad c = 1, \ldots, C. \tag{4}$$

In the implementation, the scaling vector $\mathbf{s}$ is reshaped to $C \times 1 \times 1$ and broadcast across the spatial dimensions of each feature channel.

2.5.2. Spatial Attention Spatial Attention (SA) generates a two-dimensional attention map that reweights each spatial position (i, j) of the feature map, directing the network to concentrate on discriminative regions such as lesion boundaries and dermoscopic structures. Two channel-wise pooling operations — average and max — are applied across the channel axis to produce complementary spatial descriptors, which are concatenated and passed through a 7×7 convolutional layer [10]:

$$\mathbf{M}_s = \sigma\big(f^{7\times7}\big(\big[\,\mathrm{AvgPool}_C(\mathbf{F})\,;\,\mathrm{MaxPool}_C(\mathbf{F})\,\big]\big)\big)\,, \tag{5}$$

where $\mathrm{AvgPool}_C$ and $\mathrm{MaxPool}_C$ denote pooling along the channel dimension (producing $1 \times H \times W$ tensors), $[\,\cdot\,;\,\cdot\,]$ denotes channel-wise concatenation, and $f^{7\times7}$ is a convolution with a 7×7 kernel, zero padding of 3, and a single output channel. The recalibrated output is

$$\widetilde{\mathbf{F}} = \mathbf{M}_s \otimes \mathbf{F}, \tag{6}$$

where $\otimes$ denotes element-wise multiplication broadcast across all C channels.

2.5.3. Convolutional Block Attention Module CBAM [10] combines channel and spatial attention sequentially, first refining *what* to attend and then *where* to attend. Given the input feature map $\mathbf{F}$, the channel attention map $\mathbf{M}_c \in \mathbb{R}^{C\times1\times1}$ is computed using both average- and max-pooled channel descriptors fed through a shared two-layer MLP:

$$\mathbf{M}_c(\mathbf{F}) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathbf{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathbf{F})))\,. \tag{7}$$

The channel-refined intermediate map is $\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}$. Spatial attention is then applied to $\mathbf{F}'$ using Eq. (5) to produce the spatial map $\mathbf{M}_s(\mathbf{F}') \in \mathbb{R}^{1\times H\times W}$, yielding the final output:

$$\widetilde{\mathbf{F}} = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}'. \tag{8}$$

2.5.4. Coordinate Attention Coordinate Attention (CoordAtt) [11] addresses the loss of precise spatial location information that occurs in standard 2D global average pooling. Instead of collapsing all spatial dimensions simultaneously, the module decomposes the encoding into two orthogonal 1D pooling operations that preserve one spatial dimension each:

$$z_c^h(h) = \frac{1}{W}\sum_{j=0}^{W-1} F_c(h,\,j), \qquad z_c^w(w) = \frac{1}{H}\sum_{i=0}^{H-1} F_c(i,\,w). \tag{9}$$

The resulting tensors $\mathbf{z}^h \in \mathbb{R}^{C\times H\times1}$ and $\mathbf{z}^w \in \mathbb{R}^{C\times1\times W}$ are concatenated along the spatial axis, transformed through a shared 1×1 convolution followed by batch normalisation and non-linearity, and then split and independently projected through two 1×1 convolutions with sigmoid activations to produce direction-aware attention maps $\mathbf{g}^h \in \mathbb{R}^{C\times H\times1}$ and $\mathbf{g}^w \in \mathbb{R}^{C\times1\times W}$. The final recalibrated output is

$$\widetilde{F}_c(i,j) = F_c(i,j) \times g_c^h(i) \times g_c^w(j), \tag{10}$$

encoding both channel dependencies and precise positional cues along both spatial axes at a computational cost comparable to a standard SE block.

2.6. Training Configuration

All ten experimental configurations share an identical training setup so that observed differences in performance can be interpreted under a controlled optimisation and hardware configuration, with the attention module and cross-validation protocol serving as the experimental factors.

Loss function. Class imbalance in HAM10000—where the dominant class (melanocytic nevi) accounts for over 66% of samples while the rarest class (dermatofibroma) represents fewer than 2%—poses a known risk of biased optimisation under standard cross-entropy loss, as the model can achieve low loss by exploiting the majority class.

Focal Loss [34] is therefore used to reduce the contribution of well-classified examples and concentrate the gradient signal on difficult or misclassified samples. Given the predicted probability p_t for the ground-truth class, Focal Loss is defined as

$$\mathcal{L}_{\text{focal}} = -\alpha\,(1-p_t)^{\gamma}\,\log(p_t), \tag{11}$$

where $\alpha \geq 0$ is a weighting factor and $\gamma \geq 0$ is the focusing parameter that controls the rate at which easy examples are down-weighted. When $\gamma = 0$, Eq. (11) reduces to standard cross-entropy scaled by α. In all experiments, we set $\alpha = 1$ and $\gamma = 2$. The setting $\alpha = 1$ applies no class-specific weighting; therefore, the loss differs from standard cross-entropy through the focusing term $(1-p_t)^{\gamma}$, which emphasises difficult examples rather than assigning different weights to individual classes. These values were fixed across all configurations to avoid introducing loss-function tuning as an additional experimental factor. The predicted probability is obtained from the softmax of the model logits, and the batch loss is averaged over all samples using `reduction='mean'`.

Optimiser and learning rate. All models are optimised using the Adam optimiser [33] with a fixed learning rate of $\eta = 1 \times 10^{-4}$ and default momentum coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$. No learning-rate scheduling or weight decay is applied. This shared optimisation setting was selected to preserve a controlled comparison across attention modules rather than to individually optimise the training dynamics of each module. The entire model—including the ConvNeXt-Tiny backbone, the appended attention module, and the classifier head—is updated at each step; the backbone weights are not frozen.

Hyperparameters and training duration. Each fold is trained for a fixed budget of 50 epochs with a batch size of 64. The model state obtained at the epoch with the lowest validation Focal Loss is saved as a `.pth` checkpoint and subsequently reloaded for evaluation, ensuring that the reported metrics reflect the selected generalisation state rather than the terminal epoch. Table 3 summarises the training hyperparameters applied uniformly across all configurations. The value `random_state=42` applies to the construction of the cross-validation partitions described in Section 2.2.

Table 3. Training hyperparameters applied uniformly across all ten experimental configurations.

Hyperparameter	**Value**
Optimiser	Adam
Learning rate	1×10^{-4}
β_1, β_2	0.9, 0.999
Weight decay	0
Learning-rate scheduler	None
Batch size	64
Training epochs	50
Loss function	Focal Loss ($\alpha = 1$, $\gamma = 2$)
Model selection	Lowest validation Focal Loss per fold
Input resolution	224×224 pixels
Backbone weights	ImageNet-1K pretrained
Backbone training	All layers trainable

All experiments were conducted using the same hardware and software environment, consisting of a single NVIDIA GeForce RTX 5070 GPU (12 GB GDDR7 VRAM), an Intel Core i7-14700F CPU, and 32 GB DDR5 RAM. The complete environment is reported in Table 4.

Resource profiling. To support the computational-efficiency analysis reported in Section 3, each training run records two resource metrics: (i) total wall-clock training time in seconds, measured from the beginning of the first epoch to completion of the final epoch, and (ii) peak allocated GPU memory in megabytes, obtained using `torch.cuda.max_memory_allocated()` after resetting the peak counter at the beginning of each fold with `torch.cuda.reset_peak_memory_stats()`. These values are logged per fold alongside the classification metrics and aggregated as mean $\pm$ standard deviation across the five folds.

Table 4. Hardware and software environment used in all experiments.

Component	Specification
Hardware	
GPU	NVIDIA GeForce RTX 5070 (12 GB GDDR7 VRAM)
CPU	Intel Core i7-14700F (20 cores, up to 5.4 GHz)
RAM	32 GB DDR5
Software	
Operating system	Windows-10-10.0.26200-SP0
Python	3.11
PyTorch	2.8.0+cu129
CUDA	12.9
torchvision	0.23.0+cu129
scikit-learn	1.7.1

2.7. Evaluation Metrics and Statistical Analysis

Given the severe class imbalance of HAM10000, accuracy alone is an insufficient performance indicator—a trivial classifier predicting only the majority class achieves over 66% accuracy. All configurations are therefore evaluated using *macro-averaged F1-score* (macro-F1) as the primary metric, which assigns equal weight to each of the seven classes regardless of their sample count:

$$F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{2\, P_c\, R_c}{P_c + R_c}, \tag{12}$$

where $C = 7$ is the number of classes, and P_c and R_c are the precision and recall for class c, respectively. Macro-F1 was selected instead of weighted F1 as the primary metric because support-based weighting would allow the dominant melanocytic nevus class to exert a disproportionate influence on the aggregate score. Equal weighting also ensures that rare classes, including dermatofibroma and vascular lesions, contribute equally to the overall evaluation.

Overall accuracy is reported as a secondary metric. Macro-averaged precision and macro-averaged recall are also reported, while per-class precision, recall, and F1-score are provided to enable a fine-grained analysis of performance across both majority and minority categories. All classification results are reported as mean $\pm$ standard deviation across the five folds of each configuration.

Computational efficiency is assessed using the two resource measurements recorded consistently for all experiments: total wall-clock training time and peak allocated GPU memory. These measurements are also summarised as mean $\pm$ standard deviation across five folds.

Fold-level statistical comparisons were performed separately within the image-level Stratified and lesion-level Group five-fold validation protocols. For each attention variant, the macro-F1 score from a given fold was paired with the macro-F1 score from the corresponding fold of the attention-free ConvNeXt-Tiny baseline. Because explicit validation identifiers were not available for every fold, pairing was based on matching fold numbers and the common splitter configuration used within each validation protocol.

The primary analysis used a two-sided paired-samples t-test. For each comparison, the paired mean difference, calculated as the attention model minus the corresponding baseline, was reported in percentage points together with its 95% confidence interval and Cohen's d_z paired effect size. An exact two-sided Wilcoxon signed-rank test was additionally conducted as a sensitivity analysis. Family-wise multiplicity across the eight baseline comparisons was controlled using the Holm procedure, applied separately to the paired t-test and Wilcoxon p-values. Because each comparison contained only five paired folds, the inferential results were treated as exploratory.

3. Results and Discussion

This section reports the empirical results of the ten experimental configurations described in Section 2. All metrics are computed per fold and reported as the mean ± standard deviation across the five folds of each configuration. We first present the overall classification performance (Section 3.1), followed by an analysis of the effect of the cross-validation protocol (Section 3.2), a per-class breakdown (Section 3.3), a computational efficiency comparison (Section 3.4), and a concluding discussion (Section 3.5).

3.1. Overall Classification Performance

Table 5 summarises the overall classification performance of all ten configurations in terms of accuracy, macro-averaged F1-score (macro-F1), macro-averaged precision, and macro-averaged recall. Given the severe class imbalance of HAM10000, macro-F1 is treated as the primary performance indicator, as it weights all seven classes equally and is therefore not dominated by the majority class.

Table 5. Overall classification performance of all ten configurations, reported as mean ± standard deviation (%) across five folds.

Protocol	Exp.	Attention	Accuracy	Macro-F1	Macro-P	Macro-R
Stratified	E1	Baseline	89.70 ± 1.12	83.42 ± 1.68	84.46 ± 1.79	82.98 ± 1.87
	E2	CA	90.17 ± 0.92	**84.68** ± 1.53	**85.99** ± 2.64	**83.70** ± 1.54
	E3	SA	**90.21** ± 0.78	83.06 ± 2.33	85.34 ± 1.48	81.92 ± 3.42
	E4	CBAM	89.36 ± 1.52	82.36 ± 3.20	84.37 ± 3.47	81.06 ± 3.67
	E5	CoordAtt	89.44 ± 1.01	81.90 ± 2.13	84.76 ± 2.55	80.20 ± 2.58
Lesion-level Group	E6	Baseline	84.83 ± 1.73	72.58 ± 2.93	**78.20** ± 1.80	69.99 ± 4.43
	E7	CA	84.27 ± 1.08	70.51 ± 3.14	75.07 ± 3.02	69.33 ± 2.48
	E8	SA	**85.15** ± 0.85	**73.96** ± 3.37	77.24 ± 3.30	**73.22** ± 5.56
	E9	CBAM	84.87 ± 0.84	72.23 ± 3.89	76.38 ± 2.07	70.99 ± 4.84
	E10	CoordAtt	84.47 ± 1.25	72.02 ± 2.87	77.35 ± 2.73	69.48 ± 3.62

Two observations emerge from Table 5. First, the differences among attention configurations within the same validation protocol are relatively modest compared with the differences observed between protocols. Under stratified cross-validation, macro-F1 ranges from 81.90% for CoordAtt to 84.68% for Channel Attention, corresponding to a range of 2.78 percentage points. Under lesion-level group cross-validation, macro-F1 ranges from 70.51% for Channel Attention to 73.96% for Spatial Attention, corresponding to a range of 3.45 percentage points. These within-protocol ranges are substantially smaller than the protocol-associated reduction examined in Section 3.2.

Second, the relative ranking of the configurations is not preserved across the two protocols. Under stratified cross-validation, Channel Attention (E2) achieves the highest macro-F1 of 84.68%, as well as the highest macro-precision and macro-recall, while Spatial Attention (E3) achieves the highest accuracy of 90.21%. Under lesion-level group cross-validation, Spatial Attention (E8) achieves both the highest accuracy of 85.15% and the highest macro-F1 of 73.96%. Spatial Attention is also the only attention mechanism to obtain a numerically higher macro-F1 than the corresponding attention-free group baseline (E6, 72.58%). The remaining attention variants obtain lower macro-F1 scores than the group baseline. These results indicate that the relative ranking of attention mechanisms depends on the evaluation protocol and should not be inferred from image-level stratified validation alone.

Accuracy remains consistently higher than macro-F1, particularly under lesion-level group validation. Across E6–E10, the difference between accuracy and macro-F1 ranges from 11.19 to 13.76 percentage points, with an average difference of approximately 12.46 points. This discrepancy reflects the strong influence of the majority melanocytic nevus class on overall accuracy. Macro-F1 therefore provides a more informative summary of performance across the imbalanced seven-class distribution.

3.2. Fold-Level Statistical Comparison withthe Baseline

Table 6 presents fold-level comparisons between each attention-augmented configuration and the corresponding attention-free baseline within the same cross-validation protocol. Positive mean differences indicate higher macro-F1 for the attention model, whereas negative values indicate lower macro-F1 than the baseline.

Table 6. Paired fold-level macro-F1 comparisons against the protocol-matched attention-free baseline. Mean differences and confidence intervals are expressed in percentage points.

Protocol	Comparison	Difference	95% CI	d_z	Raw p	Holm p
Stratified	E2 vs E1	1.26	$[-0.16, 2.68]$	1.10	0.070	0.560
Stratified	E3 vs E1	-0.36	$[-3.02, 2.31]$	-0.17	0.730	1.000
Stratified	E4 vs E1	-1.06	$[-4.75, 2.63]$	-0.36	0.471	1.000
Stratified	E5 vs E1	-1.52	$[-3.48, 0.44]$	-0.96	0.098	0.687
Lesion Group	E7 vs E6	-2.07	$[-6.37, 2.23]$	-0.60	0.253	1.000
Lesion Group	E8 vs E6	1.38	$[-3.16, 5.92]$	0.38	0.447	1.000
Lesion Group	E9 vs E6	-0.35	$[-3.98, 3.28]$	-0.12	0.803	1.000
Lesion Group	E10 vs E6	-0.57	$[-6.44, 5.31]$	-0.12	0.803	1.000

Difference denotes attention minus baseline. Raw p-values are from two-sided paired-samples t-tests. Holm adjustment was applied across all eight comparisons. Because each comparison contains five paired folds, the analysis is exploratory.

None of the eight fold-level comparisons remained statistically significant after Holm correction. Under image-level Stratified validation, Channel Attention (E2) was numerically higher than the baseline E1 by 1.26 percentage points (95% CI $[-0.16, 2.68]$; Cohen's $d_z = 1.10$; raw $p = 0.070$; Holm-adjusted $p = 0.560$). E2 produced a positive paired difference in all five folds, but the confidence interval included zero.

Under lesion-level Group validation, Spatial Attention (E8) was the only attention configuration with a mean macro-F1 above the baseline. Its paired mean difference relative to E6 was 1.38 percentage points (95% CI $[-3.16, 5.92]$; Cohen's $d_z = 0.38$; raw $p = 0.447$; Holm-adjusted $p = 1.000$). Three folds showed a positive difference and two showed a negative difference. The exact Wilcoxon sensitivity analysis produced the same overall conclusion: none of the attention models showed statistically significant improvement after multiplicity correction. These findings support describing E2 and E8 as numerically higher than their corresponding baselines rather than statistically superior.

3.3. Impact of the Cross-Validation Protocol

The principal result of this study is the consistent performance gap between the two cross-validation protocols. Although the same architectures, dataset, training configuration, and hyperparameters are used, lesion-level group cross-validation produces lower performance than image-level stratified cross-validation across all five model configurations. Table 7 compares the corresponding attention variants under the two protocols, while Figure 4 visualises the differences in macro-F1 and accuracy.

Table 7. Performance gap between the image-level Stratified and lesion-level Group cross-validation protocols for each attention variant. The gap is calculated as the Stratified value minus the corresponding Group value in percentage points.

Attention	Macro-F1 (%)			Accuracy (%)		
	Strat.	Lesion Group	Gap	Strat.	Lesion Group	Gap
Baseline	83.42	72.58	10.84	89.70	84.83	4.87
CA	84.68	70.51	14.17	90.17	84.27	5.90
SA	83.06	73.96	9.10	90.21	85.15	5.06
CBAM	82.36	72.23	10.13	89.36	84.87	4.49
CoordAtt	81.90	72.02	9.88	89.44	84.47	4.97
Mean	**83.08**	**72.26**	**10.82**	**89.78**	**84.72**	**5.06**

The protocol-associated macro-F1 gap ranges from 9.10 percentage points for Spatial Attention to 14.17 percentage points for Channel Attention, with an average reduction of 10.82 percentage points across the five

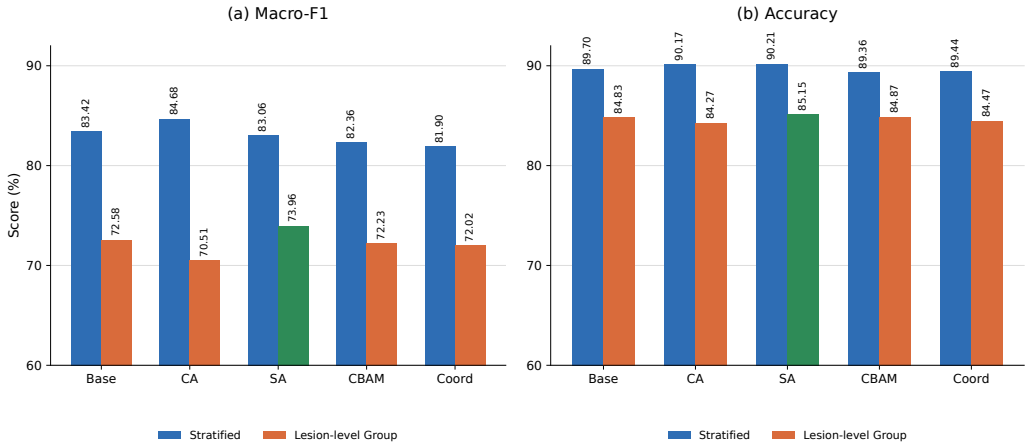

Figure 4. Comparison of macro-F1 and accuracy under image-level Stratified and lesion-level Group cross-validation.

configurations. By comparison, the accuracy gap ranges from 4.49 to 5.90 percentage points, with an average of 5.06 percentage points. Thus, the average macro-F1 gap is approximately twice the average accuracy gap. The corresponding average macro-recall gap is 11.37 percentage points. The larger reductions in macro-F1 and macro-recall indicate that the protocol change affects class-balanced performance more strongly than majority-class-dominated overall accuracy.

The magnitude of the gap has important implications for the interpretation of HAM10000 benchmarks. Under image-level stratified cross-validation, the evaluated configurations achieve accuracies of 89.36–90.21% and macro-F1 scores of 81.90–84.68%. Under lesion-level group cross-validation, the corresponding accuracies fall to 84.27–85.15%, while macro-F1 decreases to 70.51–73.96%. Image-level stratification permits different images associated with the same `lesion_id` to appear in both the training and validation subsets. Consequently, validation performance may benefit from lesion-specific visual similarities that are unavailable when the model is evaluated on unseen lesions. Because the reduction occurs across all five model configurations, the observed gap is associated with the evaluation protocol rather than with a single attention mechanism.

The evaluation protocol also affects the relative ranking of the model configurations. Under image-level stratified cross-validation, Channel Attention achieves the highest observed macro-F1 of 84.68%, whereas Spatial Attention achieves the highest accuracy of 90.21%. Under lesion-level group cross-validation, Spatial Attention achieves both the highest accuracy of 85.15% and the highest macro-F1 of 73.96%. It is also the only attention mechanism to obtain a numerically higher macro-F1 than the corresponding attention-free baseline, which achieves 72.58%. Spatial Attention exhibits the smallest protocol gap at 9.10 percentage points, whereas Channel Attention exhibits the largest gap at 14.17 percentage points. These results demonstrate that model rankings obtained under image-level stratification are not necessarily preserved under lesion-level evaluation. Lesion-level group partitioning should therefore be preferred when HAM10000 is used to estimate generalisation to unseen lesions, while results obtained from image-level stratification should be interpreted as potentially optimistic estimates.

3.4. Per-Class Performance

To identify which classes contribute to the aggregate performance differences, we examine the per-class F1-scores. Table 8 presents the mean per-class F1-score for every configuration under both cross-validation protocols. The classes follow the fixed label ordering used throughout the experiments, while the support row reports the total number of images in each class.

The per-class results show that the effect of the validation protocol is not uniform across diagnostic categories. The majority class, melanocytic nevi (*nv*), remains comparatively stable, with F1-scores of 94.53–95.22% under Stratified validation and 92.27–92.79% under lesion-level Group validation. In contrast, several less frequent

Table 8. Mean per-class F1-score (%) across five folds. Support denotes the total number of images per class, and bold values indicate the highest score within each protocol.

Protocol	Exp.	Attn.	akiec	bcc	bkl	df	mel	nv	vasc
Support			*327*	*514*	*1099*	*115*	*1113*	*6705*	*142*
Stratified	E1	Base	74.23	85.08	80.77	82.76	73.81	94.80	92.48
	E2	CA	**76.72**	**86.70**	**82.27**	**84.30**	**74.87**	94.95	**92.91**
	E3	SA	75.16	84.99	82.23	76.99	74.81	**95.22**	92.03
	E4	CBAM	73.41	85.70	79.98	80.32	71.51	94.76	90.83
	E5	Coord	74.87	85.21	80.56	75.83	72.42	94.53	89.88
Lesion-level Group	E6	Base	62.89	76.37	70.11	60.23	61.37	92.78	84.31
	E7	CA	57.46	73.54	71.91	54.86	57.76	92.38	85.69
	E8	SA	**63.55**	**78.85**	71.76	**63.63**	**61.40**	92.78	**85.74**
	E9	CBAM	58.57	75.15	**72.48**	61.63	59.58	**92.79**	85.42
	E10	Coord	61.72	77.15	71.56	60.08	58.94	92.27	82.38

classes exhibit substantially larger reductions. For the attention-free baseline, the F1-score of dermatofibroma (*df*) decreases from 82.76% to 60.23%, corresponding to a 22.53-point reduction. Melanoma (*mel*) and actinic keratosis (*akiec*) decrease by 12.44 and 11.34 percentage points, respectively. These results indicate that image-level stratification produces a larger optimistic difference for several minority and clinically challenging categories than for the majority *nv* class.

When averaged across the five model configurations, the largest class-level protocol gap is observed for *df* at 19.95 percentage points, followed by *akiec* at 14.04 points and *mel* at 13.67 points. The majority *nv* class exhibits the smallest average gap at 2.25 points, whereas the rare *vasc* class shows a comparatively moderate gap of 6.92 points. Thus, class rarity alone does not completely explain the observed pattern.

An exploratory Pearson correlation across the seven classes indicates a negative association between class support and the average protocol gap ($r = -0.68$, $p = 0.093$, 95% CI $[-0.95, 0.15]$). The images-per-lesion ratio is positively associated with the protocol gap ($r = 0.62$, $p = 0.135$, 95% CI $[-0.25, 0.94]$). Neither association reaches conventional statistical significance, partly because the analysis contains only seven class-level observations. The directions of these associations nevertheless suggest that both class rarity and the availability of repeated views may contribute to performance inflation, while neither factor alone fully accounts for the class-specific differences.

The per-class results help explain why Spatial Attention achieves the highest observed macro-F1 under lesion-level Group validation. Among the five configurations evaluated under this protocol, Spatial Attention obtains the highest F1-score for five of the seven classes: *akiec* (63.55%), *bcc* (78.85%), *df* (63.63%), *mel* (61.40%), and *vasc* (85.74%). CBAM obtains the highest F1-scores for *bkl* (72.48%) and *nv* (92.79%). A possible explanation is that Spatial Attention constructs a location-sensitive mask from channel-wise average and maximum feature summaries. At the post-backbone 7×7 feature map, this mechanism may emphasise spatially localised dermoscopic structures, such as lesion boundaries, pigment patterns, streaks, and vascular regions, while suppressing less informative surrounding areas.

For dermatofibroma, Spatial Attention decreases from 76.99% under Stratified validation to 63.63% under lesion-level Group validation, corresponding to a 13.36-point gap. This reduction is smaller than the 22.53-point gap observed for the attention-free baseline, which decreases from 82.76% to 60.23%. This comparison indicates that the baseline dermatofibroma result is more sensitive to the validation protocol. However, this interpretation remains mechanistic rather than direct evidence of improved lesion localisation, because no localisation annotations or quantitative localisation metrics were used.

3.5. Computational Efficiency

Beyond classification performance, the practical use of an attention mechanism also depends on its computational cost. Table 9 reports two training-resource metrics: peak allocated GPU memory and total wall-clock training time per fold. Both metrics are reported as the mean $\pm$ standard deviation across the five folds of each configuration. All

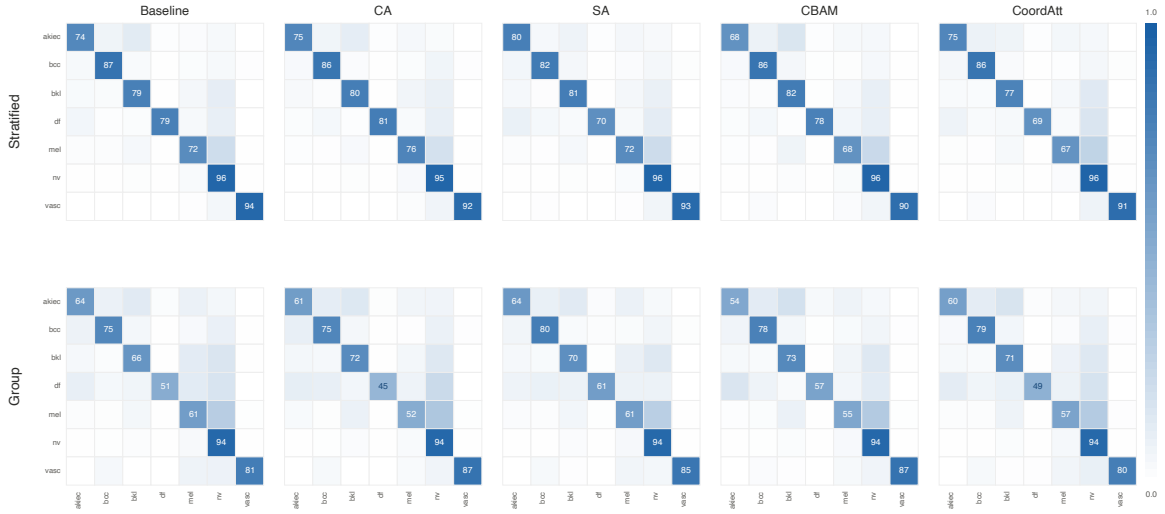

Figure 5. Row-normalised confusion matrices for E1–E10 under image-level Stratified and lesion-level Group cross-validation. Diagonal entries represent per-class recall.

experiments used an identical batch size of 64 and were executed using the hardware and software environment described in Section 2.

Table 9. Computational efficiency of all ten configurations, reported as mean ± standard deviation across five folds. Peak VRAM denotes the maximum GPU memory allocated during training, while training time denotes the wall-clock time per fold.

Protocol	Exp.	Attention	Peak VRAM (MB)	Training time (min)
Stratified	E1	Baseline	7331.0 ± 89.4	108.4 ± 0.4
	E2	CA	7269.2 ± 59.3	106.6 ± 1.3
	E3	SA	7266.5 ± 122.4	109.6 ± 1.2
	E4	CBAM	7225.4 ± 89.8	108.7 ± 0.6
	E5	CoordAtt	7182.2 ± 89.7	108.6 ± 0.4
Lesion-level Group	E6	Baseline	7203.4 ± 1.1	109.5 ± 0.5
	E7	CA	7204.7 ± 0.9	110.0 ± 0.5
	E8	SA	7245.8 ± 96.5	108.7 ± 0.5
	E9	CBAM	7096.9 ± 1.1	109.1 ± 0.7
	E10	CoordAtt	7119.7 ± 48.2	115.1 ± 3.6

Across all ten configurations, peak allocated GPU memory ranges from approximately 7097 to 7331 MB. This narrow range indicates that none of the four attention modules introduces a substantial memory overhead relative to the ConvNeXt-Tiny backbone. This finding is consistent with the experimental design, because each attention module is inserted only once at the post-backbone location and operates on the compact $768 \times 7 \times 7$ feature map. Following the uniform batch-size configuration, Channel Attention under Stratified validation records 7269.2 ± 59.3 MB of peak VRAM, which is comparable to the remaining configurations.

Training time is also broadly comparable across most configurations. Experiments E1–E9 require approximately 106.6–110.0 minutes per fold. E10, which applies Coordinate Attention under lesion-level Group validation, records the longest mean training time at 115.1 ± 3.6 minutes. Its higher fold-to-fold variability suggests that this runtime difference should be interpreted cautiously rather than as definitive evidence of an inherent computational disadvantage.

Overall, CA, SA, and CBAM exhibit similar peak-memory requirements and training durations when implemented as post-backbone plug-ins. Coordinate Attention also has comparable memory requirements, although its lesion-level configuration shows a moderately longer training time. These measurements characterise training-resource requirements only; inference latency is not reported because it was not measured consistently across all configurations.

3.6. Discussion

Three main findings emerge from the results. First, **the validation protocol has a larger observed effect on reported performance than the choice of attention mechanism**. Replacing image-level Stratified cross-validation with lesion-level Group cross-validation reduces macro-F1 by an average of 10.82 percentage points across the five configurations. The reduction is not distributed uniformly across classes. For the attention-free baseline, the F1-score of dermatofibroma decreases by 22.53 points, whereas the majority melanocytic-nevus class decreases by only 2.02 points. Across all five configurations, class support shows a negative exploratory association with the average class-level protocol gap ($r = -0.68$, $p = 0.093$, 95% CI $[-0.95,\ 0.15]$). Although this association does not reach conventional statistical significance, its direction indicates that several less frequent classes are more sensitive to the validation protocol. Results obtained using image-level stratification should therefore be interpreted as potentially optimistic estimates of generalisation to unseen lesions. Lesion-level group partitioning provides a more appropriate evaluation when multiple images of the same lesion are present, consistent with broader concerns regarding data leakage in machine-learning-based research [6, 28].

Second, **the relative performance of the attention mechanisms depends on the evaluation protocol**. Under Stratified validation, Channel Attention achieves the highest observed macro-F1 of 84.68%, whereas Spatial Attention achieves the highest accuracy of 90.21%. Under lesion-level Group validation, Spatial Attention achieves both the highest macro-F1 of 73.96% and the highest accuracy of 85.15%. It is also the only attention mechanism to obtain a numerically higher macro-F1 than the corresponding attention-free baseline. Nevertheless, the within-protocol macro-F1 ranges are limited to 2.78 points under Stratified validation and 3.45 points under lesion-level Group validation. These differences are considerably smaller than the average 10.82-point difference between protocols and should not be interpreted as broad evidence that one attention mechanism is universally superior. This caution is supported by the fold-level statistical analysis. None of the eight comparisons with the protocol-matched baseline remained statistically significant after Holm correction. Although Channel Attention showed a large paired effect size under Stratified validation ($d_z = 1.10$), its 95% confidence interval included zero and its Holm-adjusted p-value was 0.560. Spatial Attention showed a smaller positive effect under lesion-level Group validation ($d_z = 0.38$), with a wide confidence interval and a Holm-adjusted p-value of 1.000. The observed rankings should therefore be interpreted as descriptive outcomes of the present five-fold benchmark, not as confirmatory evidence of superiority.

The class-level results provide a possible explanation for the observed performance of Spatial Attention under lesion-level evaluation. Spatial Attention obtains the highest F1-score for five of the seven classes under this protocol, including *akiec*, *bcc*, *df*, *mel*, and *vasc*. Its location-sensitive mask, constructed from channel-wise average and maximum feature summaries, may emphasise spatially localised dermoscopic structures, such as lesion boundaries, pigment patterns, streaks, and vascular regions. These cues may remain useful when evaluating previously unseen lesions. However, this explanation is mechanistic rather than direct evidence of improved localisation, because the study does not include lesion-localisation annotations or a quantitative localisation metric. The remaining attention mechanisms show no consistent advantage over the baseline in the evaluated post-backbone, single-insertion setting.

Third, **accuracy alone provides an incomplete assessment of this imbalanced classification problem**. Under lesion-level Group validation, the difference between accuracy and macro-F1 averages approximately 12.46 percentage points. This divergence occurs because overall accuracy is strongly influenced by the dominant melanocytic-nevus class, which maintains an F1-score above 92% across all Group configurations, while several minority classes perform substantially worse. Macro-F1, supported by macro-precision, macro-recall, and per-class metrics, should therefore remain the primary performance indicator for comparative studies on imbalanced dermoscopic datasets.

Limitations. The findings are subject to several limitations. First, each attention module was evaluated only as a single post-backbone plug-in operating on a 7×7 ConvNeXt-Tiny feature map. The relative rankings may differ when attention is inserted within intermediate backbone stages or at multiple feature resolutions. Second, the study uses one backbone, one input resolution, and one dataset. Validation with other architectures and independent dermoscopic cohorts is therefore needed before the findings can be generalised beyond the present experimental setting.

Third, all configurations use the same fixed learning rate, optimiser, Focal Loss parameters, and ImageNet-1K initialisation. This design supports a controlled comparison but does not guarantee that each attention mechanism was evaluated under its individually optimal training configuration. Module-specific optimisation, learning-rate scheduling, weight decay, or domain-specific pretraining may alter the absolute results or relative rankings.

Fourth, each statistical comparison was based on only five paired folds. Cross-validation folds are not equivalent to independently sampled external datasets, and the resulting confidence intervals are wide and sensitive to individual folds. The inferential results should therefore be interpreted as exploratory quantification of fold-level consistency rather than definitive evidence of model superiority.

4. Conclusion

This study presented a controlled methodological benchmark of four canonical attention mechanisms—Channel Attention, Spatial Attention, CBAM, and Coordinate Attention—implemented as post-backbone modules on a shared ConvNeXt-Tiny feature extractor for seven-class dermoscopic classification on HAM10000. Each configuration was evaluated under image-level Stratified and lesion-level Group five-fold cross-validation. The results show that the validation protocol has a larger observed effect on reported performance than the choice of attention mechanism. Replacing image-level stratification with lesion-level group partitioning reduced macro-F1 by an average of 10.82 percentage points and accuracy by 5.06 points. The larger macro-F1 reduction was concentrated mainly in several minority and clinically challenging classes, while the majority melanocytic-nevus class remained comparatively stable. Results obtained under image-level stratification should therefore be interpreted as potentially optimistic estimates of generalisation to unseen lesions.

The relative ranking of the attention mechanisms was not preserved across protocols. Channel Attention achieved the highest observed macro-F1 under Stratified validation, whereas Spatial Attention achieved the highest macro-F1 under lesion-level Group validation and was the only attention mechanism to obtain a numerically higher score than the corresponding attention-free baseline. Nevertheless, the within-protocol differences among configurations were modest and do not establish universal superiority of any attention mechanism. Consistent with this interpretation, none of the eight fold-level comparisons against the corresponding attention-free baseline remained statistically significant after Holm correction. Training-resource measurements also showed comparable peak GPU-memory requirements across the evaluated modules, although Coordinate Attention exhibited a moderately longer training time under lesion-level validation.

These findings support lesion-level group partitioning as the preferred evaluation protocol for HAM10000 when multiple images share the same `lesion_id`. Future work should examine alternative attention placements, additional backbones and input resolutions, independently optimised training configurations, inference latency, and external validation on independent dermoscopic cohorts. More broadly, adopting leakage-aware evaluation practices is necessary to ensure that reported progress in skin-lesion classification reflects generalisation to unseen lesions rather than overlap between training and validation samples.

Acknowledgement

The authors gratefully acknowledge Universitas PGRI Ronggolawe Tuban for the support and research facilities provided during this study. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Code Availability

The source code, experiment notebooks, statistical-analysis scripts, environment specifications, and reproducibility documentation associated with this study are publicly available at: https://github.com/asfanme/benchmark-attention. The HAM10000 dataset is not redistributed and must be obtained separately from its official source.

Conflict of Interest

Authors declare that there is noconflict of interests regarding the publication of the paper.

REFERENCES

1. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, *Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*, CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209–249, 2021.
2. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, *Dermatologist-level classification of skin cancer with deep neural networks*, Nature, vol. 542, no. 7639, pp. 115–118, 2017.
3. K. Ramamurthy, I. Thayumanaswamy, M. Radhakrishnan, D. Won, and S. Lingaswamy, *Integration of localized, contextual, and hierarchical features in deep learning for improved skin lesion classification*, Diagnostics, vol. 14, no. 13, p. 1338, 2024.
4. P. Tschandl, C. Rosendahl, and H. Kittler, *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, Scientific Data, vol. 5, p. 180161, 2018.
5. S. Mustafa, A. Jaffar, M. Rashid, S. Akram, and S. M. Bhatti, *Deep learning-based skin lesion analysis using hybrid ResUNet++ and modified AlexNet-Random Forest for enhanced segmentation and classification*, PLOS ONE, vol. 20, no. 1, p. e0315120, 2025.
6. S. Kapoor and A. Narayanan, *Leakage and the reproducibility crisis in machine-learning-based science*, Patterns, vol. 4, no. 9, p. 100804, 2023.
7. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An image is worth 16×16 words: Transformers for image recognition at scale*, in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
8. Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, *A ConvNet for the 2020s*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.
9. J. Hu, L. Shen, and G. Sun, *Squeeze-and-excitation networks*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.
10. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, *CBAM: Convolutional block attention module*, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, vol. 11211, pp. 3–19, 2018.
11. Q. Hou, D. Zhou, and J. Feng, *Coordinate attention for efficient mobile network design*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13713–13722, 2021.
12. Z. Ullah, M. Hong, T. Mahmood, and J. Kim, *Systematic integration of attention modules into CNNs for accurate and generalizable medical image classification*, Mathematics, vol. 13, no. 22, p. 3728, 2025.
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, *ImageNet classification with deep convolutional neural networks*, in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.
14. K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
15. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.
16. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
17. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700–4708, 2017.
18. M. Tan and Q. V. Le, *EfficientNet: Rethinking model scaling for convolutional neural networks*, in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.
19. A. Adebiyi, N. Abdalnabi, E. H. Smith, J. Hirner, E. J. Simoes, M. Becevic, and P. Rao, *Accurate skin lesion classification using multimodal learning on the HAM10000 dataset*, Journal of Pathology Informatics, vol. 16, p. 100455, 2025.
20. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, *Swin Transformer: Hierarchical vision transformer using shifted windows*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.
21. N. Dağoğlu Hark, A. Kılıç, and H. İ. İçen, *CBAM-ConvNeXt: Enhancing ConvNeXt with channel and spatial attention for accurate brain tumour classification*, IET Image Processing, vol. 19, 2025.
22. S. M. Alvi and C. Era, *SkinAACN: An efficient skin lesion classification based on attention augmented ConvNeXt with hybrid loss function*, in *Proceedings of the 7th International Conference on Computer Science and Artificial Intelligence (CSAI)*, pp. 127–133, 2023.

23. E. Dağoğlu Hark, M. Mir, and S. Rizvi, *Dual channel-spatial attention with EfficientNet for accurate medical image classification and Grad-CAM interpretability*, IET Image Processing, vol. 19, 2025.
24. J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, *A simple and light-weight attention module for convolutional neural networks*, International Journal of Computer Vision, vol. 128, pp. 783–798, 2020.
25. S. Islam and R. Roy, *MedNet: a lightweight attention-augmented CNN for medical image classification*, Scientific Reports, vol. 15, p. 41936, 2025.
26. H. Liu, Y. Zhang, and Y. Chen, *Channel-spatial attention modules in convolutional neural networks for image classification*, Scientific Reports, vol. 15, 2025.
27. A. Dawood, S. Alsubai, A. Alqahtani, N. Alotaibi, and I. Mehmood, *Accurate skin lesion classification using multimodal learning on the HAM10000 and ISIC 2017 datasets*, medRxiv, 2025.
28. M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J. H. F. Rudd, E. Sala, and C.-B. Schönlieb, *Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans*, Nature Machine Intelligence, vol. 3, no. 3, pp. 199–217, 2021.
29. G. Varoquaux and V. Cheplygina, *Machine learning for medical imaging: methodological failures and recommendations for the future*, npj Digital Medicine, vol. 5, p. 48, 2022.
30. L. A. van de Mortel and G. A. van Wingen, *Data leakage in machine learning studies creep into meta-analytic estimates of predictive performance*, Molecular Psychiatry, 2025.
31. A. Apicella, F. Isgrò, and A. Pollastro, *Don't push the button! Exploring data leakage risks in machine learning and transfer learning*, Artificial Intelligence Review, vol. 58, 2025.
32. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
33. D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
34. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal loss for dense object detection*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.
35. E. Aniq, F.-A. El Ghanaoui, and M. Chakraoui, *Vision transformers for breast cancer mammographic image classification*, Statistics, Optimization & Information Computing, vol. 15, no. 2, pp. 1087–1098, 2025, doi: 10.19139/soic-2310-5070-2539.
36. D. A. Ali and H. T. Sadeeq, *An interpretable deep learning framework for multi-class dental disease classification from intraoral RGB images*, Statistics, Optimization & Information Computing, vol. 14, no. 6, pp. 3380–3397, 2025, doi: 10.19139/soic-2310-5070-2880.
37. M. Jetti, E. Aniq, M. Chakraoui, and Y. Khourdifi, *JetNet: An effective deep learning model for histopathological lung cancer classification and diagnosis*, Statistics, Optimization & Information Computing, vol. 15, no. 2, pp. 1206–1223, 2025, doi: 10.19139/soic-2310-5070-2726.