Beyond Accuracy: Cross-Validated and Threshold-Optimized Deep Learning for Primary and Metastatic Melanoma Classification from Histopathological Patches

Raden Rara Kartika Kusuma Winahyu[1], Lathifah Alfat[2*], Deyana Kusuma Wardani[3]

[1,3]Informatics Department, Astra Polytechnic, Bekasi, West Java, Indonesia
[2]Informatics Department, Faculty of Technology and Design, Universitas Pembangunan Jaya, South Tangerang, Banten, Indonesia
E-mail: [1]raden.rara@polytechnic.astra.ac.id, [2*]lathifah.alfat@upj.ac.id, [3]deyana.kusuma@ polytechnic.astra.ac.id

(Received: 26 Jan 2026, revised: 2 Mar 2026, accepted: 3 Mar 2026)

Abstract

Accurate differentiation between primary and metastatic melanoma in histopathological assessment is critical for staging and therapeutic decision-making. Although deep learning models often report high classification accuracy, their robustness and threshold-dependent clinical behavior remain insufficiently examined. We propose a cross-validated and threshold-optimized deep learning framework for classifying 206 histopathological regions of interest (ROIs), partitioned in an 80:20 split into training (n = 164) and evaluation (n = 42) subsets, using a ResNet-18 backbone. On the hold-out evaluation set, the model achieved an AUC of 0.922. To evaluate generalization stability, stratified 5-fold cross-validation was conducted across all ROIs, yielding fold AUCs ranging from 0.904 to 0.973 and a mean AUC of 0.938 ± 0.024, with a pooled out-of-fold AUC of 0.916. At a decision threshold of 0.5, the model achieved 78.6% accuracy (macro F1 = 0.7846). Increasing the threshold to 0.8 improved accuracy to 85.7% (macro F1 = 0.856), accompanied by higher precision for metastatic melanoma (0.894) and improved recall for primary melanoma (0.904), underscoring clinically meaningful sensitivity–specificity trade-offs beyond AUC alone. Grad-CAM analysis demonstrated spatially coherent activation concentrated within tumor-dense regions in true positives, minimal activation in true negatives, and intermediate activation in a borderline false negative case (probability = 0.75), linking prediction confidence to morphologically relevant evidence. Collectively, these findings highlight the importance of cross-validation rigor, threshold calibration, and interpretability in advancing clinically reliable deep learning systems for melanoma classification.

Keywords: Melanoma Metastasis Classification, Histopathological Image Analysis, Deep Learning in Digital Pathology, Threshold Optimization, Cross-Validation Robustness

I. INTRODUCTION

The rapid advancement of artificial intelligence (AI) and deep learning has significantly transformed medical image analysis, particularly in digital pathology and histopathological interpretation [1], [2]. Convolutional Neural Networks (CNNs) have demonstrated strong capability in melanoma detection by learning discriminative morphological patterns directly from digitized tissue images [3], [4]. These systems have shown potential to assist pathologists by improving diagnostic efficiency and reducing workload in routine clinical practice [5].

Despite these advancements, many deep learning studies continue to report overall accuracy or single-threshold metrics as the primary indicator of model performance [6]. While intuitive, accuracy alone may be insufficient for medical decision-making, particularly when the clinical consequences of misclassification are asymmetric [7]. In melanoma diagnosis, false-negative predictions may delay critical treatment, whereas false positives may result in unnecessary procedures and patient burden [8]. Therefore, evaluating diagnostic AI systems solely based on accuracy may obscure clinically meaningful trade-offs between sensitivity and specificity.

Recent research has emphasized the importance of decision threshold selection and calibration in medical AI applications [9], [10], [11]. In addition, evaluation metrics such as precision and recall provide more clinically meaningful insights than accuracy alone [12]. The choice of operating threshold directly influences the balance between sensitivity and specificity, and fixing the threshold arbitrarily at 0.5 may not reflect optimal clinical deployment scenarios [13], [14]. However, threshold optimization and operating-point analysis remain

underreported in histopathological melanoma classification studies.

Another limitation in existing literature is the lack of systematic error analysis. High aggregate performance metrics often obscure the clinical relevance of false positives and false negatives [15]. In addition, prior studies have shown that deep learning models may be affected by spurious correlations and dataset-specific artifacts [16], [17] which can further complicate the interpretation of misclassifications.

Interpretability techniques, such as Grad-CAM, have been widely used to enhance transparency in medical imaging models by visualizing attention patterns [18], [19], [20], [21], [22]. However, their systematic integration with threshold-optimized evaluation strategies in melanoma histopathology remains limited.

Deep neural networks are known to exhibit miscalibration and overconfidence in complex prediction tasks [23] which may reduce the reliability of probability estimates in medical applications. Strategies such as hard negative mining have been proposed to improve model robustness by emphasizing difficult samples during training [24], while post hoc calibration and threshold refinement approaches aim to optimize clinical operating points after model development [25]. Nevertheless, the combined application of threshold-optimized evaluation, systematic hard-negative analysis, and interpretability assessment remains relatively underexplored in histopathological melanoma classification.

Based on these considerations, this study argues that accuracy alone is insufficient for evaluating deep learning models in melanoma histopathology. We therefore propose a threshold-optimized and cross-validated evaluation framework that integrates operating-point analysis, pooled confusion matrix assessment, systematic hard-negative characterization, and attention-based interpretability. The objective is to enhance transparency and clinical alignment in confirmatory diagnostic scenarios

The main contributions of this study are as follows:

1. A threshold-optimized and cross-validated evaluation framework for primary and metastatic melanoma histopathology classification.
2. Explicit comparison of sensitivity-oriented and precision-oriented operating points using pooled out-of-fold predictions.
3. Systematic hard-negative and confusion-matrix-based error analysis.
4. Integration of Grad-CAM visualization to analyze attention patterns in correctly and incorrectly classified regions.

II. MATERIALS AND METHODS

The overall methodological framework adopted in this study is illustrated in Figure 1. In contrast to conventional pipelines that apply a fixed decision threshold (commonly 0.5) to network outputs, the proposed framework explicitly separates probabilistic prediction, threshold-based decision-making, and clinical interpretation. This design reflects a confirmatory diagnostic perspective in which operating thresholds and model confidence are treated as adjustable parameters rather than implicit defaults.

The pipeline consists of the following sequential stages: histopathological image acquisition, preprocessing and data augmentation, deep learning model training using stratified 5-fold cross-validation, probabilistic inference, pooled out-of-fold prediction aggregation, threshold optimization, performance evaluation across multiple operating points, and structured error analysis prior to clinical interpretation. This structured separation enables systematic investigation of sensitivity–specificity trade-offs and misclassification patterns beyond single-metric reporting.

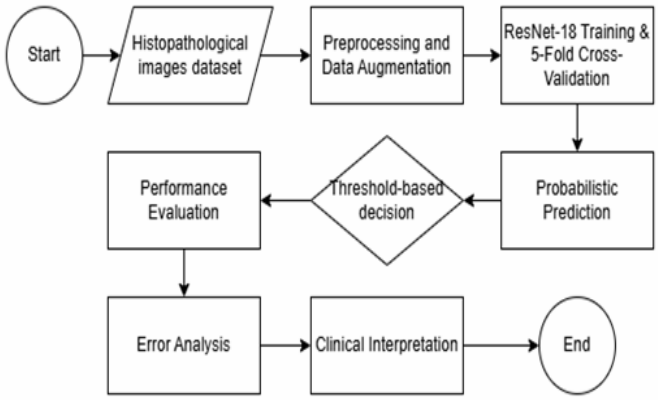

Figure 1. Research Pipeline for Cross-Validated and Threshold-Optimized Deep Learning Melanoma Classification Using Histopathological Patches

A. Dataset

This study utilized a publicly available histopathological image dataset introduced by Schuiveling [26] entitled *Melanoma Histopathology Dataset with Tissue and Nuclei Annotations*. The dataset is archived with a persistent DOI (10.5281/zenodo.15050523). It consists of pre-extracted regions of interest (ROIs) derived from whole-slide images (WSIs), provided in TIFF format. Each ROI is annotated as either primary melanoma (class 0) or metastatic melanoma (class 1).

After verifying file integrity and label consistency, a total of 206 ROI patches were included in this study, comprising 103 primary and 103 metastatic samples, resulting in a class-balanced dataset. Each ROI was treated as an independent analytical unit and uniquely identified using its ROI identifier.

B. Data Partitioning Strategy and Leakage Considerations

Data leakage is a recognized concern in digital pathology, particularly when multiple image tiles originating from the same whole-slide image appear across training and validation subsets, potentially leading to inflated performance estimates [27]. Because the released dataset provides ROI-level image patches without explicit patient-level or slide-level identifiers, strict patient-level separation could not be enforced.

To mitigate optimistic bias under this constraint, model evaluation was conducted using stratified 5-fold cross-validation across the entire dataset. In each fold, ROI

identifiers were mutually exclusive between training and validation subsets, and data augmentation was applied only within the training partition. Final performance metrics were computed using pooled out-of-fold (OOF) probabilistic predictions to provide a more robust and less split-dependent estimate of model discrimination.

C. Ethical Considerations

The dataset is publicly available and contains de-identified histopathological image patches without personal identifiers or associated clinical metadata. No patient-identifiable information was accessed or processed in this study. All analyses were conducted in accordance with the data-sharing terms specified by the dataset provider.

Given the relatively limited dataset size, performance reporting emphasizes threshold-dependent operating characteristics (e.g., sensitivity, specificity, precision) and confidence interval estimation rather than relying solely on point accuracy metrics. Representative examples of primary and metastatic melanoma ROIs after preprocessing and resizing to 224×224 pixels are shown in Figure 2.

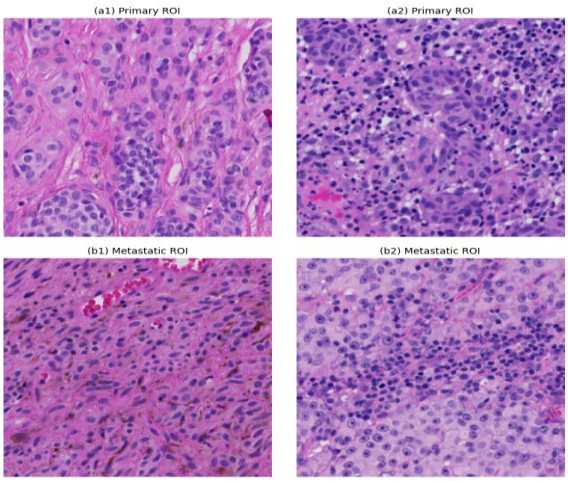

Figure 2. Representative Examples of Histopathological ROI Images Used in This Study. (a1–a2) Primary Melanoma Samples and (b1–b2) Metastatic Melanoma Samples.

D. Preprocessing and Data Augmentation

Prior to model training, all ROI images were standardized to ensure consistent input dimensions and intensity distributions. Each TIFF image was converted to a three-channel RGB format and resized to 224 × 224 pixels to match the input specification of the pretrained ResNet-18 architecture employed in this study. Pixel intensities were normalized using ImageNet mean and standard deviation values to maintain compatibility with the pretrained backbone.

To improve model robustness against staining variability and morphological heterogeneity commonly observed in histopathological images, several data augmentation strategies were applied during training. These included random horizontal flipping, 90° rotations, and additive Gaussian noise. Data augmentation has been shown to enhance generalization performance in computational pathology by reducing overfitting and mitigating domain-specific biases [28]

Importantly, augmentation operations were applied exclusively to the training subset within each cross-validation fold. Validation samples were left unaltered to ensure fair evaluation and to avoid artificially inflated performance estimates, a well-documented issue in medical imaging machine learning studies when evaluation protocols are not carefully controlled [29].

E. Model Architecture

The classification backbone employed in this study was ResNet-18, a residual convolutional neural network architecture known for its balance between representational capacity and computational efficiency. Its relatively moderate depth makes it suitable for medical image classification tasks with limited sample sizes, where excessively deep architectures may increase overfitting risk [30]. Residual connections facilitate gradient propagation during training and have demonstrated strong performance in dermatological image analysis tasks [31].

The final fully connected layer of the pretrained ResNet-18 model was replaced to produce two output logits corresponding to primary and metastatic melanoma. Class probabilities were obtained via softmax activation.

Model training was conducted using cross-entropy loss and optimized using the Adam optimizer with mini-batch updates. The same architecture and training configuration were applied consistently across all cross-validation folds to ensure fair performance comparison.

F. Experimental Configurations

Two experimental configurations were evaluated in this study to assess both baseline discriminative performance and the potential impact of targeted retraining strategies. The primary configuration, referred to as the baseline model, consisted of a standard supervised training pipeline using an ImageNet-pretrained ResNet-18 backbone optimized with binary cross-entropy loss, a strategy widely adopted in medical image classification tasks to improve generalization on limited datasets [32], [33]. This configuration served as the main reference model for all threshold-dependent and cross-validation analyses.

In addition, a secondary configuration incorporating a simplified hard negative mining (HNM) strategy was explored. Hard example mining approaches have been investigated as a means to emphasize visually ambiguous or difficult samples during training in medical imaging applications [34]. In this variant, misclassified and near-threshold samples identified within each training fold were assigned increased sampling weight during retraining to refine decision boundary discrimination [35]. Importantly, hard samples were mined strictly within the training partitions of each fold to prevent information leakage and maintain cross-validation integrity, reflecting best practices for within-fold preprocessing and evaluation to avoid optimistic bias [36]. The baseline model represents the primary analytical framework of this study,

while the HNM configuration is reported as an exploratory robustness experiment.

G. Training Protocol

All experiments were implemented in PyTorch using a ResNet-18 backbone initialized with ImageNet pre-trained weights. The final fully connected layer was replaced with a single-neuron output to produce a binary logit for metastatic probability. The model was trained using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. Binary cross-entropy with logits (BCEWithLogitsLoss) was used as the optimization objective. Training was conducted with a batch size of 32.

Model weights were updated for a fixed number of epochs without dynamic learning rate scheduling. To ensure reproducibility, all random seeds were fixed across NumPy, Python, and PyTorch, and deterministic behavior was enforced during data loading and model training.

H. Threshold Optimization and Decision Criteria

Although a default decision threshold of 0.5 is commonly adopted in binary classification tasks, such a choice may not necessarily correspond to the empirically optimal operating point [37]. To explicitly examine threshold sensitivity, two predefined thresholds (0.5 and 0.8) were evaluated to assess their impact on precision, recall, specificity, and overall predictive behavior. To move beyond accuracy-based evaluation, threshold optimization was explicitly incorporated to analyze the clinical implications of probability-based decision-making. In addition to Youden's threshold, two clinically interpretable decision thresholds (0.5 and 0.8) were evaluated to examine sensitivity–specificity trade-offs. The optimal operating threshold was estimated using Youden's J statistic, defined as sensitivity + specificity − 1.

I. Error Analysis and Hard Negative Mining

To better understand model behavior and limitations, a structured error analysis was conducted using confusion matrices derived from pooled out-of-fold predictions. Misclassified samples were examined in relation to their predicted probabilities and proximity to the decision threshold. Particular attention was given to high-confidence errors, defined as misclassifications associated with predicted probabilities far from the threshold, as such errors may carry disproportionate clinical risk and reflect systematic model bias rather than boundary uncertainty [38].

J. Evaluation Metrics

Model performance was evaluated using standard classification metrics, including accuracy, precision, recall (sensitivity), specificity, and F1-score. Confusion matrices were constructed to quantify true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), enabling structured analysis of diagnostic error patterns.

Given the asymmetric clinical implications of false-positive and false-negative predictions in melanoma diagnosis, performance interpretation extended beyond overall accuracy to include class-specific metrics [39]. This approach aligns with current recommendations for evaluating predictive AI models in clinical decision-support settings.

To assess threshold sensitivity, evaluation was performed at two predefined decision thresholds (0.5 and 0.8). This enabled explicit examination of how operating-point selection influences precision–recall trade-offs. In addition, Receiver Operating Characteristic (ROC) analysis and area under the curve (AUC) were computed to provide threshold-independent assessment of model discrimination [40].

K. Model Interpretability via Grad-CAM

To enhance interpretability and examine spatial attention patterns, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to selected validation samples. Grad-CAM produces class-specific localization maps by propagating gradients to the final convolutional layer, enabling visualization of image regions that contribute most strongly to model predictions [22]. Such visualization techniques are increasingly recommended in medical imaging to improve transparency and support clinical trust in AI systems [18], [20].

Representative samples were selected from both correctly classified true positives (TP) and misclassified false negatives (FN). Specifically, three TP and three FN cases were analyzed to compare attention distributions between confident correct predictions and clinically critical errors. Visual inspection aimed to determine whether the model focused on morphologically relevant tumor regions or on potentially spurious tissue patterns, which may indicate shortcut learning behavior [16].

Grad-CAM analysis complements threshold-sensitive evaluation by providing qualitative insight into model failure modes beyond numerical metrics. Integrating interpretability within the evaluation pipeline supports more transparent diagnostic risk assessment in computational pathology [18], [22].

L. Cross-Validation Framework

To ensure robust generalization assessment and minimize partition-specific bias, stratified 5-fold cross-validation was performed across all 206 histopathological ROIs, consistent with contemporary validation recommendations in medical AI research [29]. Fold assignments were fixed prior to training to prevent information leakage. Model discrimination was evaluated using ROC analysis and AUC for each fold, and out-of-fold predictions were aggregated to compute a pooled ROC curve and pooled AUC, providing an unbiased estimate across the dataset. All threshold-dependent analyses were conducted exclusively on validation predictions in accordance with best-practice reporting standards for clinical AI systems [41].

III. RESULTS AND DISCUSSION

A. Overall Classification Performance

The proposed ResNet-18–based framework demonstrated strong discriminative performance in distinguishing primary from metastatic melanoma on histopathological ROIs. On the 80:20 hold-out split (n = 42 evaluation ROIs), the model

DOI: 10.34148/teknika.v15i1.1454

achieved a ROC-AUC of 0.922, indicating robust ranking capability independent of decision threshold.

To assess generalization stability, stratified 5-fold cross-validation was performed across all 206 ROIs. Fold-wise AUC values ranged from 0.904 to 0.973, yielding a mean AUC of 0.938 ± 0.024. The pooled out-of-fold (OOF) AUC was 0.916, further supporting consistent performance across data partitions. The relatively low standard deviation suggests that the learned feature representations generalize reliably across different training-validation splits rather than being driven by a particular subset configuration.

Importantly, the consistency between hold-out AUC (0.922) and cross-validation mean AUC (0.938) indicates minimal performance inflation and limited overfitting. This alignment reinforces the robustness of the model and highlights the importance of cross-validation in small-scale histopathological datasets.

Beyond threshold-independent evaluation, threshold-sensitive analysis revealed clinically meaningful shifts in operating characteristics. Increasing the decision threshold from 0.5 to 0.8 improved overall accuracy from 78.6% to 85.7% and macro F1-score from 0.784 to 0.856. This improvement was accompanied by enhanced precision for metastatic melanoma and increased recall for primary melanoma, reflecting a stricter decision boundary and reduced false positive rate.

These findings demonstrate that while AUC provides a strong summary of ranking performance, clinically relevant operating behavior is substantially influenced by threshold selection. Accordingly, performance interpretation should extend beyond accuracy metrics to include calibration-driven decision analysis.

B. Threshold-Dependent Performance

Although the baseline model demonstrated strong threshold-independent discrimination (AUC = 0.922), clinical deployment necessitates the selection of a specific operating threshold. To avoid arbitrary decision boundaries, the optimal operating point was first determined using Youden's J statistic (sensitivity + specificity − 1), which identified a threshold of 0.878 as the point maximizing balanced classification performance.

In addition to this data-driven estimate, two clinically interpretable fixed thresholds (0.5 and 0.8) were evaluated alongside the Youden-optimized threshold (0.878) to systematically examine sensitivity–specificity trade-offs. A threshold of 0.5 represents the conventional default decision boundary for probabilistic classifiers, whereas a higher threshold of 0.8 imposes stricter confidence requirements for metastatic classification.

At a threshold of 0.5, the model achieved an overall accuracy of 78.6% with a macro F1-score of 0.784. This operating point favored higher sensitivity for metastatic detection but resulted in a greater number of false positive predictions. Increasing the threshold to 0.8 improved overall accuracy to 85.7% and increased the macro F1-score to 0.856. This adjustment reduced false positives while improving metastatic precision and primary recall, reflecting a more conservative classification strategy.

The Youden-optimized threshold (0.878) yielded the highest overall accuracy (88.1%) and macro F1-score (0.880). At this operating point, primary melanoma recall increased to 0.952, while metastatic precision reached 0.944, indicating an improved balance between sensitivity and specificity. Compared with the fixed thresholds of 0.5 and 0.8, the Youden threshold provided the most favorable overall operating characteristics, supporting its use as a principled, data-driven decision criterion rather than an arbitrary cutoff. Table 1 below summarizes the key quantitative results.

Table 1. Threshold-Dependent Classification Performance on Hold-Out Set (n = 42)

Threshold	Accuracy (%)	Macro F1-score
0.5	78,60	0.784
0.8	85,68	0.856
0.878	88,1	0.880

C. Confusion Matrix Analysis

To provide an intuitive understanding of threshold-dependent behavior, confusion matrices were examined for the baseline model at decision thresholds of 0.5, 0.8, and 0.878 (Youden's optimal threshold), as shown in Figs. 3–5.

At the conventional threshold of 0.5, the baseline model, as shown in Figure 3, achieved an overall accuracy of 78.6% (33/42 correct predictions). The confusion matrix indicates 15 true negatives and 18 true positives, with 6 false positives and 3 false negatives. This operating point favored higher sensitivity for metastatic melanoma (recall = 0.857) but allowed a relatively higher number of false positive predictions for metastatic classification. The macro F1-score at this threshold was 0.784, reflecting moderate balance between the two classes.

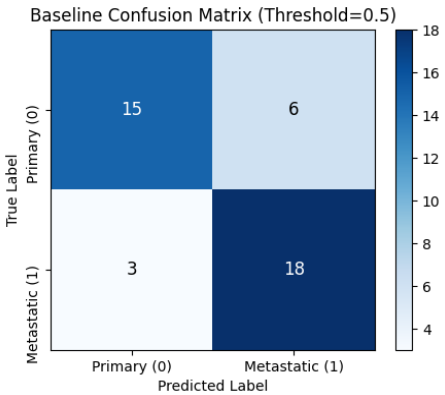

Figure 3. Confusion Matrices at Thresholds of 0.5 (Baseline Model)

Under a stricter threshold of 0.8, the classification behavior shifted noticeably (Figure 4). Overall accuracy increased to 85.7% (36/42 correct predictions), with 19 true negatives and 17 true positives. False positives were reduced from 6 to 2, while false negatives slightly increased from 3 to

4. This adjustment improved metastatic precision (0.894) and primary recall (0.904), yielding a higher macro F1-score of 0.856. The stricter confidence requirement therefore reduced overcalling of metastasis while maintaining balanced discrimination.

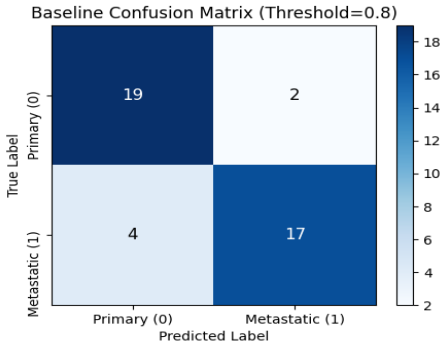

Figure 4. Confusion Matrices at Thresholds of 0.8 (Baseline Model)

When the Youden-optimized threshold (0.878) was applied, performance improved further. The resulting confusion matrix as shown in Figure 5 (20 true negatives, 17 true positives, 1 false positive, and 4 false negatives) produced the highest overall accuracy of 88.1% and a macro F1-score of 0.880. Notably, primary melanoma recall increased to 0.952, while metastatic precision reached 0.944. Compared to the fixed thresholds of 0.5 and 0.8, the Youden threshold provided the most favorable balance between minimizing false positives and preserving high sensitivity.

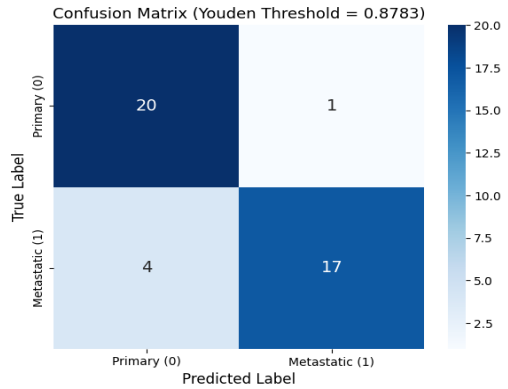

Figure 5. Confusion Matrices at Thresholds of 0.878 (Baseline Model)

Collectively, these results demonstrate that threshold selection materially alters clinical operating characteristics, despite stable AUC. The data-driven Youden threshold achieved the strongest overall performance, supporting its use as a principled alternative to arbitrary decision boundaries.

D. Class-wise Performance

To examine how training strategy influences class-level diagnostic behavior, precision and recall were analyzed separately for primary and metastatic melanoma across operating thresholds.

At the conventional decision threshold of 0.5, the baseline model achieved an overall accuracy of 78.6% with a macro F1-score of 0.784 (Table 2), demonstrating asymmetric class-wise behavior. For primary melanoma (Class 0), precision was 0.833 and recall was 0.714 (F1 = 0.7692), indicating moderate misclassification of primary cases as metastatic. Conversely, metastatic melanoma (Class 1) exhibited higher recall (0.857) but lower precision (0.750) (F1 = 0.800), reflecting stronger sensitivity for metastatic detection at the expense of increased false positive predictions. Overall, the default threshold favored metastatic sensitivity over specificity, underscoring the importance of threshold calibration in balancing diagnostic performance. The macro average is computed as the unweighted mean of class-wise metrics.

Table 2. Class-wise Performance on Validation Set (n = 42) at Decision Threshold = 0.5 (Baseline Model)

Class Label	Precision	Recall	F1-score
Primary (Class 0)	0.833	0.714	0.769
Metastatic (Class 1)	0.750	0.857	0.80
Macro Average	-	-	0.784
Overall Accuracy	-	-	0.785

At a stricter decision threshold of 0.8, class-wise evaluation demonstrated improved balance between sensitivity and precision. As summarized in Table 3, the model achieved an overall accuracy of 85.7% with a macro F1-score of 0.856. For primary melanoma (Class 0), precision was 0.8261 and recall increased to 0.904, resulting in an F1-score of 0.863, indicating stronger detection of primary cases with fewer false negatives. For metastatic melanoma (Class 1), precision improved to 0.894 while recall was 0.809, yielding an F1-score of 0.850. Compared to the conventional threshold of 0.5, this operating point reduced false positive predictions and enhanced overall discriminative consistency, reflecting the impact of stricter confidence requirements on class-wise diagnostic behavior. The macro average is computed as the unweighted mean of class-wise metrics.

Table 3. Class-wise Performance on Validation Set (n = 42) at Decision Threshold = 0.8 (Baseline Model)

Class Label	Precision	Recall	F1-score
Primary (Class 0)	0.826	0.904	0.863
Metastatic (Class 1)	0.894	0.809	0.850
Macro Average	-	-	0.856
Overall Accuracy	-	-	0.857

DOI: 10.34148/teknika.v15i1.1454

At the Youden-optimized threshold of 0.8783, class-wise evaluation demonstrated the most balanced discriminative behavior across both melanoma subtypes. As summarized in Table 4, the model achieved an overall accuracy of 88.1% with a macro F1-score of 0.880, representing the highest performance among the evaluated operating points. For primary melanoma (Class 0), recall increased to 0.952, indicating strong sensitivity in correctly identifying primary cases, while precision was 0.833, yielding an F1-score of 0.888. For metastatic melanoma (Class 1), precision reached 0.944, reflecting a substantial reduction in false positive predictions, with recall of 0.809 and an F1-score of 0.871. This operating point provides the most favorable balance between sensitivity and specificity, supporting the use of a data-driven threshold selection strategy rather than reliance on arbitrary cutoff values.

Table 4. Class-wise Performance on Validation Set (n = 42) at Youden Optimal Threshold = 0.8783 (Baseline Model)

Class Label	Precision	Recall	F1-score
Primary (Class 0)	0.833	0.952	0.888
Metastatic (Class 1)	0.944	0.809	0.871
Macro Average	-	-	0.880
Overall Accuracy	-	-	0.881

E. Decision Boundary and Visual Interpretability Analysis

To further elucidate the relationship between probabilistic confidence and spatial feature attribution, decision boundary behavior was examined in conjunction with Grad-CAM visualization (Figure 6). True positive cases exhibited spatially coherent activation concentrated within tumor-dense cellular regions, consistent with high-confidence predictions (probability ≈ 1.00). Conversely, true negative samples demonstrated minimal and diffuse activation, aligning with low predicted probabilities (≈ 0.00). These patterns indicate that model confidence corresponds to localized morphological evidence rather than diffuse background artifacts, supporting the structural validity of the learned feature representations.

Misclassified samples were predominantly located near the probabilistic decision boundary. In particular, the borderline false negative case (probability = 0.75) showed intermediate activation distributed across partially tumor-dense regions, suggesting morphological ambiguity rather than arbitrary misclassification. Similarly, the high-confidence false positive (probability = 0.99) displayed concentrated activation within visually atypical regions, highlighting potential histopathological overlap between primary and metastatic patterns. Collectively, these findings demonstrate that model errors arise primarily from boundary-adjacent uncertainty and intrinsic morphological heterogeneity, reinforcing that reliable melanoma classification requires not only high AUC but also calibrated decision thresholds and spatially interpretable evidence-thereby substantiating the central premise of moving beyond accuracy toward clinically meaningful and explainable AI deployment.

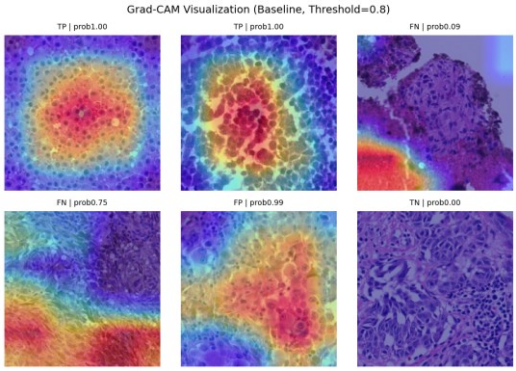

Figure 6. Grad-CAM Visualization of Baseline Model Predictions at Decision Threshold = 0.8.

F. ROC and AUC Analysis

Receiver Operating Characteristic (ROC) analysis was conducted to provide a threshold-independent evaluation of model discrimination performance. On the independent hold-out validation set (n = 42), the baseline model achieved an AUC of 0.922, indicating strong separability between primary and metastatic melanoma across all possible decision thresholds (Figure 7). The ROC curve demonstrated a consistent deviation from the diagonal reference line, reflecting high true positive rates achieved at relatively low false positive rates.

To further assess generalization stability, stratified 5-fold cross-validation was performed across all 206 ROIs. Fold-wise AUC values ranged from 0.9048 to 0.973, yielding a mean AUC of 0.938 ± 0.024. The pooled out-of-fold AUC was 0.916, closely aligned with the independent hold-out performance. The consistency between cross-validation and hold-out AUC values supports the robustness of the learned representation and suggests that performance gains are not attributable to partition-specific effects. Importantly, the Youden-optimized threshold was derived directly from the ROC curve, reinforcing the integration between global discrimination analysis and calibrated decision boundary selection.

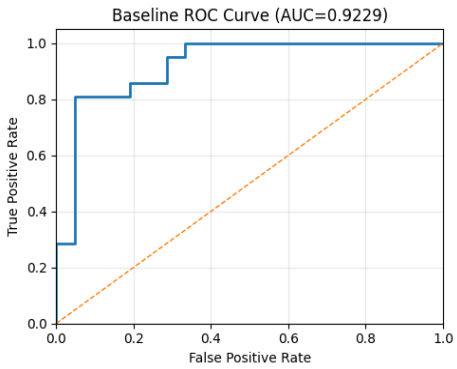

Figure 7. ROC Curve of the Baseline Model on the Validation Set (n = 42).

DOI: 10.34148/teknika.v15i1.1454

IV. CONCLUSIONS

This study demonstrates that reliable melanoma classification requires methodological rigor beyond reporting accuracy alone. Using a ResNet-18 backbone trained on 206 histopathological ROIs (80:20 split), the baseline model achieved strong discriminative performance on the hold-out evaluation set (AUC = 0.922). Stratified 5-fold cross-validation further yielded a mean AUC of 0.938 ± 0.024 with a pooled out-of-fold AUC of 0.916, confirming stable generalization and reducing concerns regarding partition-specific bias.

Threshold-dependent analysis revealed clinically meaningful shifts in model behavior. While the conventional threshold of 0.5 achieved 78.6% accuracy, increasing the threshold to 0.8 improved accuracy to 85.7%, and the Youden-optimized threshold (0.8783) achieved 88.1% accuracy with the highest macro-F1 performance. These findings emphasize that deployment-ready AI systems must incorporate data-driven threshold calibration to balance sensitivity and specificity in accordance with diagnostic priorities, rather than relying on default probability cutoffs.

Class-wise analysis further demonstrated that sensitivity–specificity trade-offs differ between primary and metastatic melanoma, underscoring the clinical consequences of threshold selection. The Youden-optimized operating point provided the most balanced discriminative behavior, reducing false positives while preserving metastatic detection capability. Such operating-point transparency is critical for translating AI models into decision-support tools in histopathological workflows.

In addition to the primary baseline configuration, a simplified hard negative mining (HNM) strategy was explored as a robustness-oriented extension. Although HNM provided incremental refinement of decision boundary discrimination, the baseline model already exhibited strong and stable performance across validation schemes. Thus, the principal contribution of this study lies in the rigorously validated and threshold-optimized baseline framework, with HNM serving as an exploratory enhancement rather than a dependency for performance.

Grad-CAM analysis demonstrated spatially coherent activation within tumor-dense regions and biologically plausible patterns across correctly and incorrectly classified cases, linking prediction confidence to morphological ambiguity. Collectively, this work advances a clinically aware evaluation paradigm for melanoma histopathology, integrating cross-validation stability, threshold optimization, and visual interpretability. By moving beyond accuracy-centric reporting, the proposed framework provides a reproducible foundation for future multi-institutional validation and real-world clinical integration of deep learning–based melanoma classification systems.

V. LIMITATIONS AND FUTURE WORK

Despite the promising results observed in this study, several limitations must be considered before clinical deployment. The dataset comprised 206 histopathological ROIs from a single-source cohort, which may not fully capture inter-institutional variability in staining protocols, scanner calibration, tissue preparation, and patient demographics. Although stratified cross-validation was implemented to strengthen internal robustness, external validation on independent, multi-center datasets is necessary to establish real-world generalizability and regulatory readiness.

Second, the present framework operates at the ROI level rather than the whole-slide image (WSI) or patient level. In clinical practice, diagnostic decisions are made based on comprehensive slide assessment rather than isolated patches. Future work should therefore incorporate slide-level aggregation strategies, such as multi-instance learning or attention-based pooling, to better align model outputs with pathologist workflows and case-level decision-making.

Third, while threshold optimization was rigorously examined using Youden's statistic and clinically interpretable cutoffs, optimal operating points in practice may vary depending on institutional priorities. For example, tertiary referral centers may prioritize higher metastatic sensitivity, whereas screening environments may emphasize specificity to reduce unnecessary downstream testing. Prospective validation studies involving practicing pathologists and real-time diagnostic scenarios are needed to determine clinically acceptable error trade-offs.

Fourth, although Grad-CAM provided biologically plausible visual explanations, interpretability remains qualitative and retrospective. For clinical integration, explanation reliability must be quantitatively validated against expert annotations or region-of-interest delineations. Establishing measurable concordance between model attention and pathologist-identified tumor structures would strengthen trust and facilitate regulatory approval.

Ultimately, translating deep learning–based melanoma classification into routine pathology practice will require multi-institutional validation, workflow integration studies, and prospective clinical trials assessing impact on diagnostic efficiency and staging accuracy. By addressing these translational steps, future research can move beyond algorithmic performance toward demonstrable clinical utility and improved patient outcomes.

REFERENCES

[1] D. Komura, M. Ochi, and S. Ishikawa, "Machine learning methods for histopathological image analysis: Updates in 2024," Jan. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.csbj.2024.12.033.

[2] X. M. Zhang *et al.*, "Artificial intelligence in digital pathology diagnosis and analysis: technologies, challenges, and future prospects," Dec. 01, 2025, *BioMed Central Ltd*. doi: 10.1186/s40779-025-00680-6.

[3] M. Kreouzi *et al.*, "Deep Learning for Melanoma Detection: A Deep Learning Approach to Differentiating Malignant Melanoma from Benign

Melanocytic Nevi," *Cancers (Basel).*, vol. 17, no. 1, Jan. 2025, doi: 10.3390/cancers17010028.

[4] M. M. Musthafa, M. T R, V. K. V, and S. Guluwadi, "Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification," *BMC Med. Imaging*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12880-024-01356-8.

[5] C. McGenity *et al.*, "Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy," Dec. 01, 2024, *Nature Research*. doi: 10.1038/s41746-024-01106-8.

[6] Z. U. Nisa *et al.*, "Beyond Accuracy: Evaluating certainty of AI models for brain tumour detection," *Comput. Biol. Med.*, vol. 193, p. 110375, Jul. 2025, doi: 10.1016/J.COMPBIOMED.2025.110375.

[7] Y. Wang *et al.*, "Economic evaluation for medical artificial intelligence: accuracy vs. cost-effectiveness in a diabetic retinopathy screening case," *NPJ Digit. Med.*, vol. 7, no. 1, Dec. 2024, doi: 10.1038/s41746-024-01032-9.

[8] J. Y. Chen *et al.*, "Skin Cancer Diagnosis by Lesion, Physician, and Examination Type: A Systematic Review and Meta-Analysis," *JAMA Dermatol.*, vol. 161, no. 2, pp. 135–146, Feb. 2025, doi: 10.1001/jamadermatol.2024.4382.

[9] C. Penso, L. Frenkel, and J. Goldberger, "Confidence Calibration of a Medical Imaging Classification System That is Robust to Label Noise," *IEEE Trans. Med. Imaging*, vol. 43, no. 6, pp. 2050–2060, Jun. 2024, doi: 10.1109/TMI.2024.3353762.

[10] J. Rudolph *et al.*, "Threshold optimization in AI chest radiography analysis: integrating real-world data and clinical subgroups," *Eur. Radiol. Exp.*, vol. 9, no. 1, Dec. 2025, doi: 10.1186/s41747-025-00632-8.

[11] A. Scarffe, A. Coates, K. Brand, and W. Michalowski, "Decision threshold models in medical decision making: a scoping literature review," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12911-024-02681-2.

[12] V. Turri, R. Dzombak, E. Heim, N. Vanhoudnos, J. Palat, and A. Sinha, "Measuring AI Systems Beyond Accuracy," 2022.

[13] S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," *PLoS One*, vol. 17, no. 1 January, Jan. 2022, doi: 10.1371/journal.pone.0262838.

[14] A. S. Sambyal, U. Niyaz, N. C. Krishnan, and D. R. Bathula, "Understanding Calibration of Deep Neural Networks for Medical Image Classification," Dec. 2023, doi: 10.1016/j.cmpb.2023.107816.

[15] H. Evans and D. Snead, "Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact," *NPJ Digit. Med.*, vol. 7, no. 1, Dec. 2024, doi: 10.1038/s41746-024-01093-w.

[16] U. Mahmood *et al.*, "Detecting Spurious Correlations With Sanity Tests for Artificial Intelligence Guided Radiology Systems," *Front. Digit. Health*, vol. 3, Aug. 2021, doi: 10.3389/fdgth.2021.671015.

[17] C. Vásquez-Venegas *et al.*, "Detecting and Mitigating the Clever Hans Effect in Medical Imaging: A Scoping Review," Aug. 01, 2025, *Springer Nature*. doi: 10.1007/s10278-024-01335-z.

[18] K. Borys *et al.*, "Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches," May 01, 2023, *Elsevier Ireland Ltd*. doi: 10.1016/j.ejrad.2023.110786.

[19] Z. Teng *et al.*, "A literature review of artificial intelligence (AI) for medical image segmentation: from AI and explainable AI to trustworthy AI," Dec. 05, 2024, *AME Publishing Company*. doi: 10.21037/qims-24-723.

[20] S. Kinger and V. Kulkarni, "A review of explainable AI in medical imaging: implications and applications," *International Journal of Computers and Applications*, vol. 46, no. 11, pp. 983–997, 2024, doi: 10.1080/1206212X.2024.2404082.

[21] T. Lai, "Interpretable Medical Imagery Diagnosis with Self-Attentive Transformers: A Review of Explainable AI for Health Care," Mar. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/biomedinformatics4010008.

[22] M. Ennab and H. Mcheick, "Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models," *Mach. Learn. Knowl. Extr.*, vol. 7, no. 1, Mar. 2025, doi: 10.3390/make7010012.

[23] M. Minderer *et al.*, "Revisiting the Calibration of Modern Neural Networks," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2106.07998

[24] W. Huang, X. Hu, S. Abousamra, P. Prasanna, and C. Chen, "Hard Negative Sample Mining for Whole Slide Image Classification," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.02212

[25] 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, Mar. 2025, doi: 10.1007/s42452-025-06587-0.

[26] M. Schuiveling, "Melanoma Histopathology Dataset with Tissue and Nuclei Annotations ," Mar. 2025, *Zenodo*. doi: 10.5281/zenodo.15050523.

[27] N. Bussola, A. Marcolini, V. Maggio, G. Jurman, and C. Furlanello, "AI slipping on tiles: data leakage in digital pathology," 2021, doi: 10.48550/arXiv.1909.06539.

[28] D. Tellez *et al.*, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," Apr. 2020, doi: 10.1016/j.media.2019.101544.

[29] M. Roberts *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest

radiographs and CT scans," *Nat. Mach. Intell.*, vol. 3, no. 3, pp. 199–217, Mar. 2021, doi: 10.1038/s42256-021-00307-0.

[30] A. Gorenshtein, T. Liba, and A. Goren, "Lightweight Transfer Learning Models for Multi-Class Brain Tumor Classification: Glioma, Meningioma, Pituitary Tumors, and No Tumor MRI Screening," *Journal of Imaging Informatics in Medicine*, 2025, doi: 10.1007/s10278-025-01686-1.

[31] R. Liu, Z. Chen, and P. Zhang, "Skin Lesion Classification Based on ResNet-50 Enhanced With Adaptive Spatial Feature Fusion," Oct. 2025, [Online]. Available: http://arxiv.org/abs/2510.03876

[32] M. Tsuneki, "Deep learning models in medical image analysis," *J. Oral Biosci.*, vol. 64, no. 3, pp. 312–320, Sep. 2022, doi: 10.1016/j.job.2022.03.003.

[33] A. W. Salehi *et al.*, "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," Apr. 01, 2023, *MDPI*. doi: 10.3390/su15075930.

[34] W. Huang, X. Hu, S. Abousamra, P. Prasanna, and C. Chen, "Hard Negative Sample Mining for Whole Slide Image Classification," 2024. [Online]. Available: https://github.com/winston52/HNM-WSI.

[35] M. Zayani, A. Toumi, and A. Khalfallah, "MHD-Protonet: Margin-Aware Hard Example Mining for SAR Few-Shot Learning via Dual-Loss Optimization," *Algorithms*, vol. 18, no. 8, Aug. 2025, doi: 10.3390/a18080519.

[36] D. Wilimitis and C. G. Walsh, "Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial," *JMIR AI*, vol. 2, no. 1, 2023, doi: 10.2196/49023.

[37] O. Rainio *et al.*, "Comparison of thresholds for a convolutional neural network classifying medical images," *Int. J. Data Sci. Anal.*, vol. 20, no. 3, pp. 2093–2099, Sep. 2025, doi: 10.1007/s41060-024-00584-z.

[38] J. M. Dolezal *et al.*, "Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology," *Nat. Commun.*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/s41467-022-34025-x.

[39] B. Van Calster *et al.*, "Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance.," *Lancet Digit. Health*, p. 100916, Dec. 2025, doi: 10.1016/j.landig.2025.100916.

[40] S. Rajaraman, P. Ganesan, and S. K. Antani, "Does deep learning model calibration improve performance in class-imbalanced medical image classification?," 2021. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-

[41] V. Sounderajah *et al.*, "Developing Specific Reporting Standards in Artificial Intelligence Centred Research," *Ann. Surg.*, vol. 275, no. 3, pp. 547–548, Mar. 2022, doi: 10.1097/SLA.0000000000005294.