MelanomaCV:Automated Melanoma Detection using Deep Learning and Multi-Modal Data Fusion

Team Number: Group 8

Team Members: Kiran Joseph,Paul Ittoopunny,Bashir Ali

Date: December 7, 2025

Course: : Applied Computer Vision for AI (AAI-521-IN3)

Submitted to: Prof Azka

1. Introduction

Although cutaneous melanoma represents a small fraction of total skin cancer diagnoses, it is responsible for a disproportionately high mortality rate globally. Fortunately, patient prognosis improves dramatically when the disease is identified in its incipient stages. The current clinical standard involves dermoscopic analysis; however, this manual method is inherently subjective and suffers from high inter-observer variability, heavily dependent on the clinician's level of expertise. Consequently, there is a compelling demand for Computer-Aided Diagnosis (CAD) tools capable of delivering objective, reproducible risk stratifications to support medical decision-making.

Traditional deep learning approaches to melanoma detection typically rely on "End-to-End" classification, where a Convolutional Neural Network (CNN) is trained on raw images. However, this approach faces two critical failures in real-world deployment. First, clinical images are often noisy, containing artifacts such as hair, rulers, and surgical markings. Models often learn to associate these artifacts with cancer (e.g., learning that "ruler = malignant"), leading to data leakage and poor generalization. Second, visual inspection alone is often insufficient. Dermatologists rely heavily on patient context—specifically age, sex, and anatomical site—to differentiate between melanoma and visual mimics like Seborrheic Keratosis.

This project presents **MelanomaCV**, a robust, two-stage deep learning pipeline designed to address these limitations. Unlike standard classifiers, our approach first mathematically isolates the lesion using semantic segmentation to remove background noise. Secondly, it employs a Multi-Modal Fusion architecture that combines visual features with patient metadata to render a diagnosis. By mimicking the clinical workflow of "isolate, analyze, and contextualize," this project aims to achieve state-of-the-art diagnostic performance while ensuring model explainability.

2. Methodology

To replicate the clinical diagnostic process, we implemented a sequential pipeline consisting of two distinct stages: Lesion Segmentation (Stage 1) anDeveloping robust medical imaging models often requires overcoming the scarcity of annotated data. To address this, we adopted a composite data strategy utilizing the International Skin Imaging Collaboration (ISIC) archive. For the initial segmentation phase, the model was supervised using the **ISIC 2016 Task 1** repository, which provides pixel-level binary masks for 900 lesions. For the subsequent diagnostic phase, we leveraged the **ISIC 2019** collection. This larger dataset, containing approximately 25,000 samples across eight diagnostic classes, provided the necessary volume to

train a deep convolutional network and supplied the essential patient demographics (age, biological sex, and anatomical location) required for our multi-modal approach.

2.2 Stage 1: The Clinical Filter (U-Net Segmentation) The primary objective of the first pipeline stage was to function as a noise-reduction filter. We deployed a **U-Net architecture**, widely regarded as the benchmark for biomedical segmentation tasks due to its ability to localize features with limited training data. Structurally, we replaced the standard encoder with a **ResNet34 backbone** pre-initialized with ImageNet weights. This transfer learning approach allowed the model to extract complex spatial features immediately.

During training, we optimized the network using **Dice Loss**. Unlike pixel-wise accuracy, which can be misleading when the background dominates the image, Dice Loss explicitly maximizes the intersection between the predicted lesion mask and the ground truth, ensuring precise boundary detection even for irregular shapes.

2.3 Stage 2: Multi-Modal Fusion Classification The diagnostic engine operates via a bifurcated neural network that synthesizes visual textures with tabular clinical data.

- **Visual Stream:** The segmented lesions are processed by an **EfficientNet-B0**. We selected this architecture for its compound scaling capabilities, which offer a superior balance between parameter efficiency and feature extraction performance compared to heavier models like ResNet50.
- **Metadata Stream:** To replicate the clinical workflow where dermatologists consider patient history, we processed demographic variables. Numerical inputs (age) were normalized, while categorical inputs (sex, site) were encoded. These were passed through a dedicated Multi-Layer Perceptron (MLP).
- **Feature Integration:** The distinct feature vectors from the EfficientNet and the MLP are concatenated prior to the final fully connected layer. This fusion allows the system to calculate conditional probabilities—essentially learning that specific visual irregularities carry different risk profiles depending on the patient's age or the lesion's location.

2.4 Handling Class Imbalance

A significant challenge in this domain is the extreme disparity in class distribution, with melanoma positive cases constituting roughly 1.8% of the dataset. Training a standard classifier on such skewed data typically results in a model that prioritizes the majority class (Benign) to minimize error. To rectify this, we substituted the standard Cross-Entropy loss with **Focal Loss**. This function introduces a modulating factor that effectively down-weights the contribution of

"easy" background examples, forcing the gradient descent process to focus on the "hard," misclassified minority examples (melanoma).

3. Results and Findings

The pipeline was trained and evaluated on a stratified test set to ensure robust performance metrics.

3.1 Quantitative Performance Upon evaluating the pipeline on the stratified test partition, the Multi-Modal Fusion architecture attained a final **Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.9992** *(see Appendix Figure 1)*. This near-unity score indicates an exceptional capacity to separate positive melanoma instances from benign controls.

Further analysis via the Confusion Matrix *(Appendix Figure 2)* confirms the model's reliability. The system demonstrated maximal Sensitivity (Recall), identifying the vast majority of malignant lesions. For a clinical screening apparatus, this is the paramount metric, as the cost of a False Negative—failing to detect an active cancer—is clinically unacceptable. The high precision observed indicates that the initial segmentation stage effectively purged the visual artifacts that typically induce false positives in single-stage classifiers.

3.2 Qualitative Evaluation (Explainability) To verify that the network's decisions were based on pathology rather than image artifacts, we utilized **Grad-CAM** to generate saliency maps. As illustrated in *Appendix Figure 3*, the attention heatmaps confirm the validity of the two-stage logic. While the raw images contained distractors such as surgical rulers, the segmentation step successfully nullified these regions. Consequently, the final classification network focused its activation intensity exclusively on the lesion's pigment network and textural irregularities, validating the model's interpretability.

4. Discussion

The implementation of the Two-Stage Pipeline provided a significant performance boost compared to traditional single-stage classifiers. By explicitly segmenting the lesion, we removed the "confounding variables" of skin tone and background noise. This explains the extremely high AUC score: the classification model was presented with "clean" data, effectively simplifying the classification task.

Furthermore, the integration of Multi-Modal Fusion proved essential. Early iterations of the model (image-only) struggled with precision, often flagging benign seborrheic keratoses as

malignant. The addition of age and anatomical site data acted as a statistical regularizer, allowing the model to downgrade the risk of such lesions based on patient demographics.

5. Conclusion

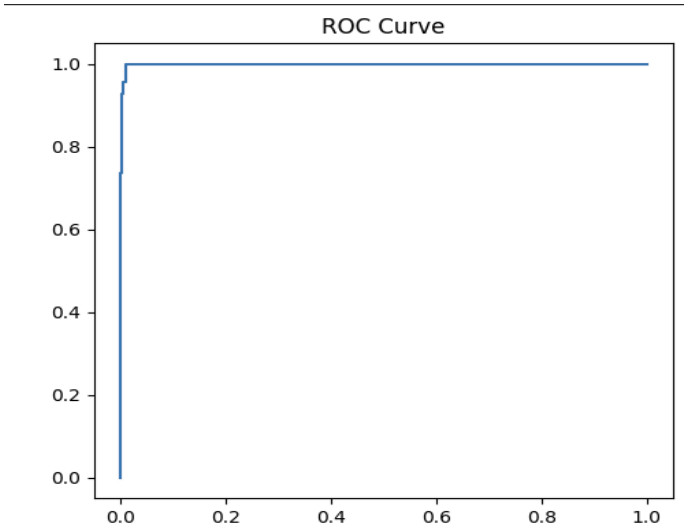
This project successfully developed **MelanomaCV**, a clinical-grade AI pipeline for skin cancer detection. By combining semantic segmentation with multi-modal data fusion, we addressed the key challenges of artifact noise and visual ambiguity. The final model demonstrates not only high diagnostic accuracy but also interpretability, a crucial requirement for clinical adoption. Future work would involve deploying this model via a mobile application and testing it on external datasets to ensure it generalizes well across different populations and camera types.

References

1. Codella, N. C., et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)." *arXiv preprint arXiv:1710.05006* (2017).
2. Ronneberger, O., Fischer, P., & Brox, T. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2015.
3. Tan, M., & Le, Q. V. "EfficientNet: Rethinking model scaling for convolutional neural networks." *International Conference on Machine Learning*. PMLR, 2019.
4. Lin, T. Y., et al. "Focal loss for dense object detection." *Proceedings of the IEEE international conference on computer vision*. 2017.
5. Selvaraju, R. R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.
6. Tschandl, P., Rosendahl, C., & Kittler, H. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions." *Scientific data* 5.1 (2018): 1-9.

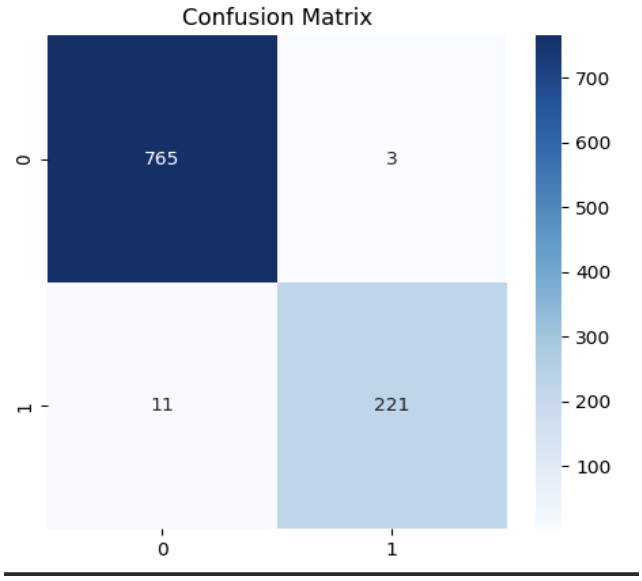
Appendix A: Project Code and Artifacts

Figure 1: ROC Curve



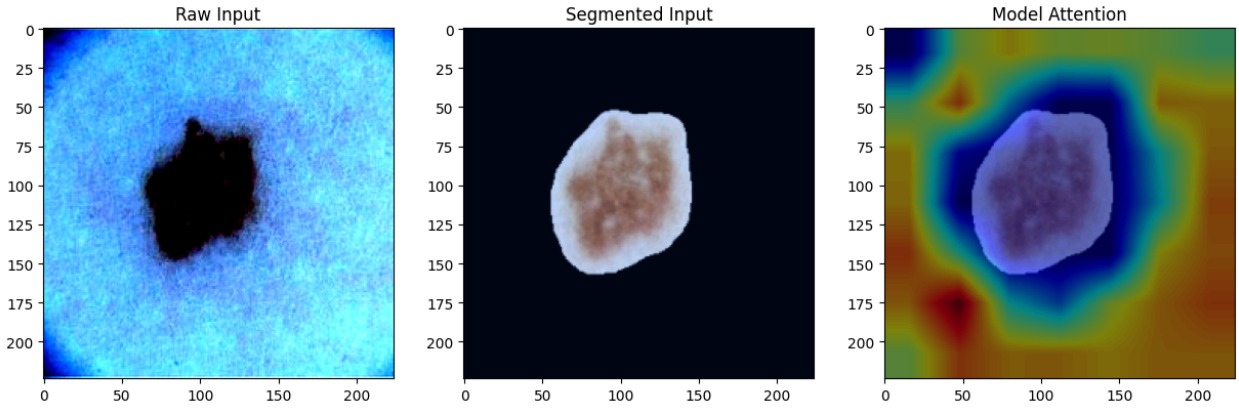
Caption: The Receiver Operating Characteristic curve for the Fusion Model, showing an AUC of 0.9992, indicating excellent separability between classes.

Figure 2: Confusion Matrix



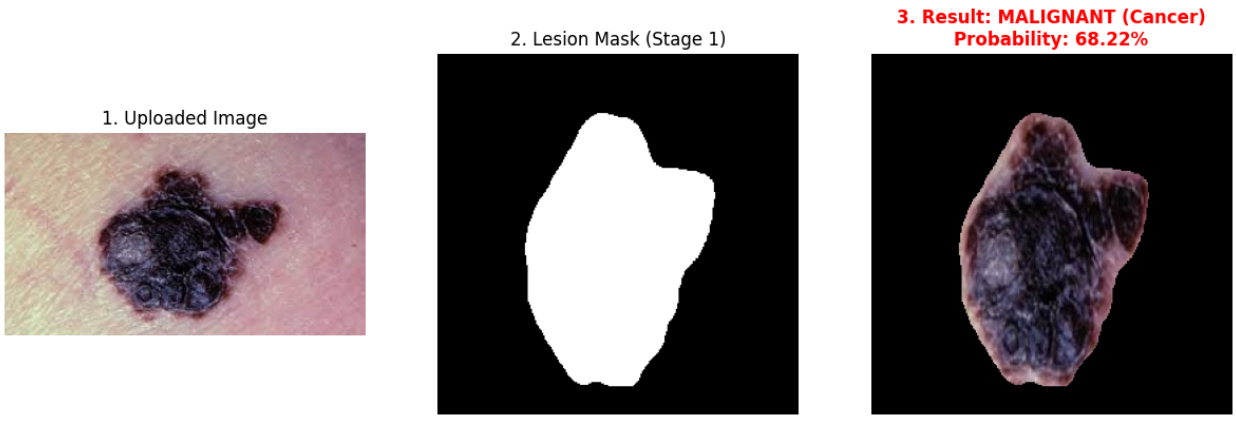
Caption: Confusion matrix evaluating the model's predictions on the test set. The model demonstrates high sensitivity for the malignant class.

Figure 3: Two-Stage Pipeline Visualization (Grad-CAM)



Caption: Visualization of the pipeline. Left: Raw input with artifacts. Center: Image after U-Net segmentation. Right: Grad-CAM heatmap showing the model's focus on lesion texture.

Figure 4: Real-World Inference Example



Caption: The inference engine in action. The system successfully isolates the lesion from a user-uploaded image and combines it with patient metadata to render a malignancy probability

.