MelanomaNet: Explainable Deep Learning for Skin Lesion Classification

Sukhrobbek Ilyosbekov
Northeastern University
ilyosbekov.s@northeastern.edu

Abstract

Automated skin lesion classification using deep learning has shown remarkable accuracy, yet clinical adoption remains limited due to the "black box" nature of these models. We present MelanomaNet, an explainable deep learning system for multi-class skin lesion classification that addresses this gap through four complementary interpretability mechanisms. Our approach combines an EfficientNet V2 backbone with GradCAM++ attention visualization, automated ABCDE clinical criterion extraction, Fast Concept Activation Vectors (FastCAV) for concept-based explanations, and Monte Carlo Dropout uncertainty quantification. We evaluate our system on the ISIC 2019 dataset containing 25,331 dermoscopic images across 9 diagnostic categories. Our model achieves 85.61% accuracy with weighted F1 score of 0.8564, while providing clinically meaningful explanations that align model attention with established dermatological assessment criteria. The uncertainty quantification module decomposes prediction confidence into epistemic and aleatoric components, enabling automatic flagging of unreliable predictions for clinical review. Our results demonstrate that high classification performance can be achieved alongside comprehensive interpretability, potentially facilitating greater trust and adoption in clinical dermatology workflows. Code is available at `https://github.com/suxrobgm/explainable-melanoma`.

1 Introduction

Skin cancer represents one of the most prevalent forms of cancer worldwide, with melanoma being the most lethal variant responsible for the majority of skin cancer deaths despite comprising only a small fraction of cases [1]. Early detection significantly improves patient outcomes, with 5-year survival rates exceeding 99% for localized melanoma but dropping below 30% for distant metastatic disease. Dermoscopy, a non-invasive imaging technique that magnifies skin lesions, has become the standard clinical tool for melanoma screening. However, accurate interpretation of dermoscopic images requires substantial expertise, and diagnostic accuracy varies considerably even among trained dermatologists.

Deep learning has emerged as a promising approach for automated dermoscopic image analysis, with convolutional neural networks achieving diagnostic accuracy comparable to or exceeding that of expert dermatologists in controlled studies [2]. Despite these impressive results, clinical adoption of AI-assisted dermatology tools remains limited. A primary barrier is the lack of interpretability inherent in deep neural networks—clinicians are understandably reluctant to trust diagnostic recommendations from systems that cannot explain their reasoning.

The ABCDE criteria (Asymmetry, Border irregularity, Color variation, Diameter, Evolution) represent the established clinical framework for melanoma assessment [3]. These criteria provide an intuitive and teachable approach that dermatologists use to communicate findings to patients and colleagues. An AI system that can relate its predictions to these familiar clinical concepts would be far more useful and trustworthy than one that simply outputs class probabilities.

In this work, we present MelanomaNet, an explainable deep learning system that addresses the interpretability gap through multiple complementary mechanisms:

- **Attention Visualization**: GradCAM++ generates heatmaps showing which image regions most influenced the prediction.
- **Clinical Criterion Extraction**: Automated analysis of ABCDE features with quantitative scores and visualizations.
- **Concept-Based Explanations**: Fast Concept Activation Vectors provide human-interpretable concept importance scores.
- **Uncertainty Quantification**: Monte Carlo Dropout decomposes uncertainty into epistemic and aleatoric components.

Our contributions include: (1) a multi-modal explainability framework that bridges deep learning predictions with clinical reasoning, (2) alignment metrics that validate

whether model attention corresponds to clinically relevant features, and (3) comprehensive uncertainty quantification that flags unreliable predictions for human review.

2 Related Work

2.1 Deep Learning for Skin Lesion Classification

The application of deep learning to dermoscopic image analysis gained significant attention following Esteva et al.'s demonstration that a CNN trained on clinical images could match dermatologist-level performance [2]. Subsequent work has explored various architectures and training strategies. The ISIC challenges have driven progress in this area by providing standardized benchmarks and large-scale annotated datasets. Recent approaches have employed transfer learning from ImageNet-pretrained models, with EfficientNet variants achieving strong results due to their favorable accuracy-efficiency tradeoff [4]. Gessert et al. [5] demonstrated that ensembles of EfficientNet models with extensive data augmentation could achieve top performance on ISIC challenges, though their work focused primarily on classification accuracy rather than interpretability.

2.2 Explainability in Medical Imaging

Explainable AI (XAI) has become increasingly important in medical applications where understanding model reasoning is critical for clinical acceptance. Gradient-weighted Class Activation Mapping (Grad-CAM) [6] and its extensions provide visual explanations by highlighting image regions that contribute most to predictions. In dermatology, attention visualization helps verify that models focus on lesion features rather than artifacts or irrelevant background regions. Chattopadhay et al. [7] proposed Grad-CAM++, which provides more accurate localization through a weighted combination of positive partial derivatives, making it particularly suitable for medical imaging where precise localization matters.

Concept-based explanations offer an alternative approach by relating model predictions to human-understandable concepts. Kim et al. [8] introduced Testing with Concept Activation Vectors (TCAV), which learns directions in a network's feature space corresponding to user-defined concepts. This approach has been applied to medical imaging to explain predictions in terms of clinically meaningful attributes. Our FastCAV implementation adapts this framework for efficient concept-based explanations aligned with dermatological concepts.

2.3 Uncertainty Quantification in Deep Learning

Reliable uncertainty estimates are essential for clinical decision support systems, as they enable appropriate human-AI collaboration by identifying cases requiring expert review. Gal and Ghahramani [9] demonstrated that dropout applied at test time approximates Bayesian inference, providing uncertainty estimates without architectural changes. This Monte Carlo Dropout approach has been widely adopted in medical imaging due to its simplicity and effectiveness. Subsequent work has decomposed uncertainty into epistemic (model uncertainty) and aleatoric (data uncertainty) components [10], providing more nuanced characterization of prediction confidence. Our system implements this decomposition to distinguish between cases where the model lacks knowledge versus cases with inherently ambiguous features.

3 Methods

3.1 Model Architecture

MelanomaNet employs EfficientNet V2-M [11] as the backbone feature extractor. EfficientNet V2 improves upon the original EfficientNet through training-aware neural architecture search and progressive learning, achieving faster training and better parameter efficiency. The medium variant provides 54 million parameters with 1280-dimensional feature outputs, balancing capacity with computational requirements.

We process dermoscopic images at 384×384 resolution, significantly higher than the standard 224×224 used in many classification tasks. This higher resolution preserves fine details crucial for dermoscopic analysis, such as subtle pigment patterns and border characteristics. The classification head consists of global average pooling followed by dropout (rate 0.3) and a linear layer mapping to 9 output classes corresponding to ISIC 2019 categories: Melanoma (MEL), Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Benign Keratosis (BKL), Dermatofibroma (DF), Vascular lesion (VASC), Squamous Cell Carcinoma (SCC), and Unknown (UNK).

Figure 1 illustrates the complete system architecture, showing how the classification pipeline integrates with four explainability modules that operate on different levels of the model's representation.

3.2 Training Configuration

We train using weighted cross-entropy loss to address significant class imbalance in the dataset, with class weights inversely proportional to class frequencies. Optimization employs AdamW with initial learning rate 10^{-4}, weight decay 10^{-4}, and cosine annealing schedule over 100 epochs. Mixed precision training accelerates computation while maintaining numerical stability. Data augmentation includes random horizontal and vertical flips, rotation (±20), affine transformations (translation 10%, scale 0.9–1.1), and color jittering (brightness, contrast, saturation 0.2, hue 0.1).

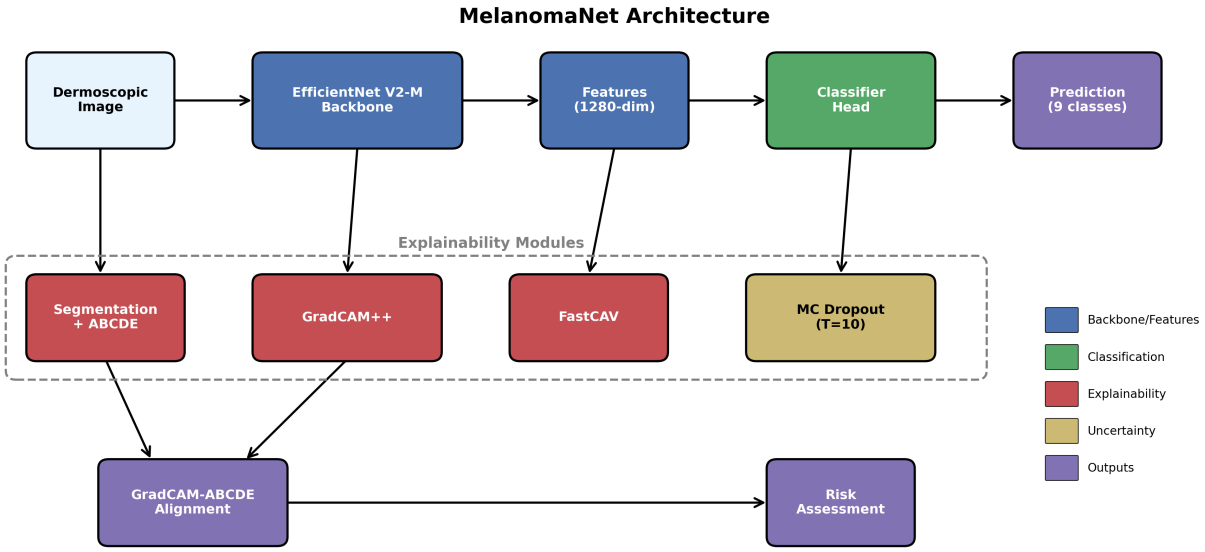

Figure 1: MelanomaNet system architecture. The main classification pipeline (top) processes dermoscopic images through EfficientNet V2-M to produce 9-class predictions. Four explainability modules (bottom, dashed box) provide complementary interpretations: ABCDE clinical criteria from image segmentation, GradCAM++ attention from feature maps, FastCAV concept scores from extracted features, and MC Dropout uncertainty estimates from stochastic forward passes. The GradCAM-ABCDE alignment module validates correspondence between model attention and clinical features.

3.3 GradCAM++ Attention Visualization

We employ GradCAM++ [7] to visualize which image regions most influence the model's predictions. GradCAM++ improves upon GradCAM by computing pixel-wise importance weights for feature map activations, providing better localization for multiple objects and sharper attention maps. The method computes gradients of the predicted class score with respect to the final convolutional layer's feature maps, then weights and combines these activations to produce a class-discriminative heatmap. The resulting visualization is upsampled to input resolution and overlaid on the original image, allowing clinicians to verify that the model attends to the lesion rather than spurious background features.

3.4 ABCDE Clinical Criterion Analysis

We implement automated extraction of ABCDE features to bridge model predictions with clinical reasoning:

Lesion Segmentation: We extract lesion masks using Otsu's thresholding with morphological refinement. Connected component analysis removes small artifacts, and hole-filling produces robust binary masks.

Asymmetry (A): We compare lesion halves along horizontal and vertical axes through the centroid. The asymmetry score quantifies the non-overlapping area between reflected halves, normalized to $[0, 1]$.

Border Irregularity (B): We analyze contour properties including compactness (perimeter^2/area) and vertex count from polygon approximation. Higher scores indicate more irregular borders.

Color Variation (C): K-means clustering ($k = 6$) identifies distinct colors within the lesion. We count colors exceeding 5% coverage and compute color standard deviation as a variation score.

Diameter (D): We compute the minimum enclosing circle and bounding box diagonal, reporting the maximum extent in pixels.

Risk Stratification: A composite risk score sums binary flags from each criterion (thresholds: asymmetry > 0.3, border > 0.4, colors > 3, diameter > 114 pixels). Scores ≥ 3 indicate high risk, ≥ 2 medium risk, and < 2 low risk.

3.5 GradCAM-ABCDE Alignment

To validate that model attention corresponds to clinically relevant features, we compute alignment metrics between GradCAM heatmaps and ABCDE analysis:

$$\text{Alignment}_{\text{border}} = \frac{\sum_{i,j} M_{\text{border}}(i,j) \cdot H(i,j)}{\sum_{i,j} M_{\text{border}}(i,j)} \quad (1)$$

where M_{border} is a dilated border mask and H is the normalized attention heatmap. Similar metrics quantify atten-

tion overlap with the overall lesion region.

3.6 Fast Concept Activation Vectors

We implement FastCAV for concept-based explanations, adapting TCAV [8] with SGD classifiers for efficiency. For each clinical concept (asymmetry, irregular border, multicolor, large diameter), we train a linear classifier to distinguish concept-positive from concept-negative examples in the model's 1280-dimensional feature space. The Concept Activation Vector (CAV) is the normal to the learned decision boundary.

The TCAV score quantifies concept influence by measuring the fraction of inputs for which the model's prediction becomes more confident when moving in the CAV direction. Positive scores indicate the concept supports the prediction; negative scores indicate opposition. This provides human-interpretable explanations showing which clinical attributes drive each classification decision.

3.7 MC Dropout Uncertainty Quantification

We employ Monte Carlo Dropout [9] with $T = 10$ stochastic forward passes to estimate uncertainty. By keeping dropout active during inference, each forward pass samples from an approximate posterior distribution over model weights, enabling Bayesian uncertainty estimation without explicit probabilistic modeling. For each input, we compute three complementary uncertainty measures:

Predictive Uncertainty captures total model uncertainty through the entropy of averaged predictions:

$$H[\bar{p}] = -\sum_{c} \bar{p}_c \log \bar{p}_c, \quad \bar{p} = \frac{1}{T}\sum_{t=1}^{T} p_t \tag{2}$$

where p_t is the softmax output from forward pass t, $\bar{p}$ is the mean prediction across all T passes, and c indexes over classes. Higher entropy indicates the model is uncertain about which class to predict.

Epistemic Uncertainty measures model uncertainty arising from limited training data:

$$\text{Var}[p] = \frac{1}{T}\sum_{t=1}^{T}(p_t - \bar{p})^2 \tag{3}$$

This variance across stochastic forward passes quantifies how much the model's predictions fluctuate when different subsets of neurons are dropped. High epistemic uncertainty suggests the model lacks sufficient training examples similar to the input, and can be reduced with more data.

Aleatoric Uncertainty captures inherent data noise that cannot be reduced:

$$\frac{1}{T}\sum_{t=1}^{T} H[p_t] \tag{4}$$

Table 1: Class distribution in the ISIC 2019 dataset.

Class	Count	Percentage
NV (Nevus)	12,875	50.83%
MEL (Melanoma)	4,522	17.85%
BCC (Basal Cell Carcinoma)	3,323	13.12%
BKL (Benign Keratosis)	2,624	10.36%
AK (Actinic Keratosis)	867	3.42%
SCC (Squamous Cell Carcinoma)	628	2.48%
VASC (Vascular)	253	1.00%
DF (Dermatofibroma)	239	0.94%

Table 2: Overall test set performance metrics.

Metric	Value
Accuracy	0.8561
Precision (weighted)	0.8600
Recall (weighted)	0.8561
F1 Score (weighted)	0.8564

This averages the entropy of individual predictions, reflecting uncertainty due to ambiguous or overlapping features in the input itself. Unlike epistemic uncertainty, aleatoric uncertainty persists regardless of training data quantity.

Predictions with predictive uncertainty exceeding threshold 0.5 are flagged as unreliable, prompting clinical review. This decomposition allows clinicians to understand whether uncertainty stems from model limitations (epistemic) or inherent case ambiguity (aleatoric).

4 Experiments and Results

4.1 Dataset

We evaluate on the ISIC 2019 Challenge dataset containing 25,331 dermoscopic images across 9 diagnostic categories. The class distribution exhibits significant imbalance: Nevus (NV) comprises 50.83% of samples while Dermatofibroma (DF) represents only 0.94%. We employ stratified splitting with 70% training (17,731 images), 15% validation (3,800 images), and 15% test (3,800 images).

4.2 Classification Performance

Table 2 presents overall classification metrics on the test set. Our model achieves 85.61% accuracy with weighted precision, recall, and F1 scores all above 0.85, demonstrating strong performance despite severe class imbalance.

Table 3 provides per-class breakdown. The model performs best on Nevus (F1=0.91) and BCC (F1=0.89), which benefit from larger sample sizes. Melanoma detection achieves F1=0.77 with 80.57% precision and 74.52% recall. Minority classes show variable performance—DF achieves

Table 3: Per-class classification performance.

Class	Precision	Recall	F1	Support
MEL	0.8057	0.7452	0.7743	679
NV	0.9255	0.9011	0.9131	1931
BCC	0.8579	0.9319	0.8934	499
AK	0.7545	0.6385	0.6917	130
BKL	0.6900	0.8270	0.7523	393
DF	0.6200	0.8611	0.7209	36
VASC	0.8947	0.8947	0.8947	38
SCC	0.8519	0.7340	0.7886	94
Macro avg	0.8000	0.8167	0.8036	3800

surprisingly strong recall (86.11%) despite having only 36 test samples, while AK struggles with 63.85% recall.

4.3 Explainability Outputs

Figures 2 and 3 demonstrate the comprehensive analysis output for test images with different risk profiles and reliability assessments.

Figure 2 shows a correctly classified nevus with 94.49% confidence. The GradCAM++ heatmap confirms the model focuses on the lesion center. ABCDE analysis shows low asymmetry (0.026), regular borders (0.116), but multiple colors (6) and larger diameter (213 pixels), yielding medium risk. Uncertainty analysis reports low predictive uncertainty (0.088), flagged as "RELIABLE." FastCAV scores show large diameter (+2.29) and multicolor (+0.53) supporting the prediction while asymmetry (-1.43) and irregular border (-1.31) oppose it.

Figure 3 presents a melanoma classified with 100% confidence. Despite high confidence, the uncertainty module flags this as "UNCERTAIN" due to high predictive uncertainty (0.76) dominated by aleatoric uncertainty (0.75), demonstrating that confidence and uncertainty provide complementary information. ABCDE analysis shows acceptable asymmetry (0.12) and regular borders (0.26), but flags multiple colors (6) and large diameter (409 pixels). FastCAV concept scores show large diameter strongly supports the melanoma prediction (+5.32) while asymmetry (-2.77), multicolor (-0.74), and irregular border (-0.33) oppose it.

4.4 Alignment Validation

The GradCAM-ABCDE alignment metrics validate that model attention corresponds to clinically relevant features. Across test samples, we observe mean lesion attention of 0.60, indicating the model predominantly focuses within the segmented lesion region rather than background. Border alignment scores (mean 0.53) suggest moderate attention to lesion boundaries, consistent with border features contributing to classification.

5 Discussion and Summary

We presented MelanomaNet, an explainable deep learning system for multi-class skin lesion classification that combines strong classification performance with comprehensive interpretability. Our approach addresses a key barrier to clinical adoption by providing multiple complementary explanation modalities that relate model predictions to familiar clinical concepts.

Clinical Relevance: The ABCDE criterion analysis bridges the gap between deep learning predictions and established dermatological assessment. By automatically extracting and quantifying these clinical features, our system generates explanations that clinicians can evaluate using their domain expertise. The alignment metrics provide validation that the model attends to clinically meaningful image regions.

Uncertainty Awareness: The MC Dropout uncertainty quantification enables appropriate human-AI collaboration by identifying predictions that warrant expert review. The epistemic/aleatoric decomposition offers additional insight—high epistemic uncertainty suggests cases outside the training distribution, while high aleatoric uncertainty indicates inherently ambiguous images requiring careful clinical evaluation.

Concept-Based Reasoning: FastCAV concept scores provide an intuitive explanation format, indicating which clinical concepts support or oppose each prediction. This approach complements attention visualization by characterizing predictions in semantic terms rather than spatial regions.

Limitations: Several limitations merit discussion. The ABCDE analysis relies on automated segmentation, which can fail for lesions with low contrast or complex backgrounds. Evolution (the E criterion) requires temporal image sequences unavailable in single-image datasets. Class imbalance remains challenging—minority classes like AK achieve lower recall despite weighted loss. Finally, while our explainability mechanisms provide useful insights, they cannot guarantee that the model's internal reasoning truly follows clinical logic.

Future Work: Potential extensions include incorporating temporal analysis for evolution tracking, expanding the concept vocabulary for FastCAV, and conducting user studies with dermatologists to evaluate clinical utility.

In conclusion, MelanomaNet demonstrates that high classification accuracy and comprehensive interpretability are not mutually exclusive. By providing GradCAM++ attention visualization, ABCDE clinical criterion analysis, concept-based explanations, and uncertainty quantification, our system offers a multi-faceted approach to explainable dermoscopic image analysis that could facilitate greater trust and adoption in clinical workflows.

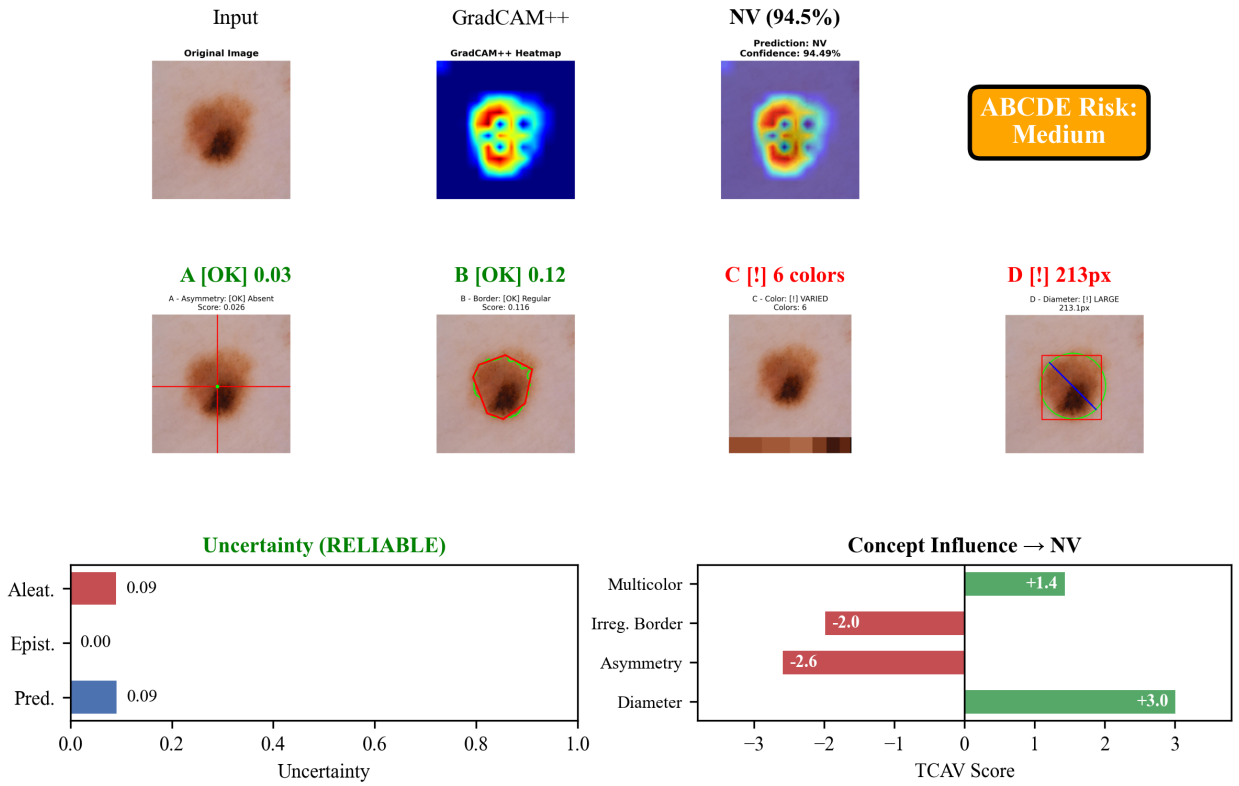

Figure 2: Analysis of a benign nevus (NV) with medium ABCDE risk. The model classifies with 94.49% confidence. Top row: original image, GradCAM++ heatmap, and overlay with prediction. Middle row: ABCDE criterion visualizations showing asymmetry axes, border contour, color palette, and diameter measurement. Bottom panels: uncertainty decomposition (left) and FastCAV concept importance scores (right).

References

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, 2023.

[2] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[3] N. R. Abbasi, H. M. Shaw, D. S. Rigel, R. J. Friedman, W. H. McCarthy, I. Osman, A. W. Kopf, and D. Polsky, "Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria," *JAMA*, vol. 292, no. 22, pp. 2771–2776, 2004.

[4] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 6105–6114.

[5] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020, pp. 616–617.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[7] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.

[8] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 2668–2677.

[9] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML)*. PMLR, 2016, pp. 1050–1059.

[10] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5574–5584.

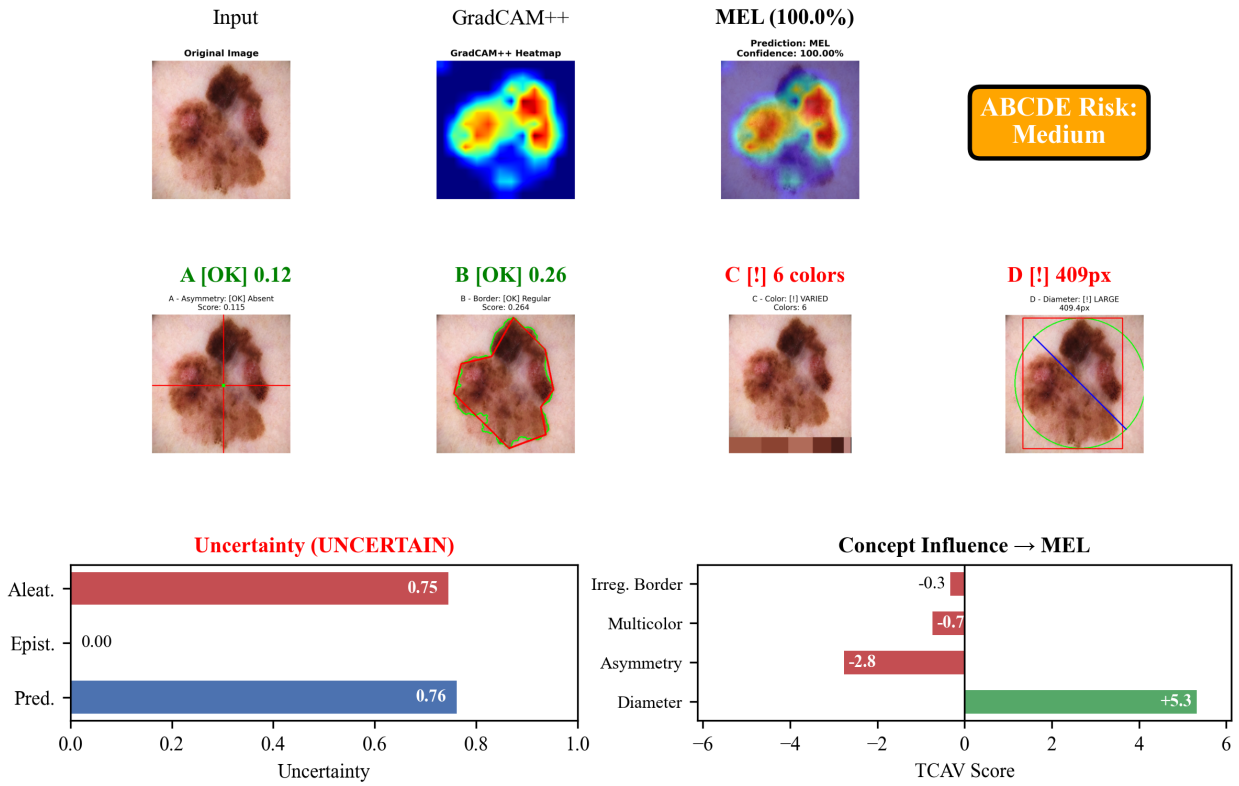

Figure 3: Analysis of a melanoma (MEL) with medium ABCDE risk. The model classifies with 100% confidence but the uncertainty module flags this as "UNCERTAIN" due to high aleatoric uncertainty (0.75). ABCDE criteria show acceptable asymmetry (0.12) and borders (0.26), but flag multiple colors (6) and large diameter (409px). FastCAV analysis reveals large diameter strongly supports the prediction (+5.32) while asymmetry opposes it (-2.77).

[11] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 10 096–10 106.