Review of Computer Engineering Research
2026 Vol. 13, No. 3, pp. 135-155
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v13i3.5143

A leakage-aware evaluation framework for reliable deep learning-based melanoma detection

Ameen Shaheen[1+]
Wael Alzyadat[2]
Aysh Alhroob[3]

[1,2,3]*Department of Software Engineering, AL-Zaytoonah University of Jordan, Amman, Jordan.*
[1]*Email: a.shaheen@zuj.edu.jo*
[2]*Email: wael.alzyadat@zuj.edu.jo*
[3]*Email: aysh@zuj.edu.jo*

(+ Corresponding author)

Article History
Received: 13 March 2026
Revised: 11 May 2026
Accepted: 5 June 2026
Published: 3 September 2026

Keywords
Deep learning reliability
Dermoscopic image analysis
Lesion-level validation
Probability calibration
Uncertainty quantification.

ABSTRACT

Deep learning has shown promising performance for automated melanoma detection; however, many published studies still rely on image-level data splits that can introduce information leakage and produce overly optimistic estimates of real-world clinical performance. This study aims to develop a leakage-aware, reliability-centered evaluation framework for deep learning–based melanoma detection from dermoscopic images that better reflects realistic deployment conditions. The study applies lesion-level GroupKFold cross-validation to a 15,000-image natural-prevalence cohort and compares two transfer-learning baselines, MobileNetV2 and EfficientNetB0, under matched preprocessing, training, and evaluation settings. Beyond conventional discrimination performance, the framework assesses probability reliability using Expected Calibration Error and Brier Score, examines the effects of leakage-free post-hoc temperature scaling, and evaluates uncertainty-aware selective prediction through Monte Carlo Dropout. The findings show consistent discrimination performance, with a mean AUC around 0.82, together with strong reliability characteristics, including an Expected Calibration Error of approximately 0.028 in the primary natural-prevalence evaluation. The results further indicate that uncertainty-based selective prediction can substantially improve safety by restricting automated decisions to higher-confidence cases, reducing the error rate to about 6.7% when retaining the most confident 70% of predictions. Overall, these findings have practical implications for trustworthy clinical AI by showing that leakage-aware validation, empirical calibration analysis, and uncertainty-guided deferral mechanisms are important for safer, more clinically meaningful, and more deployable melanoma screening systems.

Contribution/Originality: This study contributes to the existing literature by introducing a leakage-aware framework for melanoma detection. This study uses a new estimation methodology based on lesion-level GroupKFold validation, calibration assessment, and Monte Carlo Dropout. The paper's primary contribution is finding that safety-oriented evaluation provides a more clinically meaningful interpretation of model performance.

1. INTRODUCTION

Skin cancer continues to be a significant public health threat globally. Although melanoma is much less frequent than other types of non-melanoma skin cancer, it has the highest mortality rate among all forms of skin cancer, which makes dermoscopy clinically useful as an improvement in the visualization of atypical skin lesions [1]. Nevertheless, even with dermoscopy, diagnostic performance can vary depending on the dermatologist, the appearance of the skin lesion, and the context of the examination, which motivates the need for AI-assisted and AI-supported diagnostic tools for screening and triage [2].

Deep learning (DL), especially convolutional neural networks (CNN), has made significant contributions to dermoscopic image classification. Recent prospective reader studies suggest that DL can achieve performance comparable to that of clinicians in dermoscopic image classification [3]. This progress has been facilitated by public repositories of benchmarks for comparison across different approaches to dermoscopic image classification and standardized benchmarks such as ISIC 2019 that allow for reproducible comparisons across methods [4]. However, successful performance on a benchmark does not necessarily equate to clinical value [5].

A major threat to study validity is information leakage. Dermoscopic datasets often contain multiple images of the same lesion as well as near-duplicate samples; therefore, random image-level division into training and test sets can introduce leakage [6]. Leakage has been identified as one of many drivers of unreliability or exaggerated claims about performance among scientific disciplines using machine learning [7], and studies examining methodology in medical imaging have found that although the appearance of some benchmark metrics is very impressive, inadequate validation and reporting continue to be commonplace [8].

In addition to discrimination, clinical use may require that the clinician has confidence in the predictive probability distributions used for risk stratification and threshold-based triage. Accuracy and AUC measures, however, are insufficient for establishing that the predicted probabilities represent the true outcome likelihoods, a requirement for risk-based decision making [9]. Other reliability metrics, such as the Brier Score and Expected Calibration Error (ECE), provide additional assessment, yet their interpretation can vary with prevalence and sample size [10]. For example, post-hoc methods such as temperature scaling can improve calibration while preserving the ranking order of cases; however, their success depends upon the specific setting and model family [11].

The use of uncertainty quantification is expected to grow in importance to allow for the development of reliable and trustworthy clinical AI systems, especially when dealing with uncertain cases for which an expert would need to review them. As noted in reviews, uncertainty must be measured and analyzed using methods that are clinically relevant and can provide actionable information regarding decision-making [12]. Uncertainty analysis could also aid in selectively predicting patients in screening-like scenarios by allowing a model to make confident predictions and withhold predictions where there is uncertainty to reduce errors and improve patient safety. Monte Carlo Dropout has been most frequently identified as a practical starting point for estimating epistemic uncertainty in deep neural networks [13].

The motivation for this study was due to these gaps, and it focused on developing a reliability and performance analysis of a melanoma detection system utilizing dermoscopic images with leakage awareness that would prevent leakage at the lesion level, utilizing robust statistical reporting that provides results beyond point estimates, and employing methods to explicitly evaluate the model's calibration (using ECE/Brier and post-hoc scaling), in addition to uncertainty-based selective prediction to provide an even closer approximation to the clinical safety standards.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the dataset, methodology, and evaluation framework. Section 4 reports the experimental results. Section 5 discusses the findings and study limitations. Section 6 concludes the paper and outlines future research directions.

The main contributions of this research are as follows:

- A leakage-aware lesion-level evaluation framework, utilizing GroupKFold cross-validation to provide a reliable estimate of how well the model will generalize.
- A controlled comparison of two transfer-learning baselines (MobileNetV2 and EfficientNetB0), conducted in an equivalent manner.
- Statistical reporting using confidence intervals and effect sizes, with proper paired or non-parametric tests used to report the variance of the results beyond single-point measurements.
- Calibration assessments for both architectures using both the ECE and the Brier Score, which include the response of each architecture to post-hoc scaling.

- Selective prediction based on uncertainty using Monte Carlo Dropout and a risk-coverage evaluation to assess the balance between safety and coverage.

An important research gap exists despite the growing use of deep learning in dermoscopic melanoma detection. While many previous studies have focused on discrimination performance using image-level validation methods, fewer studies have evaluated the effects of leakage-aware validation, probability calibration, and uncertainty-based safety evaluations under clinically relevant conditions. This may lead to differences in how well previously reported performances will generalize to the real world, as well as the ability to deploy such systems in a trusted clinical deployment. In response to this gap, this study is guided by the following three research questions: What effect does applying a leakage-aware assessment at the lesion-level in comparison with image-level splitting of data have on the interpretation of melanoma classification performance? How do different transfer-learning baselines perform regarding discrimination and probability calibration under controlled leakage-aware validation protocols, and will uncertainty estimation enable safer selective predictions through identification of those cases that need to be deferred for specialist review?

2. RELATED WORKS

Deep learning for dermoscopic melanoma detection is developing rapidly because of the availability of public datasets and benchmarks, especially the ISIC Archive, that are helping to establish standard task definitions and allow for comparable evaluations of different methods to be performed with reproducibility [4, 14]. The large curated datasets like HAM10000 have also supported the use of transfer learning techniques in the classification of skin lesions [15]. At the same time, many recent studies have concentrated mainly on improving performance through stronger architectures, data augmentation, metadata fusion, or ensemble strategies, while giving less attention to whether these gains remain stable under stricter and more clinically realistic validation settings [16, 17]. More recent evidence also shows that deep learning systems can perform strongly in melanoma diagnosis and may assist dermatologists in decision-making, but substantial heterogeneity across datasets, study designs, and validation practices still limits direct comparison across studies [18, 19].

A persistent methodological concern in dermoscopic image analysis is information leakage. Dermoscopic datasets often contain multiple images of the same lesion, near-duplicate samples, or strongly correlated images, and image-level random splitting can therefore place related samples in both the training and test sets [20]. This can produce overly optimistic estimates of generalization and reduce the clinical credibility of reported performance. Broader machine learning methodology has similarly identified leakage as a major source of exaggerated or irreproducible findings [2]. Although this issue is increasingly acknowledged in the literature, relatively few melanoma studies make lesion-level grouping a central design requirement when evaluating model performance. This limitation supports the use of lesion-level GroupKFold validation to better simulate the clinically relevant setting of previously unseen lesions.

Another important line of research concerns probability calibration and model reliability. High accuracy or AUC does not guarantee that predicted probabilities correspond to true outcome likelihoods, which is especially important in screening and triage workflows where threshold-based decisions may be made [9]. Foundational work has shown that modern neural networks can be systematically miscalibrated, and a range of post-hoc calibration methods, including temperature scaling and its variants, have been proposed to improve probability estimates [21-24]. In medical imaging AI, recent methodological studies have further emphasized that reliability metrics must be interpreted carefully because their behavior can vary with prevalence, sample size, and evaluation design [25]. However, calibration is still not routinely examined in melanoma classification studies under leakage-aware evaluation settings, and this makes it difficult to judge whether strong discrimination performance is accompanied by clinically meaningful probability estimates.

Uncertainty quantification has also become increasingly important in trustworthy medical AI, particularly for identifying cases that should be deferred for expert review rather than automatically classified. Recent reviews emphasize that uncertainty is most useful when it is linked to clinical decision-making rather than treated as a purely statistical add-on [12, 26]. Monte Carlo Dropout remains one of the most practical baseline methods for estimating predictive uncertainty in deep neural networks [27], and selective prediction approaches provide a mechanism for reducing risk by abstaining on low-confidence cases [28, 29]. Nevertheless, relatively few studies in melanoma detection evaluate discrimination, calibration, and uncertainty together within a unified leakage-aware framework. Recent reviews of uncertainty in medical image analysis further reinforce the need for such integrated evaluation designs [30].

Increased attention on reporting and evaluating methods used with medical imaging AI has created a growing demand for validation transparency, performance variability documentation, and reproduction of study designs. The recent guidelines CLAIM, for example, emphasize the need to provide a detailed description of how data is split, what types of validations were performed, and how performance was assessed clinically [31]. Recent prospective and multi-center studies show that AI can assist in improving dermoscopic melanoma diagnosis in practice; however, these studies highlight the need for reasonable deployment assumptions and robustness in the design of studies [32].

3. MATERIALS AND METHODS

This study used an experimental quantitative approach to assess a deep learning–based melanoma detection system within a leakage-aware and reliability-centered framework. Dermoscopic images were analyzed through lesion-level cross-validation for classification purposes as well as for a comparative study with established baseline models that utilize transfer learning methods. Additionally, this study assessed the performance of the developed model from multiple dimensions. Besides assessing how well the developed model discriminates lesions, this study also evaluated calibration, post-hoc temperature scaling, and uncertainty-based selective prediction to provide a more clinically meaningful evaluation of model behavior.

Compared with many previous studies, this methodology places greater emphasis on leakage control and model reliability. It uses lesion-level GroupKFold validation instead of conventional image-level splitting, evaluates calibration in addition to discrimination, and incorporates leakage-free temperature scaling and uncertainty-based selective prediction. These elements allow the study to provide a more robust and clinically relevant assessment of melanoma detection performance.

3.1. Dataset and Task Definition

The data used in this research are based on a cohort of dermoscopic images that were obtained from the ISIC 2019 Training Collection [4]. The original data provided by the ISIC contains multi-class diagnostic labels, which are then converted into binary melanoma detection for screening/triage purposes, where melanoma (MEL) is the positive class and all other classes are considered negatives. The samples that have been labeled as unknown/ambiguous (UNK) are excluded from the analysis. All records that contain missing images are also removed in order to create a reproducible dataset.

To prevent information leaks, lesion identity is made explicit by employing the available metadata. Each image is assigned a grouping identifier, group_id, defined as lesion_id where available and image_id otherwise. The grouping identifiers are employed to ensure lesion-level separations between training, validation, and test split data sets across all folds.

Two cohorts were developed from the filtered dataset. The first was a balanced 5,000 image dataset that included (2,500 MEL / 2,500 non-MEL). This 5k dataset was designed as a sanity check to evaluate the sensitivity of the protocol and its calibration behavior in the case of equal class proportions. The second dataset was a 15,000-

image dataset created through stratified sampling to simulate a natural prevalence of melanoma (approximately 19%) and a fixed random seed. The 5k dataset was considered supplemental. The experimental cohorts and evaluation protocols are listed in Table 1 unless otherwise indicated. All other parameters were common across both cohorts and all folds including: resizing all images to 224 × 224 RGB pixels and normalizing them to [0,1]; transferring knowledge from pre-trained ImageNet backbones; optimizing the models using the Adam optimizer with a learning rate of 10-3 and a batch size of 16; early stopping based on performance on an internal validation split within each fold; and evaluating each test fold only one time.

Table 1. Experimental configuration and evaluation outputs.

Component	Sanity cohort	Natural-prevalence cohort
Cohort size	5,000 (Balanced)	15,000 (~19% MEL)
Split protocols	GroupKFold + StratifiedKFold	GroupKFold only
Purpose	Protocol sensitivity	Primary evaluation
Models	MobileNetV2 + EfficientNetB0	EfficientNetB0 (Primary)

3.2 Preprocessing and Fold-Safe Augmentation

All images are resized to a fixed size (224 x 224) RGB to match ImageNet-pre-trained models that have been used for transfer learning. Normalized pixel intensity values are then converted into floating-point values within the range [0, 1]. For each original image, denoted as Ii for the i sample with pixel values in [0, 255]. The normalized input for Ii computed as follows:

$$\tilde{I}_i = \frac{Resize(I_i,224,224)}{255} \quad (1)$$

In Equation 1, $Resize(\cdot)$ denotes resizing the image to 224 × 224 pixels; The image intensity is scaled using division by 255 so that it is mapped to the range [0,1]. Since this scale is normalized for each image individually, and not as a function of the global dataset mean and standard deviation, it thereby avoids leakage dataset-wide preprocessing. In order to reduce overfitting, data augmentation is performed exclusively on the training set for each fold. Examples of augmentations include random horizontal or vertical flips and minor photometric disturbances (e.g., brightness, contrast adjustments). No augmentations are made on either validation or test sets, thus when model performance is evaluated on unaltered held-out data. Figure 1 illustrates the fold-safe preprocessing pipeline and training-only augmentations.

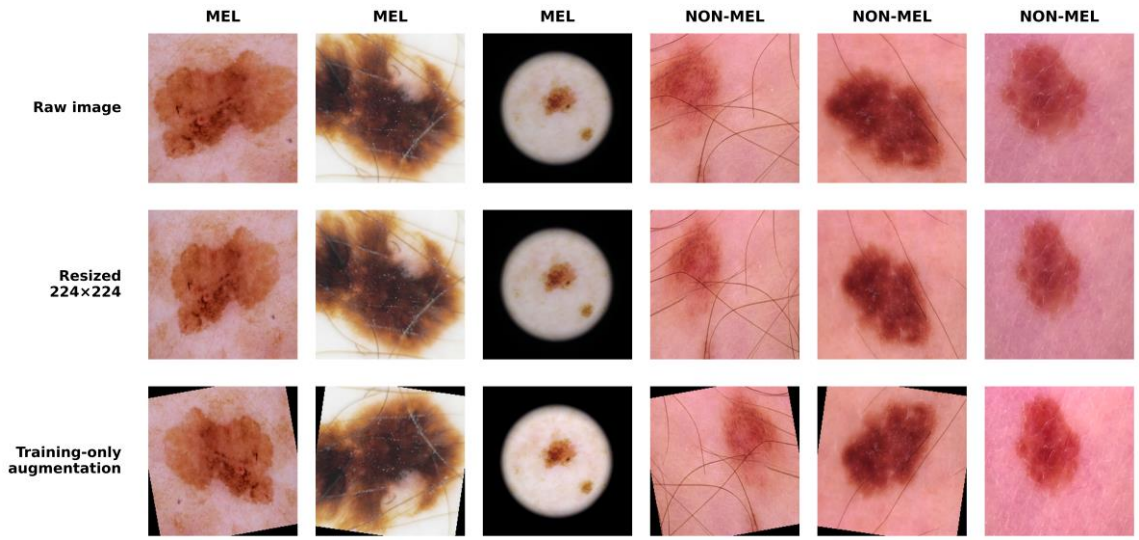

Figure 1. Representative dermoscopic images illustrating fold-safe preprocessing and training-only augmentation for melanoma and non-melanoma samples.

3.3. Leakage-Aware Grouping and Evaluation Protocols

Dermatological datasets generally include many images for each lesion; therefore, when randomly splitting images to test models, leakage of lesion-specific information can occur. This issue is mitigated by implementing an image-to-lesion group assignment method that is aware of potential leakage due to lesion identifiers.

$$g_i = \begin{cases} lesion\_id_i, if\ lesion\_id_i\ is\ available \\ image\_id_i, Otherwise \end{cases} \quad (2)$$

In Equation 2 gi is the group identifier for sample i; $lesion\_id\_i$ denotes the lesion-level identifier provided by the dataset, and $image\_id\_i$, is used as a singleton fallback when lesion identifiers are missing. This definition guarantees that all images belonging to the same lesion are treated as an inseparable group during splitting and that ungrouped images cannot create hidden overlaps.

The five-fold GroupKFold cross-validation was employed as the primary leakage-aware protocol, using the group vector as the basis for splits, while employing a stratified five-fold KFold protocol at the image-level as the secondary protocol to assess protocol sensitivity, specifically on the 5K balanced cohort. The two protocols utilize the same architecture and hyperparameters to guarantee that differences in performance are due solely to how data was split, and not to other training parameters. Algorithm 1 illustrates the leakage-aware fold generation that was used in all analyses of this work.

Algorithm 1. Leakage-aware fold construction (lesion-level GroupKFold)

Input:

Dataset D = {(I_i, y_i, lesion_id_i, image_id_i)}_{i=1..N}

Number of folds K = 5

Output:

$\{Train\_k, Val\_k, Test\_k\}\_\{k = 1..K\}$

$For\ each\ sample\ i\ =\ 1..N\ do$

$If\ lesion\_id\_i\ exists\ then\ set\ g\_i\ \leftarrow\ lesion\_id\_i$

$Else\ set\ g\_i\ \leftarrow\ image\_id\_i$

$End\ for$

$Apply\ GroupKFold\ with\ K\ folds\ on\ grouping\ vector\ g\ =\ \{g\_i\}$

$For\ each\ fold\ k\ =\ 1..K\ do$

$Test\_k\ \leftarrow\ \{\ (I\_i, y\_i)\ :\ g\_i\ assigned\ to\ fold\ k\ \}$

$TrainFull\_k \leftarrow\ D \setminus Test\_k$

$Split\ TrainFull\_k\ into\ Train\_k\ and\ Val\_k\ (internal\ split\ from\ training)$

$End\ for$

$Return\ \{Train\_k, Val\_k, Test\_k\}\_\{k = 1..K\}$

3.4. Model Architecture and Probabilistic Output

Two pre-trained (ImageNet) baseline models using transfer learning are evaluated: MobileNet-V2 and EfficientNet-B0. Each of these models has an identical lightweight classification module added to it, which contains an average pooling layer, a dropout layer, and a single fully connected (dense) layer that produces a single logit zi for each input sample. The probability of a melanoma is then determined using the sigmoid function:

$$p_i = \sigma(z_i) = 1/\left(1 + e^{(-z_i)}\right) \quad (3)$$

In Equation 3, z_i is the raw logit output for sample $i, \sigma(.)$ denotes the sigmoid transformation, and $p\_(i) \in$ (0,1) represents the model's estimated probability that sample i belongs to the positive (melanoma) class.

Unless explicitly stated otherwise, training is performed in a feature-extractor setting (frozen backbone with a trainable head) to ensure stable optimization under CPU-only execution while preserving consistent comparisons across protocols and architectures.

3.5. Training Procedure and Leakage-Free Validation Structure

The training of the model is performed in a fold-wise manner. In each fold k, $Train\_k$ is used to learn the parameters of the model; Val_k is used for early stopping and calibration fitting and $Test\_k$ is reserved for final testing purposes only. As such, this nested design ensures that there is no leakage from hyperparameter tuning or calibration into the test evaluation.

Optimization employs the Adam optimizer using a learning rate of 10-3, a batch size of 16, and an early stopping criterion based on validation area under the ROC curve (AUC). Under natural prevalence conditions, class-balanced sampling is employed during training to ensure stability in the optimization process; however, the test fold distribution remains unaltered to provide clinically relevant results. Algorithm 2 depicts the fold-wise training approach.

Algorithm 2. Fold-wise training with early stopping

Input:

Train_k, Val_k, backbone ∈ {MobileNetV2, EfficientNetB0},

Learning rate $\alpha = 10^{-3}$, batch size B = 16, patience P

Output:

Trained parameters θ_k (best checkpoint on Val AUC)

Initialize model with ImageNet − pretrained backbone + binary head
Freeze backbone; keep head trainable (feature − extractor setting)
Initialize Adam optimizer with learning rate α
bestAUC ← −∞ ; patienceCounter ← 0
While patienceCounter < P do
    Sample minibatch of size B from Train_k
    Apply fold − safe augmentation to minibatch (TRAIN ONLY)
    Forward pass → logits z, probabilities p = sigmoid(z)
    Compute binary cross − entropy loss
    Update trainable weights using Adam
    Evaluate AUC on Val_k
    If AUC improves bestAUC then
        Save checkpoint; bestAUC ← AUC; patienceCounter ← 0
    Else
        patienceCounter ← patienceCounter + 1
End while
Load best checkpoint and set θ_k accordingly
Return θ_k

3.6. Post-Hoc Calibration Via Temperature Scaling

To support reliable probabilities in clinical thresholding and risk communication, the reliability of the model's probabilities needs to be quantified; this can be done with two metrics: ECE and Brier Score. Post-hoc temperature scaling was also explored [33] which is a leak-free calibration technique that scales the logits of the model by a positive scalar T and then calculates the calibrated probabilities:

$$p_i^{(T)} = \sigma(z_i/T) \quad (4)$$

As mentioned in Equation 4, the temperature parameter is $T > 0$. Higher temperatures cause lower confidence in the probabilities (i.e., they are shifted toward 0.5), but they will not alter the ranking of the items. Lower temperatures will increase the confidence of the probabilities. Importantly, the temperature T is calculated using only the validation data and then applied to the test data without any "leakage" into the test data for the purpose of calibration. To be more specific, during cross-validation, each training fold is further split internally into $train\_sub$ or val_sub subsets within each training fold, learn the temperature T from the logits generated by val_sub by minimizing the negative log-likelihood, and apply this one time to the test fold that was previously held out. The test data is never used for selecting or tuning the temperature T or the hyperparameters for T. The details of the

algorithm for learning the temperature T on the validation set and applying it to the test set can be seen in Algorithm 3.

Algorithm 3. Leakage-free temperature scaling (validation-only fitting)

Input:
Validation logits $\{z\_i\}\_\{i \in Val\_k\}$, *labels* $\{y\_i\}\_\{i \in Val\_k\}$,
Test logits $\{z\_j\}\_\{j \in Test\_k\}$
Output:
$Temperature T_k$, $calibrated test probabilities p_j^{(T_k)}$
Initialize temperature $T \leftarrow 1.0$
Define calibrated prob on validation: $p\_i\hat{}(T) = sigmoid(z\_i / T)$
Fit T by minimizing validation NLL:
$NLL(T) = \Sigma\_\{i \in Val\_k\} [-y\_i \log p\_i\hat{}(T) - (1 - y\_i) \log(1 - p\_i\hat{}(T))]$
Obtain optimal temperature: $T\_k \leftarrow argmin\_\{T > 0\} NLL(T)$
For each test sample $j \in Test\_k$ *do*
Compute calibrated probability: $p\_j\hat{}(T\_k) \leftarrow sigmoid(z\_j / T\_k)$
End for
Return T_k *and* $\{p\_j\hat{}(T\_k)\}$

3.7. Uncertainty Estimation using Monte Carlo Dropout

To enable safety-aware behavior, predictive uncertainty is estimated using Monte Carlo Dropout at inference time [34]. For each test sample i, dropout is kept active, and the model is evaluated for M=30 stochastic forward passes. The predictive mean probability is:

$$\bar{p}_i = (1/M) \sum (m = 1)^M p_i^{(m)} \quad (5)$$

In Equation 5, M denotes the number of stochastic passes, $p_i^{(m)}$ is the predicted probability at pass m, and $\bar{p}_i$represents the ensemble-averaged predictive probability.

Uncertainty is summarized using predictive entropy:

$$H(\bar{p}_i) = -\bar{p}_i log(\bar{p}_i) - (1 - \bar{p}_i) log(1 - \bar{p}_i) \quad (6)$$

The value in Equation 6 $H(\bar{p}_i)$ is higher when the model's confidence grows closer to 0.5; conversely, it will grow smaller when the model is highly confident (near 0 or 1). This scalar-based method represents an uncertainty score that can relate directly to error rates and be selectively used as a basis for making predictions. To ensure the uncertainty analysis remains computationally feasible with the requirement that the full dataset be strictly evaluated by a holdout portion of the dataset, a single representative subset of a held-out portion of the dataset was used as the basis for the uncertainty-driven selective prediction. Specifically, uncertainty-driven selective prediction was performed on a single predetermined subset of a 15k lesion-level GroupKFold evaluation (Fold 2). The Monte Carlo Dropout Inference was also used on this held-out subset's test partition. Fold 2 was randomly selected before looking at test results and uncertainty metrics, thereby eliminating the possibility that we were optimizing the uncertainty metric based on how well it predicted the test set. Algorithm 4 illustrates the Monte Carlo Dropout procedure.

Algorithm 4. Monte Carlo Dropout uncertainty estimation (M = 30)

Input:
Trained model θ_k, *test set* $Test\_k$, *MC passes* $M = 30$
Output:
Predictive mean $\{\bar{p}\_i\}$, *entropy uncertainty* $\{H(\bar{p}\_i)\}$
Enable dropout during inference (*stochastic forward passes*)
For each test sample $i \in Test\_k$ *do*
For $m = 1..M$ *do*
Run forward pass $\rightarrow$ *probability* $p\_i\hat{}(m)$
End for
Compute predictive mean: $\bar{p}\_i \leftarrow (1/M) \Sigma\_(m = 1)\hat{}M\ p\_i\hat{}(m)$
Compute predictive entropy:
$H(\bar{p}\_i) \leftarrow -\bar{p}\_i \log(\bar{p}\_i) - (1 - \bar{p}\_i) \log(1 - \bar{p}\_i)$
End for
Return $\{\bar{p}\_i\}$ and $\{H(\bar{p}\_i)\}$.

3.8. Selective Prediction and Risk–Coverage Analysis

When uncertainty becomes usable in clinical decision-making, it will be because of the use of uncertainty to determine when the model can stop predicting. Therefore, uncertainty-driven selective predictive performance using risk/coverage analysis was used. Test cases are sorted based on increasing entropy. Only the highest-confidence portion of test data is selected for use in automated prediction. Selective risk at a desired coverage level $c \in (0,1]$.The selective risk is defined as the error rate on the retained set as:

$$Risk(c) = (1/|S_c|) \sum(i \in S_c) \mathbb{1}(\hat{y}_i \neq y_i) \quad (7)$$

In Equation 7, $\hat{y}\_i$ is the predicted class label (e.g., thresholding pi at 0.5), $1[\cdot]$ is the indicator function, and $Risk(c)$ is the error rate among retained cases. By sweeping c from 1.0 downwards, a risk–coverage curve is obtained. Algorithm 5 shows the full selective prediction workflow.

Algorithm 5. Risk–coverage curve for uncertainty-driven selective prediction

Input:
  Test set Test_k, uncertainties $\{H(\bar{p}\_i)\}$, *predictions* $\{\hat{y}\_i\}$, *labels* $\{y\_i\}$
  Coverage grid $C = \{1.0, 0.95, \ldots, 0.10\}$
Output:
  $Risk - coveragepairs$ {(c, Risk(c))}{c∈C}
Sort test samples by increasing uncertainty $H(\bar{p}\_i)$
Let the sorted index list be π
For each coverage value $c \in C$ *do*
  $n \leftarrow \lfloor c \cdot N \rfloor$
  $S\_c \leftarrow \{\pi 1, \pi 2, \ldots, \pi n\}$ *(retain n most confident samples)*
  Compute selective risk:
  $Risk(c) \leftarrow (1/|S\_c|)\ \Sigma\_(i \in S\_c)\ \mathbb{1}(\hat{y}\_i \neq y\_i)$
End for
Return $\{(c, Risk(c))\}\_\{c \in C\}$

3.9. Evaluation Metrics and Statistical Robustness

Accuracy and the Area Under the Curve of the Receiver Operator Characteristic (ROC Curve) are used to assess discrimination. In addition to the data provided, unless otherwise indicated, the accuracy is determined by using a fixed decision threshold of 0.5 on the probability estimates. Since the optimal decision threshold for screening/triaging will depend on the particular application, the area under the curve is treated as the primary threshold-independent discrimination metric. Reliability is assessed with the Brier Score and ECE. The Brier Score is calculated as:

$$Brier = (1/N) \sum(i = 1)^N (p_i - y_i)^2 \quad (8)$$

In Equation 8, N is the number of evaluated samples, p_i is the predicted probability, and y_i is the binary ground-truth label.

ECE is computed via confidence binning. Let the predictions be partitioned into K. For bin B_k, empirical accuracy is acc(Bk) and mean confidence is $conf(B_k)$. Then:

$$ECE = \sum(k = 1)^K (|B_k|/N)|acc(B_k) - conf(B_k)| \quad (9)$$

In Equation 9, $|B_k|$denotes the number of samples in bin k, and the term $|acc(B_k) - conf(B_k)|$measures bin-wise miscalibration. All metrics are calculated on separate test sets that were withheld from training and are expressed as mean ± standard deviation across folds. For paired studies (e.g., comparing lesion-based and image-based protocols; baseline vs. temperature scaling), this study reports both the results of paired t-tests/Wilcoxon signed rank tests, in addition to paired effect sizes (Cohen's d) and the bootstrap confidence intervals for the average difference of each pair, because p-values should be used with caution.

3.10. Implementation Environment

All experiments are implemented in Python using TensorFlow and executed on a Windows 64-bit CPU workstation (13th Gen Intel® Core™ i7-13620H, 32 GB RAM) to maintain reproducibility and stable

benchmarking across protocols and architectures. The same hyperparameter configuration was applied to each protocol and fold of data for each experiment. For example, transfer learning was performed using the Adam Optimizer with a learning rate of 10-3, a batch size of 16, and an early stop on the validation area under the receiver operating characteristic curve (AUC) was employed. In addition, the calibration was performed as temperature scaling solely on validation datasets that were sampled from the training dataset for the current fold, and a Monte Carlo Dropout-based uncertainty estimate with M = 30 was used for prediction. Training time on the CPU system was generally in the range of 1–2 hours per fold, depending on the backbone and cohort size, and early stopping generally occurred in 8 – 12 epochs.

4. RESULTS

This section presents the findings in two parts to align with the research's contribution to the field. First, the 15,000-image natural prevalence cohort was analyzed using a leakage-aware five-fold GroupKFold Cross-Validation at the lesion level. Second, a sanity check for protocol sensitivity was provided with respect to the type of split (lesion vs image) when there is no leakage by examining the 5,000-image balanced cohort, where the proportion of classes is controlled. Then, this study examined how well MobileNetV2 performs compared to EfficientNetB0; whether it can be calibrated with leakage-free temperature scaling; and whether safety can be improved through uncertainty-driven selective prediction (risk-coverage), and qualitative visual inspection of errors. All of these results will be reported across all folds with variability, effect size, and confidence interval information to emphasize the reliability of the findings.

4.1. Protocol Sensitivity Under Controlled Conditions (5k Balanced)

To determine how the evaluation methodology affects the reporting of results, two protocols were compared using a well-balanced dataset of 5,000 images (2,500 melanoma and 2,500 non-melanoma). The two methods were tested with the same training method using MobileNetV2. These two methods differ in that one uses a GroupKFold for the lesion level with 5 folds, while the other uses a StratifiedKFold at the image level with 5 folds. This allows us to isolate the effect of the splitting method from changes in class distribution and facilitates comparison based on the split used to make each fold.

Table 2 shows the mean ± standard deviation of five-fold results for both discrimination/reliability metrics. The AUC values were greater when random image-level splitting was performed compared to random lesion-level splitting (0.8206 ± 0.0084 vs 0.8086 ± .0179) although both methods of splitting had similar values in terms of accuracy. This relationship was likely due to an optimistic bias in the model as a result of the image-level splitting that utilized correlated images from the same lesion that appeared in both the training set and the testing set. The results from the lesion level splitting exhibited much greater variability than the image-level splitting which may have resulted from greater heterogeneity of lesions that models had to generalize to.

Table 2. Protocol comparison on the 5k balanced cohort (MobileNetV2, mean ± std across folds).

Protocol	ACC (mean ± std)	AUC (mean ± std)	Brier (mean ± std)
GroupKFold (Lesion-level)	0.7224 ± 0.0243	0.8086 ± 0.0179	0.1809 ± 0.0090
StratifiedKFold (Image-level)	0.7340 ± 0.0059	0.8206 ± 0.0084	0.1754 ± 0.0034

To prevent over-interpretation of minimal mean differences, protocol-specific mean differences were also evaluated in a fold-by-fold comparison as shown in Table 3. Only five folds were available for hypothesis testing; therefore, the statistical power is limited, and the p-values should be interpreted with caution. Accordingly, they are treated as supporting evidence rather than as a decision-making criterion. The overall findings show that protocol choice impacts the average performance and variability in the reports and emphasize the necessity for an evaluation at the lesion level that considers leakage if lesion-level correlation exists.

Table 3. Paired statistical comparison: lesion-level vs random (MobileNetV2, 5 folds).

Metric	Paired t-test p-value	Interpretation
ACC	0.298	The difference is small relative to fold-to-fold variability
AUC	0.070	Directionally consistent with higher AUC under image-level splitting; limited power (n=5)
Brier	0.167	Slightly higher (worse) Brier under lesion-level evaluation; limited power (n=5)

The fold-by-fold variability in performance metrics for ACC, AUC, and Brier Score is shown in Figure 2 for the case where GroupKFold was used to define the lesion-level folds versus StratifiedKFold was used to create the image-level folds. GroupKFold creates more variability across the folds and shows a slight shift toward lower discrimination and a higher Brier value, as would be expected in a more difficult and therefore clinically relevant setting when leakage is controlled for.

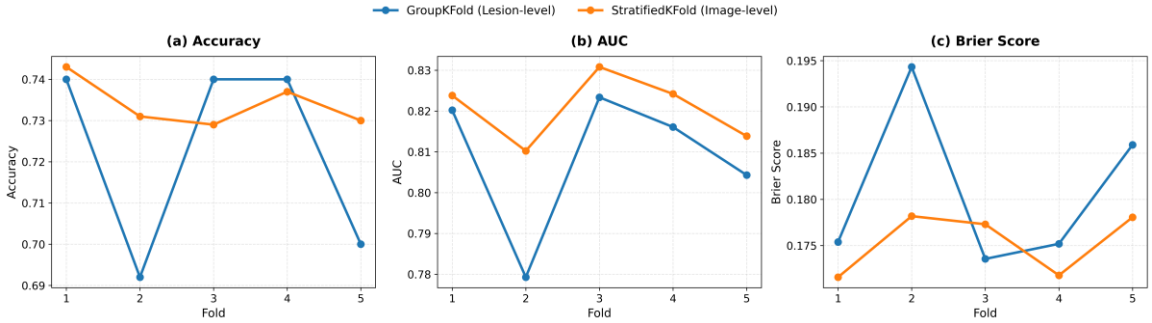

Figure 2. Protocol sensitivity (5k balanced, MobileNetV2): GroupKFold vs. StratifiedKFold across (a) ACC, (b) AUC, and (c) Brier.

4.2. Robustness Across Transfer-Learning Baselines (5k Balanced, Lesion-Level)

To assess whether conclusions depend on the chosen backbone, two transfer-learning baselines were evaluated under the same leakage-aware lesion-level protocol: MobileNetV2 and EfficientNetB0. This comparison helps separate protocol-driven effects from architecture-dependent behavior.

Table 4 shows that EfficientNetB0 has similar discrimination as MobileNetV2 but with greater baseline calibration. Specifically, EfficientNetB0 produces a lower ECE (0.0420±0.0057) compared to MobileNetV2 (0.0699 ± 0.0310), which means that its probability predictions are, on average, better calibrated against actual correctness before any additional calibration is applied to those predictions. These findings indicate that the reliability of predictions based on probabilities can depend on the type of architecture used; therefore, evaluating architectures may be necessary to evaluate overall prediction reliability as opposed to just using one model.

Table 4. Model robustness under leakage-aware lesion-level evaluation (5k balanced).

Model	ACC (mean ± std)	AUC (mean ± std)	Brier (mean ± std)	ECE (mean ± std)
MobileNetV2	0.7160 ± 0.0191	0.8016 ± 0.0136	0.1868 ± 0.0087	0.0699 ± 0.0310
EfficientNetB0	0.7244 ± 0.0121	0.8114 ± 0.0206	0.1778 ± 0.0089	0.0420 ± 0.0057

A real-world consequence of this is that while different models may have very similar AUC values, they may still have substantial differences in their ability to be calibrated at a given level of confidence. Reliability metrics thus offer complementary evidence about model performance beyond discrimination for use in workflows where decisions are made based on thresholds or levels of confidence.

4.3. Reliability and Post-Hoc Calibration (Leakage-Free Temperature Scaling) (5k Balanced)

Next, post-hoc temperature scaling was evaluated using a leakage-free procedure: the temperature parameter is fitted only on an internal validation split derived from the training fold and is never tuned on the test fold.

MobileNetV2 is calibrated in terms of temperature scaling as well as in terms of lesion level using Table 5. The average ECE is reduced by 0.0102 or 14.6% from 0.0699 ± 0.0310 to 0.0597 ± 0.0245. Although the Wilcoxon p-value was marginally significant at .0625 when n = 5, there is a large Cohen's d = 1.1459, and the 95% confidence interval for the mean improvement of ΔECE is strictly positive [0.0035, 0.0171] based on the bootstrap. Additionally, although the Brier Score indicates an even less dramatic improvement than does ECE, it also appears to indicate a statistically significant improvement with a positive confidence interval, indicating an increase in calibration performance when no leakage is present during model training.

However, EfficientNetB0 does not get an advantage in terms of temperature scaling: mean ECE goes up from 0.0420 ± 0.0057 to 0.0488 ± 0.0074. The confidence interval crosses zero, and there is a negative effect size, which indicates that the use of temperature scaling is not always beneficial when the initial calibration of models is strong; this motivates the evaluation of calibration interventions based on empirical evidence rather than assuming that all model families will benefit equally from these methods.

Table 5. Temperature scaling effects (5k balanced, lesion-level; leakage-free validation-only fitting).

Model	Metric	Baseline (mean ± std)	Temp-scaled (mean ± std)	Paired t p	Wilcoxon p	Cohen's d	95% CI of mean improvement (baseline − TS)
MobileNetV2	ECE	0.0699 ± 0.0310	0.0597 ± 0.0245	0.0625	0.0625	1.1459	[0.0035, 0.0171]
MobileNetV2	Brier	0.1868 ± 0.0087	0.1850 ± 0.0071	0.0971	0.1250	0.9650	[0.00033, 0.00317]
EfficientNetB0	ECE	0.0420 ± 0.0057	0.0488 ± 0.0074	0.1997	0.3125	−0.6863	[−0.01426, 0.00104]
EfficientNetB0	Brier	0.17785 ± 0.00889	0.17799 ± 0.00928	0.7553	0.8125	−0.1493	[−0.00081, 0.00069]

Figure 3 shows the reliability diagram (or calibration curve) for a pre-determined held-out fold (Fold 2) from the 15k natural prevalence, lesion-level GroupKFold cross-validation evaluation. The observed empirical accuracy in each probability bin follows the identity line closely; therefore, it appears as though the confidence levels have been calibrated well in this particular fold. This type of qualitative relationship between the observed empirical accuracy across all probability bins and the identity line for the reliability diagram is similar to what was found for the 15k cohort and gives a visual way to understand how the model was calibrated with respect to the leakage-aware cross-validation evaluation method.

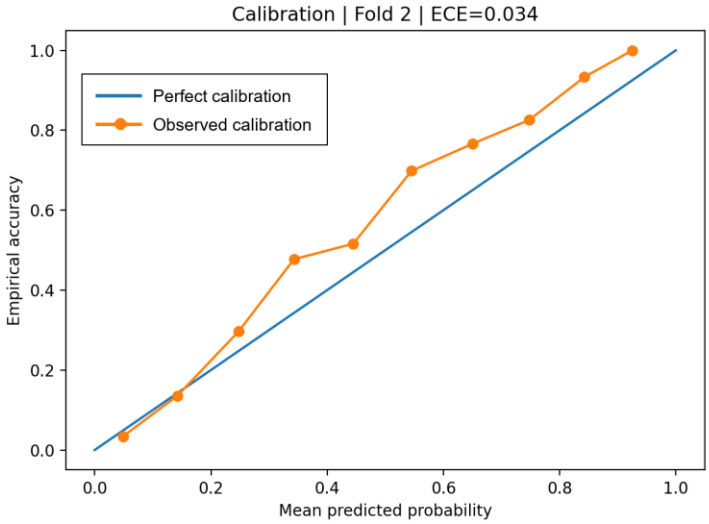

Figure 3. Calibration curve (reliability diagram).

4.4. Scale-Up Evaluation Under Natural Prevalence (15k, Lesion-Level)

To evaluate in a way that better reflects real-world prevalence and to minimize balanced-cohort bias in the testing, EfficientNetB0 was used on a 15,000-image natural-prevalence data set (using lesion-level 5-Fold

GroupKFold). This study used this as the main evaluation scenario since it provides clinically realistic class prevalence (approximately 19% for melanoma overall) while also being aware of potential leakage from splits.

Fold-by-fold results are reported in Table 6, including test-fold prevalence, accuracy (ACC), area under the ROC curve (AUC), Brier Score, and ECE. The mean discrimination for the model was AUC = 0.8224 ± 0.0242, and the reliability of the model was very good (strong) with a Brier Score of 0.1099 ± 0.0059 and an ECE of 0.0276 ± 0.0121. Fold-by-fold variability remained significant (range: 0.8065 – 0.8634); thus, multi-fold reporting will be valuable to provide more comprehensive summaries than single-split reporting.

Table 6. 15k natural-prevalence lesion-level GroupKFold results (EfficientNetB0, fold-wise).

Fold	Test MEL prevalence	ACC	AUC	Brier	ECE
0	0.1553	0.8673	0.8074	0.1019	0.0170
1	0.1900	0.8433	0.8255	0.1145	0.0266
2	0.1903	0.8527	0.8634	0.1057	0.0344
3	0.1807	0.8427	0.8093	0.1157	0.0443
4	0.1763	0.8533	0.8065	0.1118	0.0156
Mean ± Std	0.1785 ± 0.0143	0.8519 ± 0.0100	0.8224 ± 0.0242	0.1099 ± 0.0059	0.0276 ± 0.0121

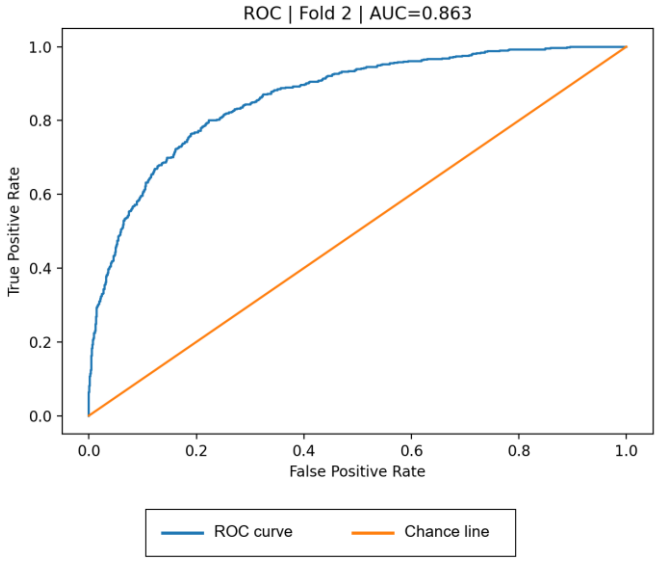

Figure 4. ROC curve of Fold 2.

The Receiver Operating Characteristic (ROC) Curve for the Pre-Specified Held-Out Test Fold (Fold 2), Which Was Evaluated at the 15K Natural-Prevalence Lesion Level, Is Displayed in Figure 4. The ROC Curve Provides a Visual Summary of the Sensitivity-Specificity Tradeoff Across Decision Thresholds. An AUC of 0.8634 Illustrates Discrimination Performance Under a Screening-Like Prevalence Setting in This Fold.

Table 7. Uncertainty and selective prediction summary (15k, Fold 2, MC Dropout M=30).

Quantity	Value
Test size	3000
MEL prevalence	0.1903
Mean entropy (Correct)	0.3337
Mean entropy (Wrong)	0.5365
Baseline error rate (≈1−ACC)	≈0.148
Risk@70% coverage	0.0648
Error rate in top-50 uncertain	0.44

Misclassified samples were significantly more uncertain than those classified correctly; the average entropy of correct predictions was 0.3337 compared to 0.5365 for misclassifications (Δ≈0.203), which suggests that uncertainty

tracks with error. What is perhaps even more important is that uncertainty has some ability to support selective prediction. For example, at 70% coverage, the risk is 0.0648 (risk@0.70 = 0.0648). In comparison to the baseline error rate of approximately 1−ACC ≈ 0.148 for the fully covered set, this represents an approximate 56% reduction in error when selecting only the retained predictions. The high-uncertainty tail is also significantly enriched with errors; the observed error rate for the top 50 most uncertain cases is 44% (≈3 times the baseline), suggesting that these cases should be selectively deferred for expert review as opposed to being automatically decided upon.

4.5. Uncertainty-Driven Selective Prediction and Safety Analysis

To establish that the uncertainty estimate will be useful for providing an actionable safety signal, the uncertainty estimate is first evaluated on the designated fold designated in advance for strict holdout testing within the15K lesion-level GroupKFold evaluation (test size = 3,000; Fold #2; MEL prevalence = 0.1903) to ensure that computational complexity does not limit the ability to perform holdout testing while utilizing CPU-only. The same MC Dropout (M=30), predictive entropy is calculated from the mean probability of each of the multiple stochastic forward passes. The same process was then repeated for each of the 5 folds of the 15K evaluation as shown in Table 8. To verify that uncertainty-driven selective prediction is not specific to a single fold in this study, MC Dropout was applied across all five folds of the 15K lesion-level validation, as shown in Table 8, and found the same trends as before: correct vs. incorrect predictive entropy and the corresponding Selective Prediction Risk @ 70%.

Table 8. Fold-wise uncertainty consistency on the 15k cohort (MC Dropout, M=30): entropy separation (correct vs. wrong) and risk at 70% coverage.

Fold	Test size	MEL prevalence	Mean entropy (Correct)	Mean entropy (Wrong)	Risk @ 70% coverage
0	3000	0.1553	0.3680	0.5425	0.0714
1	3000	0.1900	0.3312	0.5173	0.0810
2	3000	0.1903	0.3676	0.5677	0.0648
3	3000	0.1807	0.3891	0.5580	0.0833
4	3000	0.1763	0.3467	0.5082	0.0829
Mean ± Std	3000	0.1785 ± 0.0143	0.3605 ± 0.0222	0.5387 ± 0.0256	0.0767 ± 0.0082

Table 8 shows that uncertainty behavior is consistently demonstrated in every single one of the cross-validation folds. The predictive entropy of wrong classifications is larger than for right classifications on average by 0.1802 ± 0.0279 (i.e., 0.3605 ± 0.0222 vs. 0.5387 ± 0.0256). More importantly, selective prediction stability has been demonstrated: the 70% risk estimate is 0.0767 ± 0.0082, which supports the idea that uncertainty-based delay provides a reliable safety margin, as opposed to an artifact of one particular fold.

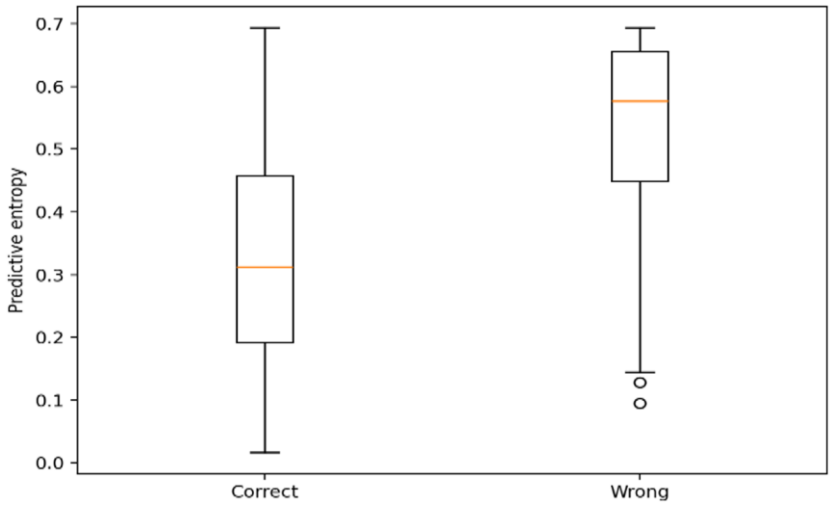

Figure 5. Uncertainty vs Error.

The distribution of predictive entropy values among the correct and incorrect classifications is compared in Figure 5, using the representative held-out test fold. The wrong classification group clearly shows an increase in entropy when compared to the "correct" classification group. Thus, misclassifications appear to generally be related to substantially increased levels of uncertainty.

This provides support for both the quantitative data found in Table 7 (higher mean entropy for errors) and demonstrates that while there was an overall greater amount of entropy for the erroneous cases, the elevation in entropy is systematic throughout the entire distribution of errors.

From a clinical perspective, this separation is useful because it indicates that uncertainty can provide a viable risk indicator. Therefore, the use of entropy as an indicator of risk could provide a means by which clinicians can identify patients whose results are likely to be unreliable, and thus better determine whether they should be automatically processed via a computerized system or deferred/subjected to a second review by a clinician.

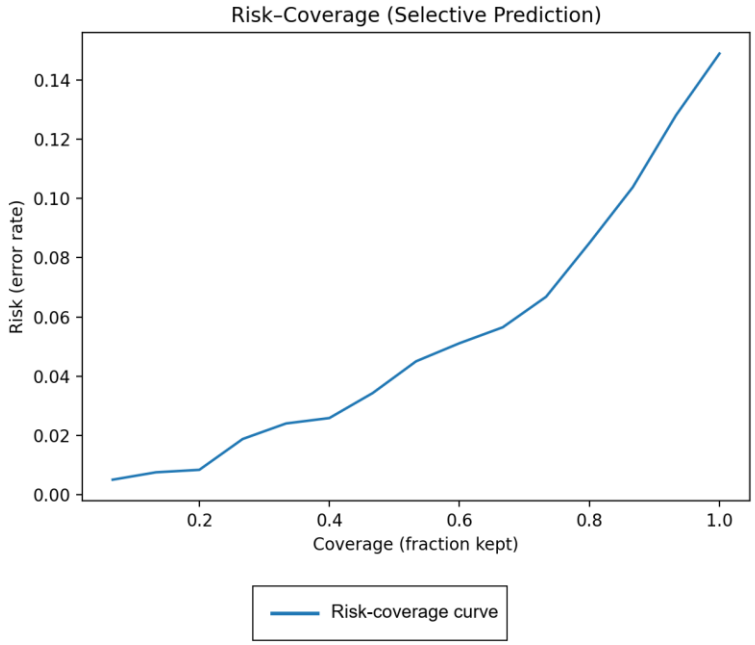

Figure 6. Risk–coverage curve.

Figure 6 shows how the Risk-Coverage tradeoff works for Selective Prediction with Uncertainty (SPU) on an example holdout test fold. As predictive entropy increases (Lower Entropy = Higher Confidence), samples are ordered from least to most uncertain, and the curve displays the error rate ("risk") of classification for the proportion of samples ("coverage") with the lowest predictive entropy (highest confidence). The monotonically decreasing relationship between risk and coverage demonstrates that SPU can successfully identify samples that will be difficult to classify; By abstaining from predicting those samples that have the greatest degree of uncertainty, SPU can significantly reduce the classification error on the retained predictions. For example, when the coverage is 70%, the risk is 6.68%, which is less than half of the baseline error rate of approximately 14.8%, which represents a 55% relative improvement in the accuracy of the retained set. This gives a clinically useful interpretation of safety: In screening or triage processes, the system can automatically process samples with high confidence, while sending samples with low confidence for an expert's evaluation, therefore reducing the potential for an incorrect automated decision.

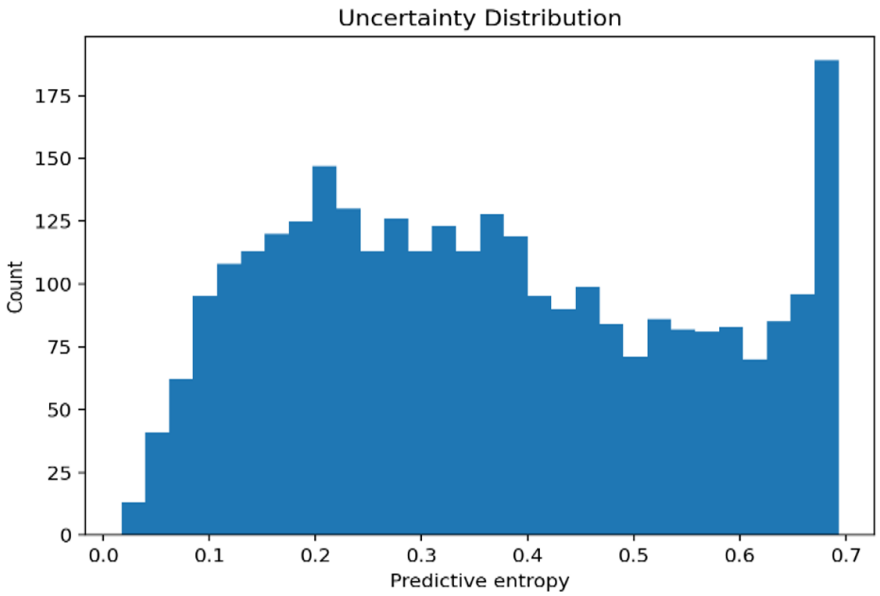

Figure 7. Uncertainty histogram.

Figure 7 illustrates the distribution of predictive entropy values as measured through MC Dropout for the representative held-out test fold. The majority of the samples are grouped in the lower-to-moderate entropy ranges. Thus, the model is very confident regarding a large proportion of the cases. Importantly, however, there is a marked high-entropy tail to this distribution. These high-entropy cases represent the most ambiguous of the predictions. The high-entropy tail is clinically important because the failure of the top 50 most uncertain samples (highest entropy) has an error rate of 44%. Thus, the uncertainty measure (entropy) identifies cases where deferral to an expert reviewer is likely to be beneficial. The distributional representation also helps provide clarity on how "high-risk" cases fall into place with respect to the overall uncertainty landscape.

4.6. Qualitative Error Analysis (Representative FP/FN/TP/TN)

Quantitative metrics can obscure clinically relevant failure patterns. Therefore, a qualitative inspection of representative false-positive cases, false negatives, and confident correct cases was performed, annotated with probability and entropy.

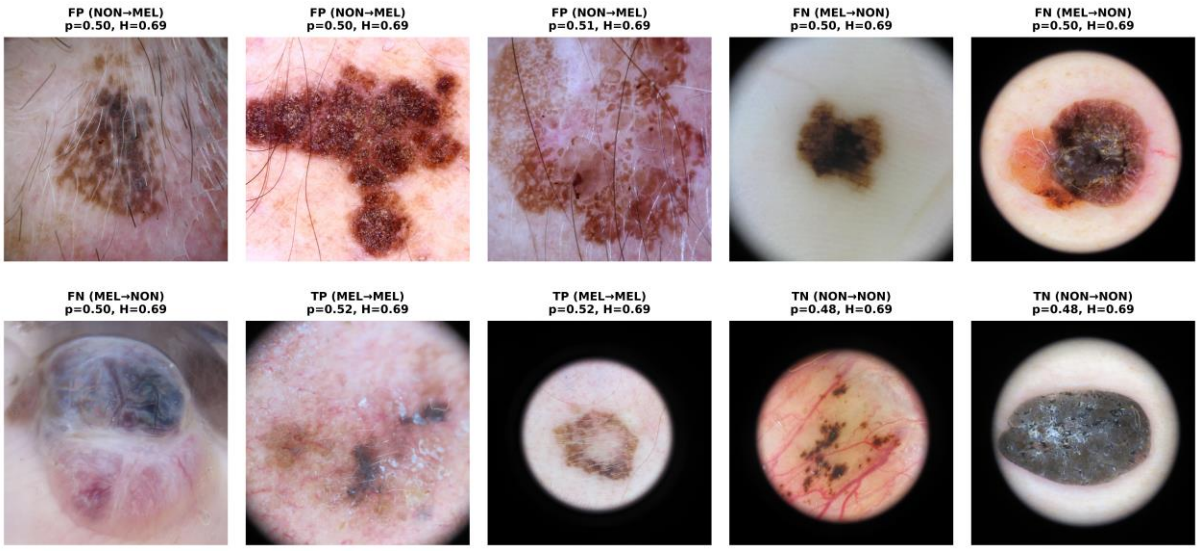

Figure 8. Error gallery FP/FN/TP/TN with p and entropy.

Figure 8 provides qualitative evidence supporting the uncertainty analysis. The presented false positives and false negatives are characterized by probabilities close to the decision threshold and elevated predictive entropy, consistent with the quantitative separation between correct and incorrect predictions observed in Figure 5. The clinical relevance of the false negative rate is significant because it reflects undiagnosed melanoma cases. Visually, the false negative examples are characterized by the presence of subtle pigment distribution and poorly defined lesion borders, which may have contributed to the models' uncertainty regarding classification and prediction. The false positive rate is associated with benign lesions having heterogeneous coloration and/or structure that can mimic malignancy. It appears that the confident correct predictions (TP/TN), as a group, have clear morphological characteristics and low entropy, consistent with the risk–coverage relationship illustrated in Figure 6, where withholding judgments for high-entropy cases resulted in fewer errors. In total, the examples within this gallery will improve interpretation and provide support for the use of uncertainty-based triage strategies. This is based upon the fact that the examples provided demonstrate that predictions that were made with a high degree of uncertainty correlated to visually ambiguous conditions.

5. DISCUSSION

This study went beyond headline accuracy and provided an assessment of how evaluation design, calibration, and uncertainty affect the interpretation of deep learning–based melanoma detection. The results from both the controlled protocol comparison on the 5k balanced cohort and the scaled-up natural-prevalence evaluation on the 15k cohort support the premise that, in clinical AI, discrimination metrics alone are insufficient and that the evaluation process itself, system calibration, and uncertainty-based safety measures should be explicitly reported. These findings are consistent with prior concerns in medical imaging AI that methodological choices can substantially influence reported performance [5].

These findings are consistent with recent evidence showing that deep learning can provide strong diagnostic performance in dermoscopic melanoma analysis while still requiring careful evaluation under realistic clinical conditions. A recent systematic review and meta-analysis [18] reported that deep learning algorithms achieved performance comparable to senior dermatologists overall and highlighted their potential to support clinical decision-making, although the authors also emphasized the need for larger and methodologically stronger validation studies. Similarly, a prospective multicenter study [19] showed that AI could improve melanoma diagnosis in heterogeneous patient-care settings, which supports the clinical relevance of AI-assisted workflows and is broadly consistent with the present study's reliability-centered evaluation perspective.

At the same time, recent prospective evidence also suggests that strong AI performance should be interpreted cautiously. In the study [18], the gain in balanced accuracy and sensitivity was accompanied by lower specificity, illustrating that improvement in one aspect of diagnostic performance may come with trade-offs in another. In addition, studies Ye et al. [18] and Marchetti et al. [35] have continued to show that real-world evaluation remains necessary because results obtained in retrospective or benchmark-driven settings may not fully reflect deployment conditions. This broader evidence aligns with the present study's emphasis on leakage-aware validation, calibration analysis, and uncertainty-guided selective prediction as necessary complements to headline discrimination metrics.

The study was not designed as a state-of-the-art benchmarking exercise. Instead, two widely accepted transfer-learning baselines, MobileNetV2 and EfficientNetB0, were evaluated under identical protocols, and the analysis focused on how validation protocol, reliability assessment, and safety-oriented analysis affect the conclusions drawn from model performance. In this respect, direct comparison with previously published melanoma classification studies should be interpreted cautiously, as differences in cohorts, preprocessing, and validation design may substantially affect the reported results.

The comparison of the lesion-level GroupKFold validation and the image-level split indicated that the method used to generate an evaluation protocol can have a material impact on reported performance. Image-level splits

generated slightly greater average AUC than did lesion-level splits, which supports the idea that images of the same lesion could be placed in both the training set and the test set and thus provide an advantage for the classifier. Though the differences in average discrimination were relatively small, they also demonstrated how image-level splits can mask variability in performance between folds that arises from heterogeneity within the lesions. As such, these findings lend support to using lesion-level, leakage-aware validation when identifiers are available for lesions [20].

MobileNetV2 and EfficientNetB0 performed similarly in terms of AUC, but EfficientNetB0 had a better baseline calibration on the natural prevalence testing, which is especially relevant as calibration has an impact on how to interpret the predicted probabilities in the clinical application using a threshold-based clinical workflow. In addition, with leakage-free post-hoc calibration, the temperature scaling helped to calibrate MobileNetV2 but was unable to improve the calibration of EfficientNetB0. As such, this suggests that calibration methods can be dependent on the architecture and context used for evaluation; therefore, it is recommended to evaluate empirically and not as routine practice. The larger cohort of patients at natural prevalence provided additional evidence of the robustness of the proposed evaluation pipeline under a variety of class distributions in the real world.

Uncertainty analysis provided an additional safety-oriented perspective. Predictive uncertainty estimated through Monte Carlo dropout was higher for misclassified cases and supported selective prediction by allowing lower-risk automated decisions to be retained while uncertain cases were deferred [27]. At reduced coverage, the observed decline in risk suggests that uncertainty estimation may be useful in assistive clinical workflows, particularly in triage settings where uncertain cases can be escalated for specialist review rather than forced into automated classification [36].

Several limitations should be acknowledged. First, although the 15k natural-prevalence cohort offers a more clinically realistic internal test environment, it remains derived from the ISIC archive and therefore does not replace external validation on independent datasets, institutions, or acquisition settings. Second, the use of frozen-backbone baselines and CPU-constrained training was intended to ensure computational stability and controlled comparison, but it may limit the absolute predictive performance achievable through more extensive fine-tuning. Third, uncertainty estimation relied on Monte Carlo dropout as a practical baseline, and its behavior may vary with the number of stochastic forward passes and the dropout configuration. These limitations should be addressed in future work through external validation and broader sensitivity analyses [37].

6. CONCLUSIONS

This study introduced a leakage-aware, reliability-centered evaluation framework for deep learning–based melanoma detection from dermoscopic images. Comparison of lesion-level GroupKFold validation with conventional random image-level splitting on a controlled balanced cohort showed that the evaluation protocol materially influences both reported discrimination performance and fold-to-fold variability. These findings indicate that lesion-level, leakage-aware validation should be preferred whenever lesion identifiers are available. To limit architecture-specific bias, two transfer-learning baselines, MobileNetV2 and EfficientNetB0, were evaluated under matched experimental conditions. In addition to discrimination, model reliability was assessed using the Brier score and expected calibration error. Under leakage-free post-hoc calibration, temperature scaling improved calibration for MobileNetV2 but not for EfficientNetB0, indicating that calibration effectiveness is model- and context-dependent and should be verified empirically rather than assumed. In the larger natural-prevalence cohort, EfficientNetB0 maintained stable discrimination and favorable reliability under lesion-level splitting, supporting the applicability of the proposed framework under more realistic class distributions. Uncertainty estimation via Monte Carlo dropout provided an additional safety-oriented component. Higher uncertainty was associated with prediction errors and enabled selective prediction, reducing risk when automated decisions were restricted to high-confidence cases. Taken together, these results support a reliability-first approach to melanoma screening AI, in which leakage-aware

validation, explicit calibration assessment, and uncertainty-informed decision strategies are treated as core requirements for trustworthy clinical deployment. The findings are also relevant to AI-assisted triage and teledermatology workflows, where deferral of uncertain cases for specialist review may improve patient safety and decision quality. Future work should prioritize external validation across independent datasets, institutions, and acquisition settings to determine generalizability. Further investigation of more expressive calibration methods, alternative uncertainty-estimation approaches, and fine-tuning strategies is also warranted, provided that leakage-free evaluation and reproducible reporting are preserved.

Funding: This research received no external funding.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229-263, 2024. https://doi.org/10.3322/caac.21834

[2] H. K. Jeong, C. Park, R. Henao, and M. Kheterpal, "Deep learning in dermatology: A systematic review of current approaches, outcomes, and limitations," *JID Innovations*, vol. 3, no. 1, p. 100150, 2023. https://doi.org/10.1016/j.xjidi.2022.100150

[3] B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhattacharya, and K. Lakshmanna, "Skin lesion classification of dermoscopic images using machine learning and convolutional neural network," *Scientific Reports*, vol. 12, no. 1, p. 18134, 2022. https://doi.org/10.1038/s41598-022-22644-9

[4] International Skin Imaging Collaboration, *ISIC 2019 challenge: Skin lesion analysis towards Melanoma detection*. Shenzhen, China: International Skin Imaging Collaboration, 2019.

[5] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," *npj Digital Medicine*, vol. 5, no. 1, p. 48, 2022. https://doi.org/10.1038/s41746-022-00592-y

[6] K. Abhishek, A. Jain, and G. Hamarneh, "Investigating the quality of DermaMNIST and Fitzpatrick17k dermatological image datasets," *Scientific Data*, vol. 12, no. 1, p. 196, 2025. https://doi.org/10.1038/s41597-025-04382-5

[7] L. Sasse *et al.*, "Overview of leakage scenarios in supervised machine learning," *Journal of Big Data*, vol. 12, no. 1, p. 135, 2025. https://doi.org/10.1186/s40537-025-01193-8

[8] R. Hammad and A. AL-Slamat, "A deep learning approach for predicting hospital length of stay for people with diabetes using electronic health records," *International Journal of Advances in Soft Computing and its Applications*, vol. 18, no. 1, pp. 192-206, 2026. https://doi.org/10.15849/ijasca.v18i1.24

[9] R. D. Riley *et al.*, "Evaluation of clinical prediction models (part 2): How to undertake an external validation study," *BMJ*, vol. 384, p. e074820, 2024. https://doi.org/10.1136/bmj-2023-074820

[10] M. Riera-Marín, J. G. López, J. Rodríguez-Comas, M. A. G. Ballester, and A. Galdran, "Multi-rater calibration error estimation," presented at the International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Springer, 2025.

[11] A.-J. Rousseau, T. Becker, S. Appeltans, M. Blaschko, and D. Valkenborg, "Post hoc calibration of medical segmentation models," *Discover Applied Sciences*, vol. 7, no. 3, p. 180, 2025. https://doi.org/10.1007/s42452-025-06587-0

[12] B. Lambert, F. Forbes, S. Doyle, H. Dehaene, and M. Dojat, "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis," *Artificial Intelligence in Medicine*, vol. 150, p. 102830, 2024. https://doi.org/10.1016/j.artmed.2024.102830

[13] A. S. Jarab, W. A. Al-Qerem, L. M. Khdour, Y. A. Mimi, and M. R. Khdour, "New emerging treatment options for metastatic melanoma: A systematic review and meta-analysis of skin cancer therapies," *Archives of Dermatological Research*, vol. 316, no. 10, p. 735, 2024. https://doi.org/10.1007/s00403-024-03467-2

[14] N. Codella *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," *arXiv preprint arXiv:1902.03368*, 2019. https://doi.org/10.48550/arXiv.1902.03368

[15] 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, no. 1, p. 180161, 2018. https://doi.org/10.1038/sdata.2018.161

[16] H. Aljawawdeh, "A deep learning-based integrative framework for cancer subtype classification leveraging multi-omics data: A software engineering approach," *International Journal of Advances in Soft Computing and its Applications*, vol. 18, no. 1, pp. 242-262, 2026. https://doi.org/10.15849/ijasca.v18i1.36

[17] K. M. Aljebory, Y. M. Jwmah, T. S. Mohammed, and A. Al Mamari, "Comparative study of LSTM, GRU, and BRNN performance: Large-scale data analytics (EMG Signal Classification)," *International Journal of Advances in Soft Computing and its Applications*, vol. 18, no. 1, pp. 98-117, 2026. https://doi.org/10.15849/ijasca.v18i1.28

[18] Z. Ye *et al.*, "Deep learning algorithms for melanoma detection using dermoscopic images: A systematic review and meta-analysis," *Artificial Intelligence in Medicine*, vol. 155, p. 102934, 2024. https://doi.org/10.1016/j.artmed.2024.102934

[19] M. Giulini *et al.*, "Combining artificial intelligence and human expertise for more accurate dermoscopic melanoma diagnosis: A 2-session retrospective reader study," *Journal of the American Academy of Dermatology*, vol. 90, no. 6, pp. 1266-1268, 2024. https://doi.org/10.1016/j.jaad.2023.12.072

[20] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Medical Image Analysis*, vol. 75, p. 102305, 2022. https://doi.org/10.1016/j.media.2021.102305

[21] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," presented at the International Conference on Machine Learning, PMLR, 2017.

[22] S. A. Balanya, J. Maroñas, and D. Ramos, "Adaptive temperature scaling for robust calibration of deep neural networks," *Neural Computing and Applications*, vol. 36, no. 14, pp. 8073-8095, 2024. https://doi.org/10.1007/s00521-024-09505-4

[23] T. Joy, F. Pinto, S.-N. Lim, P. H. S. Torr, and P. K. Dokania, "Sample-dependent adaptive temperature scaling for improved calibration," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14919-14926, 2023. https://doi.org/10.1609/aaai.v37i12.26742

[24] C. Tomani, D. Cremers, and F. Buettner, "Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration," in *Computer Vision–ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Lecture Notes in Computer Science, vol. 13673. Cham, Switzerland: Springer, 2022, pp. 555–569. https://doi.org/10.1007/978-3-031-19778-9_32

[25] B. Kocak *et al.*, "Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations," *European Journal of Radiology Artificial Intelligence*, vol. 3, p. 100030, 2025. https://doi.org/10.1016/j.ejrai.2025.100030

[26] L. Gerischer, P. Doksani, S. Hoffmann, and A. Meisel, "New and emerging biological therapies for myasthenia gravis: A focussed review for clinical decision-making," *BioDrugs*, vol. 39, no. 2, pp. 185-213, 2025. https://doi.org/10.1007/s40259-024-00701-1

[27] S. H. Ahn *et al.*, "Uncertainty quantification in automated detection of vertebral metastasis using ensemble Monte Carlo dropout," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 5, pp. 2700-2715, 2025. https://doi.org/10.1007/s10278-024-01369-3

[28] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *Journal of Machine Learning Research*, vol. 11, no. 53, pp. 1605–1641, 2010.

[29] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," presented at the International Conference on Machine Learning, PMLR, 2019.

[30] L. Huang, S. Ruan, Y. Xing, and M. Feng, "A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods," *Medical Image Analysis*, vol. 97, p. 103223, 2024. https://doi.org/10.1016/j.media.2024.103223

[31] J. Mongan, L. Moy, and C. E. Kahn Jr, "Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers," *Radiology: Artificial Intelligence*, vol. 2, no. 2, p. e200029, 2020. https://doi.org/10.1148/ryai.2020200029

[32] L. Heinlein *et al.*, "Prospective multicenter study using artificial intelligence to improve dermoscopic melanoma diagnosis in patient care," *Communications Medicine*, vol. 4, no. 1, p. 177, 2024. https://doi.org/10.1038/s43856-024-00598-5

[33] S. McKenna and J. Carse, "Calibrating where it matters: Constrained temperature scaling," *arXiv preprint arXiv:2406.11456*, 2024. https://doi.org/10.48550/arXiv.2406.11456

[34] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," presented at the International Conference on Machine Learning, PMLR, 2016.

[35] M. A. Marchetti *et al.*, "Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study)," *npj Digital Medicine*, vol. 6, no. 1, p. 127, 2023. https://doi.org/10.1038/s41746-023-00872-1

[36] J. K. Winkler *et al.*, "Assessment of diagnostic performance of dermatologists cooperating with a convolutional neural network in a prospective clinical study: Human with machine," *JAMA Dermatology*, vol. 159, no. 6, pp. 621-627, 2023. https://doi.org/10.1001/jamadermatol.2023.0905

[37] Y. Odeh, M. Al-Ruzzieh, A. Al Rifai, H. Mustafa, and M. Odeh, "Towards digital readiness of evidence-based practice in a regional cancer center using role-based business process and data modeling," *Digital Health*, vol. 10, p. 20552076241281193, 2024. https://doi.org/10.1177/20552076241281193